Published in final edited form as:
Nature. 2018 August ; 560(7718): 325–330. doi:10.1038/s41586-018-0409-3.

Genetic and transcriptional evolution alters cancer cell line drug response

Uri Ben-David[1], **Benjamin Siranosian**[1], **Gavin Ha**[1,2], **Helen Tang**[1], **Yaara Oren**[1,3], **Kunihiko Hinohara**[1,2], **Craig A. Strathdee**[1], **Joshua Dempster**[1], **Nicholas J. Lyons**[1], **Robert Burns**[2], **Anwesha Nag**[2], **Guillaume Kugener**[1], **Beth Cimini**[1], **Peter Tsvetkov**[1], **Yosef E. Maruvka**[1], **Ryan O'Rourke**[1,2], **Anthony Garrity**[1], **Andrew A. Tubelli**[1], **Pratiti Bandopadhayay**[1,2,3], **Aviad Tsherniak**[1], **Francisca Vazquez**[1], **Bang Wong**[1], **Chet Birger**[1], **Mahmoud Ghandi**[1], **Aaron R. Thorner**[2], **Joshua A. Bittker**[1], **Matthew Meyerson**[1,2,3], **Gad Getz**[1,5], **Rameen Beroukhim**[1,2,3,4,#], and **Todd R. Golub**[1,2,3,6,#]

[1]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

[2]Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[3]Harvard Medical School, Boston, Massachusetts, USA

[4]Brigham and Women's Hospital, Boston, Massachusetts, USA

[5]Massachusetts General Hospital, Boston, Massachusetts, USA

[6]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA

Abstract

Human cancer cell lines are the workhorse of cancer research. While cell lines are known to evolve in culture, the extent of the resultant genetic and transcriptional heterogeneity and its functional consequences remain understudied. Here, genomic analyses of 106 cell lines grown in two laboratories revealed extensive clonal diversity. Follow-up comprehensive genomic characterization of 27 strains of the common breast cancer cell line MCF7 uncovered rapid genetic diversification. Similar results were obtained with multiple strains of 13 additional cell lines. Importantly, genetic changes were associated with differential activation of gene expression programs and marked differences in cell morphology and proliferation. Barcoding experiments

Correspondence should be addressed to R.B. (rameen_beroukhim@dfci.harvard.edu) or T.R.G. (golub@broadinstitute.org).
#These authors jointly supervised this work.

Contributions
U.B.-D. conceived the project, collected the data, performed the experiments and carried out the analyses; B.S. assisted with computational analyses and figure preparation; G.H. assisted with ichorCNA, PyClone analyses and building the Cell STRAINER portal; H.T. assisted with cell culture and experiments; N.J.L. assisted with the L1000 assay; R.Burns, A.N. and A.R.T. assisted with the OncoPanel assay and analysis; B.C. assisted with the Cell Painting analysis; J.A.B. assisted with the chemical screens and their analysis; P.T. assisted with the proteasome activity assay and with the Western blots; P.B., R.O. and A.G. assisted with the DNA barcoding experiment; C.A.S. and K.H. derived MCF7 single-cell clones; Y.O. assisted with single-cell RNAseq data analysis; J.D. and F.V. assisted with the analysis of CRISPR screens; M.G., G.K. and A.T. assisted with the comparison of the Broad and Sanger whole-exome sequencing; Y.E.M. assisted with compiling the RPE1 RNAseq dataset; C.B., and G.G. assisted with building a web portal; B.W. and A.A.T. assisted with figure design and preparation; R.Beroukhim and T.R.G. directed the project. U.B.-D., G.H., M.M., R.Beroukhim and T.R.G. wrote the manuscript.

Competing financial interests
The authors declare no competing financial interests.

showed that cell line evolution occurs as a result of positive clonal selection that is highly sensitive to culture conditions. Analyses of single cell-derived clones demonstrated that ongoing instability quickly translates into cell line heterogeneity. Testing of the 27 MCF7 strains against 321 anti-cancer compounds uncovered strikingly disparate drug response: at least 75% of compounds that strongly inhibited some strains were completely inactive in others. This study documents the extent, origin and consequence of genetic variation within cell lines, and provides a framework for researchers to measure such variation in efforts to support maximally reproducible cancer research.

Human cancer cell lines have facilitated fundamental discoveries in cancer biology and translational medicine[1]. An implicit assumption has been that cell lines are clonal and genetically stable, and hence results obtained in one study can be readily extended to another. Yet findings involving cancer cell lines are often difficult to reproduce[2,3], leading investigators to conclude that the findings were either weak or the studies not carefully conducted. For example, while pharmacogenomic profiling of large collections of cancer cell lines have proven largely reproducible, some discrepancies in drug sensitivity remain unexplained[4–11]. We hypothesized that cancer cell lines are neither clonal nor genetically stable, and that this instability can generate variability in drug sensitivity.

Cross-laboratory comparisons

To test the hypothesis that clonal variation exists within established cell lines, we re-analyzed whole-exome sequencing data from 106 cell lines generated by both the Broad Institute (the Cancer Cell Line Encyclopedia (CCLE)) and the Sanger Institute (the Genomics of Drug Sensitivity in Cancer (GDSC)), using the same analytical pipeline for both datasets (**Methods**).

As expected, estimates of allelic fraction (AF) for germline variants were nearly identical across the two datasets (median r=0.95), indicating that sequencing artifacts do not substantially contribute to the erroneous appearance of low AF calls. However, the degree of agreement in AF for somatic variants was substantially lower (median r=0.86; $p<2*10^{-16}$; Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). Moreover, a median of 19% of the detected non-silent mutations (range, 10% to 90%) were identified in only one of the two datasets (Extended Data Fig. 1b). Likewise, 26% of genes altered by copy number alterations (CNAs) (range, 7% to 99%) were discordant (Extended Data Fig. 1c–e). These results indicate that genetic variability across versions of the same cell line is common. Indeed, a median of 22% of the genome was estimated to be affected by subclonal events across 916 CCLE cell lines (Extended Data Fig. 1f), suggesting that subclonality may underlie the observed differences.

Genetic variation across 27 MCF7 strains

We performed extensive genomic characterization of 27 versions (hereafter called “strains”) of the commonly used estrogen receptor (ER)-positive breast cancer cell line MCF7 (ref [12–14]; **Methods**, Extended Data Fig. 1g–n, 2a–b and Supplementary Table 2), including 19 strains that had not undergone drug treatment or genetic manipulation, 7 strains that carried a genetic modification generally considered to be neutral (e.g., introduction of a reporter gene, Cas9 or a DNA-barcode), and one strain (MCF7-M) that had been expanded *in vivo* in mice following anti-estrogen therapy. Strain M was found to be an outlier, consistent with having been through strong bottlenecks, and was therefore excluded from downstream quantitative analyses.

Ten chromosome arms (25% of the genome) were differentially gained or lost in a pairwise comparison of strains (Supplementary Table 3). We detected 283 genes with copy-gains and 405 genes with copy-losses (compared to basal ploidy) in at least one strain. Only a small minority of these (13% of gains and 21% of losses) were detected in all strains. 7% of gains and 13% of losses were detected in only a single strain, and the remaining events were observed variably across strains (Fig. 1b and Supplementary Table 4). The differential events included genes commonly gained or lost in breast cancer (e.g. *TP53*, *PTEN*, *EGFR*, *PIK3CA* and *MAP2K4*; Extended Data Fig. 3a). For example, *PTEN* was deleted in 17 strains and retained in the other 10 (Fig. 1c). Similarly, the estrogen receptor gene *ESR1* was gained in 12 strains, lost in 6, and unaltered in 9 (Fig. 1c), and this correlated with differential expression of ERα (p=0.009; Extended Data Fig. 3b–c and Supplementary Discussion).

Genetic variation was similarly observed at the level of point mutations, small insertions/deletions (indels) and chromosomal translocations. Only 35% of 95 non-synonymous single nucleotide variants (SNVs) and indels affecting coding sequence or splicing were shared by all strains. 29% were unique to a single strain, and the remaining were present in a subset of strains (Fig. 1d–e, Extended Data Fig. 3d–j, Supplementary Tables 5–6 and Supplementary Discussion). Similar, albeit lower, variability was observed among mutations listed as recurrent in the COSMIC database[15], consistent with COSMIC mutations tending to be clonal founder mutations (Extended Data Fig. 3f).

Unsupervised hierarchical clustering, where genetic distance is reflected by branch lengths of the dendrogram, generated branch structure that accurately reflected the strains' history. For example, strain M, which had been subjected to *in vivo* passaging and drug treatment, was the most genetically distinct; the 11 strains used by the Connectivity Map project[16] over a 10 year period clustered tightly together; and sibling strains D and E, merely a few passages apart, were the closest to each other (Fig. 1f–g and Extended Data Fig. 3g). The genetic distance between strains appeared to be affected more by passage number and genetic manipulation than by freeze-thaw cycles (Fig. 1h and Extended Data Fig. 4).

Sources of variation

Analysis of variant AFs revealed extensive subclonality across strains (Fig. 2a–b and Extended Data Fig. 5a). For example, all 27 strains harbored the *PIK3CA*-activating mutation c.1633G>A, but the AF varied from 0.21 to 0.70 (Extended Data Fig. 5b). Based on AFs and copy number status, 45% of all observed mutations were determined to be subclonal (p<0.01 in a binomial test). PyClone[17,18], which reconstructs subclonal structure by clustering mutations with similar cellular prevalence (CP), indicated multiple subclones within each MCF7 strain, with varying abundance across strains (Fig. 2c). Indeed, for 43%

of the non-silent SNVs, CP differed by >50% across strains (Extended Data Fig. 5c–d and Supplementary Table 7).

We next asked whether clonal dynamics were stochastic or the product of selection. We barcoded MCF7 cells (strain D) and evaluated the change in barcode representation over time under five culture conditions, each in five replicates. We reasoned that if clonal dynamics were stochastic, distinct barcoded populations would emerge in independent replicates. In contrast, if pre-existing subclones were selected under different conditions, enrichment of the same barcodes would be observed in replicate cultures[19]. Unsupervised hierarchical clustering by barcode representation revealed that biological replicates clustered together (Fig. 2d and Supplementary Table 8), indicating that pre-existing clones are indeed selected by changes in culture conditions.

Next, we characterized the genetic stability of three wild-type (WT) single cell-derived MCF7 clones and five single cell-derived clones with a “neutral” genetic manipulation (stable expression of a luciferase reporter; **Methods**, Extended Data Fig. 5e and Supplementary Tables 9–10). Clones derived from the same parental population differed in their mutational landscapes: a median of 15% of the non-silent SNVs detected in the WT parental population (range, 13% to 16%), were not observed in their single cell-derived progeny, or vice versa (Extended Data Fig. 5f–g).

Moreover, the single-cell clones continued to evolve into heterogeneous populations. We propagated two clones for 8–14 months, and sequenced their DNA at multiple time points (Supplementary Tables 9–10). A median of 13% of the non-silent SNVs (range, 8% to 16%) were not shared between time points (Extended data Fig. 5g). Similar results were observed based on cytogenetic analysis (Extended Data Fig. 5h–k and Supplementary Table 11), indicating that even single cell-derived clones are genomically unstable.

Gene expression variation

We next measured transcriptomic variation across the MCF7 strains using the L1000 assay[16,20,21] (Supplementary Table 12). Despite an overall similarity in their global gene expression profiles (Fig. 3a and Extended Data Fig. 6a), the 27 strains also showed extensive expression variation: a median of 654 genes (range, 10–1,574) were differentially expressed by at least two-fold between pairs of strains ($p<0.05$, $q<0.05$), and the differentially expressed genes converged on important biological pathways (Extended Data Fig. 6b–d and Supplementary Table 13). Importantly, the 27 strains clustered similarly in the space of mutations and expression profiles, and the expected downstream consequences of genetic mutations were observed in the gene expression variation (Fig. 1f–g, Fig. 3b–g, Extended Data Fig. 6e–i and Supplementary Table 14). For example, strains with inactivating *PTEN* mutation or activating *PIK3CA* mutation exhibited decreased *PTEN* and increased mTOR gene expression signatures, respectively (Fig. 3e–f and Extended Data Fig. 6g–i). Similarly, copy loss of *ESR1* was associated with reduced estrogen signaling (Fig. 3g and Extended Data Fig. 6g).

We further explored gene expression heterogeneity by single-cell RNA sequencing (scRNA-seq) of 26,465 individual cells from two parental and four single cell-derived clones (**Methods**, Extended Data Fig. 6j–r and Supplementary Discussion). Unsupervised clustering showed that cells from the single cell-derived clones did not cluster independently, but were mixed with the parental population, indicating high similarity in overall gene expression (Fig. 3h and Extended Data Fig. 6o). Interestingly, the extent of expression heterogeneity among the single cell-derived clones was not substantially lower than that seen in the parental population (Extended Data Fig. 6p), and increased with time in culture (Extended Data Fig. 6q–r, Supplementary Table 15 and Supplementary Discussion). These results suggest that variation in gene expression arises *de novo*, in addition to reflecting selection of pre-existing clones[22].

Verification in additional cell lines

To exclude the possibility that the variation observed across MCF7 strains was unique to that cell line, we repeated genomic analyses on 23 strains of the commonly used lung cancer cell line A549 (ref. [23]) (Extended Data Fig. 2c–d and Supplementary Tables 16–20). We observed a similar level of molecular variation across these strains (Extended Data Fig. 7). For example, loss of *CDKN2A*, the most significantly deleted gene in lung adenocarcinomas[24], was detected in 5 strains, but normal copy number was retained in the other 18 (Extended Data Fig. 7f). Whereas transcriptome analyses showed that estrogen signaling was the most variable pathway in MCF7 cells (Extended Data Fig. 6c and Supplementary Table 13), *KRAS* signaling was the most variable pathway in A549 (Extended Data Fig. 7n and Supplementary Table 20), a commonly used model of *KRAS*-dependent cancer.

The generalizability of our findings was further confirmed by deep targeted sequencing of multiple strains from 11 additional cell lines (Supplementary Tables 21–24 and Extended Data Fig. 8). Notably, genomic instability was not limited to transformed cancer cell lines (Supplementary Discussion). For example, the variation across 15 strains of MCF10A[25], a non-transformed human mammary cell line, was as high as that seen in MCF7 cancer cells (median discordance, 26%; range, 17% to 40%; Extended Data Fig. 8a,h).

Functional consequences of genomic variation

The extensive genomic variation across strains was associated with variation in biologically meaningful cellular properties. We examined several measures of basic cellular function, including doubling time and cell morphology, using quantitative live cell imaging[26] (**Methods**). MCF7 strains varied in doubling times by as much as 3.5-fold (median, 31h; range, 22–78h) (Extended Data Fig. 9a–b). Similarly, cell size and shape were highly variable across strains (Extended Data Fig. 9c–f and Supplementary Table 25). Clustering based on morphological traits echoed that based on genomics or transcriptomics (Extended Data Fig. 9g), and genomic features correlated with proliferation (Extended Data Fig. 9h–i and Supplementary Discussion).

Genomic instability also had major impact on drug response. We measured cell viability following treatment with 321 drugs at a single concentration (5μM) across the 27 MCF7 strains (Supplementary Table 26). 55 compounds had strong activity (>50% growth inhibition) against at least one strain. However, at least one strain was entirely resistant (<20% growth inhibition) to 48 of 55 (87%) active compounds (Fig. 4a–b and Extended Data Fig. 10a). The same phenomenon was observed at a more stringent threshold: of 42 compounds with strong activity in at least two strains, 33 (79%) were inactive in at least two strains (Extended Data Fig. 10b,c,d,j and Supplementary Discussion). All 33 differentially active compounds validated in an 8-point dose-response testing of each of the 27 strains (median Spearman's rho=0.42 between screens, $p=3*10^{-9}$; Extended Data Fig. 10k, Supplementary Table 27 and Supplementary Discussion).

The high degree of variability in drug response cannot be explained by irreproducibility of the assay. First, replicate treatments yielded highly concordant results (median Pearson's r=0.97, $p<2*10^{-16}$) (Extended Data Fig. 10l). Second, compounds with the same mechanism-of-action had similar patterns of activity across strains (Fig. 4a,c; $p=3*10^{-7}$). For example, the same activity pattern was observed for three proteasome inhibitors (bortezomib, MG-132 and carfilzomib) (Fig. 4d), and was associated with biochemically-measured differential proteasome activity (Extended Data Fig. 10m–o). Third, for 82% of differentially active compounds, we found differential gene expression signatures of compound mechanism-of-action [27] between sensitive and insensitive strains ($p=2*10^{-5}$; Fig. 4e–h, Extended Data Fig. 10p–u and Supplementary Tables 28–29).

Indeed, drug response was associated with transcriptional differences in relevant pathways. For example, strains sensitive to CDK inhibitors had an upregulated cell cycle signature, and strains sensitive to PI3K inhibitors had an upregulated mTOR signature (Fig. 4f–g and Extended Data Fig. 10p–q). Interestingly, the strains most resistant to treatment in general (strains M and Q) downregulated a signature of drug metabolism (Extended Data Fig. 10v). Differences in proliferation rate did not explain the majority of the observed differential drug activity (median Spearman's rho=0.017; p=0.60; Supplementary Table 30).

Genetic variation could be linked directly to differential drug response. For example, genetic inactivation of *PTEN* was associated with decreased *PTEN* and increased *AKT* expression signatures (Fig. 1c,e and Fig. 3e–f), and increased sensitivity to the AKT inhibitor IV (Fig. 4h–i). Similarly, *ESR1* loss was associated with reduced estrogen signaling (Fig. 1c and 3g), which was in turn associated with reduced sensitivity to tamoxifen or estrogen depletion (Fig. 4j and Extended Data Fig. 10w–x). More broadly, clustering of the MCF7 strains based on their drug response was highly similar to clustering based on genetics or gene expression (Fig. 1g, 2a, 3b, 4a, Extended Data Fig. 11a and Supplementary Discussion). Genome-wide CRISPR screens revealed that genetic dependencies were affected by genomic variation similarly to pharmacological dependencies (Extended Data Fig. 11b–f, Supplementary Table 31 and Supplementary Discussion), and functional analyses revealed that single cell-derived clones remained phenotypically unstable (Extended Data Fig. 11g–i and Supplementary Discussion).

We thus hypothesized that variation across otherwise isogenic strains might be harnessed to discover mechanisms of drug sensitivity and resistance. Indeed, we found that basal gene expression profiles across the 27 MCF7 strains could be more readily connected to mechanism-of-action of active drugs than did larger panels of breast cancer cell lines derived from different patients[5,8] (Fig. 4k, Supplementary Table 32 and Supplementary Discussion).

Discussion

Our results show that established cancer cell lines, generally thought to be clonal, are in fact highly genetically heterogeneous across strains. This heterogeneity results both from clonal dynamics (i.e., changes in the abundance of pre-existing clones) and from ongoing instability (i.e., the appearance of new genetic variants). Moreover, genetic heterogeneity yields varying patterns of gene expression, which in turn result in differential drug sensitivity. These findings have a number of important implications summarized in Extended Data Table 1.

We found that changes in clonal composition underlie much of the observed variability in cell line behavior. Such clonal composition changes follow selection by particular conditions (e.g., growth media), or by genetic manipulations associated with a population bottleneck. The genetic distance between cell line strains was strongly correlated with their gene expression distance and with their drug response distance. Cell line diversification can therefore be estimated using inexpensive profiling methods (Extended Data Fig. 11j). To facilitate routine assessment of cell line diversification, we have created the Cell STRAINER (STRAin INstability profilER) portal (https://cellstrainer.broadinstitute.org), where users can upload cell line genomic data and measure the strain's genetic distance from a reference.

Variation within cancer cell lines can also be useful in at least two ways. First, deeper characterization (e.g., by single-cell sequencing) of the heterogeneity within cultures of common cell lines could enable the study of cooperative and competitive interactions between cancer cell populations[28,29], and mechanisms of pre-existing drug resistance[19]. Second, due to their matched genetic background, naturally-occurring "isogenic-like" strains could help uncover the association between molecular features and phenotypes such as drug response.

We conclude that cancer cell lines remain a powerful tool for cancer research, but the high degree of variation across cell line strains must be considered in experimental design and data interpretation.

Online Methods

Cell culture

MCF7, HT29, MDAM453 and A375 cell line strains were cultured in RPMI-1640 (Life Technologies), with 10% Fetal Bovine Serum (Sigma-Aldrich) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies). A549, VCaP, PC3, HCC515, HepG2, HeLa and Ben-Men-1 cell line strains were cultured in DMEM (Life Technologies), with 10% Fetal Bovine Serum (Sigma), 2mM Glutamine (Sigma-Aldrich), and 1% Penicillin-

Streptomycin-Glutamine (Life Technologies). HA1E cell line strains were cultured in MEMα (Life Technologies), with 10% Fetal Bovine Serum (Sigma), 2mM Glutamine (Sigma-Aldrich), and 1% Penicillin-Streptomycin-Glutamine (Life Technologies). MCF10A cell line strains were cultured in MEGM Mammary Epithelial Cell Growth Medium (Lonza) supplemented with the MEGM Bulletkit (Lonza). The single cell-derived clones scWT3, scWT4 and scWT5, as well as their parental MCF7 population, were cultured in DMEM (Life Technologies), with 10% Fetal Bovine Serum (Sigma), 2mM Glutamine (Sigma-Aldrich), 1% Penicillin-Streptomycin-Glutamine (Life Technologies), and 10μg/mL Insulin (Sigma-Aldrich). Cells were incubated at 37°c, 5% CO_2, and passaged twice a week using Trypsin-EDTA (0.25%) (Life Technologies). All strains of the same cell line were cultured under the same conditions, cell identity was confirmed and they were confirmed to be mycoplasma-free. Cells were tested for mycoplasma contamination using the MycoAlert™ Mycoplasma Detection Kit (Lonza), according to the manufacturer's instructions. Cell line identity was confirmed using SNP-based DNA fingerprinting (see below).

Derivation of single-cell clones

The WT single cell-derived MCF7 clones were generated by cell sorting. Single cells were sorted into individual wells of 96-well plates, using BD FACSAriaII SORP Cell Sorter. Three resultant clones were expanded for a period of ~3 months before prior to the experiments. The genetically-manipulated single cell-derived MCF7-*GREB1* and MCF7-*ESR1* clones were generated using CRISPR/Cas9 mediated genome engineering to insert a NanoLuciferase reporter gene into the 3'-UTR of the respective genes. Briefly, a selectable reporter gene cassette was engineered using the EMCV IRES element to drive expression of the destabilized NLucP reporter gene (Promega) fused to the N-terminus of the Bsr blasticidin-resistance gene (Invivogen) containing a P2A self-cleaving peptide element. For targeting *GREB1*, the reporter gene cassette was subcloned into a construct containing ~2 kb of GREB1 gene surrounding the termination codon in exon 33, such that reporter gene cassette is located 9 bp downstream of the *GREB1* termination codon in the resulting mRNA hybrid transcript. A *GREB1*-specific sgRNA was generated recognizing the sequence GCTGACGGGACGACACATCTG on the sense strand, and utilizing a PAM site that is adjacent to the *GREB1* termination codon. For targeting *ESR1*, the reporter gene cassette was subcloned into a construct containing ~2 kb of *ESR1* gene surrounding the termination codon in exon 8, such that reporter gene cassette is located 21 bp downstream of the *ESR1* termination codon in the resulting hybrid mRNA transcript. An *ESR1*-specific sgRNA was generated recognizing the sequence GTCTCCAGCAGCAGGTCATAG on the anti-sense strand, and utilizing a PAM site that is 160 bp upstream of the *ESR1* termination codon. Corresponding Cas9-sgRNA and targeting construct pairs were transiently co-transfected into MCF7 cells using the LipofectAMINE 2000 reagent (Thermo-Fisher Scientific). After outgrowth for 7 days, the cells were cultured in media containing 5 μg/ml blasticidin to select for the desired recombinants. Single-cell clones were then isolated by a limiting dilution single-cell cloning in 96-well plates.

Growth rate analysis

Cells were seeded in triplicates in white, clear bottom, 96-well plates (Corning #3903), at a density of 5,000 cells/well. Plates were incubated in an IncuCyte ZOOM instrument (Essen

Bioscience) at 37°C, 5% CO_2. Four non-overlapping phase-contrast images (10X) were taken every 2 hours for a total of 160 hours. The IncuCyte ZOOM Software (version 2015A) was used to calculate the mean confluence per well at every time point (filtered to exclude objects smaller than 100 μm^2), and averaged across wells to calculate the mean confluence per strain. Doubling times were calculated for each strain, using the formula $T_{doubling} = (\log2*\Delta T)/(\log(c_2)-\log(c_1))$, where c_1 and c_2 were the minimum and maximum percent confluence during the linear growth phase, respectively, and ΔT was the time elapsed between c_1 and c_2. To account for potential differences in cell recovery following seeding, t=0 h was defined as the first time point in which the mean strain confluence surpassed a threshold of 15%. To examine the effect of estrogen-depletion on the growth of MCF7 strains, cells were cultured either in standard conditions (described above) or in estrogen-depleted conditions: RPMI-1640 without Phenol Red (Life Technologies), with 10% Charcoal Stripped Fetal Bovine Serum (Life Sciences) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies). Comparison between standard and estrogen-depleted conditions was performed by calculating the fold-change in doubling time between the two conditions.

Cell Painting

Cells were plated in triplicate at a density of 1,000 cells per well, and then stained and fixed as previously described[26,30]. Images were taken on a Perkin-Elmer Opera Phenix microscope with a 20X/1.0NA water immersion lens. Image quality control was carried out as previously described[31], using CellProfiler[30] and CellProfiler-Analyst[31]. For all 27 MCF7 strains, the majority of images in all three wells passed quality control, and therefore all strains were further considered. Image illumination correction and analysis were performed in CellProfiler. For each of the 27 MCF7 strains, the median value of the 1,784 measured features was computed and used for hierarchical clustering.

DNA & RNA extraction

Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer’s protocol. Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen), according to the manufacturer’s protocol.

DNA fingerprinting

Fingerprinting analysis was performed using 44 polymorphic loci. Picard Tools “GenotypeConcordance” was used to calculate the concordance between every pair of samples (for the MCF7 and A549 cohorts, separately). Samples with >95% concordance were called a match.

Ultra low pass whole-genome DNA sequencing (ULP-WGS)

Copy number characterization was performed using low-pass (~0.2x coverage) whole-genome sequencing. Libraries were prepared from 50ng of DNA using ThruPLEX-DNAseq sample preparation kits (Rubicon Genomics), according to the manufacturer’s protocol. The resultant libraries were quantified by Qubit fluorometer, Agilent TapeStation 2200, and RT-qPCR using the Kapa Library Quantification kit (Kapa Biosystems), according to the manufacturer's protocol. Uniquely indexed libraries were pooled in equimolar ratios and sequenced on a single Illumina NextSeq500 run with paired-end 35bp reads, at the Dana-Farber Cancer Institute Molecular Biology Core Facilities. The reads were aligned to the UCSC hg19 reference genome, using "BWA-MEM" (v0.07.15), with default parameters.

ULP-WGS data analysis

The ichorCNA algorithm[32] was applied to identify copy number alterations (CNAs) of large genomic segments, chromosome arms and whole chromosomes. First, the genome was divided into 1Mb bins and read counts were generated for each bin using the HMMcopy Suite (http://compbio.bccrc.ca/software/hmmcopy/). The raw read counts were then normalized to correct for GC-content and mappability biases using the HMMcopy R package[33], generating corrected read counts for each 1Mb bin. Segmentation and copy number prediction for each sample were performed using ichorCNA v0.1.0 (https://github.com/broadinstitute/ichorCNA), which is optimized for low coverage whole-genome sequencing. Parameters were initialized based on prior knowledge: --normal=0.01, --ploidy=c(3, 3.5, 4), --txnE=0.99999 --txnStrength=10,000, --maxCN=8. Remaining parameters were set to the default. For bin-level comparison between strains, we used the log2-transformed corrected read counts and determined gain and loss status using thresholds of 0.1 and −0.1, respectively. For arm-level calls, the copy number status was determined based on the largest overlapping segment.

Deep targeted sequencing (Profile OncoPanel v3)

Deep (~250x coverage) targeted exon sequencing of 447 genes commonly mutated in cancer was performed. Prior to library preparation, DNA was fragmented (Covaris sonication) to 250bp and further purified using Agentcourt AMPure XP beads. Size-selected DNA was ligated to sequencing adaptors with sample-specific barcodes during automated library preparation (SPRIworks, Beckman-Coulter). Libraries were pooled and sequenced on an Illumina Miseq to estimate library concentration based on the number of index reads per sample. Library construction was considered to be successful if the yield was ≥250ng, and all samples yielded sufficient library. Normalized libraries were pooled in batches, and hybrid capture was performed using the Agilent Sureselect Hybrid Capture kit with the POPv3_824272 bait set[34]. The list of 447 genes included in POPv3_824272 is provided as Supplementary Table 2. Captures were then pooled and sequenced on one HiSeq3000 lane. Pooled sample reads were de-convoluted and sorted using the Picard tools (http://broadinstitute.github.io/picard). The reads were aligned to the reference sequence b37 edition from the Human Genome Reference Consortium using "bwa aln" (http://bio-bwa.sourceforge.net/bwa.shtml), with the following parameters: "-q 5 -l 32 -k 2 -o 1", and duplicate reads were identified and removed using the Picard tools[35]. The alignments were further refined using the GATK tool for localized realignment around indel sites (https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php). Recalibration of the quality scores was also performed using GATK tools (http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr)[36,37]. Metrics for the representation of each sample in the pool were generated on the unaligned reads after sorting on the barcode (http://broadinstitute.github.io/picard/picard-metric-definitions.html). All samples achieved the CCGD recommended threshold of >30x coverage for >80% of the targeted bases. Average mean exon target coverage was 251.5x (range: 171.5x-336.7x) for the MCF7 samples, 288.9x (range: 208.2x-398.9x) for the A549 samples, and 257.32 (range 211.7x-442.68x) for the additional cell line samples.

Targeted sequencing data analysis

Mutation analysis for single nucleotide variants (point mutations, or SNVs) was performed using MuTect v1.1.4[38]. Indel calling was performed using the SomaticIndelDetector tool in GATK ((http://www.broadinstitute.org/cancer/cga/indelocator). Consecutive variants in the same codon were re-annotated to maximize the effect on the codon and marked as "Phased" variants. MuTect was run in paired mode, pairing the MCF7/A549 samples to a normal sample, CEPH1408. Mutations were called if detected in >2% of the reads (AF>0.02). All SNVs, indels, and phased variants were annotated with Variant Effect Predictor (VEP)[39]. Variants were filtered against the 6,500 exome release of the Exome Sequencing Project (ESP) database. Variants represented more than once in either the African- or European-American populations and less than twice in COSMIC were considered to be germline (given that no matched normal samples were available). A germline filter was not applied to the downstream analyses, however, as changes in such mutations between strains of the same cell line would have to arise in culture and may be functionally relevant. Non-silent mutations were considered to be those with the following BestEffect Variant Classification: missense, initiator codon, nonsense, splice acceptor, splice donor, splice region, frameshift, inframe insertion or inframe deletion. Mutations that appeared more than once in COSMIC were regarded as COSMIC mutations. The complete list of variants (SNVs, indels, and phased) for MCF7, A549 and additional cell lines are provided as Supplementary Tables 5, 17 and 23, respectively.

Copy number variants (CNVs, or CNAs) were identified using RobustCNV, an algorithm that relies on localized changes in the mapping depth of sequenced reads in order to identify changes in copy number at the sampled loci (Ducar et al. Manuscript in preparation). Systematic bias in mapping depth was reduced using robust regression, fitting the observed tumor mapping depth against a panel of normal samples (PON) captured using the same bait set. Observed values were then normalized against predicted values and expressed as log2ratios. A second normalization step was performed to remove GC bias, using a loess fit. Log2ratios were centered on segments determined to be diploid based on the allele fraction of heterozygous SNPs in the targeted panel. Normalized coverage data were next segmented using Circular Binary Segmentation[40] with the "DNAcopy" Bioconductor package. Finally, segments were assigned gain, loss, or normal-copy calls using a cutoff derived from the within-segment standard deviation of post-normalized mapping depths. Due to the high data quality and low within-segment standard deviation, a cutoff of ~0.1 was applied to all samples. Segment calls were summarized to gene calls by assigning them to capture intervals, and then counting the interval calls for each gene. Gene level calls were determined according to the following rules: "gain" = "+" calls >50%; "loss" = "-" calls >2 or in 100%; "gain+loss" = "-" calls >2 times and "+" calls <50%; "mixed" = "+" and "-" calls in the same gene, but below threshold; "Normal+" = "+" calls, but below threshold; "Normal-" = "-" calls, but below threshold; "Normal" = no "+" or "-" calls. The complete list of CNAs for MCF7, A549 and additional cell lines are provided as Supplementary Tables 4, 16 and 22, respectively.

For a subset of 60 genes (listed in Supplementary Table 2), rearrangements (structural variants, or SVs) were detected using BreaKmer[41], which is designed to detect larger genomic structural variations from single sample aligned short read target-captured high-throughput sequence data. Briefly, the method extracts ’misaligned’ sequences from a targeted region, such as split-reads and unmapped mates, assembles a contig from these reads, and re-aligns the contig to make a variant call. It classifies detected variants as ”insertions/deletions”, ”tandem duplications”, ”inversions”, and ”translocations”. The complete list of structural variants for MCF7 and A549 are provided as Supplementary Tables 6 and 18, respectively. Rearrangements were visualized using the “Circos” visualization tool[69].

Clonality analysis

To resolve clonal dynamics/composition we applied the PyClone algorithm v0.13.0 (https://bitbucket.org/aroth85/pyclone/wiki/Home) to the measured allelic fractions, accounting for copy number, LOH and cellularity[17]. PyClone enabled us to follow clonal dynamics throughout the evolution of cell populations[17,18]. For copy number input, we used results from ichorCNA segmentation and copy number predictions. Mutations with <50 read depth were excluded. The following parameters were used for PyClone: 10,000 iterations, 1,000 burn-in, “total_copy_number” for the prior. We also performed multi-sample analysis using PyClone, to determine the changes in clonal composition across strains. For the multi-sample analysis, mutations were selected as the union set across all 27 strains. The same parameters were used for PyClone multi-sample analysis as for the individual-sample runs.

DNA Barcoding experiment

Degenerate oligos for sgRNA-barcode library construction were synthesized by IDT and cloned into lentiGuide-Puro[42] by Gibson assembly, as describe in Joung et al.[43]. Approximately 300μg of Gibson product was transformed into 25μL of Endura electrocompetent cells (Lucigen). After a 1 hour recovery period, 0.1% of transformed bacteria were plated in a 10-fold dilution series on ampicillin plates to determine the number of successful transformants. The remainder of the transformed bacteria were cultured in 50mL of LB with 50ug/mL ampicillin for 16 hours at 30°c. Plasmid libraries were extracted using Plasmid MidiPlus kit (Qiagen) and sequenced to a depth of 6.2 million reads on Illumina Miniseq, corresponding to 6X coverage of >1 million barcodes. Lentivirus was prepared by transfecting a total of 10 million HEK 293FT cells, as described in Joung et al.[43]. The MCF7-D strain was cultured in standard conditions (described above), and four million cells were infected with a low multiplicity of infection (20–30%) to reduce the probability of each cell being infected with more than one barcode. Cells underwent puromycin selection, and the final cell pool contained ~160,000 unique barcodes. Cells were expanded for the experiment, and five million cells were then plated into each of 25 tissue culture flasks. Five culture conditions were then applied (with five replicates per condition): 1) RPMI-1640 (Life Technologies) with 10% Fetal Bovine Serum (Sigma-Aldrich) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies); 2) DMEM (Life Technologies) with

10% Fetal Bovine Serum (Sigma-Aldrich) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies); 3) RPMI-1640 without Phenol Red (Life Technologies), with 10% Charcoal Stripped Fetal Bovine Serum (Life Sciences) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies); 4) RPMI-1640 (Life Technologies) with 10% Fetal Bovine Serum (Sigma-Aldrich), 1% Penicillin-Streptomycin-Glutamine (Life Technologies) and 0.05% DMSO (Sigma-Aldrich); 5) RPMI-1640 (Life Technologies) with 10% Fetal Bovine Serum (Sigma-Aldrich) and 1% Penicillin-Streptomycin-Glutamine (Life Technologies), supplemented for the first 48 hours with 500nM bortezomib (Selleckchem S1013). After five weeks of culture, DNA was extracted and barcode abundance was assessed by DNA sequencing, as described in Joung et al.[43]. Libraries were sequenced to a median depth of 4.2 million reads, corresponding to a barcode coverage of >26X.

Transcriptional profiling with L1000

The L1000 expression-profiling assay was performed as previously described[16]. First, mRNA was captured from cell lysate using an oligo dT coated 384 well Turbocapture plate. The lysate was then removed, and a reverse transcription mix containing MMLV was added. The plate was washed and a mixture containing both upstream and downstream probes for each gene was added. Each probe contained a gene specific sequence, along with a universal primer site. The upstream probe also contained a microbead-specific barcode sequence. The probes were annealed to the cDNA over a 6-hour period, and then ligated together to form a PCR template. After ligation, Hot Start *Taq* and universal primers were added to the plate. The upstream primer was biotinylated to allow later staining with strepdavodin-phycorethrin. The PCR amplicon was then hybridized to Luminex microbeads via the complimentary, probe-specific barcode on each bead. After overnight hybridization the beads were washed and stained with strepdavodin-phycorethrin to prepare them for detection in Luminex FlexMap 3D scanners. The scanners measured each bead independently and reported the bead color/identity and the fluorescence intensity of the stain. A deconvolution algorithm converted these raw fluorescence intensity measurements into median fluorescence intensities for each of the 978 measured genes, producing the GEX level data. This GEX data was then normalized based on an invariant gene set, and then quantile normalized to produce QNORM level data. An inference model was applied to the QNORM data to infer gene expression changes for a total of 10,174 features. Per-strain gene expression signatures were calculated using a weighted average of the replicates, where the weights are proportional to the Spearman correlation between the replicates.

Transcriptional profiling data analysis

To examine how newly profiled MCF7 and A549 cells compared in gene expression to a previously acquired collection of cell line profiles (untreated samples that served as controls for Connectivity Map perturbational experiments), we used t-distributed stochastic neighbor embedding (t-SNE). Profiles were restricted to untreated profiles from the nine core Connectivity Map cell lines, and to batches with multiple untreated profiles. As samples first clustered based on their project codes, batch effect was next removed using the COMBAT algorithm[44]. t-SNE was applied on the batch-corrected data and visualized using a scatter plot. Analysis was completed using the “Rtsne” R package version 0.13. For the comparison of transcriptional variation across the nine core Connectivity Map cell lines, the collection of untreated profiles generated with the L1000 assay was used. Five profiles from each cell line were randomly chosen, and the expression variance of the 978 L1000 "landmark" genes was calculated for each cell line. For the comparison of L1000 gene expression data to the Cancer Cell Line Encyclopedia (CCLE) gene expression profiles, RNAseq and Affymetrix gene expression profiles were downloaded from the CCLE website (https://portals.broadinstitute.org/ccle/data). Data within each platform were processed using invariant set scaling, which adjusts profiles according the expression of 80 "invariant" genes, followed by quantile normalization[16]. The ranked gene expression order of the 978 "landmark" genes was compared using Spearman's correlation.

Chemical screening

MCF7 strains were tested against a small molecule Informer Set library of 321 anti-cancer compounds, assembled by the Cancer Target Discovery and Development (CTD^2) (https://ocg.cancer.gov/programs/ctd2/data-portal), using the same principles as those described in the Cancer Therapeutics Response Portal[8,45]. The list of screened compounds is detailed in Supplementary Table 26. Cells were seeded in in their culture media in white, 384 well plates (Corning #3570) at an initial density of 2,500 cells per well and incubated overnight at 37°c, 5% CO_2. The next day, 25nL (for primary screen) or 100nL (for confirmation dose response screen) of compound stocks in DMSO were added by pin transfer. Plates were incubated for 72 hours, cooled at RT for 10 minutes, and viability was measured using the CellTiter-Glo luminescent cell viability assay (Promega), according the manufacturer's protocol. After 10 minutes of incubation, luminescence was read on a Perkin Elmer Envision reader, at a speed of 0.1s/well.

Chemical screening data analysis

Data were analyzed in Genedata Screener version 13.0, using the normalization method "neutral controls", where the median of 32 DMSO wells on each plate was set to 0% activity and 0 raw signal was set to −100%. Positive controls (20μM MG-132 or 20μM bortezomib) were included on all plates (16 wells each) but were not used for normalization due to variability of response across cell lines. Dose response curves were fit using the "Smart Fit" strategy in Genedata. The % effect was defined as the high-concentration asymptote (Sinf) and qEC50 was the concentration at which the fitted curve crossed the inhibitory value representing half the maximal % effect. In addition, parameters were calculated at which the curve crossed absolute inhibitory values of 30% or 50% regardless of maximal effect (AbsEC30 and AbsEC50, respectively). AUC calculations were performed as previously described[8]: curves were fit with nonlinear sigmoid functions, forcing the low concentration asymptote to 1 using a 3-parameter sigmoidal curve fit. The AUC for each compound-strain pair was calculated by numerically integrating under the 8-point concentration-response curve. For visualization purposes, drug response curves were fit with a 4-parameter log-logistic function, based on normalized viability data from which the lowest dose viability had been subtracted. Plots were generated using the "LL.4" function in the "drc" R package (https://cran.r-project.org/web/packages/drc/). To examine a potential link between proliferation rate and differential drug response, doubling times were compared against the AUC values of the 33 differentially-active compounds.

Gene Set Enrichment Analysis (GSEA)

GSEA was performed using the 10,147 genes best inferred from the Connectivity Map linear model[33], also known as the BING gene set. Samples were divided into two classes depending on the comparison being made: samples with a genetic alteration vs. samples without it; samples sensitive to a drug (>50% inhibition) vs. samples insensitive to the same drug (<20% inhibition). Differential expression was calculated using the signal-to-noise (S2N) metric[46]. A ranked gene list and S2N values served as the input for the GSEA pre-ranked module of Gene Set Enrichment Analysis, using the Java app version 3.0. The analysis was run using the 'hallmark', 'KEGG', 'positional' and 'oncogenic' signature collections from MsigDB. To compare between our MCF7 panel, CTD[2] and GDSC, drug response were downloaded from the CTRP website (https://ocg.cancer.gov/programs/ctd2/data-portal; "v20.data.curves_post_qc" file, updated October 14th 2015) and from the GDSC website (http://www.cancerrxgene.org/downloads; "log(IC50) and AUC values" file, updated July 4th 2016). Expression profiles were downloaded from the CCLE website to match the CTD2 drug response data (https://portals.broadinstitute.org/ccle/data; "CCLE_Expression_Entrez_2012–09-29.gct"; updated October 17th 2012), and from the GDSC website to match with the GDSC drug response data (http://www.cancerrxgene.org/downloads; "RMA normalized expression data for cell lines", updated March 2nd 2017). Expression profiles were filtered to include only the genes that belong to the L1000 BING set. GSEA compared the expression patterns of the 5 strains/cell lines with the highest AUC values for each matched drug with the 5 strains/cell lines with the lowest AUC values for that drug. As the robustness of gene expression signatures varies, this quantitative analysis was restricted to the 50 well-defined hallmark GSEA gene sets[27].

Single-cell RNA sequencing

MCF7 cells were cultured as described above. For following transcriptional changes post drug treatment, MCF7-AA cells were exposed to 500nM of bortezomib (Selleckchem S1013) and harvested before treatment, after 12 hours of exposure (t12), after 24 hours of exposure (t48), or after 72 hours of exposure followed by drug wash and 24 hours of recovery (t72+24). Cells were washed, trypsinized, passed through a 40μM cell strainer, centrifuged at 400g, and resuspended at a concentration of 1,000 cells/μL in PBS + 0.5% BSA. Single cells were processed through the Chromium Single Cell 3′ Solution platform using the Chromium Single Cell 3' Gel Bead, Chip and Library Kits (10X Genomics), as per the manufacturer's protocol. Briefly, 7,000 cells were added to each channel, and were then partitioned into Gel Beads in emulsion in the Chromium instrument, where cell lysis and barcoded reverse transcription of RNA occurred, followed by amplification, shearing and 5′ adaptor and sample index attachment. Libraries were sequenced on an Illumina NextSeq 500.

Single-cell RNA sequencing data analysis

Reads were mapped to the GRCh38 human transcriptome using cell ranger 2.1.0, and transcript-per-million (TPM) was calculated for each gene in each filtered cell barcodes sample. TPM values were then divided by 10, since the complexity of single-cell libraries is estimated to be on the order of 100,000 transcripts. For each cell, we quantified the number of genes expressed and the proportion of the transcript counts derived from mitochondrial encoded genes. Cells with either <1,000 detected genes or >0.15 mitochondrial fraction were excluded from further analysis. Finally, the resulting expression matrix was filtered to remove genes detected in <3 cells. We focused on highly variable genes for downstream principal component analysis (PCA). For each dataset, we used the Seurat R package to detect variable genes based on fitting a relationship between the mean and the dispersion of each gene. We next scaled the data and regressed out UMI number and mitochondrial gene fraction to remove technical noise. The resulting scaled data were used as an input for PCA. Top significant PCs, estimated by a manual inspection of the PCA standard deviations elbow plots, were used to generate tSNE plots. For each dataset, we used Seurat[47] (http://satijalab.org/seurat/) to identify genes that vary between samples. To detect differentially-active pathways, gene ontology (GO) enrichment analysis was performed with MSigDB[27] (http://software.broadinstitute.org/gsea/msigdb) using the differentially expressed genes that passed the following thresholds: |log2FC|>0.25, Bonferroni corrected p-value<0.01, the gene was detected in >10% of the cells in each of the compared groups. Expression signatures for selected pathways were downloaded from MSigDB[27]. We evaluated the degree to which individual cells express a certain expression signature by using a procedure that takes into account the variability in signal-to-noise ratio, as previously reported[48]. To calculate pairwise cell distances, variable genes were detected, and the cell embedding matrix for the top significant PCs was used to calculate the Euclidean distance between every two cells within each sample.

Analysis of genome-wide CRISPR screens

CERES dependency scores[49] were obtained from the Broad Institute Achilles website (https://portals.broadinstitute.org/achilles/datasets/18/download). Due to an unusually large difference in screen quality between MCF7 and KPL1, the subtle differences in dependency status between these lines were dominated by effects related to screen quality. To remove these uninteresting sources of variation, we corrected CERES gene scores by removing their first six principal components. These components were well-explained by experimental batch effects related to screen performance and pDNA pool. Corrected dependency scores < −0.5 were defined as dependencies. Genes listed as "pan_dependent" in the original dependency dataset were excluded from further analysis. For a more stringent overlap comparison, genes with CERES scores between −0.4 and −0.6 in MCF7 or KPL1 were further excluded. To implement the force directed layout, described in Extended Data Fig. 11b, the full corrected dependency matrix was reduced to its top 100 principle components and a k-means clustering algorithm was run repeatedly on cell lines. Here, k is the number of clusters, and the mean cluster size (number of cell lines) / k is a parameter similar to perplexity in tSNE, set to 6 for our data. Edges between cells were weighted according to the frequency with which they co-clustered, with edges appearing less than 30% of the time ignored. Cells were then laid out using the SFPD spring-block algorithm[50]. Cell line RNAseq gene expression data and RPPA protein expression data were obtained from the CCLE website (https://portals.broadinstitute.org/ccle/data). Single sample GSEA was calculated using the ssGSEA algorithm[51].

Chymotrypsin-like activity

MCF7 cells were plated in triplicates in 96 well plates at a density of 20,000 cells per well. 24 hours later, chymotrypsin-like activity of the proteasome was assayed, using the Proteasome-Glo™ assay (Promega), according to manufactures protocol. The activity levels were normalized to the relative cell number that was measured using the fluorescent detection of resazurin dye reduction (544-nm excitation and 590-nm emission).

Western blots

For PSMC2 and PSMD2 immunoblotting, cells were lysed in HENG buffer (50mM Hepes-KOH pH 7.9, 150mM NaCl, 2mM EDTA pH 8.0, 20mM sodium molybdate, 0.5% Triton X-100, 5% glycerol), with protease inhibitor cocktail (Roche Diagnostics #11836153001). Protein concentration was determined by the BCA assay (Thermo-Fisher Scientific #23227), and proteins were resolved on SDS-PAGE for immunoblot analysis. Antibodies against the following human proteins were used: alpha-Tubulin (ab80779; Abcam), PSMC2 (MSS1–104; Enzo Life Sciences) and PSMD1 (C-7; Santa-Cruz). Visualization was performed using the ChemiDoc MP System (Bio-Rad), and ImageLab Software (Bio-Rad) was used to quantify relative band intensities. For ERα immunonblotting, cells were lysed with a mix of 4X protein loading buffer (Li-Cor 928–40004) and 10X NuPAGE sample reducing agent (Life Technologies NP0009). Protein concentration was normalized by cell counting, and proteins were resolved on SDS-PAGE. Antibodies against the following human proteins were used: beta-Actin (N-21; Santa Cruz), ERα (F-10; Santa Cruz). Visualization was performed using the Odyssey CLx imaging machine (Li-Cor), and Image Studio Software (Li-Cor) was used to quantify the relative intensities.

Generation and comparison of dendrograms

Dendrograms were constructed using Euclidean distances for continuous measures and Manhattan distances for discrete measures. Complete linkage hierarchical clustering was performed in all cases. The mutation status dendrogram was based on mutations with AF>0.05. The gene expression dendrogram was based on the 978 “landmark” genes directly measured by the L1000 assay. The copy number dendrograms were based on discrete calls (loss, normal or gain) assigned to each event based on its log2 copy number ratio, using a cutoff value of +/−0.1. The drug response dendrogram was based on normalized viability values. The cell morphology dendrogram was based on the full list of 1,784 cellular features measured. The barcode representation dendrogram was based on the log2 transformed number of reads, including only barcodes with >1,000 reads in at least one sample. To understand how dendrograms from different sources compared, the Fowlkes-Mallows index was used, as it could capture similarities in global clustering while ignoring within-group variance[52]. The “Bk” function in the “dendextend” R package was used for computations and visualizations. We compared dendrograms from different sources with k values ranging from 5 to 26. A background distribution was calculated by randomly shuffling the labels of the trees a 1,000 times, and calculating Bk values. The 95% upper quantile of the randomized distribution for each k was plotted. The maximum Bk value was used to estimate the degree of similarity between the compared pair of dendrograms.

Calculation of the distances between strains based on their genomic features

CNA distance based on LP-WGS was determined by the fraction of the genome affected by discordant CNA calls. CNA and SNV distances based on targeted sequencing were determined by Jaccard indices, defined as the number of shared events between strains (intersection) divided by the total number of evens in these strains (union). For SNVs, both the mutated gene and the exact amino acid change had to be identical to be counted as a shared event. Gene expression distances were defined as the Euclidean distances between L1000 expression profiles. Drug response distances were defined as the Euclidean distances between drug response profiles, after limiting the drug set to active drugs only (i.e., drugs that reduced the viability of at least one strain by >50%) and thresholding viability values to +/−100.

Comparisons across CCLE cell lines

Gene-level mRNA expression, copy number and mutation status data were downloaded from the CCLE website (https://portals.broadinstitute.org/ccle/data; “CCLE_Expression_Entrez_2012–09-29.gct”, updated October 17th 2012; CCLE_copynumber_byGene_2013–12-03.txt, May 27th 2014; CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct, updated February 29th 2016). The total number of point mutations and copy number changes were counted for each cell line. Chromosome arm-level events in CCLE samples were generated as described in Ben-David et al.[53], and the number of arm-level events was counted for each cell line. The fraction of the genome affected by subclonal events was estimated using ABSOLUTE[54]. Combined CNA-SNV genomic instability scores were calculated as described in Zhang et al.[55]. The DNA repair gene set was derived from the Molecular Signature Databse (http://software.broadinstitute.org/gsea/msigdb), using the “DNA_Repair” GO signature[56]. The CIN70 gene set was derived from the publication by Carter et al.[57]. For each gene set, genes not expressed at all in the CCLE dataset were removed, and the remaining gene expression values were log2-transformed and scaled by subtracting the gene expression means. The signature score was defined as the sum of these scaled gene expression values.

Comparison of Broad (CCLE) and Sanger (GDSC) genomic features

Whole-exome sequencing data for 107 matched cell lines were downloaded from the Sanger Institute (http://cancer.sanger.ac.uk/cell_lines, EGA accession number: EGAD00001001039) for the GDSC cell lines, and from the GDC portal (https://portal.gdc.cancer.gov/legacy-archive) for the CCLE cell lines. For copy number analysis, For copy number analysis, the GATK4 somatic copy number variant pipeline was applied (https://gatkforums.broadinstitute.org/gatk/discussion/9143/how-to-call-somatic-copy-number-variants-using-gatk4-cnv)36,37. Gene-level copy number calls were generated by mapping genes from segment calls using the Consensus Coding Sequence database[58]. The gene-level values were log2 transformed, and converted to discrete values using pre-defined thresholds (+/−0.1, +/−0.3 and +/−0.5). To determine the % of discordance for each cell line, the number of discordant CNA calls between each pair of strains was divided by the total number of genes (excluding genes with a neutral copy number call in both data sets). For analysis of somatic variants, the CCLE/Sanger merged mutation calls were downloaded from the CCLE portal (https://portals.broadinstitute.org/ccle/data), and target interval list files were generated for each of the 107 matched cell lines in CCLE. Mutation calling was performed using MuTect[38], with default parameters and "--force_output" enabled, to count the number of reads supporting the reference and alternate allele for each variant in each cell line. For analysis of germline variants, a common target interval list file consisting of a panel of 105,995 SNPs was generated, based on common SNVs found in 1,019 CCLE RNAseq samples, and Mutect was applied with the same parameters as described above. Comparison of allelic fractions was performed using the subset of variants with minimum depth of coverage of 10 in both Sanger and CCLE datasets and with minimum of allelic fraction of 0.1 in at least one dataset. Out of the 107 cell lines, one cell line (Dov13) lacked any germline concordance and was thus excluded from all analyses.

Cytogenetic analysis

Karyotyping was performed by KaryoLogic,Inc. (www.karyologic.com/) on 50 G-banded metaphase spreads per sample. Every spread displayed multiple chromosomal rearrangements with many marker chromosomes. A marker was defined as "a structurally abnormal chromosome that cannot be unambiguously identified by conventional banding cytogenetics." The analysis was performed according to the International System for Human Cytogenetic Nomenclature (ISCN) 2016 guidelines. Rare metaphases with >100 chromosomes were excluded from further analysis.

E-karyotyping analysis

RNAseq data from non-manipulated/non-treated samples of the near-diploid human cell line RPE1 were downloaded from the NCBI SRA website (https://www.ncbi.nlm.nih.gov/sra). STAR– paired aligner was used to align paired-end samples, and STAR –non-paired aligner was used to align the non-paired samples[59]. The STAR to RSEM tool[60] was used to generate the gene-level expression values using the gtex pipeline (https://github.com/broadinstitute/gtex-pipeline). To infer arm-level copy-number changes from gene expression profiles, the RSEM values were subjected to the e-karyotyping method[61]. Briefly, RSEM values were log2-converted, genes that were not expressed (log2RSEM<1) in >20% of the samples were excluded, and expression levels of the remaining genes were floored to RSEM=1. The median expression value of each gene across all samples was subtracted from the expression value of that gene, in order to obtain comparative values. The 10% most variable genes were removed from the dataset to reduce transcriptional noise. The relative gene expression data were then subjected to a CGH-PCF analysis, with a stringent set of parameters: Least allowed deviation = 0.5; Least allowed aberration size = 30; Winsorize at quantile = 0.001; Penalty = 18; Threshold = 0.01. CNAs exceeding 80% of the length of a chromosome arm were called arm-level CNAs.

Comparison of arm-level CNAs between cell line propagation and tumor progression

Recurrence of chromosome arm-level CNAs during breast cancer progression was determined by their frequency in TCGA samples, as previously described[53]. Recurrence of chromosome arm-level CNAs during cell line propagation was determined by comparing the arm-level calls of the strains directly separated by extensive passaging (strain D vs. strain L vs. strain AA, strain B vs. strains I/P), as shown in Extended Data Fig. 2a. Only arms that are recurrently gained or lost (but not both) in TCGA (q-value<0.05), and that have variable copy number status across the MCF7 panel, were considered for the comparison.

Statistical analysis

The significance of the difference between genomic instability associated with different sources of genetic variation, and that of the difference in chromosome number between two time points of single cell-derived clones, were determined using the two-tailed Wilcoxon rank-sum test. The significance of the difference in the euclidean distance between compounds that work through the same MoA and compounds that work through different MoA's, that of the difference in the discordance of non-silent SNVs at different stages of transformation, that of the difference in CIN70 and wGII scores between cell lines derived from primary tumors and those derived from metastases, that of the difference between the somatic and germline SNV Pearson correlations of the Broad-Sanger cell lines, and that of the differences in the Broad-Sanger somatic SNV concordance between MSI and MSS cell lines and between primary-derived and metastasis-derived cell lines, were determined using the one-tailed Wilcoxon rank-sum test. The significance of the difference in mutation cellular prevalence across strains was determined by a Kruskal-Wallis test. The significance of the difference in AKT inhibitor IV sensitivity between *PTEN*-wt and *PTEN*-mut strains, that of the difference in the relative growth effect of ER-depletion between *ESR1*-loss and no-*ESR1*-loss strains, that of the difference in proteasome activity between bortezomib-sensitive and bortezomib-insensitive strains, that of the difference in ERα protein expression levels between strains, and that of the difference in the number of arm-level CNAs between matched early-late MCF7 strains were determined using the one-tailed Student's t-test. The significance of the difference in doubling times, and that of the difference in sensitivity to estrogen depletion, was determined using the two-tailed Student's t-test. The significance of the correlation between the two replicates of the primary screen was determined using Pearson's correlation. The significance of the correlation between doubling time and the number of protein coding mutations, that of the correlation between doubling time and the fraction of subclonal mutations, that of the correlation between doubling time and drug response, were determined using a Spearman's correlation, excluding the broadly resistant strains Q and M. The significance of the correlation between *ESR1* CERES dependency scores and estrogen signaling, and that of the correlation between *GATA3* CERES dependency scores and GATA3 protein expression levels were determined using a Spearman's correlation. The deviation of the doubling time-drug response correlations from a hypothetical mean value of 0 was determined using a two-tailed one-sample t-test. The significance of the difference between the emergence and disappearance of recurrent arm-level CNAs during cell line propagation was determined using McNemar's test. The significance of the correlation between the primary and secondary drug screens was determined using a Spearman's correlation (including only compounds that were active in both screens). The significance of the directionality of drug-pathway association, and the likelihood that a mutation would be clonal given the number of reads that detected it, were determined using a binomial test. The significance between the fraction of pathways correctly identified between the MCF7 panel, CTD[2] and GDSC, was determined using a two-tailed Fisher's exact test. GSEA p-values and FDR-corrected q-values are shown as provided by the default analysis output. For the comparison of pathway prediction shown in

Supplementary Table 32, FDR q-values were re-calculated using only the pre-selected pathways. Thresholds for significant associations were determined as: $p<0.05$; $q<0.25$. The significance of the difference in the karyotypic variation between parental and single cell-clone derived cultures was determined using the Levene’s test. The significance of differentially-expressed genes in the single-cell RNAseq data was determined by an analysis of variance (ANOVA) followed by a Games-Howell post hoc test and a Bonferroni correction. Box plots show the median, 25th and 75th percentiles, lower whiskers show data within 25th percentile −1.5 times the IQR, upper whiskers show data within 75th percentile +1.5 times the IQR, and circles show the actual data points. Statistical tests were performed using the R statistical software (http://www.r-project.org/), and the box plots and violin plots were generated using the “boxplot” and “vioplot” R packages, respectively.

Code availability

The code used to generate and/or analyze the data during the current study are publicly available, or available from the authors upon request.

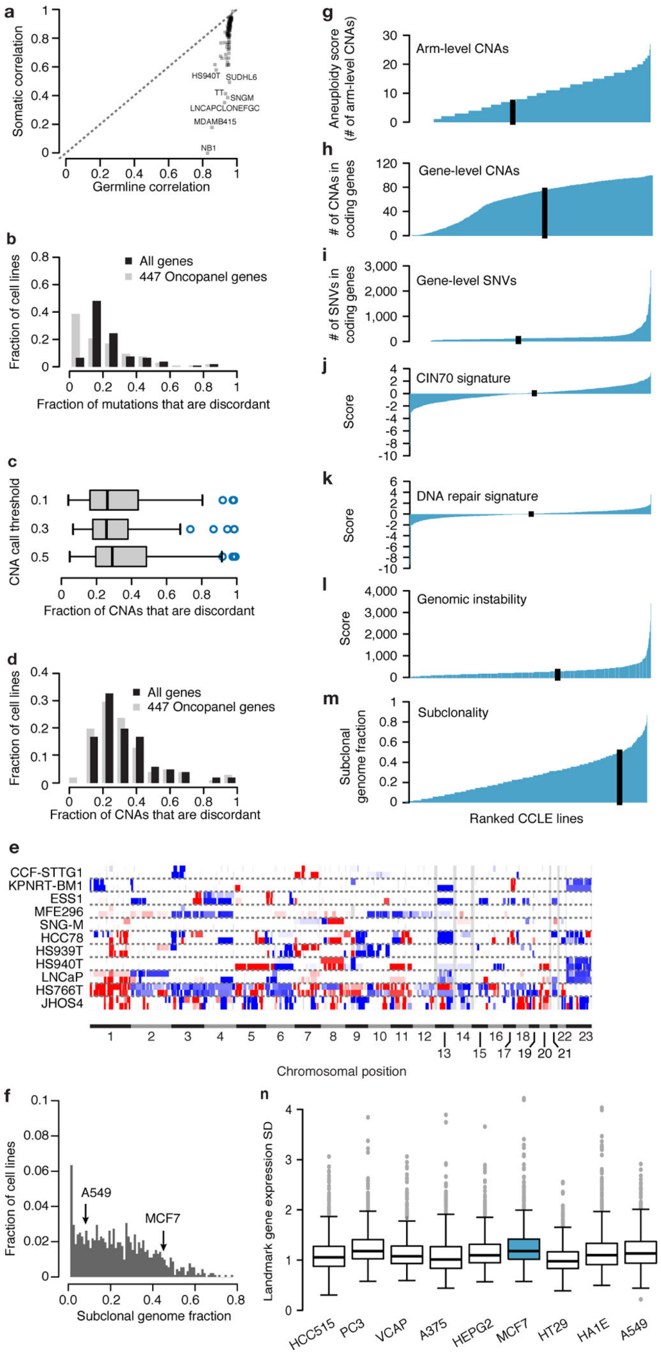
Data availability

The datasets generated during and/or analyzed during the current study are available within the article, its supplementary information files, or available from the authors upon request. DNA sequencing data were deposited to SRA with the BioProject ID PRJNA398960. Single-cell RNA sequencing data were deposited to the Gene Expression Omnibus (GEO, accession number GSE114462). Source Data of all immunostaining blots are available in the online version of this paper.

URLs

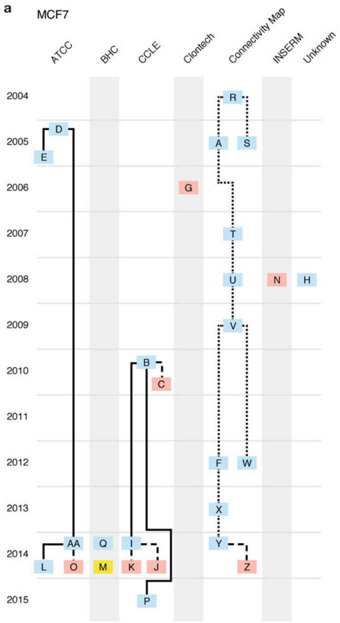
The cell divergence portal is accessible at: https://cellstrainer.broadinstitute.org.

Extended Data



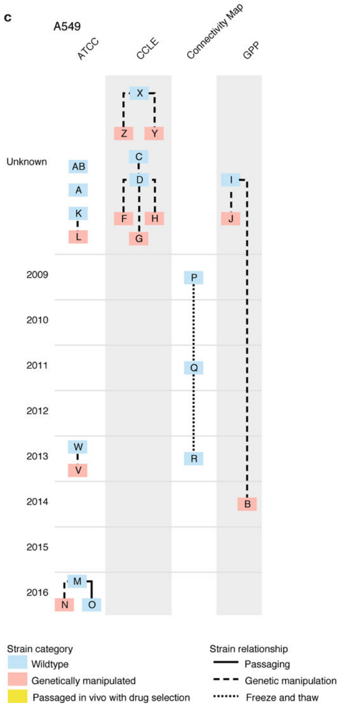
Extended Data Figure 1: Comparison of Broad and Sanger genomic features across 106 cell lines

(**a**) Comparison of the Pearson correlations of germline vs. somatic SNVs across 106 paired cell lines. (**b**) A histogram presenting the distribution of mutation discordance fractions across cell lines. The distribution of all non-silent SNVs is shown in black, and that of the 447 genes included in the Oncopanel is shown in gray. (**c**) Comparison of the fraction of discordant gene-level CNAs between the Broad and the Sanger (n=106 cell lineS), using three different threshold for CNA calling. Bar, median; box, 25th and 75th percentiles; whiskers, 1.5*IQR of lower and upper quartile; circles: data points. (**d**) A histogram presenting the distribution of mutation discordance fractions across cell lines. Mutations are colored as in (**c**). (**e**) CNA landscapes of 11 paired cell lines. For each cell line, the upper row presents the CNA landscape of the Broad strain, and the lower row presents that of the Sanger strain, Copy number gains are shown in red, and copy number losses are shown in blue. CNAs <10Mb in size are not presented. (**f**) A histogram presenting the fraction of the genome affected by subclonal events across 916 cell lines from the Cancer Cell Line Encyclopedia. MCF7 and A549 are denoted by arrows. (**g**) All CCLE cell lines ranked by their aneuploidy scores. (**h**) All CCLE cell lines ranked by the number of gene-level CNAs that they harbor. (**i**) All CCLE cell lines ranked by the number of gene-level SNVs that they harbor. (**j**) All CCLE cell lines ranked by their chromosomal instability (CIN70) signature scores[47]. (**k**) All CCLE cell lines ranked by their DNA repair signature scores[80]. (**l**) All CCLE cell lines ranked by their genomic instability scores[79]. (**m**) All CCLE cell lines ranked by their subclonal genome fraction[78]. A vertical black line represents the rank of MCF7 in each comparison. (**n**) Comparison of gene expression variation across multiple strains of nine cell lines, including MCF7. Box plots present the standard deviations of the expression levels for the 978 landmark genes directly measured in L1000. Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile; circles: all data points.



b

Strain_ID	Origin	Year	Passage	Manipulations	Remarks
MCF7-A	Connectivity Map	2005		Freeze & thaw	
MCF7-B	CCLE	2010			Parental of MCF7-I
MCF7-C	CCLE	2010		EGFP reporter	
MCF7-D	ATCC	2005	p+5		Parental of MCF7-E
MCF7-E	ATCC	2005	p+12		Derived directly from MCF7-D
MCF7-F	Connectivity Map	2012		Freeze & thaw	
MCF7-G	Clontech	2006		Tet-off	
MCF7-H	unknown	2008			
MCF7-I	CCLE	2014		Extensive passaging	Derived directly from MCF7-B; parental of MCF7-K
MCF7-J	CCLE	2014		DNA-barcoded (PRISM)	
MCF7-K	CCLE	2014		Cas9-expressing	Derived directly from MCF7-I
MCF7-L	ATCC	2014	High passage	Extensive passaging	Derived directly from MCF7-AA
MCF7-M	BHC	2014		*In vivo* tamoxifen treatment	Persistent cells: passaged in xenografts --> treated with tamoxifen --> passaged in culture again; derived from same parental as MCF7-Q
MCF7-N	INSERM	2008		YFP reporter	
MCF7-O	ATCC	2014		Luciferase reporter	
MCF7-P	CCLE	2015		Extensive passaging	Derived directly from MCF7-B
MCF7-Q	BHC	2014			Parental of MCF7-M (with continued passaging)
MCF7-R	Connectivity Map	2004			Parental of MCF7-A/F/S/T/U/V/W/X/Y/Z
MCF7-S	Connectivity Map	2005		Freeze & thaw	
MCF7-T	Connectivity Map	2007		Freeze & thaw	
MCF7-U	Connectivity Map	2008		Freeze & thaw	
MCF7-V	Connectivity Map	2009		Freeze & thaw	
MCF7-W	Connectivity Map	2012		Freeze & thaw	
MCF7-X	Connectivity Map	2013		Freeze & thaw	
MCF7-Y	Connectivity Map	2014		Freeze & thaw	
MCF7-Z	Connectivity Map	2014		Cas9-expressing	
MCF7-AA	ATCC	2014	Low passage		Parental of MCF7-L

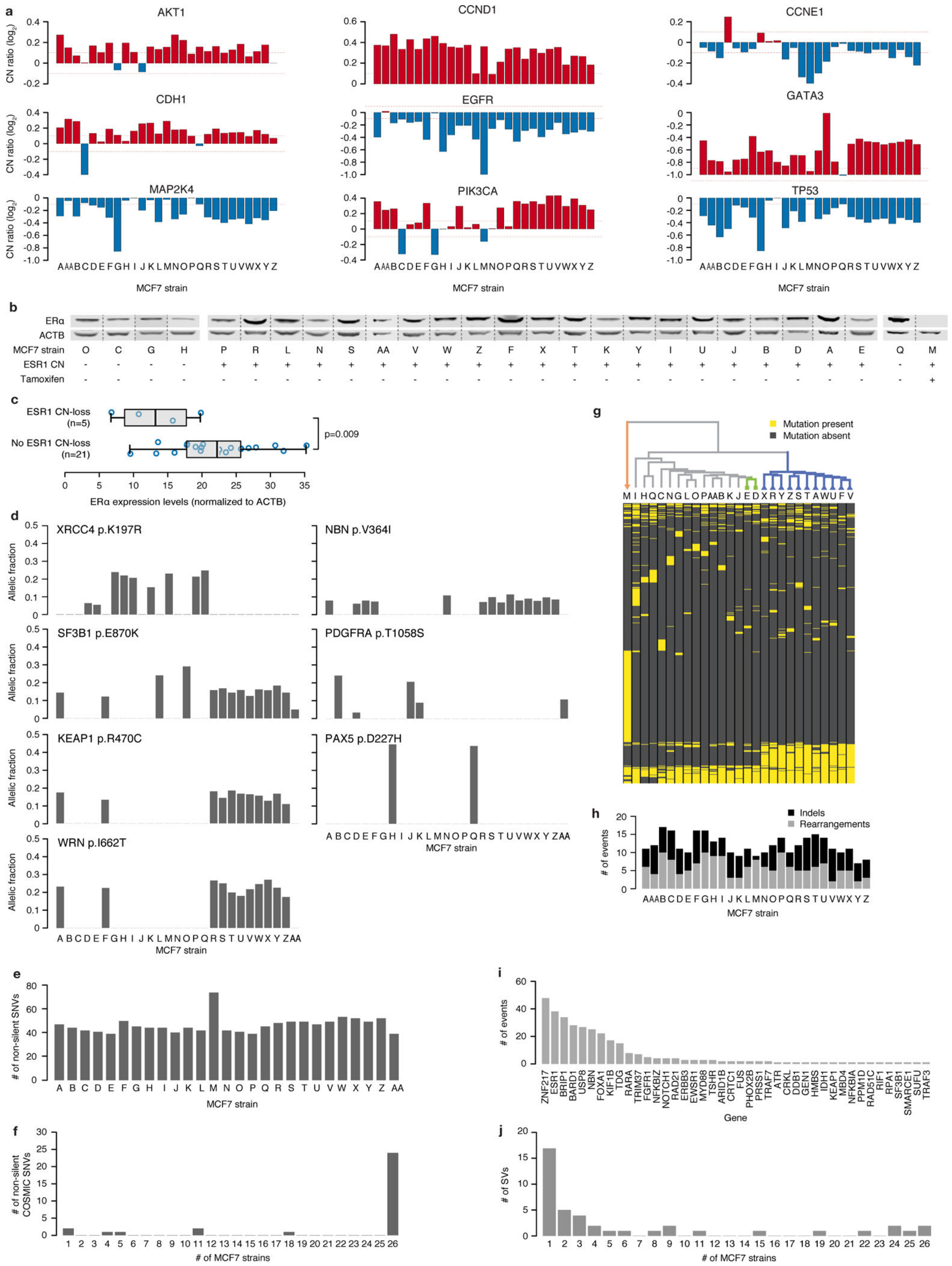


d

Strain_ID	Origin	Year	Passage	Manipulations	Remarks
A549-A	ATCC/Meyerson				
A549-B	GPP	2014		Cas9-expressing	Derived directly from A549-I
A549-C	CCLE/Hahn	2017	Early passage		Parental of A549-D
A549-D	CCLE/Hahn				Parental of A549-F/G/H
A549-F	CCLE/Hahn			sgRNA against Chr26 (intergenic)	Derived directly from A549-D
A549-G	CCLE/Hahn			sgRNA against TRIB (intergenic)	Derived directly from A549-D
A549-H	CCLE/Hahn			pLX313_Renilla-expressing	Derived directly from A549-D
A549-I	GPP				Parental of A549-B/J
A549-J	GPP			Cas9-expressing	Derived directly from A549-I
A549-K	ATCC/Meyerson				
A549-L	ATCC/Meyerson			DX-HPRT1 (inducible degradation of HPRT1)	
A549-M	ATCC/Amon	2016	Early passage		Parental of MCF7-N/O
A549-N	ATCC/Amon	2016		GFP reporter	Derived directly from A549-M
A549-O	ATCC/Amon	2016	Late passage	Extensive passaging	Derived directly from A549-M
A549-P	Connectivity Map	2009			Parental of A549-Q/R
A549-Q	Connectivity Map	2011		Freeze & thaw	
A549-R	Connectivity Map	2013		Freeze & thaw	
A549-V	ATCC/Hahn	2013	p+11	Cas9-expressing	Derived directly from A549-W
A549-W	ATCC/Hahn	2013	p+10		Parental of A549-V
A549-X	CCLE/PRISM				Parental of A549-Y/Z
A549-Y	CCLE/PRISM			DNA-barcoded (PRISM)	Derived directly from A549-X
A549-Z	CCLE/PRISM			DNA-barcoded (PRISM)	Derived directly from A549-X
A549-AB	ATCC/Brugge				

Extended Data Figure 2: Schematic representation of the MCF7 and A549 strains included in the current study

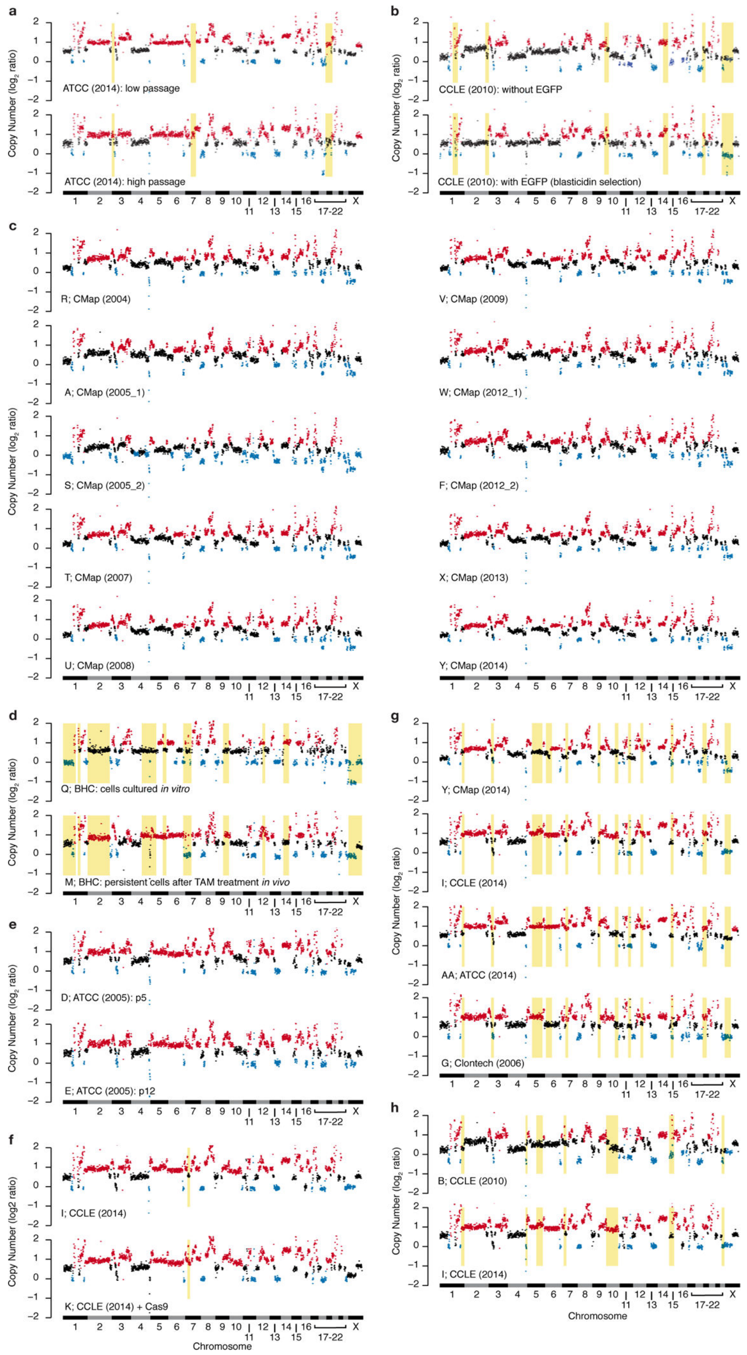
(**a**) MCF7 strains included in this study, presenting their origins (columns), years of acquisition (rows), manipulations (color) and progeny relationships (arrows). (**b**) A table of the MCF7 strains included in this study, presenting their origins, years of acquisition, passage numbers, and genetic manipulations. (**c**) A549 strains included in this study, presenting their origins (columns), years of acquisition (rows), manipulations (color) and progeny relationships (arrows). (**d**) A table of the A549 strains included in this study, presenting their origins, years of acquisition, passage numbers, and genetic manipulations.



Extended Data Figure 3: Genetic variation across 27 MCF7 strains.

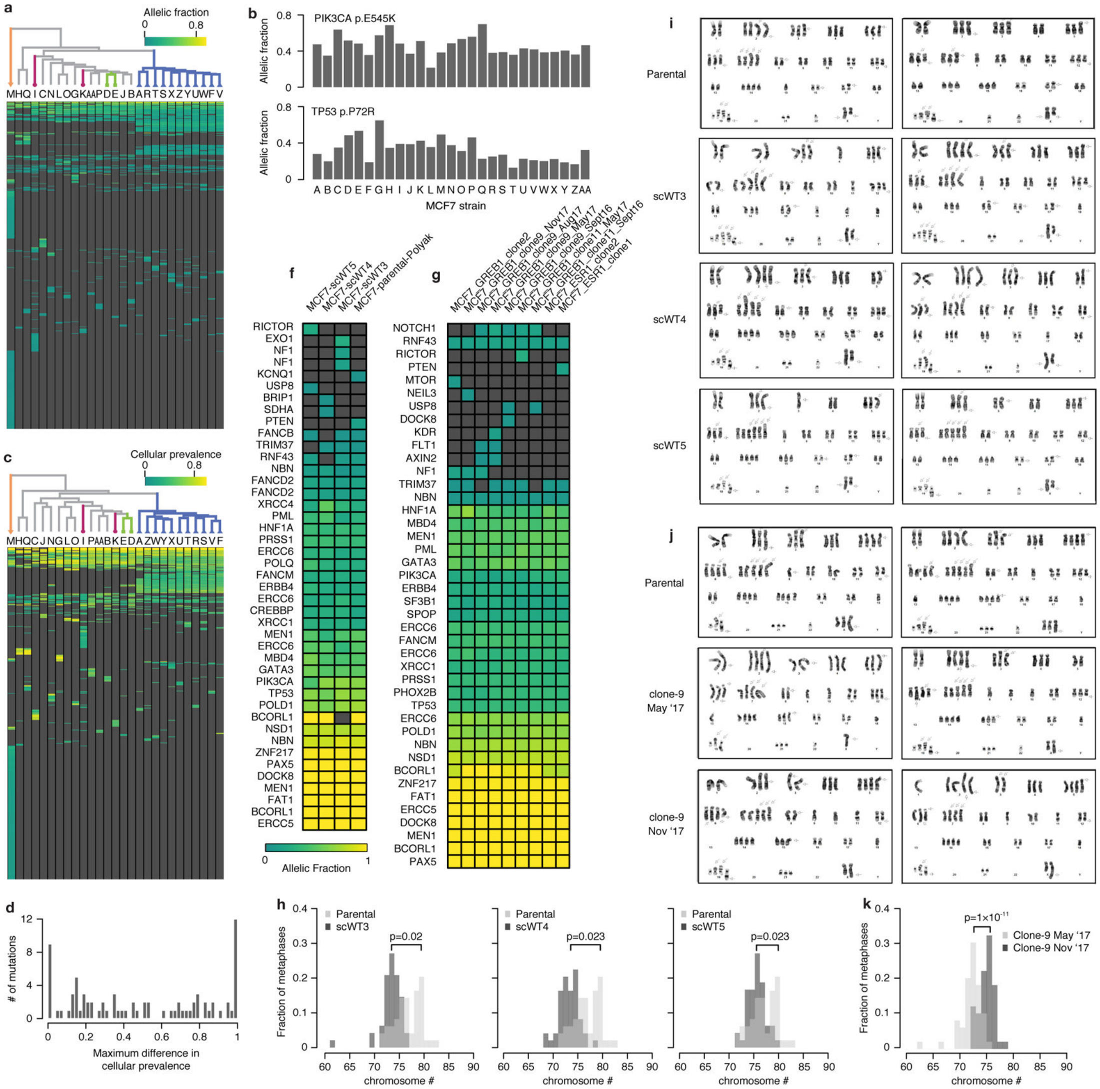
(**a**) Variation in the copy number status of nine selected genes across 27 MCF7 strains. Copy number gains are shown in red, and copy number losses in blue. Thresholds for relative gains/losses were set at ±0.1. (**b**) Western blots presenting the relative protein expression levels of ERα across strains. The expression of β-actin was used for normalization. For gel source data, see Supplementary Fig. 1. The experiment was repeated twice with similar results. (**c**) Quantification of the relative expression of ERα. Strains Q and M were excluded from the comparison. Bar, median; box, 25th and 75th percentiles; whiskers, data within

1.5*IQR of lower or upper quartile; circles: all data points. One-tailed t-test. (**d**) The allelic fractions of non-silent mutations in seven selected genes across 27 MCF7 strains. (**e**) The number of non-silent point mutations (SNVs) across the 27 MCF7 strains. (**f)** The number of COSMIC non-silent point-mutations shared by each number of MCF7 strains. (**g**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on all their SNVs. Groups of strains expected to cluster together based on their evolutionary history are highlighted, as in Fig. 1. Bottom: a corresponding heatmap, showing the mutation status of all mutations across the 27 MCF7 strains. Shown are mutations identified only in a subset of the strains, which were detected in above 5% of the reads (allelic fraction>0.05). The presence of a mutation is shown in yellow, and its absence in gray. (**h**) The number of large (>15bp) indels and rearrangements across the 27 MCF7 strains. Indels are shown in gray, and rearrangements in black. (**i**) The recurrence of SVs in each of the 42 (out of 60) genes for which at least one event was detected. (**j**) The number of SVs shared by each number of MCF7 strains. Note that this analysis is limited to the 60 genes listed in Supplementary Table 2.

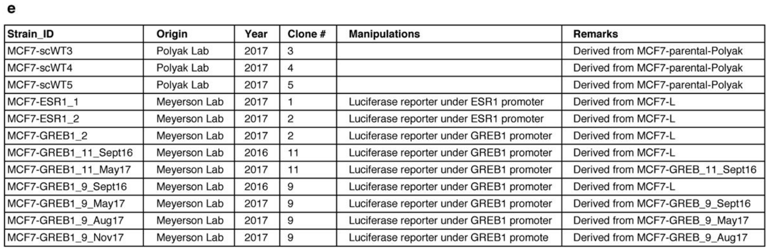


Extended Data Figure 4: Comparison of CNA landscapes between MCF7 strains.

(**a**) CNA landscapes of a pair of MCF7 strains separated from each other by extensive passaging. (**b**) CNA landscapes of three pairs of MCF7 strains separated from each other by a genetic manipulation (introduction of a GFP reporter). (**c**) CNA landscapes of 10 MCF7 strains separated by multiple freeze-thaw cycles, with little passaging in between. (**d**) CNA landscapes of a pair of MCF7 strains that were either cultured *in vitro* (top) or cultured *in vivo* and treated with tamoxifen (bottom). (**e**) CNA landscapes of a pair of MCF7 strains separated by merely seven passages from each other. (**f**) CNA landscapes of a pair of MCF7 strains before (top) or after (bottom) introduction of Cas9. (**g**) CNA landscapes of a pair of MCF7 strains obtained from four different sources. (**h**) CNA landscapes of a pair of MCF7 strains separated from each other by extensive passaging. Data points represent 1Mb bins throughout the genome. Gains are shown in red, losses in blue. Yellow backgrounds highlight differential CNAs between the compared strains.



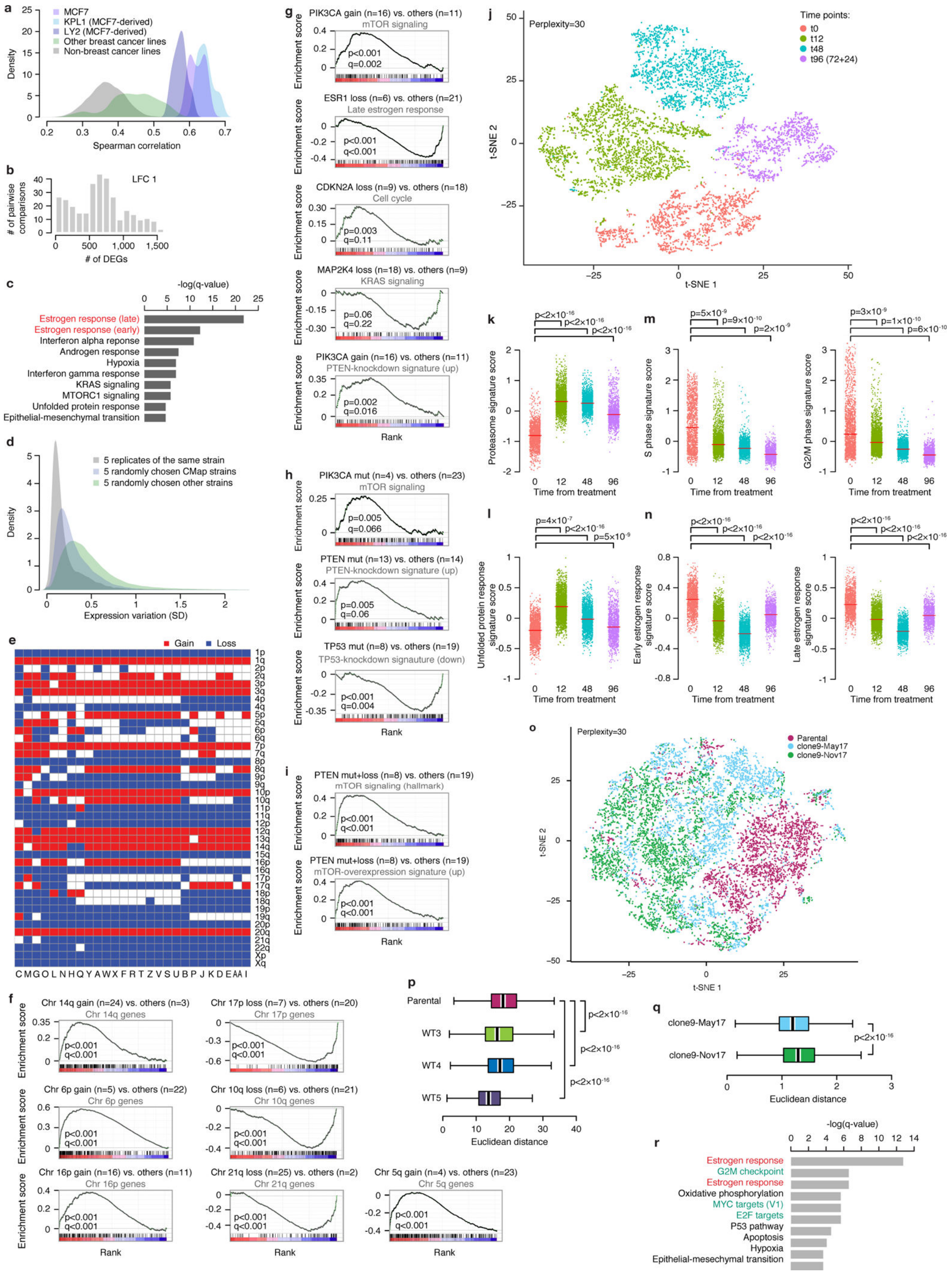
e



Strain_ID	Origin	Year	Clone #	Manipulations	Remarks
MCF7-scWT3	Polyak Lab	2017	3		Derived from MCF7-parental-Polyak
MCF7-scWT4	Polyak Lab	2017	4		Derived from MCF7-parental-Polyak
MCF7-scWT5	Polyak Lab	2017	5		Derived from MCF7-parental-Polyak
MCF7-ESR1_1	Meyerson Lab	2017	1	Luciferase reporter under ESR1 promoter	Derived from MCF7-L
MCF7-ESR1_2	Meyerson Lab	2017	2	Luciferase reporter under ESR1 promoter	Derived from MCF7-L
MCF7-GREB1_2	Meyerson Lab	2017	2	Luciferase reporter under GREB1 promoter	Derived from MCF7-L
MCF7-GREB1_11_Sept16	Meyerson Lab	2016	11	Luciferase reporter under GREB1 promoter	Derived from MCF7-L
MCF7-GREB1_11_May17	Meyerson Lab	2017	11	Luciferase reporter under GREB1 promoter	Derived from MCF7-GREB_11_Sept16
MCF7-GREB1_9_Sept16	Meyerson Lab	2016	9	Luciferase reporter under GREB1 promoter	Derived from MCF7-L
MCF7-GREB1_9_May17	Meyerson Lab	2017	9	Luciferase reporter under GREB1 promoter	Derived from MCF7-GREB_9_Sept16
MCF7-GREB1_9_Aug17	Meyerson Lab	2017	9	Luciferase reporter under GREB1 promoter	Derived from MCF7-GREB_9_May17
MCF7-GREB1_9_Nov17	Meyerson Lab	2017	9	Luciferase reporter under GREB1 promote	Derived from MCF7-GREB_9_Aug17

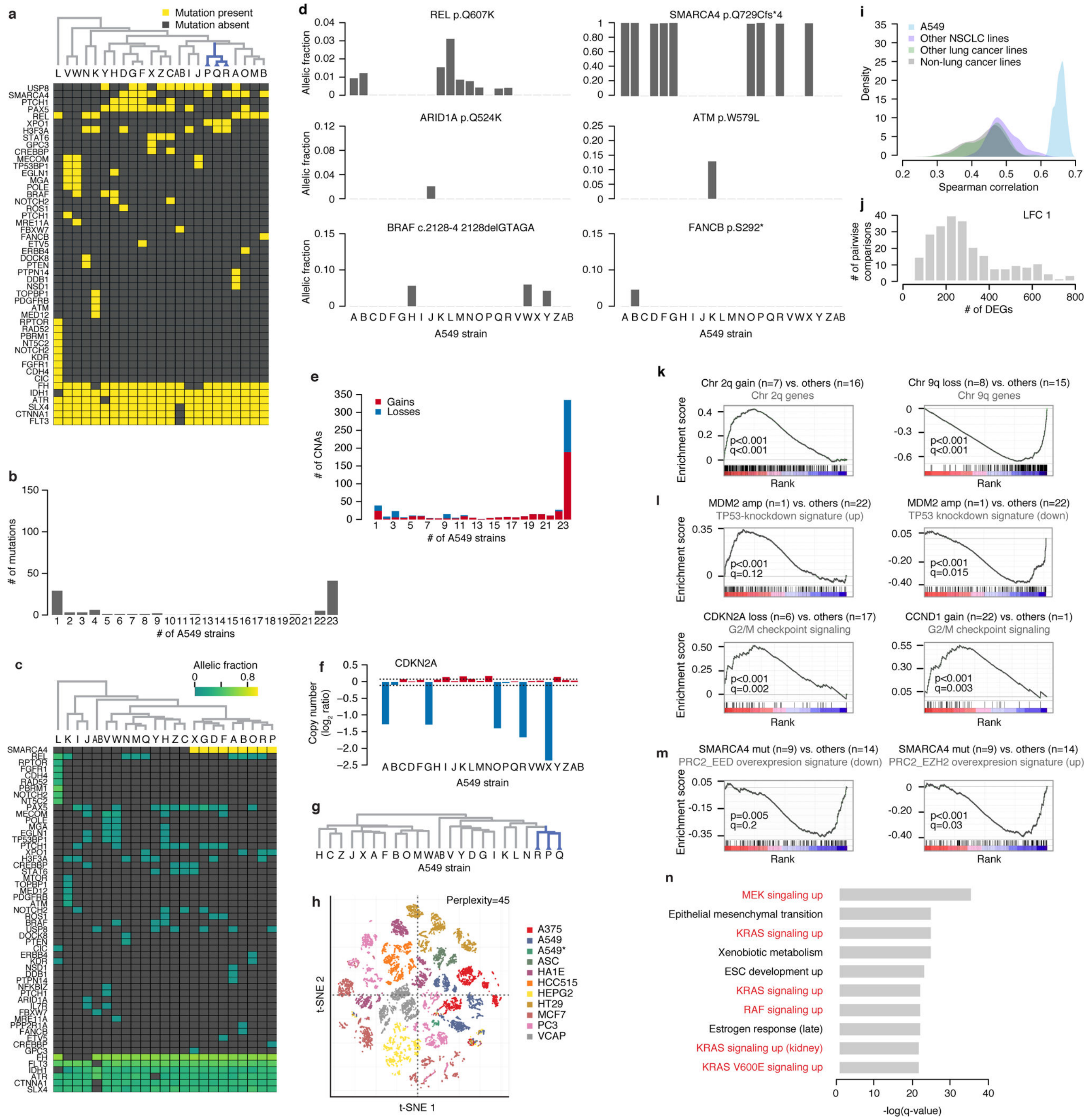
Extended Data Figure 5: Characterization of the variation in allelic fraction and cellular prevalence of SNVs across 27 MCF7 strains and their single cell-derived clones.

(**a**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on the allelic fractions of all their SNVs. Groups of strains expected to cluster together based on their evolutionary history are highlighted, as in Fig. 1. Bottom: a corresponding heatmap, showing the allelic fractions of all mutations across the 27 MCF7 strains. Shown are mutations identified only in a subset of the strains. The presence of a mutation is shown in color according to its allelic fraction. (**b**) The AF of an activating PIK3CA mutation (top) and an inactivating TP53 mutation (bottom) across strains. (**c**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on their SNV cellular prevalence. Groups of strains expected to cluster together based on their evolutionary history are highlighted, as in Fig. 1. Bottom: a corresponding heatmap, showing the cellular prevalence of all mutations across the 27 MCF7 strains. Shown are mutations identified only in a subset of the strains. The presence of a mutation is shown in color according to its cellular prevalence. (**d**) The distribution of the maximal differences in cellular prevalence (CP) of non-silent mutations, across 27 MCF7 strains. The peak at maxΔCP=1 represents SNVs that are clonal in at least one strain but are nearly or completely absent in at least one other strain; the peak at maxΔCP=0 represents SNVs that are detected at similar prevalence across all 27 strains; and the peak at maxΔCP=~0.1 represents a group of SNVs present at CP=~0.1 only in strain M. (**e**) A table of the MCF7 single cell-derived clones included in this study, presenting their parental cell line, genetic manipulations and relationship to one another. (**f**) A heatmap presenting the allelic fractions of non-silent mutations in three WT single cell-derived MCF7 clones and its parental population. The presence of a mutation is shown in color according to its allelic fraction. (**g**) A heatmap presenting the allelic fractions of non-silent mutations in five genetically-manipulated single cell-derived MCF7 clones. For two of the clones, samples were passaged for a prolonged time and sequenced at multiple time points. The presence of a mutation is shown in color according to its allelic fraction. **(h)** Comparison of the karyotypic variation between parental and single cell-derived cell populations. Histograms present the distribution of chromosome numbers from the parental (light gray) and single cell-derived (dark gray) populations. P-values indicate the significance of the difference between the variations (rather than the means) of the populations from a one-tailed Levene's test (n=50 metaphases per group). (**i**) Two representative karyotypes from each sample. Note that all single cell-derived clones are karyotipically heterogeneous. Marker chromosomes are not shown. Arrows point to partially aberrant chromosomes. Images are representative of 50 metaphases counted per sample. (**j**) Two representative karyotypes from two cell populations of the same single cell-derived clone, separated by 6 months of culture propagation. Marker chromosomes are not shown. Arrows point to partially aberrant chromosomes. Images are representative of 50 metaphases counted per sample. (**k**) Comparison of the karyotypic variation between two cell populations of the same single cell-derived clone, separated by 6 months of culture propagation. Histograms present the distribution of chromosome numbers from the early (light gray) and late (dark gray) populations. 50 metaphases were counted per sample. P-value indicates the significance of the difference between the means of the populations from a two-tailed Wilcoxon rank-sum test.



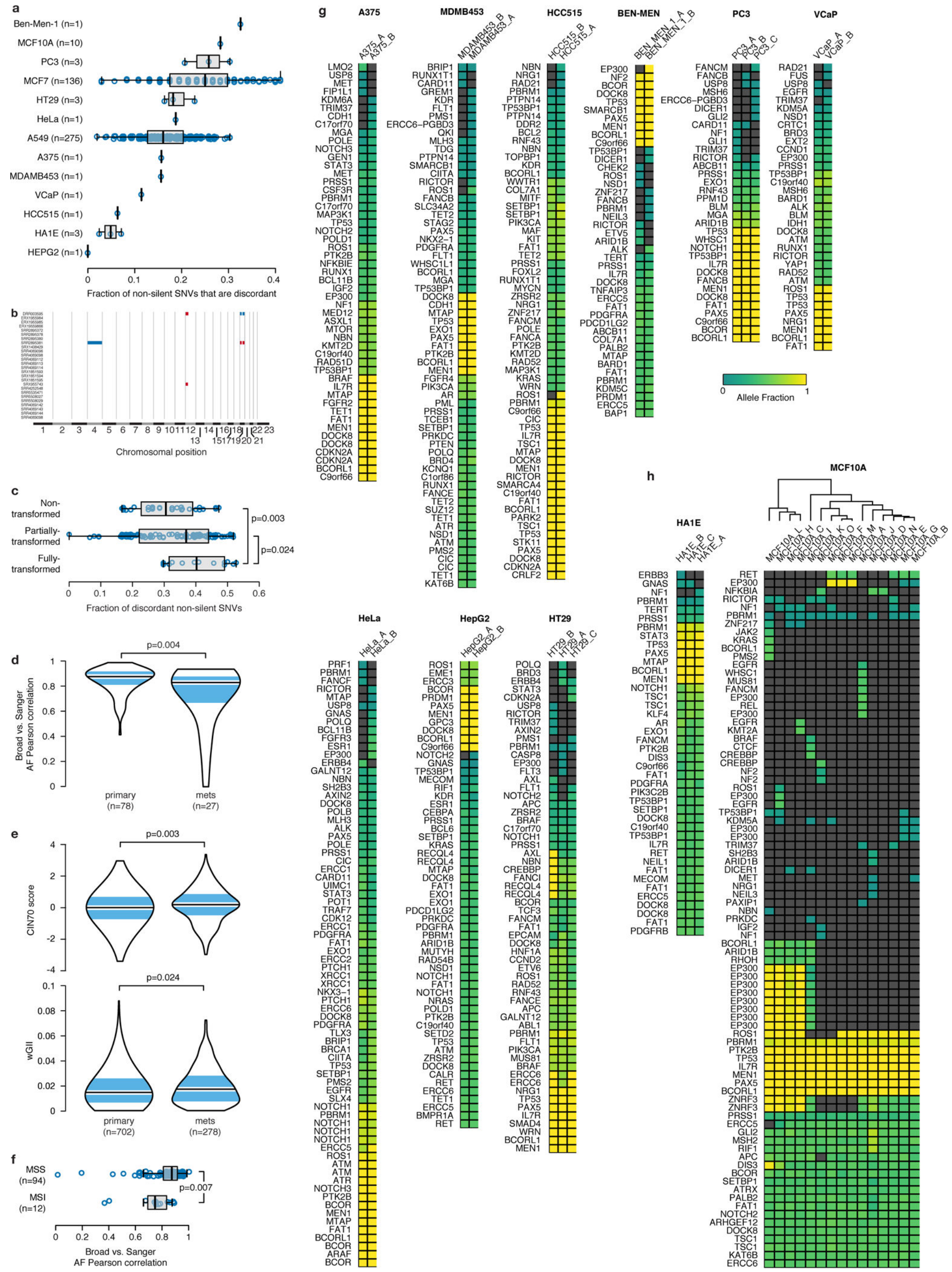
Extended Data Figure 6: Transcriptomic variation across 27 MCF7 strains and their single cell-derived clones.

(**a**) Comparison of the L1000-based MCF7 expression profiles to microarray-based expression profiles from CCLE. Histograms present the distributions of the Spearman correlations between the 27 MCF7 strains and either MCF7 (light purple), two MCF7 derivatives (dark purple and blue), other breast cancer cell lines (green) or non-breast cancer cell lines (gray). The comparison is based on the 978 “landmark” genes directly measured in L1000. (**b**) The number of differentially expressed genes (DEGs) identified in all possible pair-wise comparisons of MCF7 strains, using a two-fold change cutoff. LFC, log fold change; DEGs, differentially expressed genes. (**c**) The 10 top “hallmark” gene sets identified by GSEA to be significantly enriched among the 100 genes that are most differentially expressed across the MCF7 strains. The two gene sets related to estrogen response are highlighted in red. (**d**) Comparison of gene expression variation within and between strains. Histograms present the distributions of gene expression variation within replicates of the same strain (gray), between closely related strains (purple), and between all strains (green). The comparison is based on the 978 “landmark” genes directly measured in L1000. (**e**) Heatmap presenting the arm-level CNA profiles of 27 MCF7 strains. Gains are shown in red, losses in blue. (**f**) GSEA reveals down-regulation of the genes on chromosomes 10q, 17q and 21q in strains that have lost copies of these arms, and up-regulation of the genes on chromosomes 5q, 6p, 14q and 16p in strains that have gained copies of these arms. (**g**) GSEA reveals up-regulation of mTOR signaling (gene set: hallmark_MTORC1_Signaling) and of genes that are up-regulated when *PTEN* is knocked-down (gene set: PTEN_DN.v2_UP; bottom) in strains that have gained *PIK3CA*, down-regulation of the estrogen response signature (gene set: hallmark_Estrogen_Response_Late) in strains that have lost *ESR1*, cell cycle signature (gene set: KEGG_cell_cycle) in strains that have lost *CDKN2A*, and down-regulation of *KRAS* signaling (gene set: hallmark_KRAS_Signaling_DN) in strains that have lost *MAP2K4*. (**h**) GSEA reveals up-regulation of mTOR signaling (gene set: hallmark_MTORC1_Signaling) in strains with high prevalence of an activating *PIK3CA* mutation, up-regulation of genes that are up-regulated when *PTEN* is knocked-down (gene set: PTEN_DN.v1_UP) in strains that harbor an inactivating *PTEN* mutation, and down-regulation of genes that are down-regulated when *TP53* is knocked-down (gene set: P53_DN.v1_DOWN) in strains with high cellular prevalence of an inactivating *TP53* mutation. (**i**) GSEA reveals up-regulation of mTOR signaling (gene sets: MTOR_UP.N4.V1_UP and hallmark_MTORC1_Signaling) in strains that have both *PTEN* copy number loss and an inactivating *PTEN* mutation. **(j)** A tSNE plot of single-cell RNA sequencing (scRNA-seq) data from MCF7-AA cells treated with bortezomib (500nM) at different time points. Each dot represents a single cell, and cells are colored by time point. (**k**) Comparison of the proteasome gene expression signature across time points. (**l**) Comparison of the unfolded protein response gene expression signature across time points. (**m**) Comparison of two proliferation gene expression signatures, S (top) and G2M (bottom), across time points. (**n**) Comparison of the early (top) and late (bottom) response to estrogen gene expression signatures across time points. Red lines denote mean values. P-values indicate significance from a one-way ANOVA followed by a Games-Howell post hoc test. n= 1,726, 2,743, 1,851 and 1,235 cells for t0, t12, t48 and t96, respectively. (**o**) A tSNE plot of scRNA-seq data from a parental population and its single cell-derived clone at two time points. Each dot represents a single cell, and cells are colored by sample. (**p**) Comparison of the transcriptional heterogeneity between a parental MCF7 population and its single cell-derived clones. n=2,904, 2,990, 3,896 and 4,583 cells for parental, WT3, WT4 and WT5, respectively. (**q**) Comparison of the transcriptional heterogeneity between two cultures of the same single-cell clone, separated by 6 months of continuous passaging. n=4,295 and 4,116 cells, for clone9-May17 and clone9-Nov17, respectively. Box plots present the Euclidean distance between the cells in each cell population. Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile. P-values indicate significance from a one-way ANOVA followed by a Games-Howell post hoc test. (**r**) The 10 top “hallmark” gene sets identified by GSEA to be significantly enriched among the top differentially expressed genes (DEGs) between the two cultures of clone MCF7_GREB1_9 (May ‘17 vs. Nov ‘17). The gene sets related to estrogen response are highlighted in red, and those related to proliferation are highlighted in green.



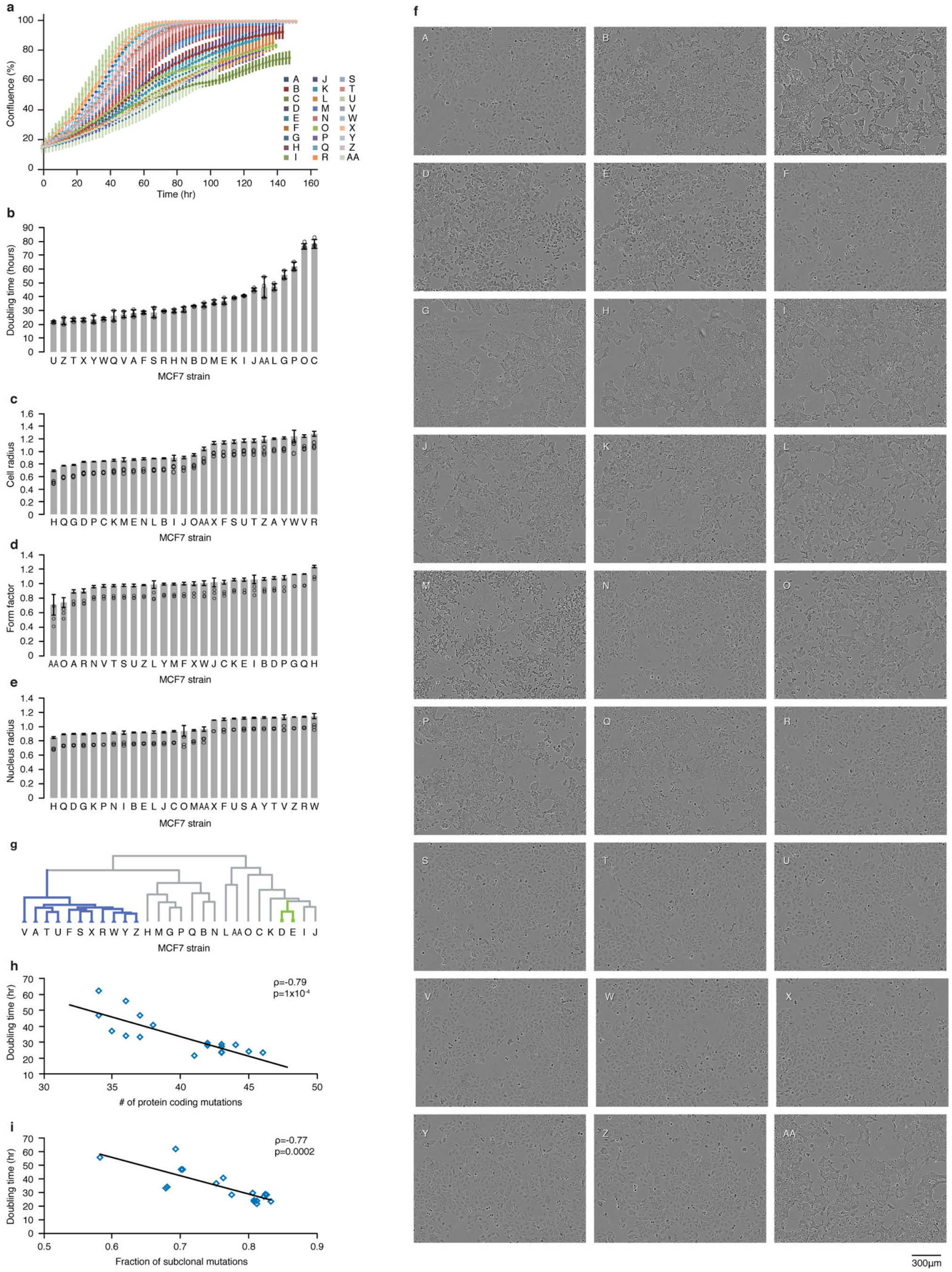
Extended Data Figure 7: Extensive genetic and transcriptional variation across 23 strains of A549.

(**a**) Top: unsupervised hierarchical clustering of 23 A549 strains, based on their non-silent single nucleotide variant (SNV) profiles derived from deep targeted sequencing. Strains expected to cluster together based on their evolutionary history are highlighted in blue. Bottom: a corresponding heatmap, showing the mutation status of non-silent mutations across the 23 A549 strains. Shown are mutations identified only in a subset of the strains, which were detected in above 5% of the reads (allelic fraction>0.05). The presence of a mutation is shown in yellow, and its absence in gray. (**b**) The number of non-silent point-mutations shared by each number of A549 strains. (**c**) Top: unsupervised hierarchical clustering of 23 A549 strains, based on the allelic fractions of their non-silent SNVs. Bottom: a corresponding heatmap, showing the allelic fractions of non-silent mutations across the 23 A549 strains. Shown are mutations identified only in a subset of the strains. The presence of a mutation is shown in color according to its allelic fraction. (**d**) The allelic fractions of non-silent mutations in six selected genes across 23 A549 strains. Note the inactivating frameshift mutation in *SMARCA4*, one of the most frequently mutated genes in lung adenocarcinoma[27,64,65], which was detected at an allelic fraction of ~1 in 9 of the strains, but was not detected at all in the other 14 strains. (**e**) The number of gene-level copy number alterations (CNAs) shared by each number of MCF7 strains. Copy number gains are shown in red, and copy number losses in blue. (**f**) CNA variation in the copy number of *CDKN2A*. Copy number gains are shown in red, and copy number losses in blue. Thresholds for relative gains/losses were set at +/−0.1, respectively. (**g**) Unsupervised hierarchical clustering of 23 A549 strains, based on their global gene expression profiles. Strains expected to cluster together based on their evolutionary history are highlighted in blue. (**h**) A tSNE plot of L1000-based gene expression profiles from multiple samples of nine cancer cell lines. The asterisk denotes the 23 A549 strains profiled in the current study. (**i**) Comparison of the L1000-based A549 expression profiles to microarray-based expression profiles from CCLE. Histograms present the distributions of the Spearman correlations between the 23 A549 strains and either A549 (light blue), other non-small cell lung cancer cell lines (purple), other lung cancer cell lines (green) or non-lung cancer cell lines (gray). The comparison is based on the 978 "landmark" genes directly measured in L1000. (**j**) The number of differentially expressed genes (DEGs) identified in all possible pairwise comparisons of A549 strains, using a two-fold change cutoff. LFC, log fold change; DEGs, differentially expressed genes. (**k**) Arm-level gains are associated with significant up-regulation, and arm-level losses are associated with significant down-regulation, of genes transcribed from the aberrant arms. For example, GSEA reveals up-regulation of the genes on chromosome 2q in strains that have gained a copy of that arm (left), and down-regulation of the genes on chromosome 9q in strains that have lost a copy of that arm (right). (**l**) Gene-level CNAs are associated with significant dysregulation of the perturbed pathways. For example, GSEA reveals up-regulation of the genes that are up-regulated, and down-regulation of the genes that are down-regulated, when *TP53* is knocked-down in strains with *MDM2* high-level copy number gain; and up-regulation or down-regulation of the G2/M cell cycle checkpoint signature in strains with *CDKN2A* copy number loss or *CCND1* copy number gain, respectively. (**m**) Point mutations are associated with significant dysregulation of the perturbed pathways. For example, GSEA reveals down-regulation of two PRC2-related expression signatures in strains with an inactivating *SMARCA4*. (**n**) The 10 top gene sets identified by GSEA to be significantly enriched among the 100 genes that are most differentially expressed across the A549 strains. The six gene sets related to *KRAS* signaling are highlighted in red.



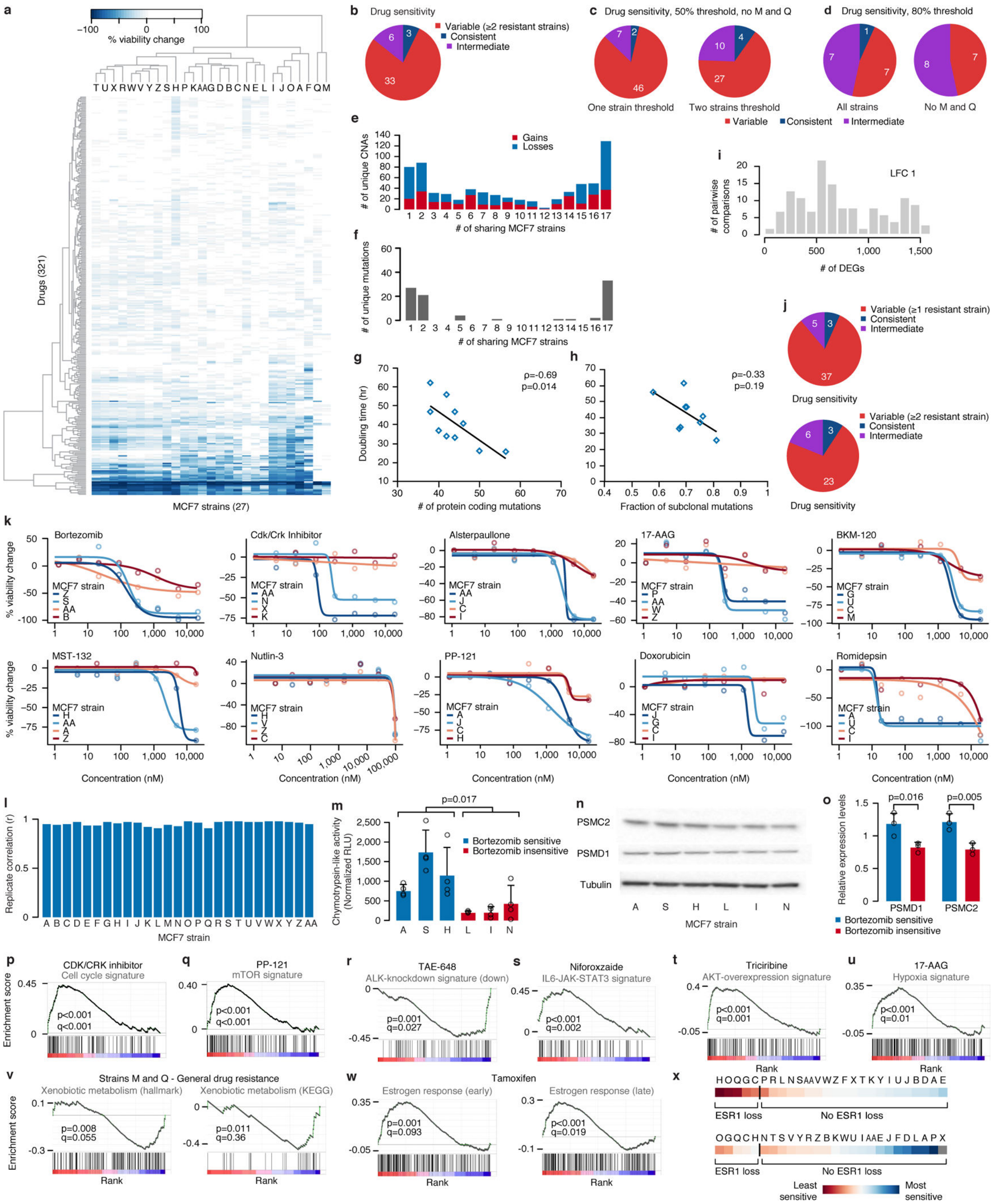
Extended Data Figure 8: Genetic variation across multiple strains of additional cancer and non-cancer cell lines

(**a**) A bar plot presenting the fraction of non-silent SNVs that are discordant between pairs of strains of the same cell line. Bars represent mean ± s.e.m. n, number of strain pairs compared. (**b**) Arm-level CNAs arise in RPE1 samples. Plots present CNAs detected by an e-karyotyping analysis of 26 RPE1 samples. Gains are shown in red, losses in blue. (**c**) Comparison of variability in non-silent SNVs between non-transformed, partially-transformed and fully-transformed MCF10A samples. Box plots present the fraction of discordant non-silent SNVs between pairs of samples within each category. Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile; circles: all data points. one-tailed Wilcoxon rank-sum test, n=28, 112 and 14 strain pairs, for the non-transformed, partially-transformed and the fully-transformed groups, respectively. (**d**) Comparison of the Broad-Sanger allelic fraction correlations of cell lines derived from primary tumors and those derived from metastases. Bar, median; colored rectangle, 25th and 75th percentiles; width of the violin indicates frequency at that value. One-tailed Wilcoxon rank-sum test. (**e**) Top: comparison of the chromosomal instability (CIN70) gene expression signature score between CCLE lines derived from primary tumors and those derived from metastases. Botton: comparison of the weighted-genomic integrity index (wGII) between CCLE lines derived from primary tumors and those derived from metastases. Bar, median; colored rectangle, 25th and 75th percentiles; width of the violin indicates frequency at that value. One-tailed Wilcoxon rank-sum test. (**f**) Comparison of the Broad-Sanger allelic fraction correlations of microsatellite-stable cell lines (MSS) and microsatellite-unstable cell lines (MSI). Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile; circles: all data points. One-tailed Wilcoxon rank-sum test. **(g)** Heatmaps presenting the allelic fractions of non-silent mutations in multiple strains of cancer cell lines. The presence of a mutation is shown in color according to its allelic fraction. (**h**) Heatmaps presenting the allelic fractions of non-silent mutations in multiple strains of the non-cancer cell lines HA1E and MCF10A. The presence of a mutation is shown in color according to its allelic fraction. Also shown is an unsupervised hierarchical clustering of the 15 MCF10A strains, which represent different degrees of cellular transformation, based on their non-silent mutation profiles.

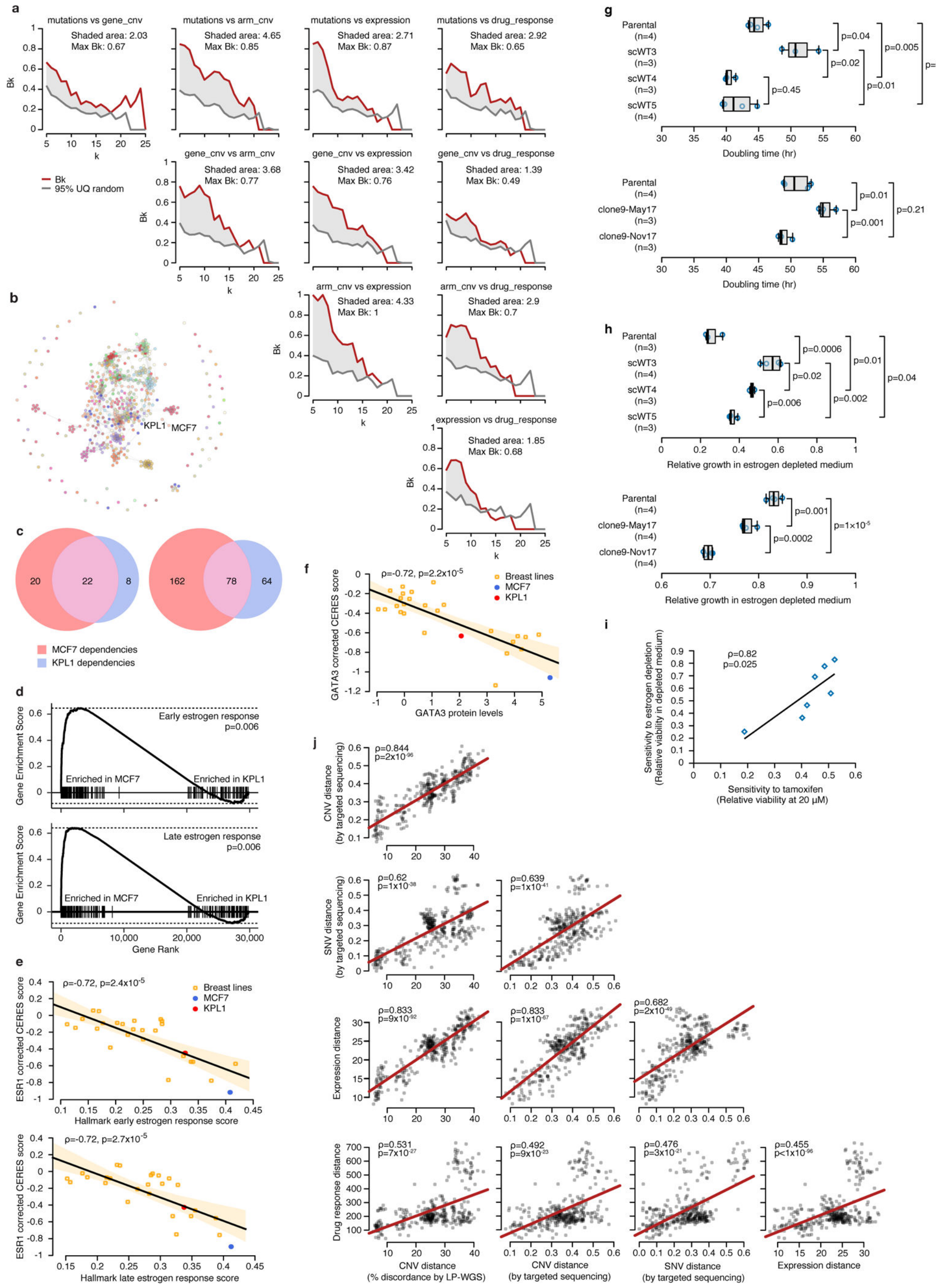


Extended Data Figure 9: Characterization of cell proliferation and morphology across 27 MCF7 strains.

(**a**) Growth response curves of 27 MCF7 strains, based on microscopy imaging. Mean ± s.d., n= 3 replicate wells per data point. (**b**) Mean ± s.d., n= 3 replicate wells per data point. (**c**) Variation in cellular radius across the 27 MCF7 strains. (**d**) Variation in form factor, a measure of circularity, across the 27 MCF7 strains. (**e**) Variation in nuclear radius across the 27 MCF7 strains. Data points represent mean values, and error bars represent standard deviations. (**f**) Microscopy imaging of the 27 MCF7 strains, showing the morphological differences between them. Scale, 300μM. Images are representative of 5 replicate wells per strain. (**g**) Unsupervised hierarchical clustering of 27 MCF7 strains, based on 1,784 morphological features. (**h**) The correlation between proliferation rate (shown as doubling time) and the number of non-silent protein coding mutations, across 18 naturally-occurring MCF7 strains (i.e., strains that have not undergone drug selection or genetic manipulation). Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively. (**i**) The correlation between proliferation rate (shown as doubling time) and the fraction of subclonal mutations, across 18 naturally-occurring MCF7 strains. Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively.



Extended Data Figure 10: Characterization of drug response variation across 27 MCF7 strains. (**a**) Unsupervised hierarchical clustering of 27 MCF7 strains, based on their response to all 321 compounds in the primary screen. Groups of strains expected to cluster together based on their evolutionary history are highlighted, as in Fig. 1. (**b**) Pie chart presenting the classification of the screened compounds based on their differential activity. The response to each active compound was defined as “consistent” if viability change was <−50% for all strains, “variable” if viability change was <−50% for some strains and >−20% for other strains, or “intermediate” if viability change was in between these values. Classification was performed using a two strain threshold. (**c**) Pie charts Pie charts as in (**b**), only excluding strains Q and M that were generally more drug resistant. Classification was performed using a one-strain or a two strain threshold (left and right charts, respectively). (**d**) Pie charts as in (**b**), only using an activity threshold of viability change <−80%. Classification was performed using a one-strain threshold, either including all strains (left) or excluding strains Q and M (right). (**e**) The number of gene-level copy number alterations (CNAs) shared by each number of MCF7 strains. Copy number gains are shown in red, and copy number losses in blue. (**f**) The number of non-silent point-mutations shared by each number of MCF7 strains. The 10 naturally-occurring CMap strains were averaged and considered as a single sample. (**g**) The correlation between proliferation rate (shown as doubling time) and the number of non-silent protein coding mutations, across naturally-occurring MCF7 strains (n=10). Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively. The 10 naturally-occurring CMap strains were averaged and considered as a single sample. (**h**) The correlation between proliferation rate (shown as doubling time) and the fraction of subclonal mutations, across naturally-occurring MCF7 strains (n=10). Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively. The 10 naturally-occurring CMap strains were averaged and considered as a single sample. (**i**) The number of differentially expressed genes (DEGs) identified in all possible pair-wise comparisons of MCF7 strains, using a two-fold change cutoff. LFC, log fold change; DEGs, differentially expressed genes. The 10 naturally-occurring CMap strains were averaged and considered as a single sample. (**j**) Pie charts presenting the classification of the screened compounds based on their differential activity. The response to each active compound was defined as "consistent" if viability change was < −50% for all strains, "variable" if viability change was <−50% for some strains and >−20% for other strains, or "intermediate" if viability change was in between these values. Classification was performed using a one-strain or a two strain resistance threshold (left and right charts, respectively). The 10 naturally-occurring CMap strains were averaged and considered as a single sample. (**k**) Shown are the dose response curves for ten compounds. For each compound, eight concentrations were tested in each strain. Two sensitive strains and two insensitive strains are plotted. Each data point represents the mean of two replicates. Nutlin-3, a compound that had no toxicity against any of the strains in the primary screen, was included as negative control. Romidepsin, a compound that killed all strains very efficiently in the primary screen was included as positive control and turned out to be differentially active at lower concentrations. (**l**) The Pearson's correlation of the two compound screen replicates across the MCF7 strains. (**m**) Strains more sensitive to proteasome inhibitors exhibit higher proteasome activity. The chymotrypsin-like activity of the proteasome was measured in three sensitive and three insensitive strains. Mean ± s.d., one-tailed t-test, n=4 replicate wells. (**n**) Western blots presenting the relative protein expression levels of the proteasome 19S complex members PSMC2 and PSMD1 in three sensitive and three insensitive strains. The expression of α-tubulin was used for normalization. The experiment was repeated once, with n=3 strains per group. For gel source data, see Supplementary Fig. 1. (**o**) Quantification of the relative expression of PSMC2 and PSMD1. Bars represent mean values, and error bars represent standard deviations. Mean ± s.d., one-tailed t-test, n=3 strains per group. (**p**) Up-regulation of the KEGG cell cycle signature in strains sensitive to the cell cycle inhibitor CDK/CRK inhibitor (n= 3), compared to insensitive strains (n=12). (**q**) Up-regulation of mTOR signaling in strains sensitive to the PI3K inhibitor PP-121 (n=11), compared to insensitive strains (n=5). (**r**) Down-regulation of the genes that are down-regulated when *ALK* is knocked-down in strains sensitive to the ALK inhibitor TAE-684 (n=4), compared to insensitive strains (n=15). (**s**) Up-regulation of IL6-JAK-STAT3 signaling in strains sensitive to the STAT inhibitor Nifuroxazide (n=9), compared to insensitive strains (n=6). (**t**) Up-regulation of the genes that are upregulated when *AKT* is over-expressed in strains sensitive to the AKT inhibitor Triciribine (n=2), compared to insensitive strains (n=8). (**u**) Up-regulation of hypoxia signaling in strains sensitive to the HSP inhibitor 17-AAG (n=3), compared to insensitive strains (n=15). (**v**) Down-regulation of xenobiotic metabolism signatures in strains M and Q (n=2), which exhibited an increased resistance to most compounds, compared to the other strains (n=25). (**w**) Up-regulation of the early and late estrogen response signatures, in strains most sensitive to the ER inhibitor tamoxifen (n=5), compared to the least sensitive strains (n=5). (**x**) Sensitivity to estrogen depletion and to tamoxifen is associated with the copy number status of *ESR1*. Heatmaps represent the relative viability in estrogen-depleted medium (top) and to tamoxifen (at 16.6µM; bottom).



Extended Data Figure 11: Comparison of genetic-, transcriptomic- and drug response-based clustering trees, genomic distances and CRISPR dependencies.

(**a**) Comparison of clustering trees using the Fowlkes-Mallows approach. The dendrograms based on SNVs, gene-level CNAs, arm-level CNAs, gene expression profiles and drug response patterns were all compared to each other. The Fowlkes-Mallows index (Bk) was computed for all the potential numbers of clusters (k values) ranging from 5 to 26. The red line represents the observed Bk values, whereas the dashed gray line represents the 95% upper quantile of the randomized distribution. The maximum Bk value represents the degree of similarity between the compared pair of dendrograms. **(b)** Force-directed layout of screened lines using a similarity matrix determined by the probability of cell lines co-clustering in dependency space. Cell lines (nodes) are colored by lineage. (**c**) Top: the overlap of dependencies in KPL1 and MCF7 using corrected CERES scores, with genes showing depletion effects in all cell lines (i.e., pan-essential genes) excluded. The threshold for dependency was set as a CERES score <−0.5. Bottom: overlap in dependency with genes of indeterminate dependency status (CERES scores between −0.4 and −0.6) in either cell line excluded. (**d**) A two-sample Gene Set enrichment analysis (GSEA) of MCF7 and KPL1 against the estrogen response gene sets (n=1 sample per group). Expression of the estrogen signaling pathway is strongly enriched in MCF7. (**e**) The correlation between ESR1 dependency values and the single-sample GSEA enrichment scores of the estrogen response hallmark gene sets (n=27 cell lines). The difference in estrogen response signaling between MCF7 and KPL1 predicts their differing levels of dependency on ESR1. (**f**) The correlation between *GATA3* dependency and GATA3 protein levels (z-scored RPPA values; n=27 cell lines). The difference in GATA3 protein levels between MCF7 and KPL1 predicts their differing levels of dependency on *GATA3*. Spearman's rho values and p-values indicate the strength and significance of the correlations, respectively. (**g**) Top: comparison of proliferation rates between a parental MCF7 population and its single cell-derived clones. Bottom: comparison of proliferation rates between two cultures of the same single-cell clone, separated by 6 months of continuous passaging. Box plots present the population doubling time of each sample. Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile; circles: all data points. Two-tailed t-test; n, replicate wells. (**h**) Top: comparison of the sensitivity to estrogen depletion between a parental MCF7 population and its single cell-derived clones. Bottom: comparison of the sensitivity to estrogen depletion between two cultures of the same single-cell clone, separated by 6 months of continuous passaging. Box plots present the relative growth rate in estrogen-depleted medium. Bar, median; box, 25th and 75th percentiles; whiskers, data within 1.5*IQR of lower or upper quartile; circles: all data points. Two-tailed t-test; n, replicate wells. (**i**) The correlation between sensitivity to tamoxifen (relative viability at 20μM) and the sensitivity to estrogen depletion (relative growth rate), across the parental MCF7 populations and their single-cell clones (n=7). Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively. (**j**) Correlation plots between various measures to estimate cell line strains (n=351 strain pairs). CNA distances (based on low-pass whole-genome sequencing or targeted sequencing), SNV distances, gene expression distances and drug response distances were compared to each other. CNV distance based on LP-WGS was determined by the fraction of the genome affected by discordant CNV calls. CNV and SNV distances based on targeted sequencing were determined by Jaccard indices. Gene expression and drug response distances were determined by Euclidean distances. Spearman's rho value and p-value indicate the strength and significance of the correlation, respectively.

Extended Data Table 1:

Implications of this study for the use of cell lines in cancer research

A summary of the main findings of this study, their practical implications, and recommendations for addressing them.

Findings	Implications	Recommendations
Given two strains, ~20% of mutations would be observed in only one of them	There is ~10% likelihood that a mutation observed in a strain would not appear in a database of cell line genomic features	• Be cautious when using published datasets of genomic features as “lookup tables”
Prolonged passaging introduces more variation than multiple freeze-thaw cycles	For most cell lines, freezing and thawing is likely to be associated with fewer changes than maintaining in culture	• Keep track of passage number • Use passage-matched controls • For large-scale screens, prepare multiple frozen vials for downstream analyses
Various genomic, transcriptomic and phenotypic assays yield highly similar clustering trees	Simple and inexpensive genome-wide assays can serve as a proxy for diversification	• Use inexpensive genome-wide assays (e.g., LP-WGS) and compare to published references using Cell STRAINER: https://cellstrainer.broadinstitute.org • Exclude strains that show extreme diversification
Genetic manipulations that are considered “neutral” can introduce genetic variation	Cell lines with fluorescent reporters, DNA barcodes or Cas9 expression are not identical to their parental cell lines	• Use efficient infection methods to reduce the bottleneck associated with antibiotic selection • Characterize manipulated strains to ensure they retain hallmark genomic features • In CRISPR screens, correct for copy number effects using the copy number landscape of the screened strain
Genetic and transcriptomic variation may affect drug response	Inconsistencies in drug response studies may be attributed to genetic and transcriptomic variability	• Genetic and transcriptomic distances should be considered when comparing drug response data • Compare drug response data to genomic data from the same strain
	Transcriptional differences between sensitive and resistant strains can elucidate compound mechanism of action	• Use characterized isogenic-like strains to uncover associations between molecular features and drug response
Pre-existing heterogeneity within culture underlies cell line instability	Single cell-derived clones differ from one another genetically, transcriptionally and phenotypically	• Confirm the genomic features of single cell-derived clones • Avoid comparisons between bottlenecked cell populations, whenever possible
	Subtle differences in culture conditions can lead to changes in cell line clonal composition	• Keep culture conditions constant
Heterogeneity keeps emerging in culture due to ongoing genomic instability	Prolonged passaging of single cell-derived clones can lead to their diversification Cell lines with deficient maintenance of genome integrity (e.g., MSI or TP53-mutant) are more prone to genomic evolution	• Re-confirm genomic features of single cell-derived clones following prolonged passaging • Apply these recommendations more stringently to genomically unstable cell lines

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Guo Wei, Weiqun Zhang, Chris Mader, Jennifer Roth, David Thomas, Andrew Kung, Desiree Davison, Candace Chouinard, Katherine Walsh, Adrija Navarro, Alice Berger, Douglas Wheeler, Xin Jin, Andrew Hong, Marianna Trakala, Patricia Cho, Johan Kuiken, Romain Boidot, Xiaodong Lu, Franklin Huang, Cory Johannessen, Xaiyun Wu, Stefano Santaguida, Nilay Sethi, Angelika Amon, Kornelia Polyak, Joan Brugge, Dihua

Yu, Juha Klefstrom, William Hahn, Ian Dunn and Yu Mei for contributing cell lines for this study; Matthew Ducar and Samantha Drinan for assistance with the OncoPanel assay; Zach Herbert for assistance with the whole-genome sequencing; John Davis, Sarah Johnson, David Lahr, Josh Gould, Max Macaluso, Xiaodong Lu and Ted Natoli for assistance with the L1000 assay; David Feldman for assistance with the barcoding experiment. James McFarland, Juliann Shih, Coyin Oh, Andrew Cherniack and Paul Clemons for computational assistance; Victor Jones and Jennifer Gale for assistance with drug screening; Kate Hartland for assistance with Cell Painting staining and imaging; Aviv Regev, Orit Rozenblatt-Rosen, Anna Neumann, Daniella Dionne and Lan Nguyen for assistance with 10X single-cell RNAseq; Elizabeth Gonzalez for assistance with cytogenetic analysis; Lee Lichtenstein, David Benjamin, Samuel K. Lee, Valentin Ruano-Rubio and Aaron Chevalier for the GATK CNV pipeline; Zuzana Tothova, Jesse Boehm, Ofir Cohen, Cory Johannessen, Aravind Subramanian, Anne Carpenter, Itay Tirosh, Yehuda Brody, Yotam Zeira and Romanos Pistofidis for discussions. U.B.-D. is supported by a HFSP postdoctoral fellowship. This work was supported by HHMI (T.R.G.), NIH (R01 CA188228; R.B.), Gray Matters Brain Cancer Foundation (R.B.), Bridge Project (P.B. and R.B.), Broad Institue SPARC award (D.F., P.B., and R.B.), and BroadNext10 trainee grant (U.B.-D.).

References

1. Sharma SV, Haber DA & Settleman J Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat Rev Cancer 10, 241–53 (2010). [PubMed: 20300105]
2. Freedman LP, Cockburn IM & Simcoe TS The Economics of Reproducibility in Preclinical Research. PLoS Biol 13, e1002165 (2015). [PubMed: 26057340]
3. Prinz F, Schlange T & Asadullah K Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10, 712 (2011). [PubMed: 21892149]
4. Barretina J et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–7 (2012). [PubMed: 22460905]
5. Garnett MJ et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–5 (2012). [PubMed: 22460902]
6. Basu A et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 154, 1151–61 (2013). [PubMed: 23993102]
7. Yang W et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 41, D955–61 (2013). [PubMed: 23180760]
8. Seashore-Ludlow B et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov 5, 1210–23 (2015). [PubMed: 26482930]
9. Haibe-Kains B et al. Inconsistency in large pharmacogenomic studies. Nature 504, 389–93 (2013). [PubMed: 24284626]
10. Cancer Cell Line Encyclopedia, C. & Genomics of Drug Sensitivity in Cancer, C. Pharmacogenomic agreement between two cancer cell line data sets. Nature 528, 84–7 (2015). [PubMed: 26570998]
11. Haverty PM et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature 533, 333–7 (2016). [PubMed: 27193678]
12. Soule HD, Vazguez J, Long A, Albert S & Brennan M A human cell line from a pleural effusion derived from a breast carcinoma. J Natl Cancer Inst 51, 1409–16 (1973). [PubMed: 4357757]
13. Brooks SC, Locke ER & Soule HD Estrogen receptor in a human cell line (MCF-7) from breast carcinoma. J Biol Chem 248, 6251–3 (1973). [PubMed: 4353636]
14. Lee AV, Oesterreich S & Davidson NE MCF-7 cells--changing the course of breast cancer research and care for 45 years. J Natl Cancer Inst 107(2015).
15. Bamford S et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91, 355–8 (2004). [PubMed: 15188009]
16. Subramanian A et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. bioRxiv (2017).
17. Roth A et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods 11, 396–8 (2014). [PubMed: 24633410]
18. Eirew P et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature 518, 422–6 (2015). [PubMed: 25470049]
19. Bhang HE et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat Med 21, 440–8 (2015). [PubMed: 25849130]

20. Berger AH et al. High-throughput Phenotyping of Lung Cancer Somatic Mutations. Cancer Cell 30, 214–228 (2016). [PubMed: 27478040]

21. Peck D et al. A method for high-throughput gene expression signature analysis. Genome Biol 7, R61 (2006). [PubMed: 16859521]

22. Gupta PB et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell 146, 633–44 (2011). [PubMed: 21854987]

23. Lieber M, Smith B, Szakal A, Nelson-Rees W & Todaro G A continuous tumor-cell line from a human lung carcinoma with properties of type II alveolar epithelial cells. Int J Cancer 17, 62–70 (1976). [PubMed: 175022]

24. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–50 (2014). [PubMed: 25079552]

25. Soule HD et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res 50, 6075–86 (1990). [PubMed: 1975513]

26. Bray MA et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 11, 1757–74 (2016). [PubMed: 27560178]

27. Liberzon A et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425 (2015). [PubMed: 26771021]

28. Janiszewska M et al. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. Nat Genet 47, 1212–9 (2015). [PubMed: 26301495]

29. Venteicher AS et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science 355(2017).

Additional Methods References

30. Bray MA, Fraser AN, Hasaka TP & Carpenter AE Workflow and metrics for image quality control in large-scale high-content screens. J Biomol Screen 17, 266–74 (2012). [PubMed: 21956170]

31. Dao D et al. CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. Bioinformatics 32, 3210–3212 (2016). [PubMed: 27354701]

32. Adalsteinsson VA et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun 8, 1324 (2017). [PubMed: 29109393]

33. Ha G et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res 22, 1995–2007 (2012). [PubMed: 22637570]

34. Sholl LM et al. Institutional implementation of clinical tumor profiling on an unselected cancer population. JCI Insight 1, e87062 (2016). [PubMed: 27882345]

35. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–60 (2009). [PubMed: 19451168]

36. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–303 (2010). [PubMed: 20644199]

37. DePristo MA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–8 (2011). [PubMed: 21478889]

38. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31, 213–9 (2013). [PubMed: 23396013]

39. McLaren W et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–70 (2010). [PubMed: 20562413]

40. Olshen AB, Venkatraman ES, Lucito R & Wigler M Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–72 (2004). [PubMed: 15475419]

41. Abo RP et al. BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. Nucleic Acids Res 43, e19 (2015). [PubMed: 25428359]

42. Sanjana NE, Shalem O & Zhang F Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11, 783–784 (2014). [PubMed: 25075903]

43. Joung J et al. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc 12, 828–863 (2017). [PubMed: 28333914]

44. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–27 (2007). [PubMed: 16632515]

45. Rees MG et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 12, 109–16 (2016). [PubMed: 26656090]

46. Golub TR et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–7 (1999). [PubMed: 10521349]

47. Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]

48. Tirosh I et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–96 (2016). [PubMed: 27124452]

49. Meyers RM et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet 49, 1779–1784 (2017). [PubMed: 29083409]

50. Hu Y Efficient, High-Quality Force-Directed Graph Drawing. The Mathematica Journal 10, 37–71 (2006).

51. Barbie DA et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–12 (2009). [PubMed: 19847166]

52. Fowlkes EB & Mallows CL A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association 78, 553–569 (1983).

53. Ben-David U et al. Patient-derived xenografts undergo mouse-specific tumor evolution. Nat Genet 49, 1567–1575 (2017). [PubMed: 28991255]

54. Carter SL et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 30, 413–21 (2012). [PubMed: 22544022]

55. Zhang S, Yuan Y & Hao D A genomic instability score in discriminating nonequivalent outcomes of BRCA½ mutations and in predicting outcomes of ovarian cancer treated with platinum-based chemotherapy. PLoS One 9, e113169 (2014). [PubMed: 25437005]

56. 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, 15545–50 (2005). [PubMed: 16199517]

57. Carter SL, Eklund AC, Kohane IS, Harris LN & Szallasi Z A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 38, 1043–8 (2006). [PubMed: 16921376]

58. Pujar S et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res 46, D221–D228 (2018). [PubMed: 29126148]

59. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]

60. Li B, Ruotti V, Stewart RM, Thomson JA & Dewey CN RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26, 493–500 (2010). [PubMed: 20022975]

61. Ben-David U, Mayshar Y & Benvenisty N Virtual karyotyping of pluripotent stem cells on the basis of their global gene expression profiles. Nat Protoc 8, 989–97 (2013). [PubMed: 23619890]

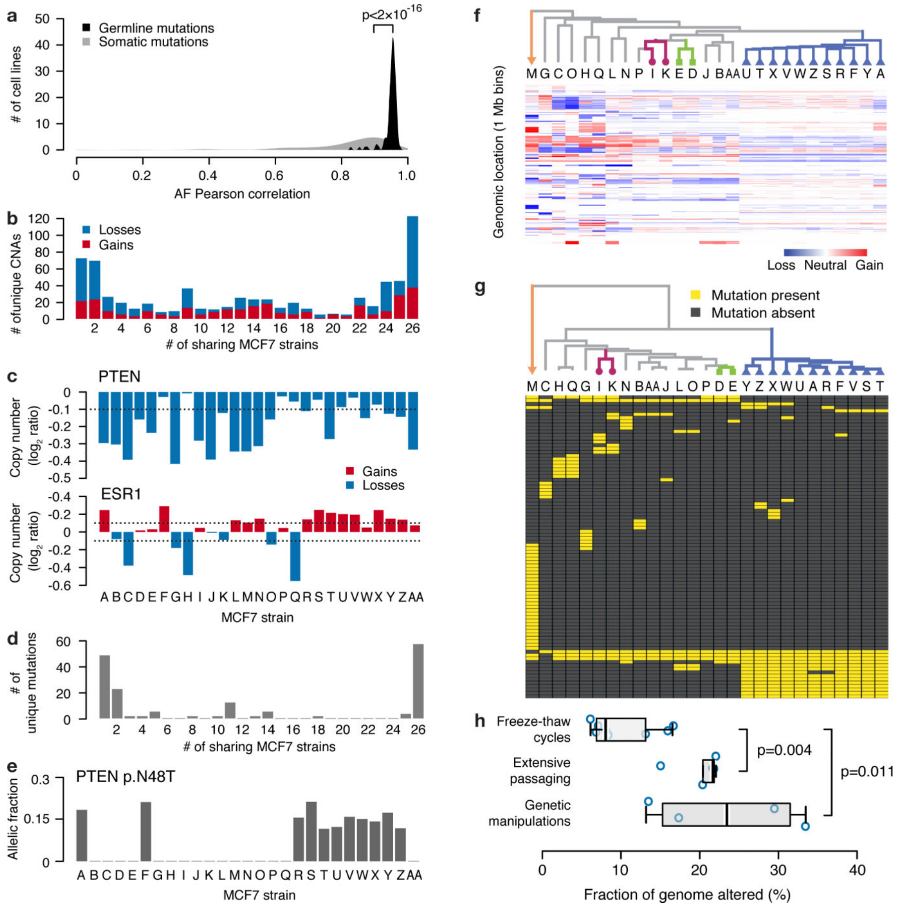


Figure 1: Extensive genetic variation across 27 strains of the cancer cell line MCF7.

(**a**) The distribution of pairwise allelic fraction (AF) correlations between the Broad and the Sanger cell lines (n=106), for germline (black) and somatic (gray) SNVs. One-tailed paired Wilcoxon rank-sum test. (**b**) The number of gene-level copy number alterations (CNAs) shared by each number of MCF7 strains. Gains, red; losses, blue. (**c**) CNAs of two genes, *PTEN* and *ESR1*. (**d**) The number of non-silent point-mutations shared by each number of MCF7 strains. (**e**) The AFs of inactivating mutations in the tumor suppressor *PTEN*. (**f**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on CNA profiles derived from low-pass whole-genome sequencing. Orange, strain M subjected to *in vivo* passaging and drug treatment; blue, 11 Connectivity Map strains cultured in the same lab without extensive passaging; green, strains D and E cultured in the same lab and separated by few passages; purple, strains I and K separated by Cas9 introduction. Bottom: a corresponding heatmap, showing the CNA landscapes of the strains, relative to the median CNA landscape. Gains, red; losses, blue. (**g**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on their non-silent SNV profiles derived from deep targeted sequencing. Colors, as in (**f**). Bottom: a corresponding heatmap, showing the mutation status of non-silent mutations across strains. Shown are mutations identified in a subset of the strains at AF>0.05. Mutation present, yellow; mutation absent, gray. (**h**) Comparison of the magnitude of CNAs observed following multiple freeze-thaw cycles (n=9; R/A/S vs. W/X/Y), extensive passaging (n=5; D vs. L vs. AA, B vs. I/P), and genetic manipulations (n=4; AA vs. O, B vs.

C, I vs. J/K). Bar, median; box, 25th and 75th percentiles; whiskers, 1.5*IQR of lower and upper quartile; circles: data points. Two-tailed Wilcoxon rank-sum test.

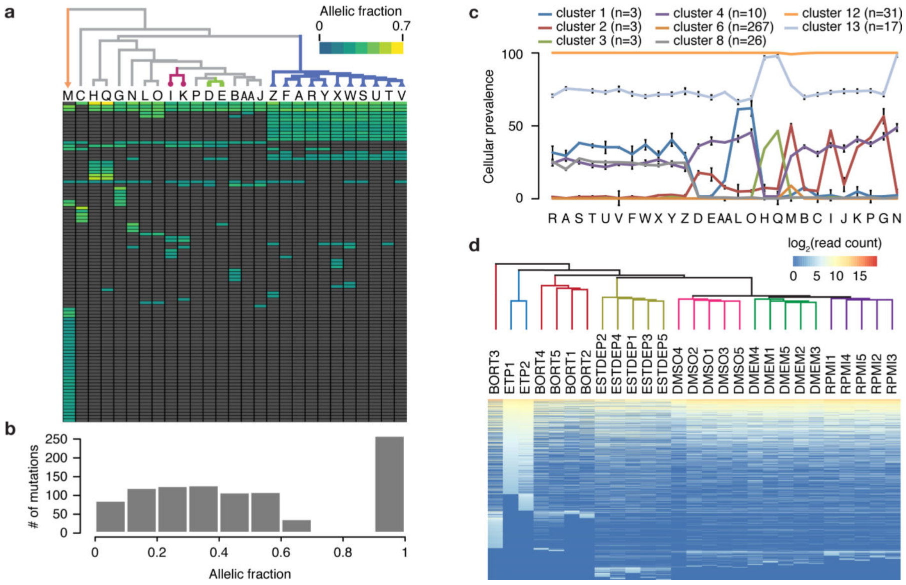


Figure 2: Genetic heterogeneity and clonal dynamics underlying genetic variation.

(**a**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on the allelic fractions of their non-silent SNVs. Colors, as in Fig. 1. Bottom: a corresponding heatmap, showing the AFs of non-silent mutations present in a subset of the strains. (**b**) The distribution of AFs of non-silent mutations across strains. (**c**) The cellular prevalence of mutation clusters across MCF7 strains, as identified by a PyClone analysis. Shown are mutation clusters with differential abundance (ΔCP>0.15), the clonal cluster (cluster #6; CP≈1 in all clones) and a cluster unique to MCF7-M (cluster #12). n = mutations per cluster, mean ± s.e.m. (**d**) Top: unsupervised hierarchical clustering of 27 samples of DNA-barcoded MCF7-D, based on barcode representation. Dendrogram branches are colored by culture condition. Bottom: a corresponding heatmap of barcode representation. ETP, early time point; RPMI, RPMI-1640 medium; DMEM, DMEM medium; DMSO, RPMI-1640 with 0.05% DMSO; ESTDEP, estrogen-depleted RPMI-1640 medium; BORT, bortezomib (500nM; 48hr exposure) followed by RPMI-1640.

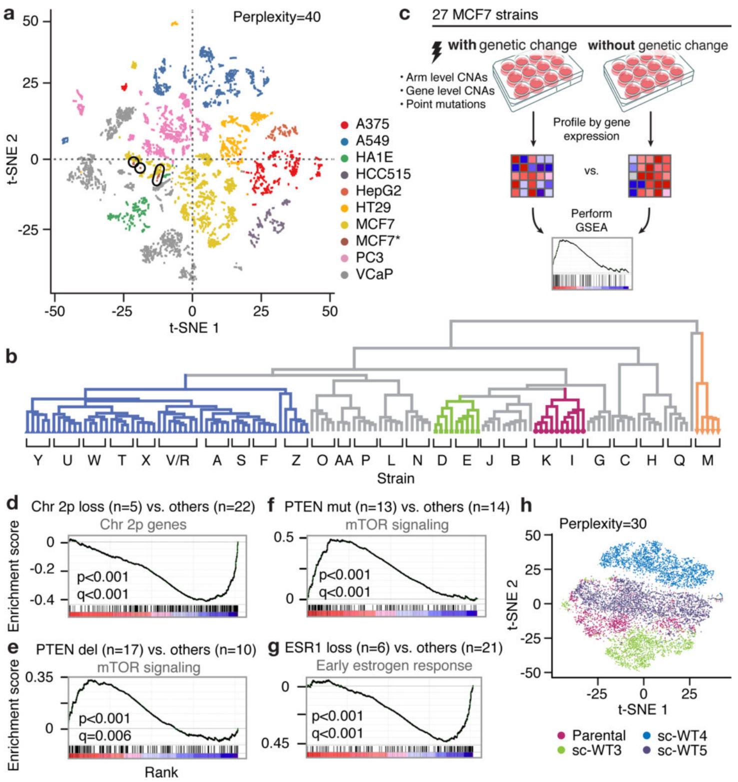


Figure 3: Extensive transcriptomic variation associated with genetic variation.

(**a**) A tSNE plot of gene expression profiles from multiple samples of nine cancer cell lines. *, the 27 MCF7 strains profiled in the current study (also encircled). (**b**) Unsupervised hierarchical clustering of the strains, based on their global gene expression profiles. Colors, as in Fig. 1. (**c**) Schematics presenting the analysis performed to evaluate the association between genetic variation and transcriptional programs. (**d**) Arm-level gains and losses are associated with significant up- and down-regulation of genes transcribed from the aberrant arms. (**e**) Gene-level CNAs are associated with significant dysregulation of the perturbed pathways. For example, up-regulation of mTOR signaling in strains that have lost a copy of *PTEN*. (**f**) Point mutations are associated with significant dysregulation of the perturbed pathways. For example, up-regulation of mTOR signaling in strains with an inactivating *PTEN* mutation. (**g**) Copy number loss of *ESR1* is associated with significant down-regulation of the estrogen response. (**h**) A tSNE plot of single-cell RNA sequencing data from a parental population and three of its single cell-derived clones.

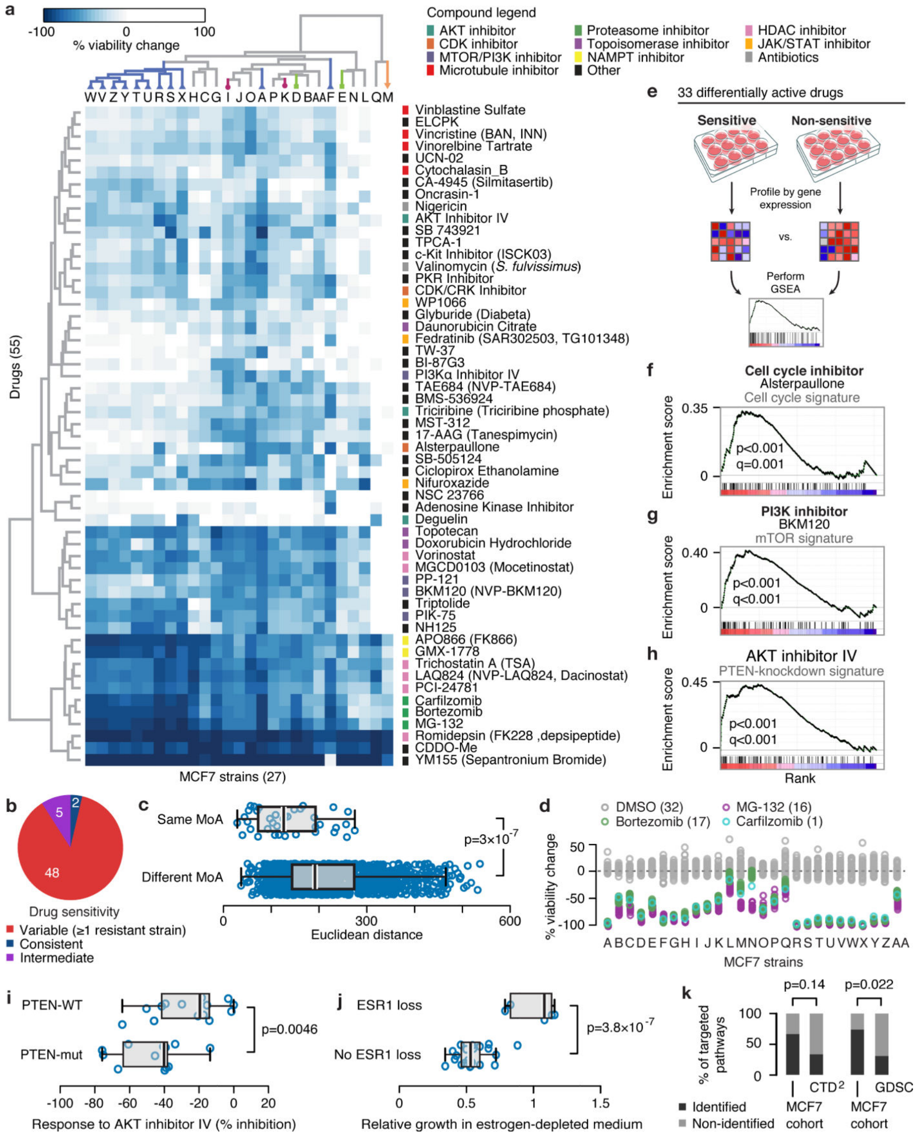


Figure 4: Drug response consequences of genetic and transcriptomic variation.

(**a**) Top: unsupervised hierarchical clustering of 27 MCF7 strains, based on their response to the 55 active compounds in the primary screen. Colors, as in Fig. 1. Bottom: a corresponding heatmap, showing the % of viability change for each compound across strains. Compounds colored based on their mechanism-of-action. (**b**) Classification of the screened compounds based on their differential activity. Consistent, viability change <−50% for all strains; variable, viability change <−50% for some strains and >−20% for other strains; intermediate, viability change in between these values. (**c**) Comparison of the similarity in drug response patterns between compounds that share the same mechanism-of-action (n=39) and compounds that work through different mechanisms (n=1,439). One-tailed Wilcoxon rank-sum test. (**d**) Highly similar differential drug response patterns for three proteasome inhibitors: bortezomib, MG-132 and carfilzomib. Each data point represents the mean of two replicates. The number of data points per strain is mentioned within parentheses. The response pattern with no drug (DMSO control) is presented for comparison. (**e**) Schematics presenting the analysis performed to evaluate the association between drug response and transcriptional variation. (**f**) Up-regulation of the KEGG cell cycle signature in strains sensitive to the cell cycle inhibitor alsterpaullone (8 sensitive vs. 15 resistant strains). (**g**) Up-regulation of mTOR signaling in strains sensitive to the PI3K inhibitor BKM-120 (8 sensitive vs. 5 resistant strains). (**h**) Up-regulation of the genes that are up-regulated when *PTEN* is knocked-down in strains sensitive to AKT inhibitor IV (6 sensitive vs. 9 resistant strains). (**i**) Strains with *PTEN* mutation (n=12) respond more strongly to AKT inhibitor IV than strains without the mutation (n=14). (**j**) Strains with *ESR1* copy number loss (n=5) grow better in estrogen-depleted medium than strains without *ESR1* loss (n=21). (**k**) Comparison of GSEA-based MoA identification between the MCF7 cohort and the CTD[2] (n=15) and GDSC (n=19) cohorts, across matched drugs. Two-tailed Fisher's exact test. For all box plots: bar, median; box, 25th and 75th percentiles; whiskers, 1.5*IQR of lower and upper quartile; circles: data points.