

Genetic and transcriptional evolution alters cancer cell line drug response

Uri Ben-David¹, Benjamin Siranosian¹, Gavin Ha^{1,2}, Helen Tang¹, Yaara Oren^{1,3}, Kunihiko Hinohara^{1,2}, Craig A. Strathdee¹, Joshua Dempster¹, Nicholas J. Lyons¹, Robert Burns², Anwasha Nag², Guillaume Kugener¹, Beth Cimini¹, Peter Tsvetkov¹, Yosef E. Maruvka¹, Ryan O'Rourke^{1,2}, Anthony Garrity¹, Andrew A. Tubelli¹, Pratiti Bandopadhyay^{1,2,3}, Aviad Tsherniak¹, Francisca Vazquez¹, Bang Wong¹, Chet Birger¹, Mahmoud Ghandi¹, Aaron R. Thorner², Joshua A. Bittker¹, Matthew Meyerson^{1,2,3}, Gad Getz^{1,4}, Rameen Beroukhi^{1,2,3,5,7*} & Todd R. Golub^{1,2,3,6,7*}

Human cancer cell lines are the workhorse of cancer research. Although cell lines are known to evolve in culture, the extent of the resultant genetic and transcriptional heterogeneity and its functional consequences remain understudied. Here we use genomic analyses of 106 human cell lines grown in two laboratories to show extensive clonal diversity. Further 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. Notably, genetic changes were associated with differential activation of gene expression programs and marked differences in cell morphology and proliferation. Barcoding experiments 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 continuous instability quickly translates into heterogeneity of the cell line. When the 27 MCF7 strains were tested against 321 anti-cancer compounds, we uncovered considerably different drug responses: at least 75% of compounds that strongly inhibited some strains were completely inactive in others. This study documents the extent, origins and consequences 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 medicine¹. An implicit assumption has been that cell lines are clonal and genetically stable, and therefore that results obtained in one study can be readily extended to another. However, 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, although pharmacogenomic profiling of large collections of cancer cell lines have proven to be mostly 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 reanalysed 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 the allelic fraction of germline variants were nearly identical across the two datasets (median $r = 0.95$), indicating that sequencing artefacts do not substantially contribute to the erroneous appearance of low allelic fraction calls. However, the degree of agreement in allelic fraction for somatic variants was substantially lower (median $r = 0.86$; $P < 2 \times 10^{-16}$; Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). Moreover, a median of 19% of the detected non-silent mutations (range, 10–90%) were identified in only one of the two datasets (Extended Data Fig. 1b). Similarly, 26% of genes that had

copy number alterations (CNAs; which are also known as copy number variants) (range, 7–99%) were discordant (Extended Data Fig. 1c–e). These results indicate that genetic variability across cultures 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 changes in subclonal composition 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 oestrogen receptor (ER)-positive breast cancer cell line MCF7^{12–14} (Methods, Extended Data Figs. 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 (for example, 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-oestrogen 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 number gains and 405 genes with copy number losses (compared to basal ploidy) in at least one strain. Only a small minority of these changes (13% of gains and 21% of losses) were detected in all strains. Of these changes, 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

¹Broad Institute of Harvard and MIT, Cambridge, MA, USA. ²Dana-Farber Cancer Institute, Boston, MA, USA. ³Harvard Medical School, Boston, MA, USA. ⁴Massachusetts General Hospital, Boston, MA, USA. ⁵Brigham and Women's Hospital, Boston, MA, USA. ⁶Howard Hughes Medical Institute, Chevy Chase, MD, USA. ⁷These authors jointly supervised this work: Rameen Beroukhi, Todd R. Golub. *e-mail: rameen_beroukhi@dfci.harvard.edu; golub@broadinstitute.org

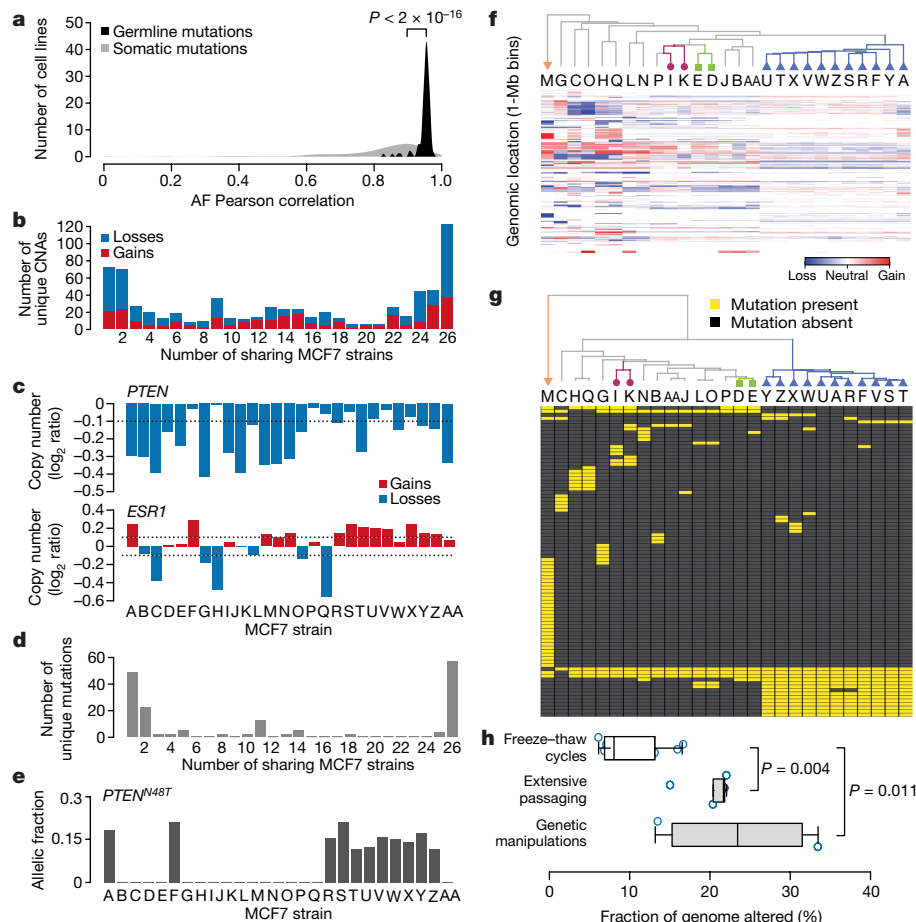


Fig. 1 | Extensive genetic variation across 27 strains of the cancer cell line MCF7. **a**, The distribution of pairwise allelic fraction (AF) correlations between the Broad and the Sanger cell lines ($n = 106$), for germline (black) and somatic (grey) SNVs. One-tailed paired Wilcoxon rank-sum test. **b**, The number of gene-level CNAs shared by each number of MCF7 strains. Red, gains; blue, losses. **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 allelic fraction of inactivating mutations in the tumour 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 laboratory without extensive passaging; green, strains D and E cultured in the same laboratory and separated by few

passages; purple, strains I and K separated by Cas9 introduction. Bottom, corresponding heat map of the CNA landscapes of the strains relative to the median CNA landscape. Red, gains; blue, losses. **g**, Top, unsupervised hierarchical clustering of 27 MCF7 strains, based on their non-silent SNV profiles derived from deep targeted sequencing. Colours as in **f**. Bottom, corresponding heat map of the mutation status of non-silent mutations across strains. Mutations that were identified in a subset of the strains at AF > 0.05 are shown. Yellow, mutation present; grey, mutation absent. **h**, Comparison of the magnitude of CNAs observed following multiple freeze-thaw cycles ($n = 9$; R, A and S versus W, X and Y), extensive passaging ($n = 5$; D versus L versus AA, B versus I and P), and genetic manipulations ($n = 4$; AA versus O, B versus C, I versus J and K). Bar, median; box, 25th and 75th percentiles; whiskers, $1.5 \times$ IQR of lower and upper quartile; circles, data points. Two-tailed Wilcoxon rank-sum test.

Supplementary Table 4). The differential events included genes commonly gained or lost in breast cancer (for example, *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 oestrogen receptor gene *ESR1* was gained in 12 strains, lost in 6 and unaltered in 9 (Fig. 1c), and this correlated with differential expression of ER α ($P = 0.009$; Extended Data Fig. 3b, c and Supplementary Discussion).

Genetic variation was similarly observed at the level of point mutations, small insertions or deletions (indels) and chromosomal translocations. Only 35% of 95 non-synonymous single nucleotide variants (SNVs) and indels that affected the coding sequence or splice regions 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, Supplementary Tables 5, 6 and Supplementary Discussion). Similar, albeit lower, variability was observed among mutations listed as recurrent in the COSMIC database¹⁵, consistent with COSMIC mutations tending to be clonal mutations of the founding populations (Extended Data Fig. 3f).

Unsupervised hierarchical clustering analysis, in which genetic distance was reflected by the branch lengths of the dendrogram, generated a branch structure that accurately reflected the history of the strains. 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 project¹⁶ over a 10-year period clustered tightly together; and sibling strains D and E, which were only 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 allelic fractions revealed extensive subclonality across strains (Fig. 2a, b and Extended Data Fig. 5a). For example, all 27 strains had a *PIK3CA*-activating mutation (G1633A), but the allelic fraction varied from 0.21 to 0.70 (Extended Data Fig. 5b). On the basis of allelic fractions and copy number status, 45% of all observed mutations were determined to be subclonal ($P < 0.01$ in a

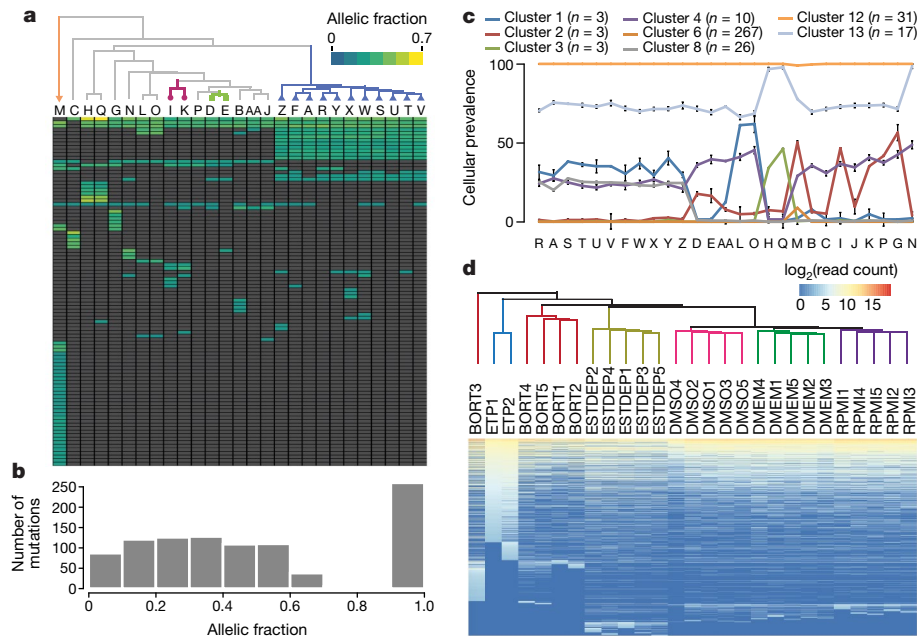


Fig. 2 | Genetic heterogeneity and clonal dynamics underlying genetic variation. **a**, Top, unsupervised hierarchical clustering of 27 MCF7 strains based on the allelic fractions of their non-silent SNVs. Colours as in Fig. 1. Bottom, corresponding heat map of the allelic fractions of non-silent mutations present in a subset of the strains. **b**, The distribution of allelic fractions of non-silent mutations across strains. **c**, The cellular prevalence of mutation clusters across MCF7 strains identified by a PyClone analysis. Mutation clusters with differential abundance (a difference in cellular prevalence (Δ CP) $>$ 0.15), the clonal cluster (cluster 6; CP \approx 1

in all clones) and a cluster unique to MCF7-M (cluster 12) are shown. n = mutations per cluster, data are mean \pm s.e.m. **d**, Top, unsupervised hierarchical clustering of 27 samples of DNA-barcoded MCF7-D based on barcode representation. Dendrogram branches are coloured by culture condition. Bottom, corresponding heat map of barcode representation. ETP, early time point; RPMI, RPMI 1640 medium; DMEM, DMEM medium; DMSO, RPMI 1640 with 0.05% DMSO; ESTDEP, oestrogen-depleted RPMI 1640 medium; BORT, bortezomib (500 nM; 48 h exposure) followed by RPMI 1640.

binomial test). PyClone^{17,18}, which reconstructs subclonal structure by clustering mutations with similar cellular prevalence, found multiple subclones within each MCF7 strain, with varying abundance across strains (Fig. 2c). Indeed, for 43% of the non-silent SNVs, cellular prevalence differed by $>$ 50% across strains (Extended Data Fig. 5c, d and Supplementary Table 7).

We next investigated 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. By contrast, if pre-existing subclones were selected under different conditions, enrichment of the same barcodes would be observed in replicate cultures¹⁹. Unsupervised hierarchical clustering by barcode representation revealed that biological replicates clustered together (Fig. 2d and Supplementary Table 8), indicating that pre-existing subclones are indeed selected by changes in culture conditions.

Next, we characterized the genetic stability of three wild-type 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 wild-type 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, 8–16%) 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: 654 genes (median; range, 10–1,574) were differentially expressed by at least twofold 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). Notably, 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 (Figs. 1f, g, 3b–g, Extended Data Fig. 6e–i and Supplementary Table 14). For example, strains with inactivating *PTEN* mutations or activating *PIK3CA* mutations had decreased *PTEN* and increased mTOR gene expression signatures, respectively (Fig. 3e, f and Extended Data Fig. 6g–i). Similarly, copy number loss of *ESR1* was associated with reduced oestrogen signalling (Fig. 3g and Extended Data Fig. 6g).

We further explored gene expression heterogeneity using single-cell RNA sequencing 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). Notably, the extent of expression heterogeneity among the single-cell-derived clones was not substantially lower than the heterogeneity of 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 indicate that variation in gene expression arises de novo, in addition to reflecting selection of pre-existing subclones²².

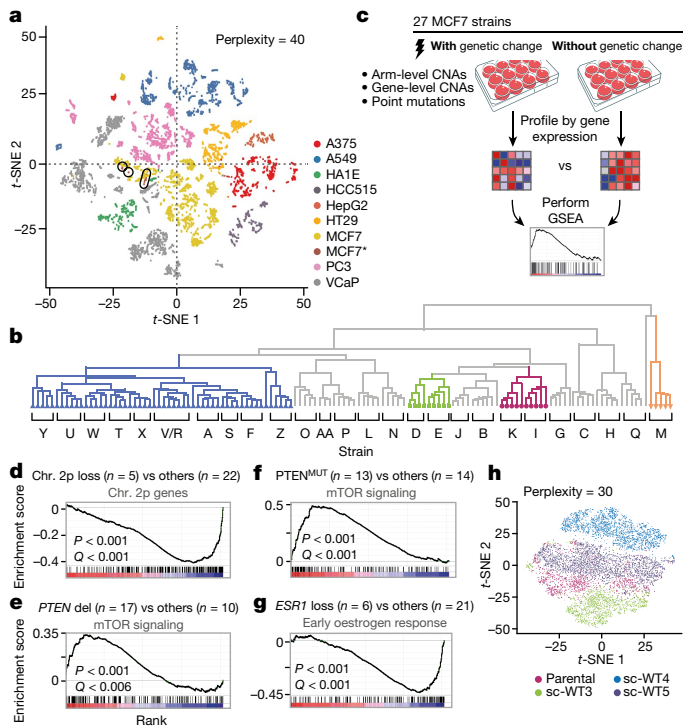


Fig. 3 | Extensive transcriptomic variation associated with genetic variation. **a**, A *t*-distributed stochastic neighbour embedding (*t*-SNE) plot of gene expression profiles from multiple samples of nine cancer cell lines. The 27 MCF7 strains profiled in the current study are indicated by an asterisk in the key and are encircled in the plot. **b**, Unsupervised hierarchical clustering of the strains, based on their global gene expression profiles. Colours as in Fig. 1. **c**, Schematics of 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 downregulation of genes transcribed from the aberrant arms. **e**, Gene-level CNAs are associated with significant dysregulation of the perturbed pathways. For example, upregulation of mTOR signalling was found in strains that had lost a copy of *PTEN*. **f**, Point mutations are associated with significant dysregulation of the perturbed pathways. For example, upregulation of mTOR signalling was found in strains with an inactivating *PTEN* mutation. **g**, Copy number loss of *ESR1* is associated with significant downregulation of the oestrogen response. **h**, A *t*-SNE plot of single-cell RNA-seq data from a parental population and three of its single-cell-derived clones, scWT3–5, single-cell wild-type clone 3–5.

Verification in additional cell lines

To exclude the possibility that the variation that we 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 A549²³ (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 adenocarcinomas²⁴, was detected in 5 strains, but a normal copy number was retained in the other 18 (Extended Data Fig. 7f). Whereas transcriptome analyses showed that oestrogen signalling was the most variable pathway in MCF7 cells (Extended Data Fig. 6c and Supplementary Table 13), *KRAS* signalling 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 (Extended Data Fig. 8 and Supplementary Tables 21–24). Notably, genomic instability was not limited to transformed cancer cell lines (Supplementary Discussion). For example, the variation across 15 strains of MCF10A²⁵, a non-transformed human mammary cell line, was as high as the variation that we found in MCF7 cancer cells (median discordance, 26%; range, 17–40%; 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 imaging²⁶ (Methods). MCF7 strains varied in doubling times by as much as 3.5-fold (median, 31 h; range, 22–78 h; 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 was similar to clustering 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 effect on drug response. We measured cell viability following treatment with 321 drugs at a single concentration (5 μ M) across the 27 MCF7 strains (Supplementary Table 26). Of these, 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 out of these 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–d, j and Supplementary Discussion). All 33 differentially active compounds were validated in an eight-point dose–response analysis of each of the 27 strains (median Spearman's $\rho = 0.42$ between screens, $P = 3 \times 10^{-9}$; 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 $r = 0.97$, $P < 2 \times 10^{-16}$; Extended Data Fig. 10l). Second, compounds with the same mechanism of action had similar patterns of activity across strains (Fig. 4a, c; $P = 3 \times 10^{-7}$). 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 the mechanism of action²⁷ of the compounds between sensitive and insensitive strains ($P = 2 \times 10^{-5}$; 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). Notably, the strains that were the most resistant to treatment in general (strains M and Q) showed downregulation of drug metabolism pathways (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 (Figs. 1c, e, 3e, f), and increased sensitivity to the AKT inhibitor IV (Fig. 4h, i). Similarly, *ESR1* loss was associated with reduced oestrogen signalling (Figs. 1c, 3g), which was in turn associated with reduced sensitivity to tamoxifen or oestrogen 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 genomics or gene expression (Figs. 1g, 2a, 3b, 4a, Extended Data Fig. 11a and Supplementary Discussion). Genome-wide CRISPR screens showed that genetic dependencies were affected by genomic variation similar to pharmacological dependencies (Extended Data Fig. 11b–f, Supplementary Table 31 and Supplementary Discussion), and functional analyses revealed that single-cell-derived clones remained

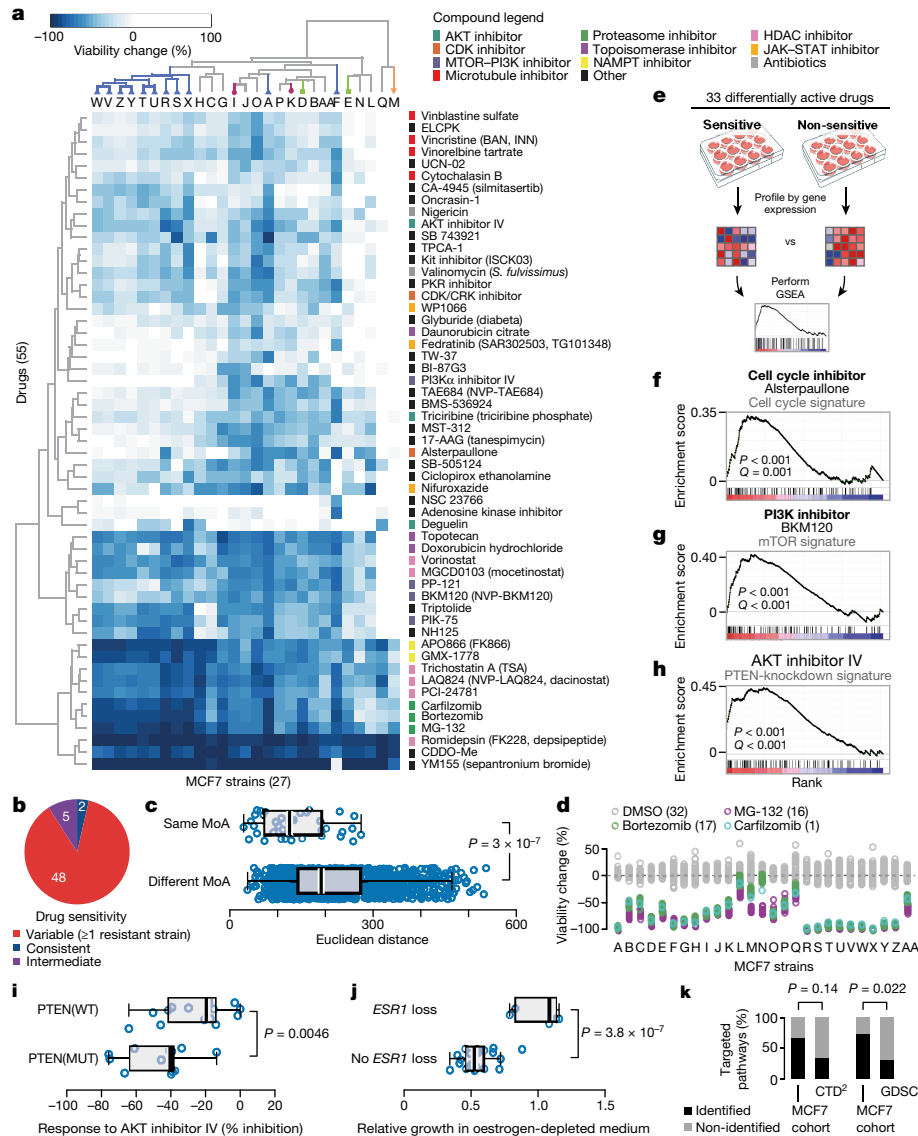


Fig. 4 | Drug-response consequences of genetic and transcriptomic variation. **a**, Top, unsupervised hierarchical clustering of 27 MCF7 strains based on their response to the 55 active compounds in the primary screen. Colours as in Fig. 1. Bottom, corresponding heatmap of the percentage of viability change for each compound across strains. Compounds are coloured 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 in parentheses. The response pattern with no drug (DMSO control)

is presented for comparison. **e**, Schematics of the analysis performed to evaluate the association between drug response and transcriptional variation. **f**, Upregulation of the KEGG cell cycle signature in strains sensitive to the cell cycle inhibitor alsterpaullone (8 sensitive and 15 resistant strains). **g**, Upregulation of mTOR signalling in strains sensitive to the PI3K inhibitor BKM-120 (8 sensitive and 5 resistant strains). **h**, Upregulation of the genes that are upregulated when *PTEN* is knocked down in strains sensitive to AKT inhibitor IV (6 sensitive and 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 oestrogen-depleted medium than strains without *ESR1* loss ($n = 21$). **k**, Comparison of gene set enrichment analysis-based MoA identification between the MCF7 cohort and the CTD² ($n = 15$) 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, $1.5 \times$ the interquartile range of the lower and upper quartile; circles, data points.

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 the 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. This heterogeneity results both from clonal dynamics (that is, changes in the abundance of pre-existing subclones) and from continuous instability (that is, the appearance of new genetic variants). Moreover, genetic heterogeneity leads to varying patterns of gene expression, which in turn result in differential drug sensitivity. These findings have a number of important implications, which are summarized in Extended Data Table 1.

We found that changes in clonal composition underlie much of the observed variability in cell line behaviour. Such clonal composition changes follow selection by particular conditions (for example, growth medium) 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 their strain's genetic distance from a reference.

Variation within cancer cell lines can also be useful in at least two ways. First, deeper characterization (for example, 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 resistance¹⁹. Second, owing to their matched genetic background, naturally occurring 'isogenic-like' strains could help to 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 their genomic evolution leads to a high degree of variation across cell line strains, which must be considered in experimental design and data interpretation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0409-3>.

Received: 29 September 2017; Accepted: 22 June 2018;

Published online 8 August 2018.

- Sharma, S. V., Haber, D. A. & Settleman, J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer* **10**, 241–253 (2010).
- Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165 (2015).
- 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).
- Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
- Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
- 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–D961 (2013).
- Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
- Haibe-Kains, B. et al. Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
- The Cancer Cell Line Encyclopedia & Genomics of Drug Sensitivity in Cancer Investigators. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
- Haverty, P. M. et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
- Soule, H. D., Vazquez, 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**, 1409–1416 (1973).
- Brooks, S. C., Locke, E. R. & Soule, H. D. Estrogen receptor in a human cell line (MCF-7) from breast carcinoma. *J. Biol. Chem.* **248**, 6251–6253 (1973).
- Lee, A. V., Oesterreich, S. & Davidson, N. E. MCF-7 cells—changing the course of breast cancer research and care for 45 years. *J. Natl Cancer Inst.* **107**, djv073 (2015).
- Bamford, S. et al. The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
- Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
- Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
- Eirew, P. et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
- Bhang, H. E. et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* **21**, 440–448 (2015).
- Berger, A. H. et al. High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* **30**, 214–228 (2016).
- Peck, D. et al. A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
- Gupta, P. B. et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633–644 (2011).
- 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).
- The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- Soule, H. D. et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer Res.* **50**, 6075–6086 (1990).
- Bray, M. A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
- Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
- 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**, 1212–1219 (2015).
- Venteicher, A. S. et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).

Acknowledgements We thank G. Wei, W. Zhang, C. Mader, J. Roth, D. Thomas, A. Kung, D. Davison, C. Chouinard, K. Walsh, A. Navarro, A. Berger, D. Wheeler, X. Jin, A. Hong, M. Trakala, P. Cho, J. Kuiken, R. Boidot, X. Lu, F. Huang, C. Johannessen, X. Wu, S. Santaguida, N. Sethi, A. Amon, K. Polyak, J. Brugge, D. Yu, J. Klefstrom, W. Hahn, I. Dunn and Y. Mei for contributing cell lines for this study; M. Ducar and S. Drinan for assistance with the OncoPanel assay; Z. Herbert for assistance with the whole-genome sequencing; J. Davis, S. Johnson, D. Lahr, J. Gould, M. Macaluso, X. Lu and T. Natoli for assistance with the L1000 assay; D. Feldman for assistance with the barcoding experiment; J. McFarland, J. Shih, C. Oh, A. Cherniack and P. Clemons for computational assistance; V. Jones and J. Gale for assistance with drug screening; K. Hartland for assistance with Cell Painting staining and imaging; A. Regev, O. Rozenblatt-Rosen, A. Neumann, D. Dionne and L. Nguyen for assistance with 10X single-cell RNA sequencing; E. Gonzalez for assistance with cytogenetic analyses; L. Lichtenstein, D. Benjamin, S. K. Lee, V. Ruano-Rubio and A. Chevalier for the GATK CNA pipeline; Z. Tothova, J. Boehm, O. Cohen, C. Johannessen, A. Subramanian, A. Carpenter, I. Tirosh, Y. Brody, Y. Zeira and R. 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.Be.), Gray Matters Brain Cancer Foundation (R.Be.), Bridge Project (P.B. and R.Be.), Broad Institute SPARC award (P.B. and R.Be.) and BroadNext10 trainee grant (U.B.-D.).

Reviewer information Nature thanks J. Yang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author 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.Bu., 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 RNA-sequencing 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 data; Y.E.M. assisted with compiling the RPE1 RNA-sequencing 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.Be. and T.R.G. directed the project. U.B.-D., G.H., M.M., R.Be. and T.R.G. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0409-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0409-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.Be. or T.R.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture. The MCF7, HT29, MDAM453 and A375 cell lines were cultured in RPMI 1640 (Life Technologies), with 10% fetal bovine serum (Sigma-Aldrich) and 1% penicillin–streptomycin–glutamine (Life Technologies). The A549, VCaP, PC3, HCC515, HepG2, HeLa and Ben-Men-1 cell lines were cultured in DMEM (Life Technologies), with 10% fetal bovine serum (Sigma-Aldrich) and 1% penicillin–streptomycin–glutamine (Life Technologies). The HA1E cell line was cultured in MEM α (Life Technologies), with 10% fetal bovine serum (Sigma-Aldrich) and 1% penicillin–streptomycin–glutamine (Life Technologies). The MCF10A cell line was 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-Aldrich), 1% penicillin–streptomycin–glutamine (Life Technologies) and 10 $\mu\text{g ml}^{-1}$ insulin (Sigma-Aldrich). Cells were incubated at 37 °C, 5% CO₂ 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 the cells 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 wild-type single-cell-derived MCF7 clones were generated by cell sorting. Single cells were sorted into individual wells of 96-well plates, using a BD FACSAriaII SORP Cell Sorter. Three resultant clones were expanded for a period of approximately three months before 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. In brief, 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 approximately 2 kb of the *GREB1* gene surrounding the termination codon in exon 33, such that the 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 that recognized the sequence GCTGACGGGACGACACATCTG on the sense strand, using a PAM site that is adjacent to the *GREB1* termination codon. For targeting *ESR1*, the reporter gene cassette was subcloned into a construct containing approximately 2 kb of *ESR1* gene surrounding the termination codon in exon 8, such that the 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 that recognized the sequence GTCTCCAGCAGCAGGTCATAG on the anti-sense strand, using a PAM site that is 160 bp upstream of the *ESR1* termination codon. Corresponding Cas9 sgRNA and targeting construct pairs were transiently transfected into MCF7 cells using Lipofectamine 2000 (Thermo Fisher Scientific). After growth for seven days, the cells were cultured in medium containing 5 $\mu\text{g ml}^{-1}$ blasticidin to select for the desired recombinants. Single-cell clones were then isolated by 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 per well. Plates were incubated in an IncuCyte ZOOM instrument (Essen Bioscience) at 37 °C, 5% CO₂. Four non-overlapping phase-contrast images (10 \times) were taken every 2 h for a total of 160 h. IncuCyte ZOOM software (version 2015A) was used to calculate the mean confluence per well at each 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 $T_{\text{doubling}} = (\log_2(\Delta T)) / (\log_2(c_2) - \log_2(c_1))$, in which c_1 and c_2 were the minimum and maximum percentage confluency 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 confluency surpassed a threshold of 15%. To examine the effect of oestrogen depletion on the growth of MCF7 strains, cells were cultured either in standard conditions (described above) or in oestrogen-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 oestrogen-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 20 \times /1.0NA water-immersion lens. Image quality control was carried out as previously described³¹, using CellProfiler³⁰ and CellProfiler-Analyst³¹. 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 and 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. 'GenotypeConcordance' (Picard tools) was used to calculate the concordance between each pair of samples (separately for the MCF7 and A549 cohorts). Samples with >95% concordance were called a match.

Ultra-low-pass whole-genome DNA sequencing. Copy number characterization was performed using low-pass (approximately 0.2 \times coverage) whole-genome sequencing. Libraries were prepared from 50 ng of DNA using ThruPLEX-DNAseq sample preparation kits (Rubicon Genomics) according to the manufacturer's protocol. The resultant libraries were quantified using a 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 35-bp reads, at the Dana-Farber Cancer Institute Molecular Biology Core Facilities. The reads were aligned to the UCSC hg19 reference genome, using 'BWA-MEM' (v.0.07.15), with default parameters.

Ultra-low-pass whole-genome DNA-sequencing data analysis. The ichorCNA algorithm³² was applied to identify CNAs of large genomic segments, chromosome arms and whole chromosomes. First, the genome was divided into 1-Mb 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 mapability biases using the HMMcopy R package³³, generating corrected read counts for each 1-Mb bin. Segmentation and copy number prediction for each sample were performed using ichorCNA v.0.1.0 (<https://github.com/broadinstitute/ichorCNA>), which is optimized for low-coverage whole-genome sequencing. Parameters were initialized based on prior knowledge: $-\text{normal} = 0.01$, $-\text{ploidy} = c(3, 3.5, 4)$, $-\text{txnE} = 0.99999$, $-\text{txnStrength} = 10,000$, $-\text{maxCN} = 8$. 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. Deep (approximately 250 \times coverage) targeted exon sequencing of 447 genes that are commonly mutated in cancer was performed (Profile OncoPanel v.3). Prior to library preparation, DNA was fragmented (Covaris sonication) to 250 bp 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 ≥ 250 ng, and all samples had sufficiently high yields. Normalized libraries were pooled in batches, and hybrid capture was performed using the Agilent Sureselect Hybrid Capture kit with the POPv3_824272 bait set³⁴. 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 deconvoluted and sorted using 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: $-\text{q} \ 5 \ -1 \ 32 \ -k \ 2 \ -o \ 1$, and duplicate reads were identified and removed using Picard tools³⁵. 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 >30 \times coverage for >80% of the targeted bases. Average mean exon target coverage was 251.5 \times (range: 171.5–336.7 \times) for the MCF7 samples, 288.9 \times (range: 208.2–398.9 \times) for the A549 samples and 257.32 (range 211.7–442.68 \times) for the additional cell line samples.

Targeted sequencing data analysis. Mutation analysis for SNVs was performed using MuTect v.1.1.4³⁸. Indel calling was performed using the SomaticIndelDetector

tool in GATK (<http://www.broadinstitute.org/cancer/cga/indelocator>). Consecutive variants in the same codon were reannotated to maximize the effect on the codon and marked as ‘Phased’ variants. MuTect was run in paired mode, pairing the MCF7 or 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 Predictor³⁹. Variants were filtered against the 6,500 exome release of the Exome Sequencing Project database. Variants that were represented more than once in either the African- or European-American populations and were found 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, 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 lists of variants (SNVs, indels and phased) for MCF7, A549 and additional cell lines are provided in Supplementary Tables 5, 17 and 23, respectively.

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 (M. Ducar et al. manuscript in preparation). Systematic bias in mapping depth was reduced using robust regression, fitting the observed tumour mapping depth against a panel of normal samples captured using the same bait set. Observed values were then normalized against predicted values and expressed as log₂ ratios. A second normalization step was performed to remove GC bias, using a LOESS fit. log₂ ratios were centred on segments that were 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 Segmentation⁴⁰ 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. Owing to the high data quality and low within-segment standard deviation, a cutoff of around 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) were detected using Breakmer⁴¹, which is designed to detect larger genomic structural variations from single-sample-aligned short-read target-captured high-throughput sequencing data. In brief, 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 lists of structural variants for MCF7 and A549 are provided in Supplementary Tables 6 and 18, respectively. Rearrangements were visualized using the ‘Circos’ visualization tool⁴⁰.

Clonality analysis. To resolve clonal dynamics and composition, we applied the PyClone algorithm v.0.13.0 (<https://bitbucket.org/aroth85/pyclone/wiki/Home>) to the measured allelic fractions, accounting for copy number, loss of heterozygosity and cellularity¹⁷. 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 oligonucleotides for sgRNA–barcode library construction were synthesized by IDT and cloned into lentiGuide–Puro⁴² by Gibson assembly, as previously described⁴³. Approximately 300 µg of Gibson product was transformed into 25 µl of Endura electrocompetent cells (Lucigen). After a 1-h recovery period, 0.1% of transformed bacteria were plated in a tenfold dilution series on ampicillin-containing plates to determine the number of successful transformants. The remainder of the transformed bacteria were cultured in 50 ml of LB with 50 µg ml⁻¹ ampicillin for 16 h at 30 °C. Plasmid libraries were extracted using a Plasmid MidiPlus kit (Qiagen) and sequenced to a depth of 6.2 million reads on an Illumina Miniseq, corresponding to 6 × coverage of >1 million barcodes.

Lentivirus was prepared by transfecting a total of 10 million HEK293FT cells, as previously described⁴³. 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 approximately 160,000 unique barcodes. Cells were expanded for the experiment, and five million cells were then plated into 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 h with 500 nM bortezomib (Selleckchem, S1013). After five weeks of culture, DNA was extracted and barcode abundance was assessed by DNA sequencing, as previously described⁴³. Libraries were sequenced to a median depth of 4.2 million reads, corresponding to a barcode coverage of >26 ×.

Transcriptional profiling with L1000. The L1000 expression-profiling assay was performed as previously described¹⁶. 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-h 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 streptavidin–phycoerythrin. The PCR amplicon was then hybridized to Luminex microbeads via the complementary, probe-specific barcode on each bead. After overnight hybridization the beads were washed and stained with streptavidin–phycoerythrin to prepare them for detection in Luminex FlexMap 3D scanners. The scanners measured each bead independently and reported the bead colour and 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. These GEX data were 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, for which 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*-SNE. Profiles were restricted to untreated profiles from the nine core connectivity map cell lines, and to batches with multiple untreated profiles. Because samples were first clustered based on their project codes, batch effect was next removed using the COMBAT algorithm⁴⁴. *t*-SNE analysis was applied to the batch-corrected data and visualized using a scatter plot. Analysis was completed using the ‘Rtsne’ R package version 0.13⁶². 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, RNA-sequencing (RNA-seq) 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 to the expression of 80 ‘invariant’ genes, followed by quantile normalization¹⁶. The ranked gene expression order of the 978 landmark genes was compared using a Spearman’s correlation.

Chemical screening. MCF7 strains were tested against a small-molecule target set library of 321 anti-cancer compounds, assembled by the Cancer Target Discovery and Development (CTD²; <https://ocg.cancer.gov/programs/ctd2/data-portal>), using the same principles as those described in the Cancer Therapeutics Response Portal^{18,45}. The list of screened compounds is included as Supplementary Table 26. Cells were seeded in their culture medium in white, 384-well plates (Corning, 3570) at an initial density of 2,500 cells per well and incubated overnight at 37 °C, 5% CO₂. The next day, 25 nl (for primary screen) or 100 nl

(for confirmation dose–response screen) of compound stocks in DMSO were added by pin transfer. Plates were incubated for 72 h, cooled at room temperature for 10 min, and viability was measured using the CellTiter-Glo luminescent cell viability assay (Promega), according to the manufacturer's protocol. After 10 min of incubation, luminescence was read on a Perkin Elmer Envision reader, at a speed of 0.1 s per well.

Chemical screening data analysis. Data were analysed in Genedata Screener version 13.0, using the normalization method 'neutral controls', for which the median of the 32 DMSO wells on each plate was set to 0% activity and 0 raw signal was set to –100%. Positive controls (20 μ M MG-132 or 20 μ M bortezomib) were included on all plates (16 wells each) but were not used for normalization owing to variability in the response across cell lines. Dose–response curves were fit using the 'Smart Fit' strategy in Genedata. The percentage of effect was defined as the high-concentration asymptote (sinf) and the qEC₅₀ was the concentration at which the fitted curve crossed the inhibitory value representing half of 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 described⁸: curves were fit with nonlinear sigmoid functions, forcing the low concentration asymptote to 1 using a three-parameter sigmoidal curve fit. The AUC for each compound–strain pair was calculated by numerically integrating under the eight-point concentration–response curve. For visualization purposes, drug response curves were fit with a four-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. Gene set enrichment analysis (GSEA) was performed using the 10,147 genes best inferred from the connectivity map linear model³³, also known as the BING gene set. Samples were divided into two classes depending on the comparisons being made: samples with a genetic alteration versus samples without it; samples sensitive to a drug (>50% inhibition) versus samples insensitive to the same drug (<20% inhibition). Differential expression was calculated using the signal-to-noise metric⁴⁶. A ranked gene list and signal-to-noise values served as the input for the GSEA preranked module of GSEA, using the Java app version 3.0. The analysis was run using the 'hallmark', 'KEGG', 'positional' and 'oncogenic' signature collections from the Molecular Signature Database (MsigDB)²⁷ (<http://software.broadinstitute.org/gsea/msigdb>). To compare between our MCF7 panel, CTD² and GDSC, drug responses were downloaded from the CTRP website (<https://ocg.cancer.gov/programs/ctd2/data-portal>; 'v20.data.curves_post_qc' file, updated 14 October 2015) and from the GDSC website (<http://www.cancerrxgene.org/downloads>; 'log(IC₅₀) and AUC values' file, updated 4 July 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 17 October 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 2 March 2017). Expression profiles were filtered to include only the genes that belong to the L1000 BING set. GSEA compared the expression patterns of the five strains or cell lines with the highest AUC values for each matched drug with the five 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 sets²⁷.

Single-cell RNA-seq. MCF7 cells were cultured as described above. To follow transcriptional changes after drug treatment, MCF7-AA cells were exposed to 500 nM of bortezomib (Selleckchem, S1013) and collected before treatment, after 12 h of exposure (*t*₁₂), after 48 h of exposure (*t*₄₈) or after 72 h of exposure followed by drug wash and 24 h of recovery (*t*₉₆). Cells were washed, trypsinized, passed through a 40- μ m cell strainer, centrifuged at 400g and resuspended at a concentration of 1,000 cells per μ l in PBS containing 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) per the manufacturer's protocol. In brief, 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-seq data analysis. Reads were mapped to the GRCh38 human transcriptome using cell ranger version 2.1.0, and transcript-per-million values were calculated for each gene in each filtered cell barcoded sample. Transcript-per-million values were then divided by 10, since the complexity of single-cell libraries is estimated to be in the order of 100,000 transcripts. For each cell, we quantified the number of expressed genes and the proportion of the transcript counts derived

from mitochondrial 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 Seurat⁴⁷ (<http://satijalab.org/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 unique molecular identification number and mitochondrial gene fraction to remove technical noise. The resulting scaled data were used as an input for PCA. Top significant principal components, estimated by a manual inspection of the PCA standard deviations elbow plots, were used to generate *t*-SNE plots. For each dataset, we used Seurat⁴⁷ (<http://satijalab.org/seurat/>) to identify genes that vary between samples. To detect differentially active pathways, gene ontology (GO) enrichment analysis was performed with MSigDB²⁷ (<http://software.broadinstitute.org/gsea/msigdb>) using the differentially expressed genes that passed the following thresholds: $|\log_2(\text{fold change})| > 0.25$, Bonferroni-corrected $P < 0.01$, the gene was detected in >10% of the cells in each of the compared groups. Expression signatures for selected pathways were downloaded from MSigDB²⁷. 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 reported⁴⁸. To calculate pairwise cell distances, variable genes were detected, and the cell embedding matrix for the top significant principal components was used to calculate the Euclidean distance between every two cells within each sample.

Analysis of genome-wide CRISPR screens. CERES dependency scores⁴⁹ were obtained from the Broad Institute Achilles website (<https://portals.broadinstitute.org/achilles/datasets/18/download>). Owing 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 plasmid DNA 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) divided by *k* is a parameter similar to perplexity in *t*-SNE, set to 6 for our data. Edges between cells were weighted according to the frequency with which they clustered together, with edges appearing less than 30% of the time ignored. Cells were then laid out using the SFPD spring-block algorithm⁵⁰. Cell line RNA-seq gene expression data and reverse-phase protein array protein expression data were obtained from the CCLE website (<https://portals.broadinstitute.org/ccle/data>). Single-sample GSEA was calculated using the ssGSEA algorithm⁵¹.

Chymotrypsin-like activity. MCF7 cells were plated in triplicates in 96-well plates at a density of 20,000 cells per well. After 24 h, chymotrypsin-like activity of the proteasome was assayed, using the Proteasome-Glo assay (Promega), according to manufacturer's 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 (50 mM HEPES-KOH pH 7.9, 150 mM NaCl, 2 mM EDTA pH 8.0, 20 mM sodium molybdate, 0.5% Triton X-100, 5% glycerol), with protease inhibitor cocktail (Roche Diagnostics, 11836153001). Protein concentrations were determined by BCA assay (Thermo Fisher Scientific, 23227) and proteins were resolved using SDS–PAGE for immunoblot analysis. Antibodies against the following human proteins were used: α -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 ER α immunoblotting, cells were lysed with a mix of 4 \times protein loading buffer (Li-Cor, 928-40004) and 10 \times NuPAGE sample reducing agent (Life Technologies, NP0009). Protein concentration was normalized by cell counting and proteins were resolved by SDS–PAGE. Antibodies against the following human proteins were used: β -actin (N-21, Santa Cruz), ER α (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 \log_2 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 the 1,784 measured cellular features. The barcode representation dendrogram was based on the \log_2 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 variance⁵². 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 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 ultra-low-pass whole-genome DNA sequencing 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 events 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 (that is, drugs that reduced the viability of at least one strain by $>50\%$) and setting the threshold for 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 17 October 2012; 'CCLE_copynumber_byGene_2013-12-03.txt', updated 27 May 2014; 'CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct', updated 29 February 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 previously described⁵³, 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 ABSOLUTE⁵⁴. Combined CNA–SNV genomic instability scores were calculated as described previously⁵⁵. The DNA repair gene set was derived from MSigDB (<http://software.broadinstitute.org/gsea/msigdb>), using the 'DNA_Repair' GO signature⁵⁶. The CIN70 gene set was derived from a previous publication⁵⁷. For each gene set, genes that were not expressed at all in the CCLE dataset were removed, and the remaining gene expression values were \log_2 -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, 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 database⁵⁸. The gene-level values were \log_2 transformed, and converted to discrete values using predefined thresholds (± 0.1 , ± 0.3 and ± 0.5). To determine the percentage 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 datasets). 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 MuTect³⁸, 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 that consisted of a panel of 105,995 SNPs was generated, based on common SNVs found in 1,019 CCLE RNA-seq 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 (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. RNA-seq data from non-manipulated and 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 samples⁵⁹. The STAR to RSEM tool⁶⁰ 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 analysed using the e-karyotyping method⁶¹. In brief, RSEM values were \log_2 -converted, genes that were not expressed (\log_2 RSEM < 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 tumour progression. Recurrence of chromosome arm-level CNAs during breast cancer progression was determined by their frequency in TCGA samples, as previously described⁵³. 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 versus strain L versus strain AA, strain B versus strains I and P), as shown in Extended Data Fig. 2a. Only arms that are recurrently gained or lost (but not both) in TCGA ($Q < 0.05$), 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 difference between chromosome numbers at two time points of single-cell-derived clones was 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 MoAs, in the discordance of non-silent SNVs at different stages of transformation, in chromosomal instability (CIN70) and weighted-genomic integrity index (wGII) scores between cell lines derived from primary tumours and those derived from metastases, between the somatic and germline SNV Pearson correlations of the Broad-Sanger cell lines, in the Broad-Sanger somatic SNV concordance between microsatellite-stable and microsatellite-unstable cell lines and between primary tumour-derived and metastasis-derived cell lines was determined using a one-tailed Wilcoxon rank-sum test. The significance of the difference in mutation cellular prevalence across strains was determined by Kruskal–Wallis test. The significance of the difference in AKT inhibitor IV sensitivity between *PTEN*^{WT} and *PTEN*^{MUT} strains, in the relative growth effect of ER depletion between *ESR1* loss and no-*ESR1* loss strains, in proteasome activity between bortezomib-sensitive and bortezomib-insensitive strains, in ER α protein expression levels between strains and in the number of arm-level CNAs between matched early-late MCF7 strains was determined using a one-tailed Student's *t*-test. The significance of the difference in doubling times and in sensitivity to oestrogen depletion was determined using a 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, the correlation between doubling time and the fraction of subclonal mutations, 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 oestrogen signalling and between *GATA3* CERES dependency scores and *GATA3* protein expression levels was 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

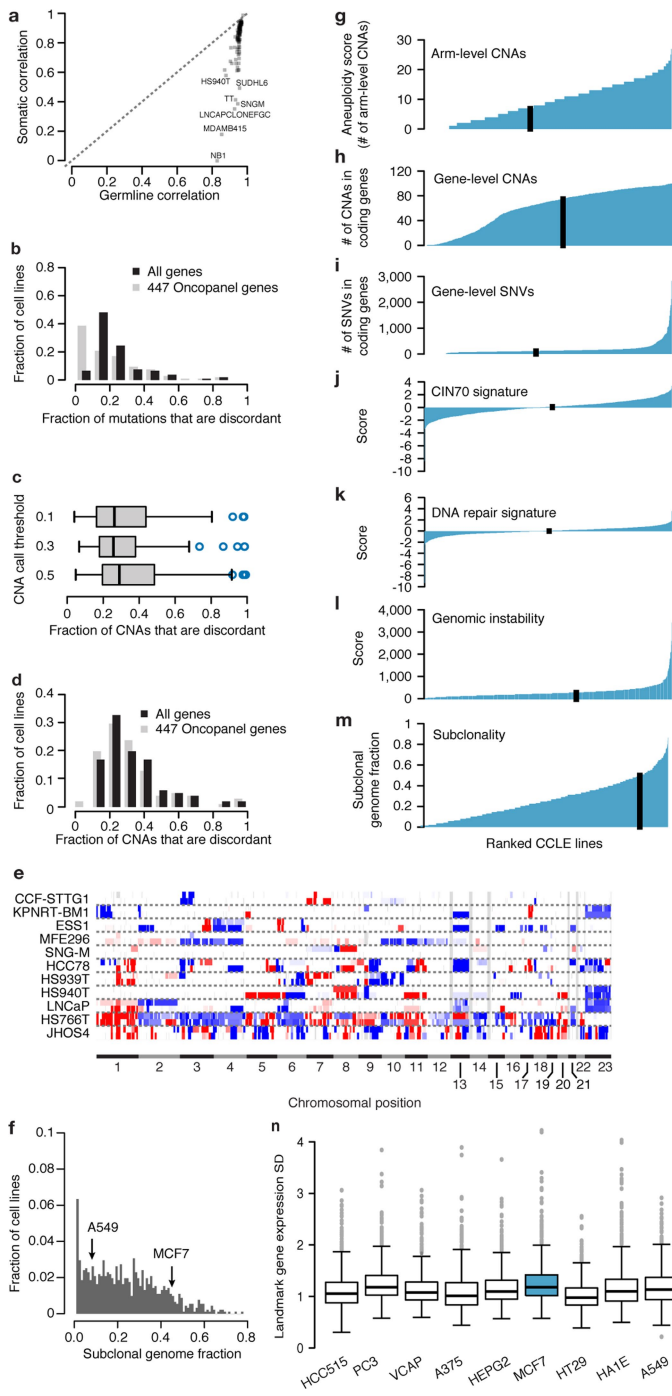
the fraction of pathways correctly identified between the MCF7 panel, CTD² and GDSC was determined using a two-tailed Fisher's exact test. GSEA *P* values and FDR-corrected *Q* values are shown as provided by the default analysis output. For the comparison of pathway prediction shown in Supplementary Table 32, FDR *Q* values were recalculated 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 RNA-seq 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 interquartile range (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.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

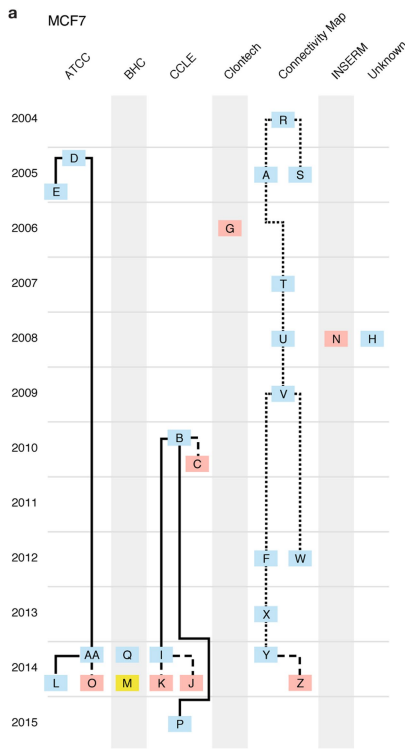
Code availability. The code used to generate and/or analyse the data during the current study are publicly available, or available from the corresponding authors upon request.

Data availability. The datasets generated during and/or analysed during the current study are available within the article, its Supplementary Information or from the corresponding authors upon request. DNA sequencing data were deposited to SRA with BioProject accession number PRJNA398960. Single-cell RNA-seq 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. The cell divergence portal is accessible at: <https://celltrainer.broadinstitute.org>.

30. Bray, M. A., Fraser, A. N., Hasaka, T. P. & Carpenter, A. E. Workflow and metrics for image quality control in large-scale high-content screens. *J. Biomol. Screen.* **17**, 266–274 (2012).
31. Dao, D. et al. CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. *Bioinformatics* **32**, 3210–3212 (2016).
32. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
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).
34. Sholl, L. M. et al. Institutional implementation of clinical tumor profiling on an unselected cancer population. *JCI Insight* **1**, e87062 (2016).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
37. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
38. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
39. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* **26**, 2069–2070 (2010).
40. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
41. Abo, R. P. et al. BreakMer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res.* **43**, e19 (2015).
42. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
43. Joung, J. et al. Genome-scale CRISPR–Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* **12**, 828–863 (2017).
44. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
45. Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
46. Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
47. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
48. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
49. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
50. Hu, Y. Efficient, high-quality force-directed graph drawing. *Math. J.* **10**, 37–71 (2006).
51. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
52. Fowlkes, E. B. & Mallows, C. L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–569 (1983).
53. Ben-David, U. et al. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* **49**, 1567–1575 (2017).
54. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
55. Zhang, S., Yuan, Y. & Hao, D. A genomic instability score in discriminating nonequivalent outcomes of BRCA1/2 mutations and in predicting outcomes of ovarian cancer treated with platinum-based chemotherapy. *PLoS ONE* **9**, e113169 (2014).
56. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
57. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
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).
59. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
60. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
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**, 989–997 (2013).
62. Krijthe, J. H. *Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation* <https://github.com/krijthe/Rtsne> (2015).

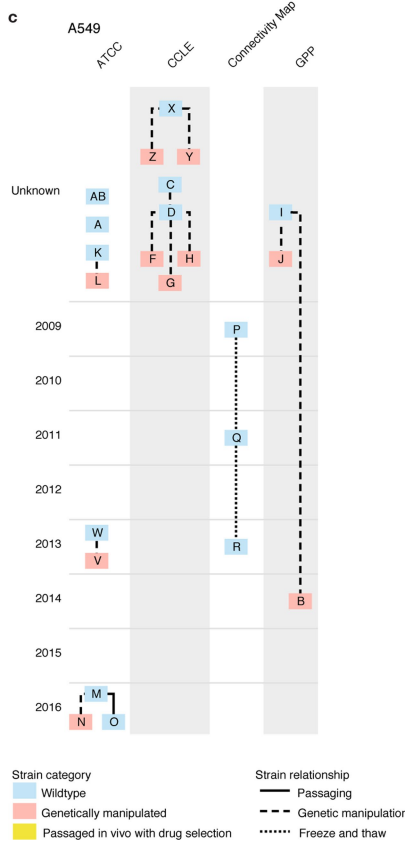


Extended Data Fig. 1 | Comparison of Broad and Sanger genomic features across 106 cell lines. **a**, Comparison of the Pearson correlations of germline versus somatic SNVs across 106 paired cell lines. **b**, A histogram of the distribution of mutation discordance fractions across cell lines. Black, the distribution of all non-silent SNVs; grey, the distribution of the 447 genes included in the Oncopanel. **c**, Comparison of the fraction of discordant gene-level CNAs between the Broad and the Sanger ($n = 106$ cell lines) datasets, using three different thresholds for CNA calling. Bar, median; box, 25th and 75th percentiles; whiskers, $1.5 \times$ IQR of lower and upper quartile; circles, data points. **d**, A histogram of the distribution of CNA discordance fractions across cell lines. Bars are coloured as in **b**. **e**, CNA landscapes of 11 paired cell lines. For each cell line, the CNA landscape of the Broad strain (top) and the Sanger strain (bottom) are shown. Red, copy number gains; blue, copy number losses. CNAs < 10 Mb in size are not presented. **f**, A histogram of the fraction of the genome affected by subclonal events across 916 cell lines from the CCLE. 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 their gene-level CNAs. **i**, All CCLE cell lines ranked by their chromosomal instability (CIN70) signature scores⁵⁷. **k**, All CCLE cell lines ranked by their DNA-repair signature scores⁵⁶. **l**, All CCLE cell lines ranked by their genomic instability scores⁵⁵. **m**, All CCLE cell lines ranked by their subclonal genome fraction⁵⁴. The vertical black line shows the rank of MCF7 in each comparison. **n**, Comparison of gene expression variation across multiple strains of nine cell lines, including MCF7. Box plots are 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 $1.5 \times$ 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		<i>In vivo</i> 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

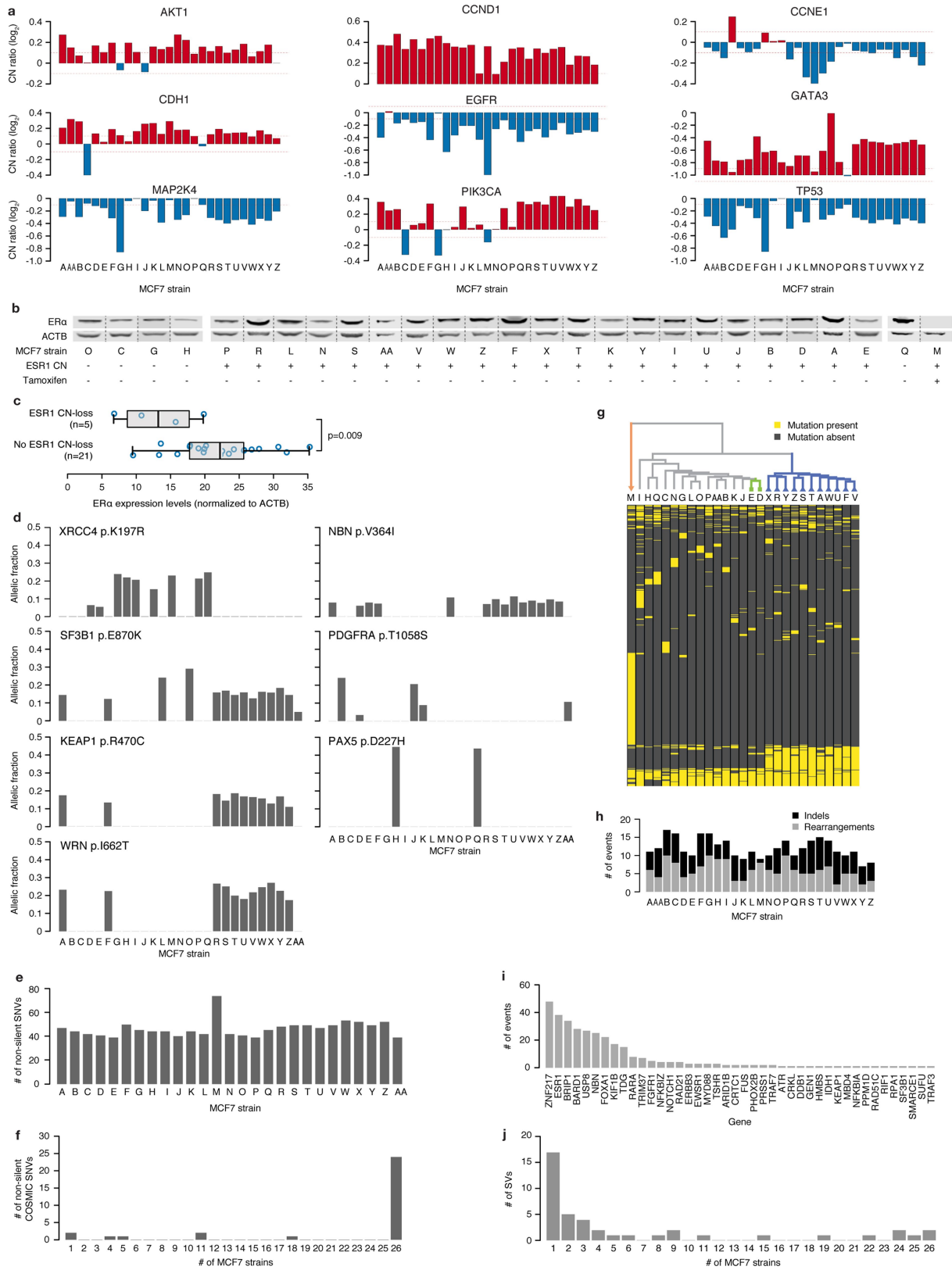


d

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 Fig. 2 | Schematic representation of the MCF7 and A549 strains included in the current study. a. MCF7 strains included in this study; their origins (columns), years of acquisition (rows), manipulations (colours) and progeny relationships (lines) are shown. **b.** A table

corresponding to **a.** c. A549 strains included in this study, their origins (columns), years of acquisition (rows), manipulations (colours) and progeny relationships (lines) are shown. **d.** A table corresponding to **c.**

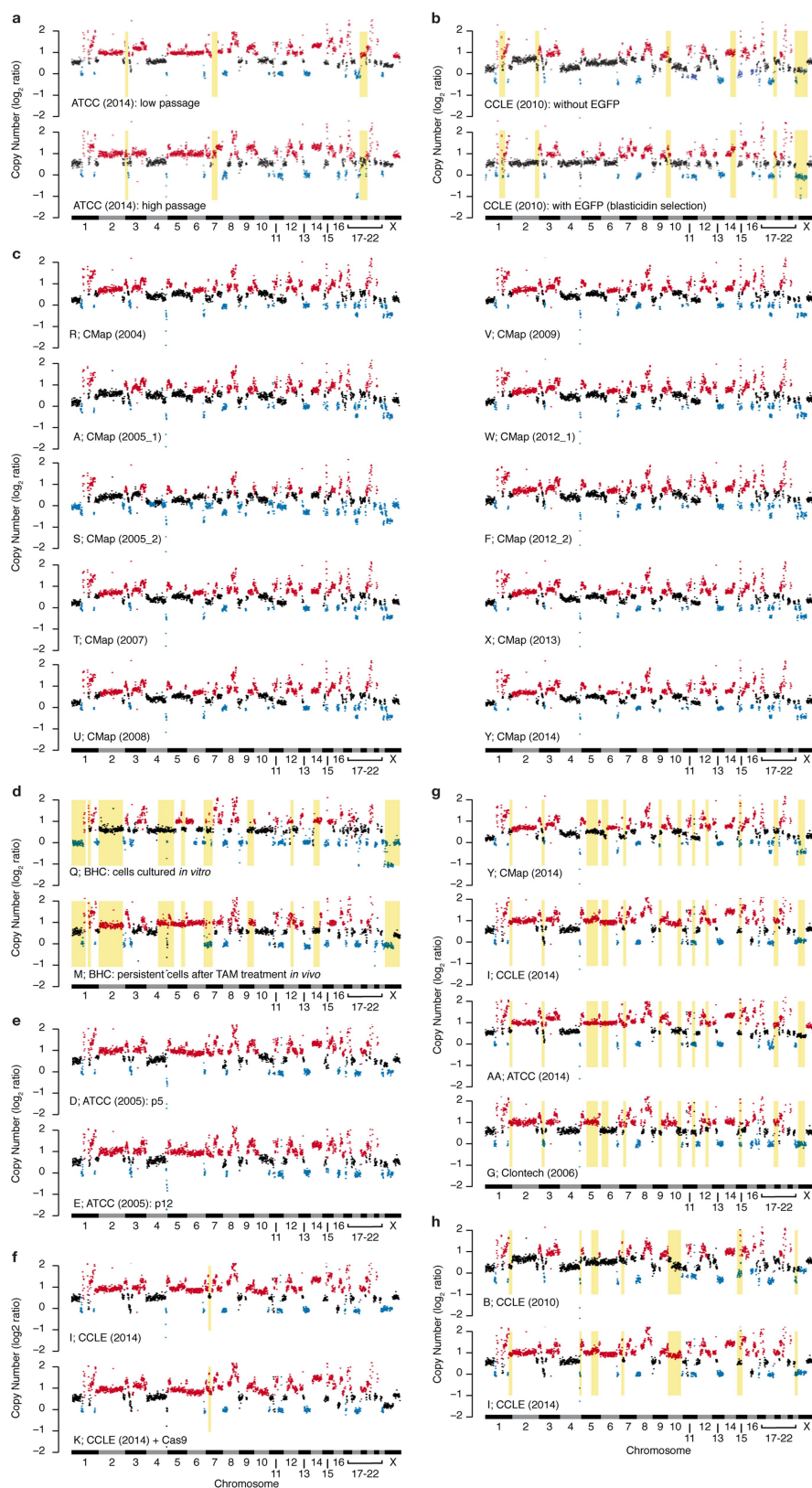


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Genetic variation across 27 MCF7 strains.

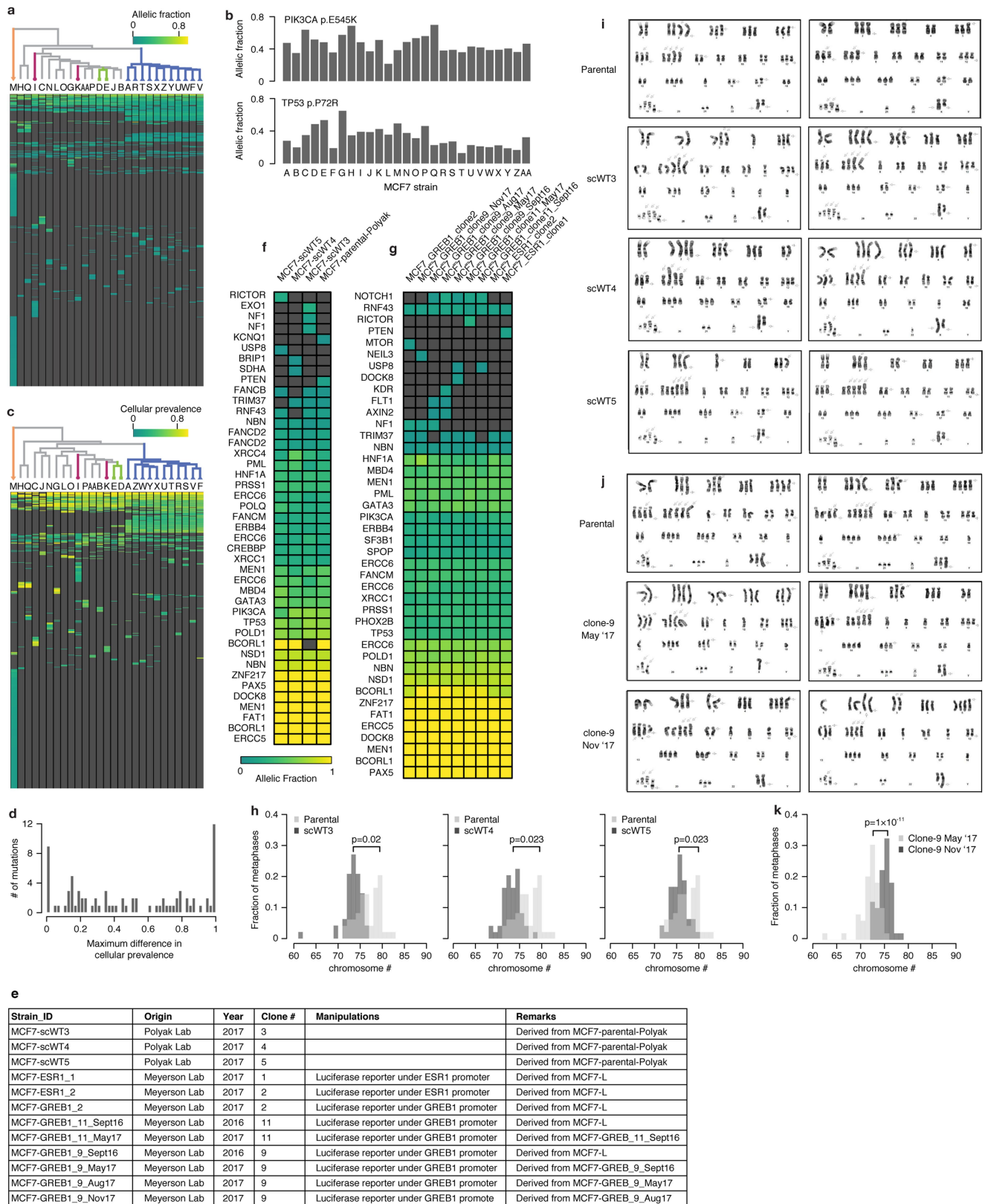
a, Variation in the copy number status of nine selected genes across 27 MCF7 strains. Red, copy number gains; blue, copy number losses. Thresholds for relative gains and losses were set at 0.1 and -0.1 , respectively. **b**, Western blots of the relative protein expression levels of ER α across strains. 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 ER α . Strains Q and M were excluded from the comparison. Bar, median; box, 25th and 75th percentiles; whiskers, data within $1.5 \times$ 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 of their SNVs. Groups of strains expected to cluster together based on their evolutionary history are highlighted, as in Fig. 1. Bottom, a corresponding heat map, showing the mutation status of all mutations across the 27 MCF7 strains. Mutations that were identified in only a subset of the strains that were detected in above 5% of the reads ($AF > 0.05$) are shown. Yellow, presence of a mutation; grey, absence of a mutation. **h**, The number of large (>15 -bp) indels and rearrangements across the 27 MCF7 strains. Grey, indels; black, rearrangements. **i**, The recurrence of structural variants in each of the 42 (out of 60) genes for which at least one event was detected. **j**, The number of structural variants shared by each number of MCF7 strains. Note that this analysis is limited to the 60 genes listed in Supplementary Table 2.



Extended Data Fig. 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 a pair 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 from each other by seven passages. **f**, CNA landscapes of a pair of MCF7 strains before (top) and after (bottom) the 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 1-Mb bins throughout the genome. Red, gains; blue, losses; black, normal copy numbers; yellow, differential CNAs between the compared strains.

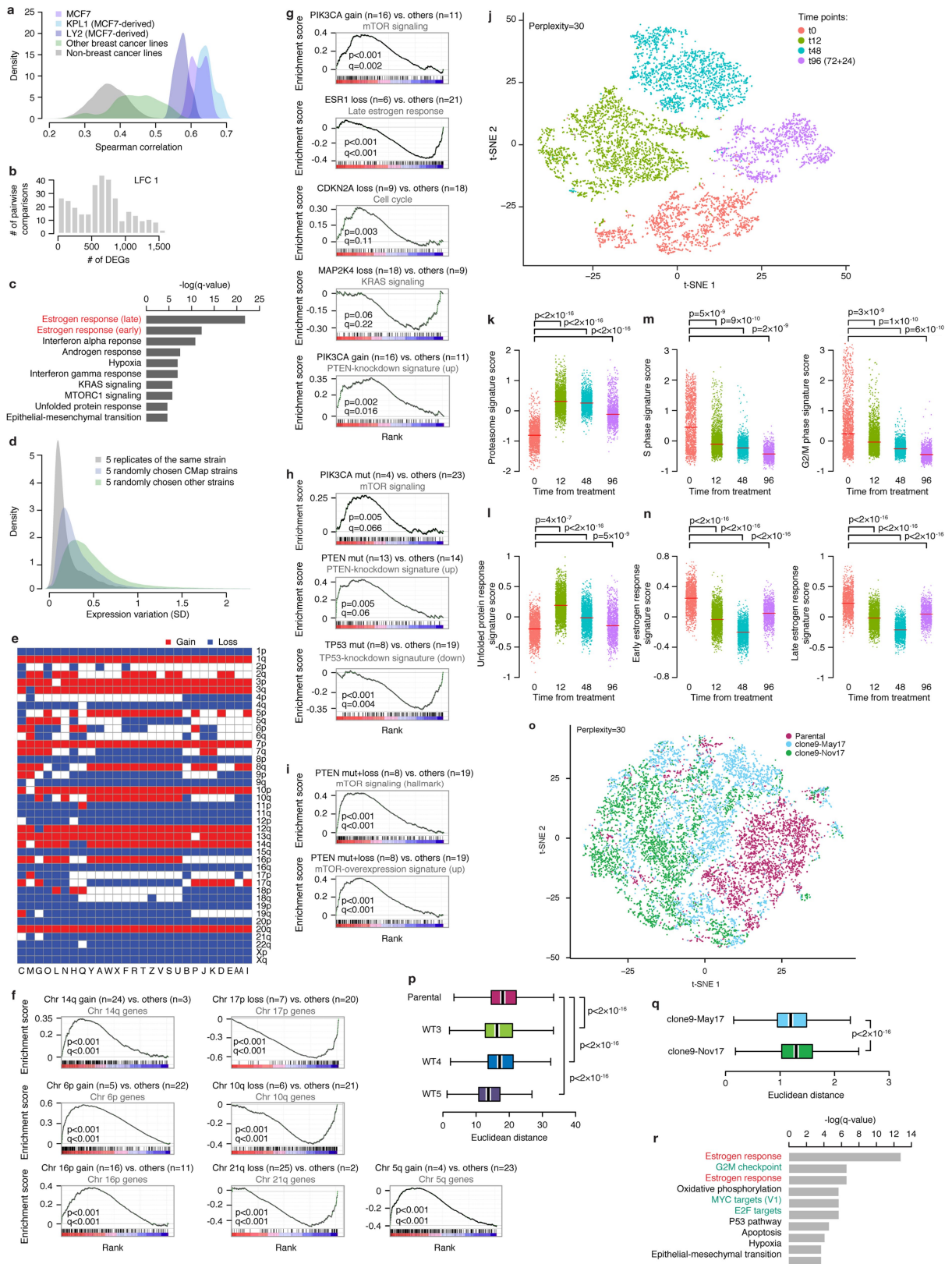


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Characterization of the variation in SNV allelic fraction and cellular prevalence 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 heat map, showing the allelic fractions of all mutations across the 27 MCF7 strains. Mutations that were identified in only a subset of the strains are shown. The presence of a mutation is shown in colour according to its allelic fraction. **b**, The allelic fractions 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 heat map, showing the cellular prevalence of all mutations across the 27 MCF7 strains. Mutations that were identified in only a subset of the strains are shown. The presence of a mutation is shown in colour 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 maximum $\Delta\text{CP} = 1$ 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 maximum $\Delta\text{CP} = 0$ represents SNVs that are detected at similar prevalence across all 27 strains; and the peak at maximum $\Delta\text{CP} \approx 0.1$ represents a group of SNVs present at $\text{CP} \approx 0.1$ only in strain M. **e**, Description of the MCF7 single-cell-derived clones included in this study, including their parental cell line, genetic manipulations and relationship to one another. **f**, A heat map showing the allelic fractions of non-silent mutations in three wild-type single

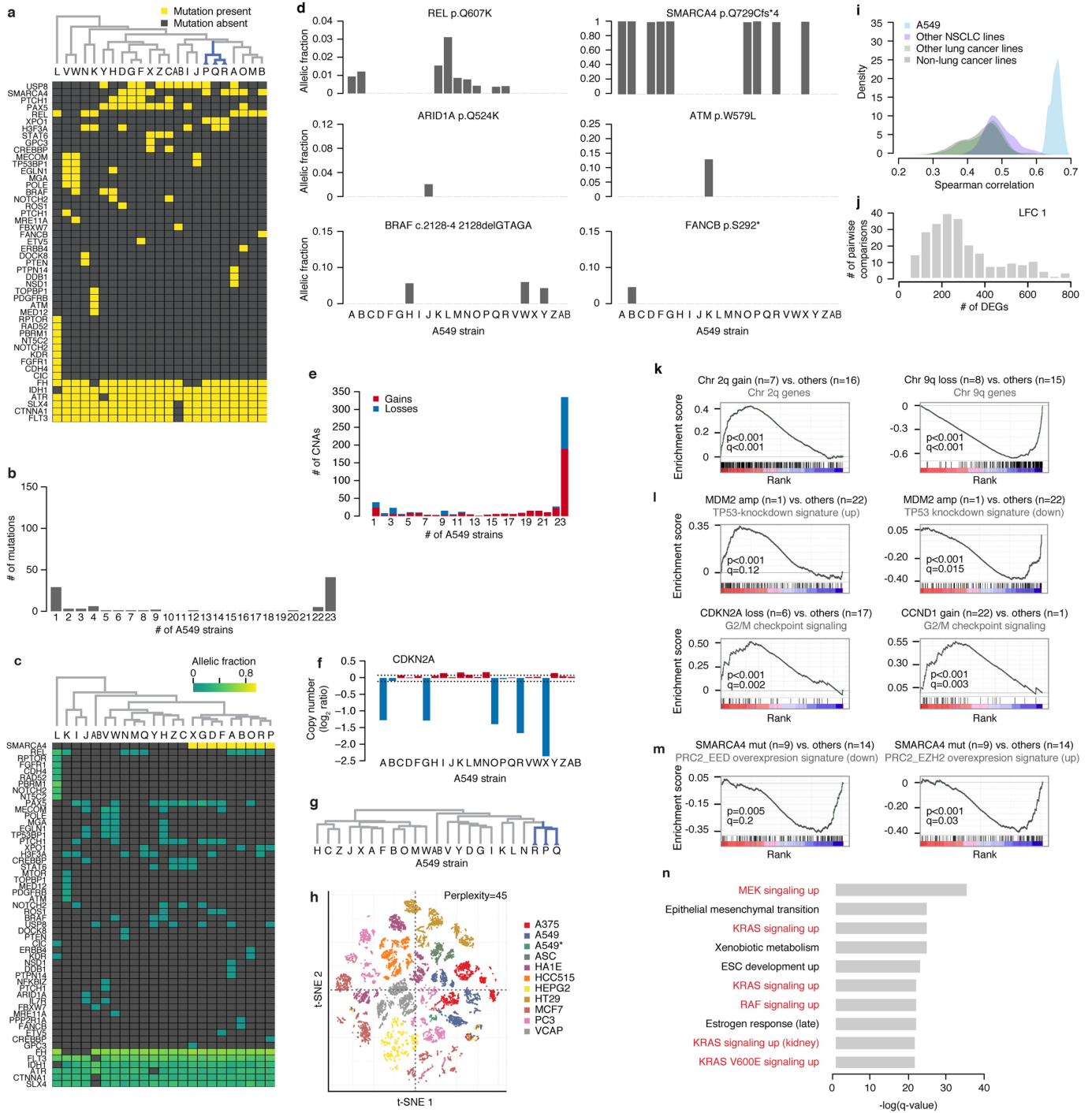
cell-derived MCF7 (scWT3–scWT5) clones and the parental population. The presence of a mutation is shown in colour according to its allelic fraction. **g**, A heat map showing the allelic fractions of non-silent mutations in five genetically manipulated single-cell-derived MCF7 clones. For two of the clones, samples were passaged for a prolonged time and sequenced at multiple time points. The presence of a mutation is shown in colour according to its allelic fraction. **h**, Comparison of the karyotypic variation between parental and single-cell-derived cell populations. Histograms show the distribution of chromosome numbers from the parental (light grey) and single-cell-derived (dark grey) populations. *P* values indicate the significance of the differences between the variations (rather than the means) of the populations using a one-tailed Levene's test ($n = 50$ metaphases per group). **i**, Two representative karyotypes of each sample. Note that all single-cell-derived clones are karyotypically 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 six 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 six months of culture propagation. Histograms show the distribution of chromosome numbers from the early (light grey) and late (dark grey) populations. Per sample, 50 metaphases were counted. The *P* value indicates the significance of the difference between the means of the populations using a two-tailed Wilcoxon rank-sum test.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 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 show 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 (grey). The comparison is based on the 978 landmark genes directly measured in L1000. **b**, The number of differentially expressed genes identified in all possible pairwise comparisons of MCF7 strains, using a twofold 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 oestrogen response are highlighted in red. **d**, Comparison of gene expression variation within and between strains. Histograms show the distributions of gene expression variation within replicates of the same strain (grey), between closely related strains (purple) and between all strains (green). The comparison is based on the 978 landmark genes directly measured in L1000. **e**, Heat map showing the arm-level CNA profiles of 27 MCF7 strains. Red, gains; blue, losses. **f**, GSEA reveals downregulation of the genes on chromosomes 10q, 17q and 21q in strains that have lost copies of these arms, and upregulation of the genes on chromosomes 5q, 6p, 14q and 16p in strains that have gained copies of these arms. **g**, GSEA of the upregulation of mTOR signalling (gene set: hallmark_MTORC1_signalling) and of genes that are upregulated when *PTEN* is knocked down (gene set: PTEN_DN.v2_UP) in strains that have gained *PIK3CA*; downregulation of the oestrogen response signature (gene set: hallmark_oestrogen_response_late) in strains that have lost *ESR1*; cell cycle signature (gene set: KEGG_cell_cycle) in strains that have lost *CDKN2A*; and downregulation of *KRAS* signalling (gene set: hallmark_KRAS_signalling_DN) in strains that have lost *MAP2K4*. **h**, GSEA of the upregulation of mTOR signalling (gene set: hallmark_MTORC1_signalling) in strains with high prevalence of an activating *PIK3CA* mutation; upregulation of genes that are upregulated when *PTEN* is knocked down (gene set: PTEN_DN.v1_UP) in strains that

have an inactivating *PTEN* mutation; and downregulation of genes that are downregulated 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 upregulation of mTOR signalling (gene sets: MTOR_UP.N4.V1_UP and hallmark_MTORC1_signalling) in strains that have both *PTEN* copy number loss and an inactivating *PTEN* mutation. **j**, A *t*-SNE plot of single-cell RNA-seq data from MCF7-AA cells treated with bortezomib (500 nM) at different time points. Each dot represents a single cell, and cells are coloured 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 (left) and G2M (right), across time points. **n**, Comparison of the early (left) and late (right) response to oestrogen 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 t_0, t_{12}, t_{48} and t_{96} , respectively. **o**, A *t*-SNE plot of single-cell RNA-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 coloured 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, scWT3, scWT4 and scWT5, respectively. **q**, Comparison of the transcriptional heterogeneity between two cultures of the same single-cell clone, separated by six months of continuous passaging. $n = 4,295$ and $4,116$ cells, for clone9-May17 and clone9-Nov17, respectively. Box plots show the Euclidean distance between the cells in each cell population. Bar, median; box, 25th and 75th percentiles; whiskers, data within $1.5 \times$ 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 between the two cultures of clone MCF7_GREB1_9 (May 2017 versus November 2017). The gene sets related to oestrogen response are highlighted in red, and those related to proliferation are highlighted in green.



Extended Data Fig. 7 | See next page for caption.

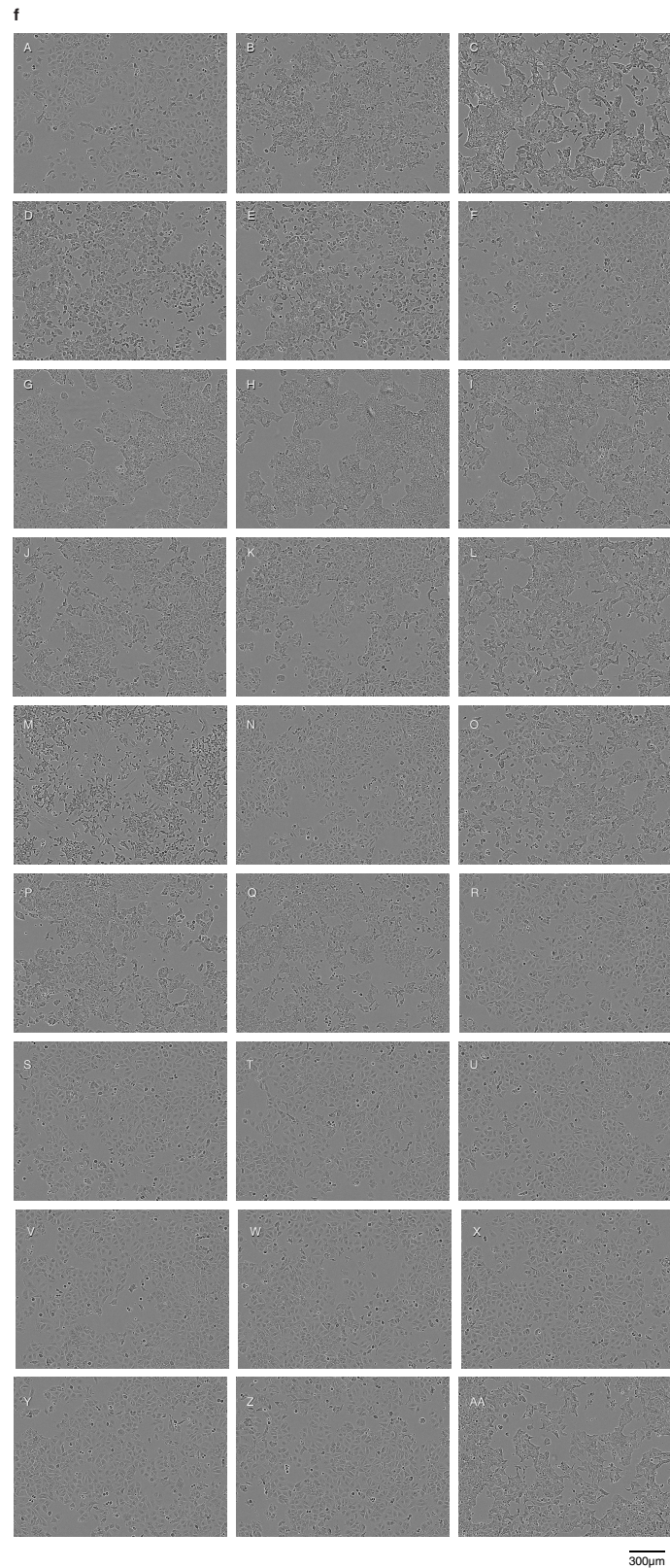
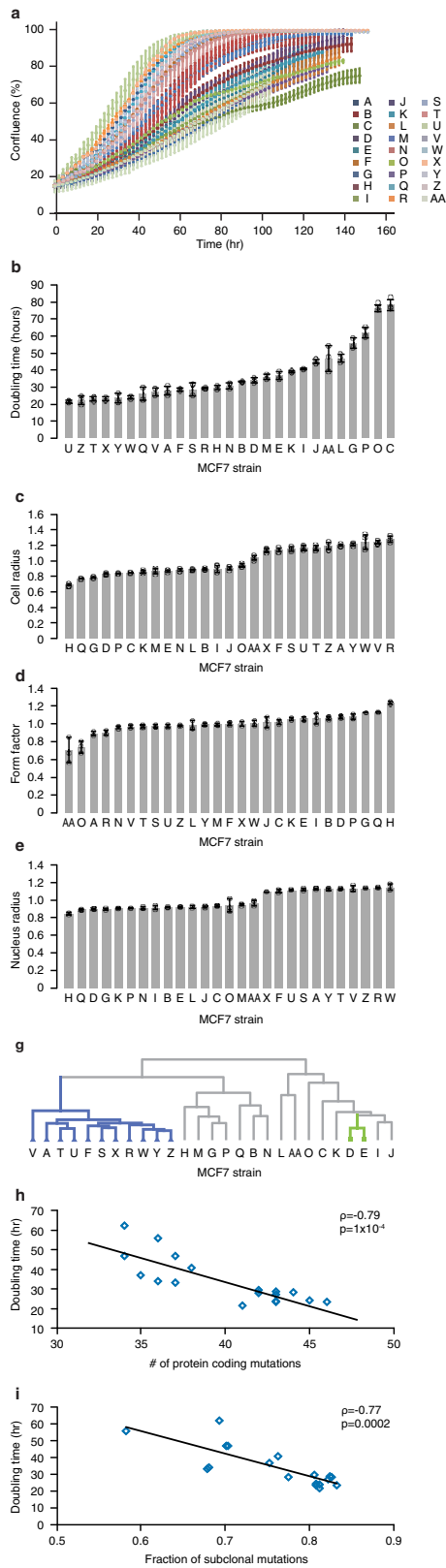
Extended Data Fig. 7 | Extensive genetic and transcriptional variation across 23 strains of A549. **a**, Top, unsupervised hierarchical clustering of 23 A549 strains, based on their non-silent SNV profiles derived from deep targeted sequencing. Strains expected to cluster together based on their evolutionary history are highlighted in blue. Bottom, a corresponding heat map, showing the mutation status of non-silent mutations across the 23 A549 strains. Mutations that were identified in only a subset of the strains, which were detected in above 5% of the reads ($AF > 0.05$) are shown. The presence of a mutation is shown in yellow, and its absence in grey. **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 heat map, showing the allelic fractions of non-silent mutations across the 23 A549 strains. Mutations that were identified in only a subset of the strains are shown. The presence of a mutation is shown in colour 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 adenocarcinoma²⁴, 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 CNAs shared by each number of MCF7 strains. Red, copy number gains; blue, copy number losses. **f**, CNA variation in the copy number of *CDKN2A*. Red, copy number gains; blue, copy number losses. Thresholds for relative gains and losses were set at 0.1 and -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 *t*-SNE 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 between the L1000-based A549 expression profiles and the microarray-based expression profiles from CCLE. Histograms show the distributions of the Spearman correlations between the 23 A549 strains and A549 (light blue), other non-small-cell lung cancer cell lines (purple), other lung cancer cell lines (green) or non-lung cancer cell lines (grey). The comparison is based on the 978 landmark genes directly measured in L1000. **j**, The number of differentially expressed genes identified in all possible pairwise comparisons of A549 strains, using a twofold change cutoff. **k**, Arm-level gains are associated with significant upregulation and arm-level losses are associated with significant downregulation of genes transcribed from the aberrant arms. GSEA showing upregulation of the genes on chromosome 2q in strains that have gained a copy of that arm (left), and downregulation 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. GSEA reveals upregulation of the genes that are upregulated, and downregulation of the genes that are downregulated, when *TP53* is knocked down in strains with *MDM2* high-level copy number gain; and upregulation or downregulation of the G2/M cell cycle checkpoint signature in strains with *CDKN2A* copy number loss or *CCND1* copy number gain. **m**, Point mutations are associated with significant dysregulation of the perturbed pathways. For example, GSEA reveals downregulation 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* signalling are highlighted in red.

Extended Data Fig. 8 | Genetic variation across multiple strains of additional cancer and non-cancer cell lines.

a, The fraction of non-silent SNVs that are discordant between pairs of strains of the same cell line. Data are mean \pm s.e.m. n , number of strain pairs compared. **b**, Arm-level CNAs arise in RPE1 samples. Plots show CNAs detected by an e-karyotyping analysis of 26 RPE1 samples. Red, gains; blue, losses. **c**, Comparison of variability in non-silent SNVs between non-transformed, partially transformed and fully transformed MCF10A samples. Box plots show the fraction of discordant non-silent SNVs between pairs of samples within each category. Bar, median; box, 25th and 75th percentiles; whiskers, data within $1.5 \times$ 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 tumours and those derived from metastases. Bar, median; coloured 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

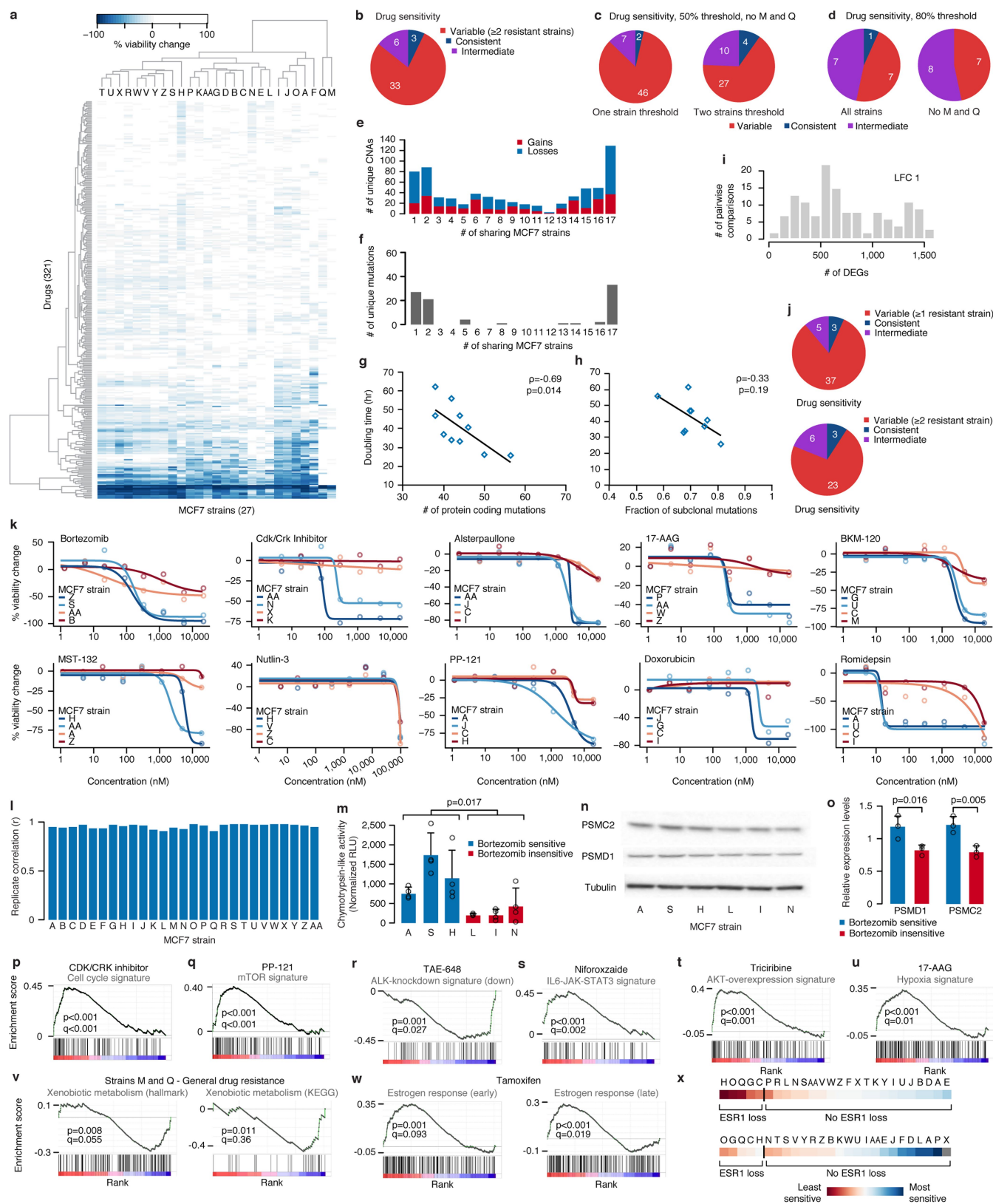
between CCLE lines derived from primary tumours and those derived from metastases. Bottom, comparison of the weighted-genomic integrity index (wGII) between CCLE lines derived from primary tumours and those derived from metastases. Bar, median; coloured rectangle, 25th and 75th percentiles; width of 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 $1.5 \times$ IQR of lower or upper quartile; circles, all data points. One-tailed Wilcoxon rank-sum test. **g**, Heat maps show the allelic fractions of non-silent mutations in multiple strains of cancer cell lines. The presence of a mutation is shown in colour according to its allelic fraction. **h**, Heat maps show 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 colour 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 Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Characterization of cell proliferation and morphology across 27 MCF7 strains. **a**, Growth response curves of 27 MCF7 strains, based on microscopy imaging. **b**, Doubling time of the 27 MCF7 strains, as measured by automatic microscopy imaging. **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. **a–e**, Data are mean \pm s.d., circles show individual values; $n = 3$ replicate wells per data point. **f**, Microscopy imaging of the 27 MCF7 strains, showing the morphological differences between them. Scale bar, 300 μm . Images are representative of five 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 (that is, strains that have not undergone drug selection or genetic manipulation). Spearman's ρ and P values 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 ρ and P values indicate the strength and significance of the correlation, respectively.

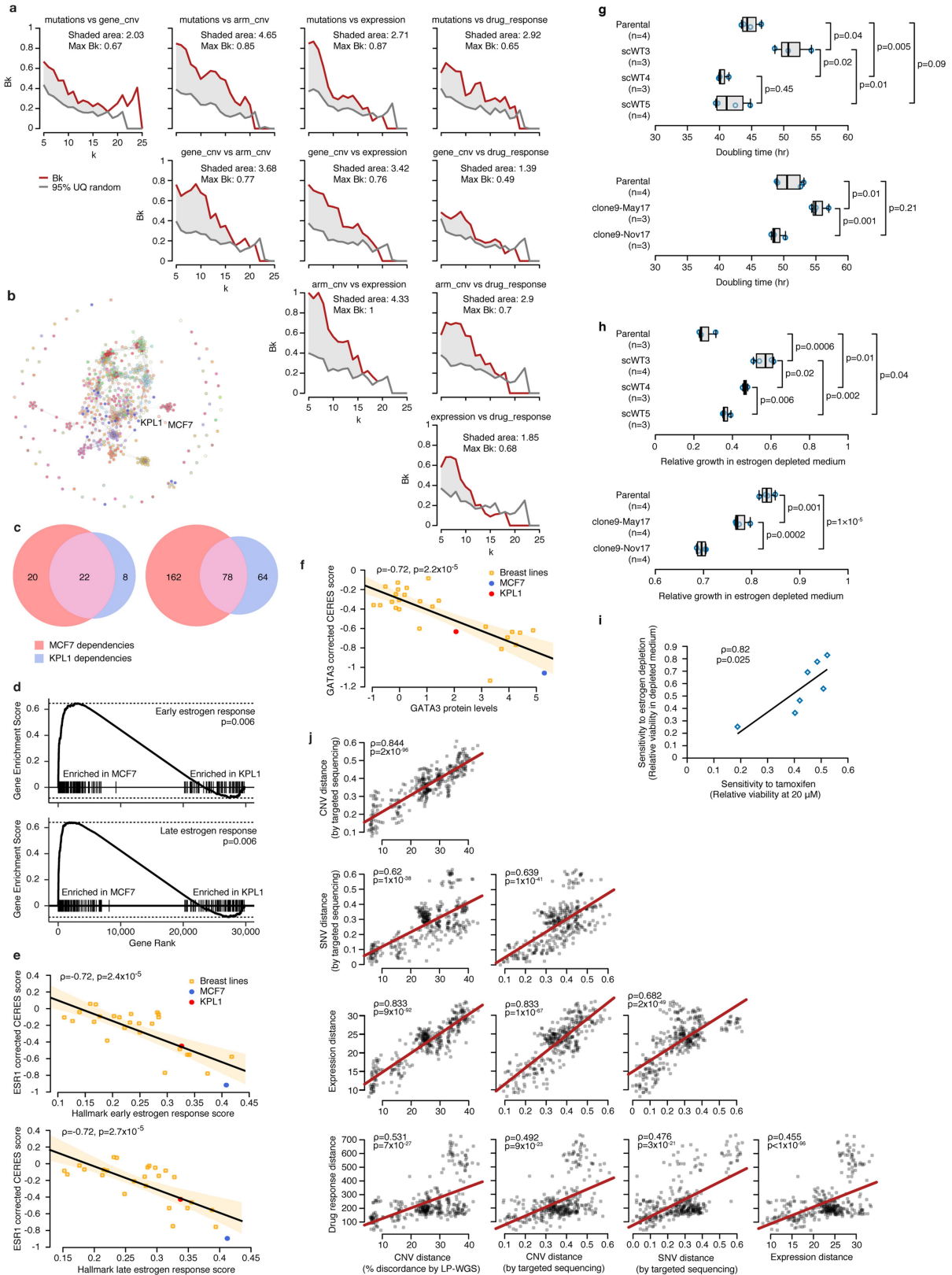


Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 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 of 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 as in **b** 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** 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 CNAs shared by each number of MCF7 strains. Red, copy number gains; blue copy number losses. **f**, The number of non-silent point mutations shared by each number of MCF7 strains. The 10 naturally occurring connectivity map 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 ρ and P values 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 ρ and P values 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 identified in all possible pairwise comparisons of MCF7 strains, using a twofold change cutoff. The 10 naturally occurring CMap strains were averaged and considered as a single sample. **j**, Pie charts of 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 (top and bottom charts, respectively). The 10 naturally occurring CMap strains were averaged and considered as a single sample.

k, The dose-response curves for ten compounds are shown. 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. Data are mean \pm s.d., one-tailed t -test, $n = 4$ replicate wells. **n**, Western blots of 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. Data are mean \pm s.d., one-tailed t -test, $n = 3$ strains per group. **p**, Upregulation 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**, Upregulation of mTOR signalling in strains sensitive to the PI3K inhibitor PP-121 ($n = 11$) compared to insensitive strains ($n = 5$). **r**, Downregulation of the genes that are downregulated when *ALK* is knocked down in strains sensitive to the *ALK* inhibitor TAE-684 ($n = 4$) compared to insensitive strains ($n = 15$). **s**, Upregulation of IL-6-JAK-STAT3 signalling in strains sensitive to the *STAT* inhibitor nifuroxazide ($n = 9$) compared to insensitive strains ($n = 6$). **t**, Upregulation of the genes that are upregulated when *AKT* is overexpressed in strains sensitive to the *AKT* inhibitor triciribine ($n = 2$) compared to insensitive strains ($n = 8$). **u**, Upregulation of hypoxia signalling in strains sensitive to the HSP inhibitor 17-AAG ($n = 3$) compared to insensitive strains ($n = 15$). **v**, Downregulation 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**, Upregulation of the early and late oestrogen response signatures, in strains most sensitive to the ER inhibitor tamoxifen ($n = 5$) compared to the least sensitive strains ($n = 5$). **x**, Sensitivity to oestrogen depletion and to tamoxifen is associated with the copy number status of *ESR1*. Heat maps represent the relative viability in oestrogen-depleted medium (top) and in response to tamoxifen (at $16.6\mu\text{M}$; bottom).



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 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 were based on SNVs, gene-level CNAs, arm-level CNAs, gene expression profiles and drug response patterns and were all compared to each other. The Fowlkes–Mallows index (Bk) was computed for all potential numbers of clusters (k values) ranging from 5 to 26. The red lines indicate the observed Bk values, whereas the grey lines represent the 95% upper quantile of the randomized distribution. The maximum Bk value represents the degree of similarity between the compared pair of dendrograms. The grey shading represents the difference between the observed Bk values and those of the 95% upper quantile of the randomized distribution. **b**, Force-directed layout of screened lines using a similarity matrix determined by the probability of cell lines clustering together in dependency space. Cell lines (nodes) are coloured by lineage. **c**, Left, the overlap of dependencies in KPL1 and MCF7 using corrected CERES scores, with genes showing depletion effects in all cell lines (that is, pan-essential genes) excluded. The threshold for dependency was set as a CERES score < -0.5 . Right, 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 GSEA of MCF7 and KPL1 against the oestrogen response gene sets ($n = 1$ sample per group). Expression of the oestrogen signalling pathway is strongly enriched in MCF7. **e**, The correlation between ESR1 dependency values and the single-sample GSEA enrichment scores of the oestrogen response hallmark gene sets ($n = 27$ cell lines). The difference in oestrogen response signalling between MCF7 and KPL1 predicts their differing levels of dependency on ESR1. **f**, The correlation between *GATA3* dependency and *GATA3* protein levels (z -scored values for reverse-phase protein arrays; $n = 27$ cell lines). The difference in *GATA3* protein levels between MCF7 and KPL1 predicts their differing levels of dependency on

GATA3. Spearman's ρ 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 six months of continuous passaging. Box plots show the population doubling time of each sample. Bar, median; box, 25th and 75th percentiles; whiskers, data within $1.5 \times$ IQR of lower or upper quartile; circles, all data points. Two-tailed t -test; n , replicate wells. **h**, Top, comparison of the sensitivity to oestrogen depletion between a parental MCF7 population and its single-cell-derived clones. Bottom, comparison of the sensitivity to oestrogen depletion between two cultures of the same single-cell clone, separated by six months of continuous passaging. Box plots show the relative growth rate in oestrogen-depleted medium. Bar, median; box, 25th and 75th percentiles; whiskers, data within $1.5 \times$ 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 $20 \mu\text{M}$) and the sensitivity to oestrogen depletion (relative growth rate), across the parental MCF7 populations and their single-cell clones ($n = 7$). Spearman's ρ value and P values 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 ultra-low-pass whole-genome sequencing or targeted sequencing), SNV distances, gene expression distances and drug response distances were compared to each other. CNA distance based on ultra-low-pass whole-genome DNA-sequencing 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. Gene expression and drug-response distances were determined by Euclidean distances. Spearman's ρ and P values 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

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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Genetic and transcriptomic distances should be considered when comparing drug response data • Compare drug response data to genomic data from the same strain
Pre-existing heterogeneity within culture underlies cell line instability	Transcriptional differences between sensitive and resistant strains can elucidate compound mechanism of action	<ul style="list-style-type: none"> • Use characterized isogenic-like strains to uncover associations between molecular features and drug response • Confirm the genomic features of single cell-derived clones • Avoid comparisons between bottlenecked cell populations, whenever possible • Keep culture conditions constant
	Single cell-derived clones differ from one another genetically, transcriptionally and phenotypically	
Heterogeneity keeps emerging in culture due to ongoing genomic instability	Subtle differences in culture conditions can lead to changes in cell line clonal composition	<ul style="list-style-type: none"> • Re-confirm genomic features of single cell-derived clones following prolonged passaging • Apply these recommendations more stringently to genomically unstable cell lines
	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	

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

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software were used to collect data in this study.

Data analysis

The following software were used to analyze data in this study: CellProfiler v3.0.0, CellProfiler Analyst v2.2.1, Picard Tools, GATK v3.8, HMMcopy R package, IncuCyte ZOOM v2015A, MuTect v1.1.4, R statistical software v3.3.0, GSEA v3.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size -- for each cell line, as many strains as possible were collected.
Data exclusions	Strain MCF7-M was found to be an outlier strain, consistent with having been through strong bottlenecks, and was therefore excluded from the quantitative analyses presented in the manuscript.
Replication	All attempts at data replication were successful.
Randomization	No randomization was performed.
Blinding	Investigators were not blinded to group allocations.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies against the following human proteins were used: alpha-Tubulin (ab80779; Abcam), PSMC2 (MSS1-104; Enzo Life Sciences) and PSMD1 (C-7; Santa-Cruz), beta-Actin (N-21; Santa Cruz), ER α (F-10; Santa Cruz).
Validation	All antibodies used in this study were validated by the manufacturers, and used in previous publications. Antibodies were validated in this study only by the size of the bands detected by them.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	The following cell lines from the following sources were used in this study: MCF7: from ATCC, CCLE, Clontech, CMap, BHC, Inserm A549: from ATCC, CCLE, CMap, Amon lab, Brugge lab, Meyerson lab, Hahn lab HT29: CMap, Amon lab, Bass lab MDAM453: CCLE, Brugge lab A375: CMap, Johannessen lab VCaP: CMap, Hahn lab PC3: CMap, Garraway lab HCC515: CMap, Minna lab HepG2: CMap, Lindquist lab HeLa: CMap, CCLE
---------------------	---

	Ben-Men-1: Dunn lab HA1E: CMap, Meyerson lab MCF10A: Brugge lab, Klefstrom lab, Yu lab, Golub lab, Horizon Discovery
Authentication	Cell line authentication was performed by DNA Fingerprinting analysis with 44 polymorphic loci.
Mycoplasma contamination	All cell line were confirmed to be negative to mycoplasma using the MycoAlert™ Mycoplasma Detection Kit (Lonza).
Commonly misidentified lines (See ICLAC register)	Cell lines are not in the list of misidentified cell lines, except for KPL-1. This cell line was used precisely because it was reported (and confirmed by us in this study) to be a strain of MCF7.