RESOURCE

nature biotechnology

A comprehensive transcriptional portrait of human cancer cell lines

Christiaan Klijn[1], Steffen Durinck[1,2], Eric W Stawiski[1,2], Peter M Haverty[1], Zhaoshi Jiang[1], Hanbin Liu[1], Jeremiah Degenhardt[1], Oleg Mayba[1], Florian Gnad[1], Jinfeng Liu[1], Gregoire Pau[1], Jens Reeder[1], Yi Cao[1,3], Kiran Mukhyala[1], Suresh K Selvaraj[3], Mamie Yu[3], Gregory J Zynda[1], Matthew J Brauer[1], Thomas D Wu[1], Robert C Gentleman[1], Gerard Manning[1], Robert L Yauch[3], Richard Bourgon[1], David Stokoe[3], Zora Modrusan[2], Richard M Neve[3], Frederic J de Sauvage[2], Jeffrey Settleman[3], Somasekar Seshagiri[2] & Zemin Zhang[1]

Tumor-derived cell lines have served as vital models to advance our understanding of oncogene function and therapeutic responses. Although substantial effort has been made to define the genomic constitution of cancer cell line panels, the transcriptome remains understudied. Here we describe RNA sequencing and single-nucleotide polymorphism (SNP) array analysis of 675 human cancer cell lines. We report comprehensive analyses of transcriptome features including gene expression, mutations, gene fusions and expression of non-human sequences. Of the 2,200 gene fusions catalogued, 1,435 consist of genes not previously found in fusions, providing many leads for further investigation. We combine multiple genome and transcriptome features in a pathway-based approach to enhance prediction of response to targeted therapeutics. Our results provide a valuable resource for studies that use cancer cell lines.

Cancer cell lines have transformed our understanding of human cancer cell biology, from oncogene function to therapeutic sensitivity[1]. They represent a wide range of tumor types and show great diversity in their response to perturbations, thereby providing an important model system to capture the disease heterogeneity that defines the cancer patient population. Deep analysis of cancer cell lines at the genomic level is critical to establishing the oncogenic driver events as well as the mechanisms underlying response to treatments[2,3]. Although copy number and genome resequencing data for a number of cancer lines have been reported[4–10], the transcriptome remains relatively understudied. Most previous efforts have used probe-based hybridization technologies[5,8,11,12], which fail to reveal certain RNA features, such as oncogenic fusion transcripts, noncoding transcripts, non-human RNAs and expressed single-nucleotide variants (SNVs). Any of these features might influence the results of experimental or preclinical studies with these cell lines and could affect their response to investigational agents. Furthermore, the identification of potential oncogenic events in a large panel of cell lines provides an opportunity for rapid characterization of their role in tumorigenesis through functional validation.

Here, we present comprehensive sequence analysis of the poly(A)$^+$ transcriptome of 675 commonly used cancer cell lines with matched SNP array data. Analysis of the RNA-seq data revealed a comprehensive portrait of expression in these cell lines. To illustrate the utility of this resource, we describe the identification of 2,200 gene fusion pairs. We also reveal new gene regulation patterns for the known oncogenes *MET* and *EGFR* and show that by combining gene copy number data, expression data, mutation status and gene fusion information, it is possible to predict the response to clinical compounds including MEK, PI3K and FGFR inhibitors in many cell lines. The availability of this transcriptome and SNP array data can greatly enhance our understanding of drug response in clinically relevant models and thereby expedite the development of effective personalized medicine.

RESULTS

Data collection and characterization

We collected RNA-sequencing and SNP array data for 675 frequently used cancer cell lines (**Supplementary Fig. 1a** and **Supplementary Tables 1**,**2**). Many cell lines, despite their unique names, share a common origin[13]. To identify and resolve genomically similar cell lines, we clustered cell lines based on SNP genotyping data (**Fig. 1a** and **Supplementary Fig. 1b**) and identified 109 cell lines with >85% SNP concordance (median: 74.7% ± 1.8% (median average deviation)) with one or more other lines. After selecting one representative line from the genomically related lines, we retained 610 distinct cell lines (**Fig. 1b** and **Supplementary Fig. 1c**) that were used for further analyses (**Supplementary Table 3**).

We compared our data set with two genomics studies of cancer cell lines—the Cancer Cell Line Encyclopedia (CCLE)[5] and a study published by the Sanger Institute[8] in which gene expression was studied using microarrays. We referred to a previous analysis[14], which directly compared the CCLE and Sanger studies, as a basis for comparison. Besides the 276 lines common to all three studies, our study included 148 additional lines that were not investigated in either the CCLE

[1]Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California, USA. [2]Department of Molecular Biology, Genentech Inc., South San Francisco, California, USA. [3]Department of Discovery Oncology, Genentech Inc., South San Francisco, California, USA. Correspondence should be addressed to J.S. (jeffrees@gene.com) or S.S. (sekar@gene.com) or Z.Z. (zemin@pku.edu.cn).

Received 11 April; accepted 15 October; published online 8 December 2014; doi:10.1038/nbt.3080

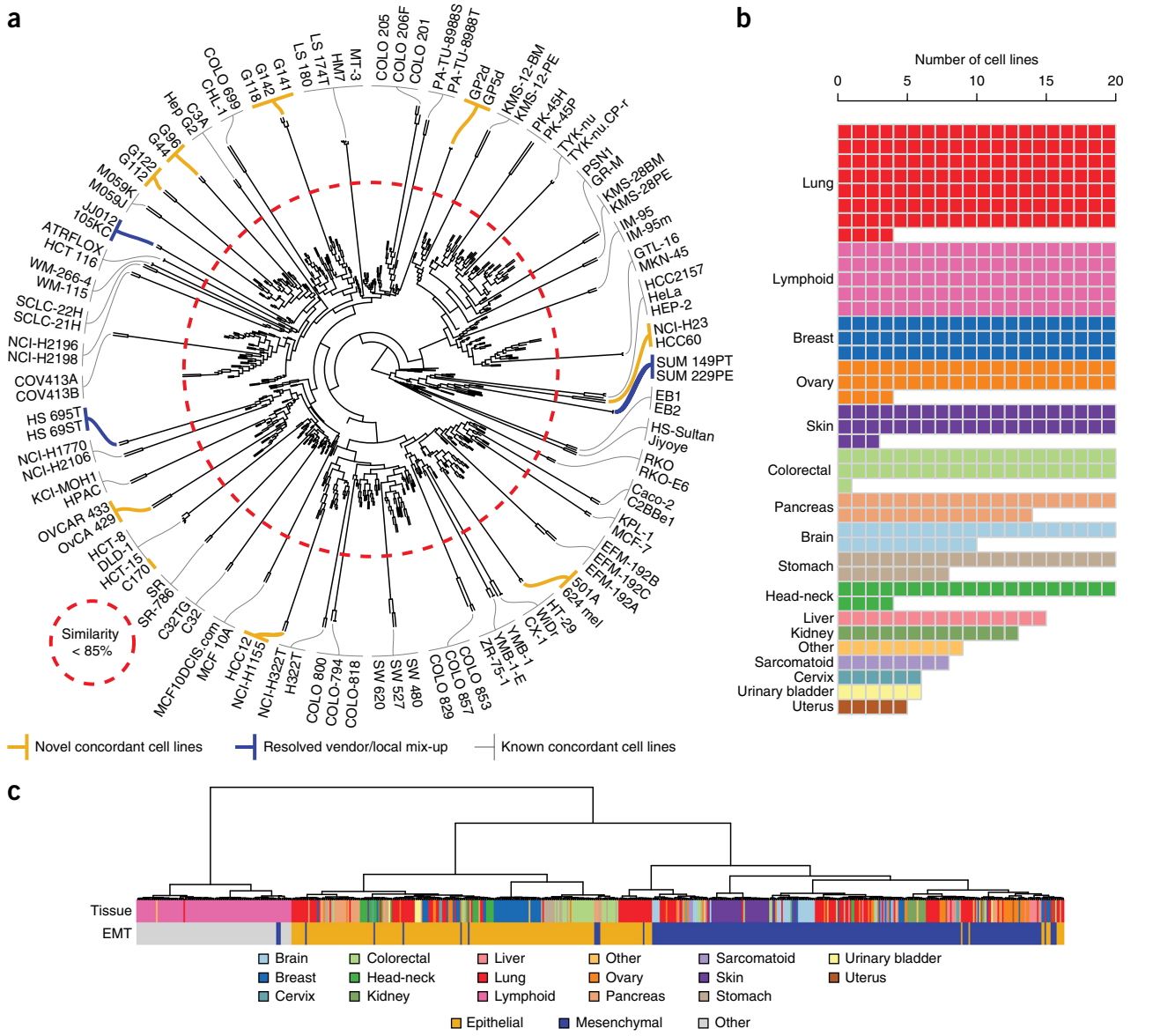

Figure 1 Data set overview. (**a**) Hierarchical clustering of 675 cancer cell lines based on SNPs, visualized in a radial projection with kappa scaling. The distance measure used was 1 – concordance, where concordance is the fraction of SNPs that were shared between cell lines. Clustering was done using complete linkage. The red circle indicates the cutoff for genomic similarity. Cell lines with >85% SNP concordance extend beyond the red dashed line and are named at the terminal leaves and colored to indicate either known cell-line concordance (black), newly described concordance (orange) or a resolved technical mix-up (blue). (**b**) Stacked grid plot showing the distribution of the final data set over the 18 tissue groups into which we divided the data. (**c**) Hierarchical clustering of 610 cell lines based on gene expression values derived from RNA-seq data. Gene expression was represented as variance-stabilized data from the DESeq package in the R programming language, established from gene-based read counts. We used the 1,000 most-variable genes as determined by interquartile range. Clustering was done using Euclidean distance and Ward linkage. The top color bar represents the cell line tissue of origin; the bottom color bar shows whether the line expresses an epithelial or a mesenchymal gene expression signature. Cell lines that could not be assigned epithelial or mesenchymal expression are shown in gray.

or the Sanger study (**Supplementary Fig. 2a**). To our knowledge no other group has published large-scale genomics data for these 148 lines. Gene expression correlation between the overlapping lines was very consistent (median correlation coefficient: CCLE-GNE 0.81, Sanger-GNE 0.79; **Supplementary Fig. 2b**,**c**).

The global analysis of DNA copy number and gene expression of these cell lines can recapitulate many previous findings and reveal novel patterns. GISTIC[15] analysis of the SNP array data identified recurrent copy number variations (CNVs) in known cancer genes, such as the amplification of *MYC* and *ERBB2* and the loss of *CDKN2A* (**Supplementary Fig. 1d**,**e** and **Supplementary Table 4**). The overall CNV patterns were consistent with those seen in primary tumors[16]. Unsupervised clustering of RNA-seq–derived gene expression showed that although cell lines from a particular tissue predominantly clustered together, the lymphoid cells formed a notably distinct cluster, as was previously observed[17] (**Fig. 1c**). Among the 1,000 most-variable genes used for the unsupervised analysis, we noted the presence of canonical epithelial-to-mesenchymal transition (EMT)–related genes such as E-cadherin (*CDH1*), Vimentin (*VIM*) and *ZEB1*. Indeed, there was a significant overlap between the 1,000 most variable genes and a published EMT-derived gene signature ($P < 0.0001$, two-sided Fisher's exact test)[18]. When we divided our cell lines based on the

EMT-derived expression signature[18] (**Supplementary Fig. 3**), we found that brain, liver and kidney cell lines were exclusively mesenchymal, and head-and-neck, bladder and colorectal cell lines were mostly epithelial. When we overlaid E/M classifications with the gene expression correlation network, most cell lines strongly clustered based on E/M status (**Fig. 1c** and **Supplementary Fig. 4**). This suggests that EMT-associated gene expression differences are a major determinant of gene expression–based stratification among cancer cell lines.

Our RNA-seq data also provide a rich source of expressed polyadenylated transcripts that may not code for proteins. Long intergenic noncoding RNAs (lincRNAs) are generally polyadenylated and can be detected in our RNA-seq data. Previously, the antisense noncoding RNA HOTAIR has been associated with EMT[19], but no other noncoding RNAs have been implicated in EMT. We tested whether the lincRNAs from the cell lines in our study were associated with the epithelial or mesenchymal states in a manner similar to what we observed in our analysis of the protein coding genes. For this analysis blood and skin were excluded as they showed confounding lineage-specific lincRNA expression. We identified a group of lincRNAs that were differentially expressed between E and M state (**Supplementary Fig. 5** and **Supplementary Table 5**). Further examination of these transcripts in cancer cell lines may uncover novel roles for noncoding transcripts in disease-relevant processes, such as EMT.

MET, *EGFR*, *ITGA3* and *EPHA2* co-regulated expression network

Our data also provide the opportunity to find relationships between genes based on correlation of their expression. When analyzing cell lines it is relatively straightforward to validate the functional relevance of findings by perturbing the genes involved. To illustrate this approach, we calculated Pearson's correlation of gene-by-gene expression over all cell lines and summarized those relationships in a correlation network. In mining the network for known cancer genes, we found that expression of *EGFR*, *EPHA2*, *ITGA3* and *CAV2* is strongly correlated with expression of the *MET* oncogene (**Supplementary Fig. 6**). Moreover, *MET* and *EGFR* expression was robustly correlated across most tissue types, with the exception of pancreas and head-and-neck cell lines (**Supplementary Fig. 7**).

To determine whether the observed co-expression reflects functional co-regulation, we performed perturbation studies involving *MET*, *EGFR* and their downstream effectors. The expression of *MET*, *EGFR*, *EPHA2* and *ITGA3* was 1.25- to 2.5-fold lower after either short hairpin RNA (shRNA)-mediated knockdown of *MET* or pharmacologic inhibition of *EGFR* (**Supplementary Fig. 8a**). Conversely, we observed a 1.5- to 4-fold increase of expression of these four genes when either *MET* or *EFGR* were stimulated by their ligands in MCF10A cells with a *PTEN* deletion. In MCF10A cells with wild-type *PTEN*, EGFR and MET ligands did not significantly stimulate gene expression (**Supplementary Fig. 8b**), indicating that PTEN acts as a repressor of the observed regulation in these cells.

To determine whether this co-regulation was associated with either the PI3K/AKT/mTOR or MAPK/ERK signaling pathways, which are both engaged downstream of active *MET* and *EGFR*, we treated MCF10A cells with a PI3K inhibitor (GDC-0941 (ref. 20)) or a MEK inhibitor (GDC-0973 (ref. 21)). Both inhibitors caused a two- to fivefold reduction of expression of the four co-regulated genes (**Supplementary Fig. 8a**), suggesting that both pathways regulate these genes. Although signaling crosstalk between these genes was previously described[22,23], our results argue that *MET* and *EGFR* signaling activates a previously undescribed positive-feedback pathway downstream of the PI3K and MEK signaling pathways, increasing

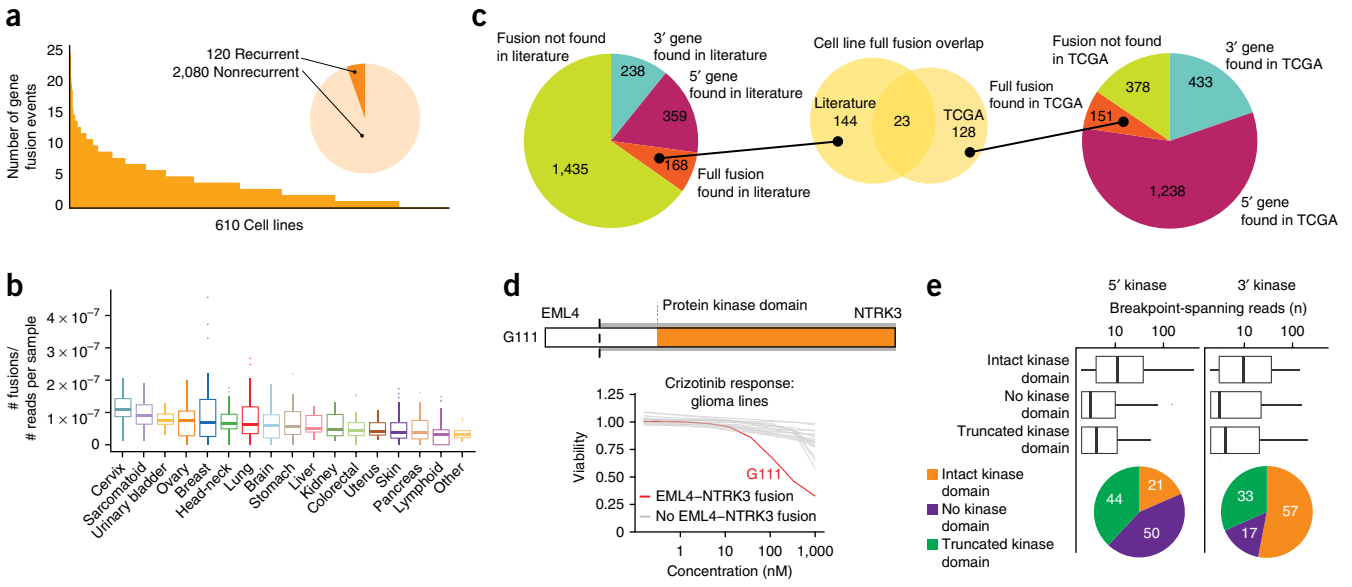

Figure 2 Detection of gene fusion events in human cell lines. (**a**) Histogram of the number of fusion genes per cell line. (inset) Distribution of recurrent (occurring in more than one cell line) and nonrecurrent gene fusions. (**b**) The distribution of fusions per sample, grouped by tissue of origin and corrected for total reads per sample. (**c**) Genes previously known to be involved in gene fusions versus novel fusion genes (left). Genes found in fusions in 6,730 TCGA samples (right). The Venn diagram shows the overlap between fusion gene pairs found in full in the Mitelman literature database (left) and TCGA samples (right). (**d**) Schematic representation of the EML4-NTRK3 fusion found in the G111 glioma cell line. The second figure shows the mean viability of all glioma cell lines included in our study over a crizotinib (PF-2341066) dose treatment range. The G111 cell line carrying the EML4-NTRK3 fusion is highlighted in red. (**e**) Boxplot of fusions with different states of kinase domain versus the number of breakpoint-spanning reads (log-scale), stratified by the position of the kinase in the fusion: as a 5′ partner (left) or a 3′ partner (right). Pie charts show the distribution of kinase domain status for kinase fusions, again stratified on the position of the kinase in the fusion.

Table 1 Fusion genes detected in cancer cell lines

Known fusion	Primary tumors found	Cell line	Cell line tissue	Fusion in cell line	Breakpoint-spanning reads	Model, new or known
BCR ABL1	Lymphoid	SUP-B15	Lymphoid	BCR ABL1	22	Known
		KU812	Lymphoid	BCR ABL1	38	Known
		K-562	Lymphoid	BCR ABL1	76	Known
		HNT-34	Lymphoid	BCR ABL1	40	Known
RUNX1 RUNX1T1	Lymphoid	Kasumi-1	Lymphoid	RUNX1 RUNX1T1	93	Known
EWSR1 FLI1	Sarcomatoid	TC-71	Sarcomatoid	EWSR1 FLI1	49	Known
		MHH-ES-1	Sarcomatoid	EWSR1 FLI1	179	Known
		A-673	Sarcomatoid	EWSR1 FLI1	73	Known
NPM1 ALK	Lymphoid	SU-DHL-1	Lymphoid	NPM1 ALK	66	Known
		SR	Lymphoid	NPM1 ALK	124	Known
MLL MLLT3	Lymphoid	THP-1	Lymphoid	MLL MLLT3	47	Known
CRTC1 MAML2	Salivary Gland	NCI-H292	Lung	CRTC1 MAML2	10	Known
PICALM MLLT10	Lymphoid	U-937	Lymphoid	PICALM MLLT10	7	Known
MLL MLLT4	Lymphoid	OCI-AML2	Lymphoid	MLL MLLT4	4	Known
	Lymphoid	ML-2	Lymphoid	MLL MLLT4	20	Known
PAX3 FOXO1	Sarcomatoid	SJCRH30	Sarcomatoid	PAX3 FOXO1	53	Known
EML4 ALK	Lung	NCI-H2228	Lung	EML4 ALK	17	Known
FGFR3 TACC3	Various	RT-112	Urinary Bladder	FGFR3 TACC3	271	Known
		Hep G2	Liver	FGFR3 TACC3	2	Known
SLC34A2 ROS1	Lung	HCC78	Lung	SLC34A2 ROS1	140	Known
TPM3 NTRK1	Thyroid/colon	KM-12	Colorectal	TPM3 NTRK1	69	Known
BRD4 C15orf55	Various	RPMI 2650	Head-neck	BRD4 C15orf55	79	New
CCT3 C1orf61	Melanoma	HOP-92	Lung	CCT3 C1orf61	2	New
CD74 ROS1	Lung	HCC1493	Breast	CD74 ROS1	59	New
KIFC3 CNGB1	Ovarian	BFTC-909	Kidney	KIFC3 CNGB1	2	New
NPLOC4 PDE6G	Breast	NCI-H716	Colorectal	NPLOC4 PDE6G	8	New
PROM1 TAPT1	Breast	NCI-H146	Lung	PROM1 TAPT1	5	New
SCAMP2 WDR72	Melanoma	A549	Lung	SCAMP2 WDR72	16	New
SLC37A1 ABCG1	Ovarian	HCC1428	Breast	SLC37A1 ABCG1	29	New
TFG ALK	Lung/lymphoma	SCC-3	Lymphoid	TFG ALK	15	New
TNS3 PKD1L1	Ovarian	NCI-H810	Lung	TNS3 PKD1L1	4	New
		NCI-H727	Lung	TNS3 PKD1L1	2	New

their own expression as well as that of *ITGA3* and *EPHA2*. *CAV2* upregulation may be explained by its genomic proximity to the *MET* gene. This example shows that by mining expression in cancer cell lines potentially many more regulation patterns can be discovered.

Expression of viral sequences in cancer cell lines

Presence of virus-derived DNA and RNA in a cell line can be an important indicator of the oncogenic pathway engaged in the associated tumor. Moreover, the presence of foreign genes can influence the results of experimental studies with such models. By mapping unmapped reads to a collection of viral genomes, we found expressed sequences from many known cancer-associated viruses including human papillomavirus and hepatitis-B virus (**Supplementary Fig. 9**). We also found evidence of mouse leukemia virus expression in multiple tissue types, consistent with a previous report[24]. Using our paired-end RNA-seq data, we detected human-virus RNA chimeras involving various human cancer viruses, that potentially result from viral integration into the host genome[25] (**Supplementary Table 6** and **Supplementary Fig. 10**). We also detected the integration of murine viruses into human cell lines (**Supplementary Table 7**), which are likely remnants of prior transfection experiments or persistent contaminants and should be taken into account when designing experiments with these lines.

Oncogenic fusions in cancer cell lines

RNA sequencing allows for the detection of aberrant transcriptomic events, including those that produce fusion transcripts. Many cancer-causing fusion genes have been identified in various tumors[26,27], and fusions such as *BCR-ABL* and *EML4-ALK* have been successfully targeted to yield clinical benefit[28,29]. We identified 2,371 unique in-frame fusion events that had at least two breakpoint-spanning reads and passed the stringent filters in our pipeline, which includes removal of fusions found in three data sets of normal tissues available to us. These fusions represented 2,200 unique pairs of genes, of which 120 were found more than once (**Fig. 2a** and **Supplementary Table 8**). Cell lines in this study carried a median of three fusions (**Fig. 2a**,**b**), similar to what was observed in tumors previously[30]. Because the number of fusions detected correlated with the total number of reads (**Supplementary Fig. 11**), we corrected for the number of reads and still observed a substantial difference among tissues for the number of fusions found per sample. We further validated our fusion study results by comparing our findings to fusions reported previously from MCF-7 and BT-474 cell lines[30,31], and found previously described fusions as well as several novel fusions (**Supplementary Fig. 12**). The most notable findings are shown in **Tables 1** and **2**, and all fusions are reported in **Supplementary Table 8**.

Of the 2,200 gene fusion pairs found, 168 had been previously reported[32] (**Fig. 2c**). We identified 21 known cell line models carrying canonical oncogenic fusions frequently described in primary tumors (**Table 1**). Most of these were associated with many breakpoint-spanning reads, indicating that they are highly expressed. We also identified 11 cell lines not typically used to study known fusions (**Table 1**). These lines have not previously been found to carry these fusions and thus present useful models in which to study their function. For example, we identified RMPI 2650 cells as a new model for the *BRD4-C15orf55* fusion (also known as the *BRD4-NUT* fusion), which is seen in aggressive midline carcinomas, a rare cancer without any effective treatment. Also, we found the *CD74-ROS1* fusion in the HCC1493 breast cancer line. Although cell lines have been identified carrying other *ROS1* fusions, the *CD74-ROS1* fusion had previously been found only in primary tumors. Additionally, we observed less characterized fusions such as *SCL37A1-ABCG1* (ref. 33) and *SCAMP2-WDR72* (ref. 34) that were previously detected in tumors only, so cell lines harboring these fusions provide systems to ascertain whether these are true oncogenic fusions.

We then determined whether fusions detected from The Cancer Genome Atlas (TCGA) consortium data could be detected in our cell line collection. We analyzed 6,730 tumor samples for which RNA-seq data were available from TCGA and found 151 fusions that were present in our cell line data set (**Fig. 2c** and **Supplementary Table 9**). Of these only 23 were previously reported, indicating that we have identified 128 cell lines with, to our knowledge, previously undescribed gene fusions also found in tumors. In total, we identified 296 fusion gene pairs in 232 cell lines that can serve as model systems for studying cell biology and drug efficacy.

Among the fusion pairs we found in cell lines, 359 5′-partners and 238 3′-partners were previously found, in the literature, to have different fusion partners (**Fig. 2c**), and a total of 1,822 of the 2,200 fusions found in cell lines (82.8%) have at least one of the gene partners

Table 2 New fusions partners for known fusion genes and new fusions with kinase genes

Known/kinase fusion gene	Literature tumor tissue	Cell line	Cell line tissue	Fusion	Breakpoint-spanning reads	Fusion gene, known or kinase
ALK	Lung/lymphoid	NCI-H2228	Lung	TRMT61B ALK	2	Known
		KELLY	Brain	TERT ALK	4	Known
BRAF	Various	SK23	Skin	CUL1 BRAF	47	Known
FGFR2	Various	NCI-H716	Colorectal	FGFR2 COL14A1	429	Known
		DMS 79	Lung	FGFR2 COL14A1	5	Known
MAML2	Salivary gland	ES2-TO	Ovary	YAP1 MAML2	4	Known
MAP3K3	Lung	NCI-H1734	Lung	MRC2 MAP3K3	44	Known
MLL	Lymphoid	BJAB	Lymphoid	MLL CLTC	10	Known
MYC	Various	SCLC-22H	Lung	SDF4 MYC	20	Known
NTRK3	Various	G111	Brain	EML4-NTRK3	50	Known
PRKG1	Lung	G59	Brain	SUPV3L1 PRKG1	41	Known
		AN3 CA	Uterus	VGLL4 PRKG1	15	Known
RAF1	Various	OCI-AML2	Lymphoid	MBNL1 RAF1	20	Known
RUNX1	Lymphoid	OCI-M2	Lymphoid	RUNX1 TSPEAR	13	Known
SET	AML	MKN7	Stomach	SET ANP32E	6	Known
		NCI-H2172	Lung	SET ANP32E	3	Known
		COR-L26	Lung	SET ANP32E	2	Known
TYK2	NA	MOLM-16	Lymphoid	ELAVL1 TYK2	54	Kinase
STK24	NA	NCI-H2795	Lung	DOCK9 STK24	35	Kinase
EPHB4	NA	MG-63	Sarcomatoid	CLIP4 EPHB4	19	Kinase
MAP2K2	NA	LS1034	Colorectal	GRHL2 MAP2K2	18	Kinase
GRK4	NA	RT-112	Urinary Bladder	WHSC1 GRK4	13	Kinase
STYK1	NA	MDA-MB-361	Breast	LRP6 STYK1	13	Kinase

NA, not available.

represented in a TCGA fusion (**Supplementary Table 10**). Some oncogenes—such as *MLL*, *ALK*, *ROS1* and the *FGFR* genes—have a wide variety of fusion partners where the partner seems to yield no additional oncogenic properties to the fusion. We identified 17 previously found oncogenic fusion genes such as *ALK*, *FGFR2* and *RAF1*, for which we identified unreported fusion partners (**Table 2**). *ALK* fusions define a clinical target for treatment with crizotinib (Xalkori)[29]. Indeed, all cell lines carrying *ALK* fusions (known and novel) were sensitive to crizotinib treatment (**Supplementary Fig. 13** and **Supplementary Table 11**). We also found two unreported fusions involving classic lymphoid drivers *MLL* and *RUNX1* (*MLL-CLTC* and *RUNX1-TSPEAR*), adding to the repertoire of partners with which these genes are fused. Additionally, we identified a fusion between *EML4* and *NTRK3* in G111, a glioma cell line. *NTRK3* has previously been found in the *ETV6-NTRK3* fusion in breast, sarcomatoid and thyroid cancers, and *EML4* is the classic 5′ fusion partner of *ALK* in lung cancer. However, they have not been previously found together in a fusion, and although *NTRK1* has been found in glioma fusions[35], no *NTRK3* fusions have been previously reported for glioma. To test whether the *EML4-NTRK3* fusion is functional, we examined crizotinib response in all glioma cell lines. Only the G111 line carrying the *EML4-NTRK3* fusion showed sensitivity to crizotinib (**Fig. 2d**), suggesting that *NTRK3* fusions may be relevant therapeutic targets in glioma, a disease with limited treatment options.

Many of the known fusion oncogenes involve protein kinases, which are crucial in many signaling pathways. In total we identified 246 fusions containing kinases, 84 of which retain an intact kinase catalytic domain. We hypothesized that many of the fusions involving an intact kinase domain are likely to be functionally relevant. Indeed, fusions carrying an intact protein kinase domain tended to exhibit relatively high expression levels compared to other fusions, as indicated by the number of breakpoint-spanning reads (**Fig. 2e**). When the kinase protein is the 3′ partner of the fusion, the kinase domain is more likely intact, indicating functional selection. Fusions lacking a complete kinase domain were more frequently found in association with DNA copy number amplifications, possibly implicating them as by-products of the amplification process (**Supplementary Fig. 14**). Because kinase fusions with a retained kinase domain may represent new driver events, we list kinase fusions with strong read support in **Table 2**. It remains to be determined whether these kinases are oncogenic.

Pathway-based drug response prediction

Many current targeted therapies aim to block signaling pathway dysregulation by inhibiting a critical protein in the pathway. As mutations in various genes can dysregulate the same pathway and a given gene can be mutated by different mechanisms, it is important to identify all types of mutation when characterizing dysregulated pathways. To illustrate the utility of our data, we integrated transcriptome and genome information derived from cancer cell lines to enhance prediction of drug response. As an example one can consider the fibroblast growth factor receptor (FGFR) genes *FGFR1*, *FGFR2* and *FGFR3*. Our data revealed that these genes are fused with multiple partners (**Supplementary Fig. 15**). These fusions have been associated with

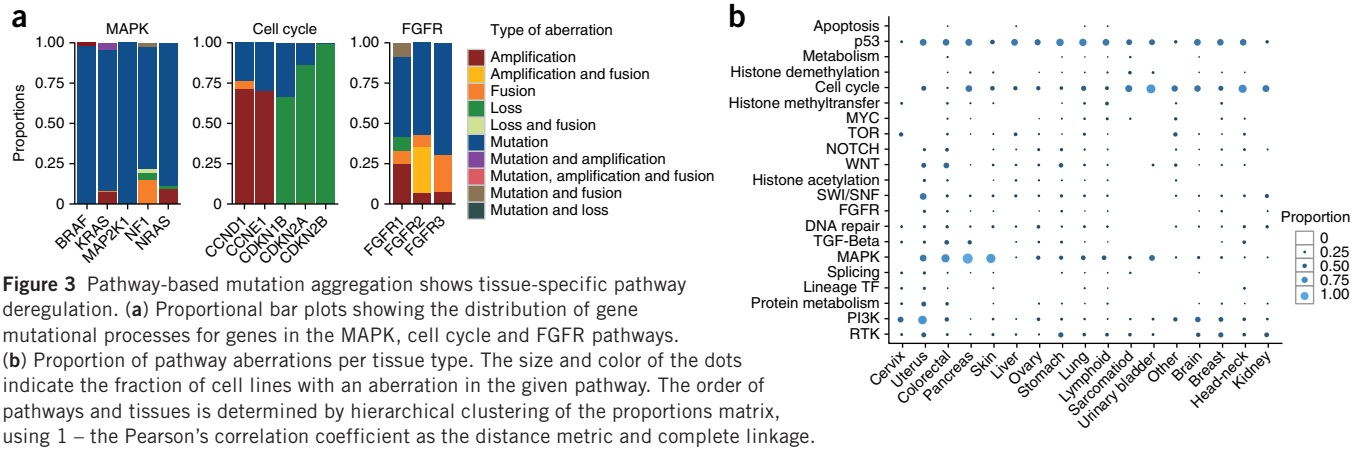

Figure 3 Pathway-based mutation aggregation shows tissue-specific pathway deregulation. (**a**) Proportional bar plots showing the distribution of gene mutational processes for genes in the MAPK, cell cycle and FGFR pathways. (**b**) Proportion of pathway aberrations per tissue type. The size and color of the dots indicate the fraction of cell lines with an aberration in the given pathway. The order of pathways and tissues is determined by hierarchical clustering of the proportions matrix, using 1 – the Pearson's correlation coefficient as the distance metric and complete linkage.

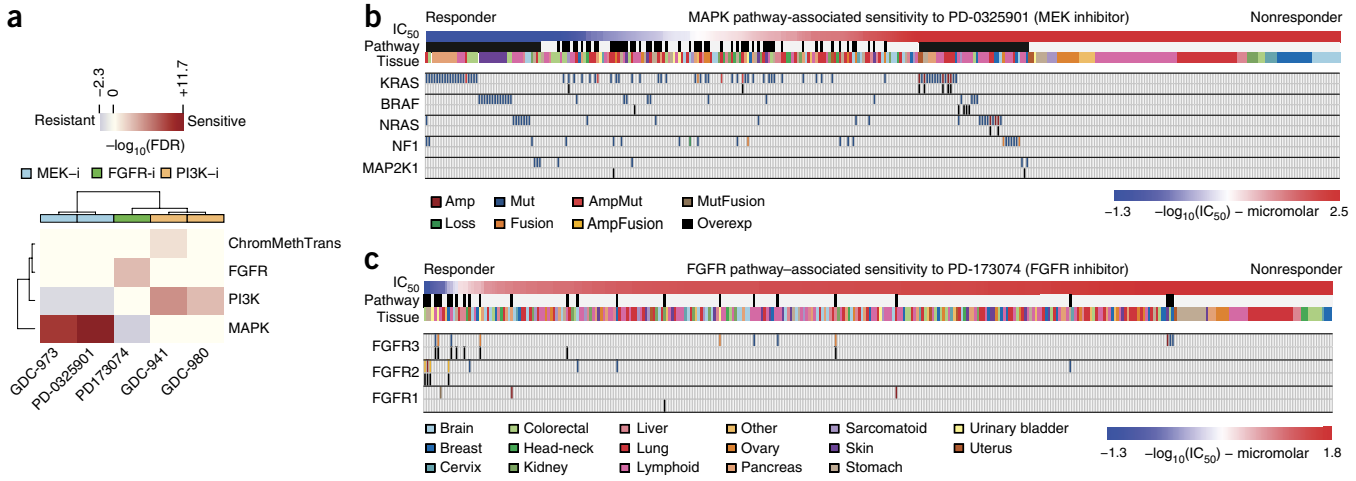

Figure 4 Pathway aggregation of cell line aberrations. (**a**) Heatmap showing the predictive value of the pathway profile (*y* axis) in response to therapeutics. The color of the heatmap shows the $-\log_{10}$ FDR of the Wilcoxon Rank Sum Test on the drug IC_{50} between pathway-aberrant and pathway wild-type cell lines. Nonsignificant pathways (FDR < 0.1) are not shown. The heatmap is constructed by hierarchical clustering using Euclidean distance and Ward's linkage. (**b**) Schematic of the IC_{50} of 350 cell lines for PD-0325901 with annotation for the tissue of origin and the mutational and expression status of the MAPK pathway genes (*KRAS*, *BRAF*, *NRAS*, *NF1* and *MAP2K1*) as determined from RNA sequencing data and Illumina 2.5 M SNP array data. For each gene the upper bar denotes genomic aberrations and the lower bar, overexpression. (**c**) Schematic of the IC_{50} of 350 cell lines for PD-173074 with annotation for the tissue of origin and the mutational and expression status of the FGFR pathway genes (*FGFR1*, *FGFR2* and *FGFR3*), as determined from RNA sequencing data and Illumina 2.5 M SNP array data.

sensitivity to FGFR kinase inhibitors for cell lines harboring such fusions[36]. However, *FGFR* genes are also known to be the target of amplification and mutation, which could also confer sensitivity to inhibition[37]. It therefore seems logical to aggregate different mutations in the different FGFR genes into one pathway-based predictor for drug response (**Supplementary Table 12**), as an alternative to traditional approaches to cancer genome analysis that treat each gene and each mutational mechanism as separate predictors[5,8].

To determine base-level mutations in our data set, we established a process for detecting and filtering variants found in RNA sequencing data that addresses the complication of mapping artifacts and the lack of matched normal samples and identified high-confidence mutations (**Supplementary Fig. 16**). We compared our mutation calls to those found in the CCLE and Sanger data sets and found strong overlap, with mutation rates that were similar to the ones published[38] and with the top COSMIC genes mutated as expected (**Supplementary Fig. 17**). Next, we integrated all mutation, copy number and gene fusion information for a set of pathways. We defined 'pathway' as a set of genes associated with a common signaling process and known to be functionally altered in cancer[39]. For every cell line, we scored each defined pathway as aberrant or normal. Different mutational processes affected the genes in the pathways, again highlighting the need to aggregate these mutations for a meaningful analysis (**Fig. 3a**). Based on the proportion of samples aberrant for each pathway, we noted clear differences in pathway deregulation between different tissues (**Fig. 3b**).

We then examined associations between our data-driven pathway predictors and the response to five targeted drug candidates—MEK inhibitors, GDC-973 (ref. 21) and PD-0325901 (ref. 40), PI3K inhibitors, GDC-980 (ref. 41) and GDC-941 (ref. 20), and the FGFR inhibitor PD-173074 (ref. 42) (**Supplementary Table 13**). Based on the Wilcoxon rank sum test on the IC_{50} values between pathway-aberrated and pathway-neutral cell lines, we found PI3K pathway aberration to be significantly associated with PI3K inhibitor-treatment response (GDC-941 false-discovery rate (FDR) < 10^{-10}, GDC-980 FDR < 0.0019), and MAPK pathway aberrations to be significantly associated with MEK inhibitor treatment (PD-0325901 FDR < 10^{-11}, GDC-973 FDR < 10^{-9}) (**Fig. 4a**). Notably, a deregulated PI3K pathway predicted a significantly worse response to MEK inhibitors. Using the pathway approach we identified almost all cell lines sensitive to MEK inhibitor PD-0325901 (**Fig. 4b**). Overall, the significance of association with IC_{50} sensitivity (the Benjamini & Hochberg–corrected *P*-value of the Wilcoxon rank-sum test) was improved by orders of magnitude compared to the individual gene-based method (**Supplementary Fig. 18**). Similarly, by integrating the various aberrations for the *FGFR* genes (*FGFR1*, *FGFR2* and *FGFR3*), we identified almost all of the FGFR inhibitor–sensitive cell lines (**Fig. 4c**).

DISCUSSION

It is becoming increasingly apparent that cell lines exhibit considerable genomic and transcriptomic diversity comparable to the diversity observed among solid tumors. Comprehensive characterization of existing cell lines will allow selection of appropriate cell line models for biological studies and drug discovery. It is worth noting that many cell lines (~10% of the cell lines reported in this study) are genetically related, despite their distinct names. Additionally, some cell lines are mislabeled or misclassified in the literature or the public databases. These artifacts could be identified only by systematic classification performed on a large scale, involving hundreds of cell lines.

Our study greatly expands the current understanding of human cancer cell lines by cataloguing coding and noncoding RNA expression, mutation, fusion, expression of viral sequences and DNA copy number changes in 675 cell lines. This analysis complements the large-scale clinical sequencing data generated by TCGA (The Cancer Genome Atlas), ICGC (International Cancer Genome Consortium) and others by enabling informed selection of cell lines most appropriate for use in follow-up analysis of discoveries made in clinical samples. The identification of many known and new oncogenic fusions in cell lines provides a rich resource of cell lines for further experimentation.

With the rapidly expanding discovery and development of "rationally targeted" therapeutics, it becomes critical to identify biomarkers

predictive of clinical benefit. We show how integration of various dimensions of genomic and transcriptomic data can improve the prediction of sensitivity to targeted therapeutics. Going forward, incorporation of the vast body of knowledge of pathway aberration into patient assessment will lead to more effective cancer treatment. As cell lines remain among the most widely used models for preclinical evaluation of candidate cancer drugs, our results provide a foundation for many additional discoveries that will enable further cancer studies and the successful development of novel therapeutics.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. European Genome-phenome Archive (EGA): EGAS00001000610. ArrayExpress accession number E-MTAB-2706.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Genentech cell line bank (gCell) and the compound screening group (gCSI) for contributing cell lines and results to this paper. We thank A. Bruce for graphical assistance.

AUTHOR CONTRIBUTIONS

C.K., F.J.d.S., J.S., S.S. and Z.Z. conceived the project. C.K., J.S., S.S. and Z.Z. wrote the manuscript. C.K., S.D., E.W.S., P.M.H., Z.J., H.L., J.D., O.M., F.G., J.L., G.P., J.R., K.M., G.J.Z., M.J.B., T.D.W., R.C.G., G.M. and R.B. performed bioinformatics data analysis or provided computational infrastructure. Y.C., S.K.S., M.Y., R.L.Y., D.S., Z.M. and R.M.N. prepared cell lines and performed biochemical experiments including drug treatments and sequencing.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. 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).
2. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
3. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **24**, 236–244 (2000).
4. Abaan, O.D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.* **73**, 4372–4382 (2013).
5. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
6. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
7. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
8. Garnett, M.J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
9. Liu, J. *et al.* Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome Res.* **22**, 2315–2327 (2012).
10. Neve, R.M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).
11. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).
12. Ross, D.T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**, 227–235 (2000).
13. American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell line misidentification: the beginning of the end. *Nat. Rev. Cancer* **10**, 441–448 (2010).
14. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
15. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
16. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
17. Lukk, M. *et al.* A global map of human gene expression. *Nat. Biotechnol.* **28**, 322–324 (2010).
18. Taube, J.H. *et al.* Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. USA* **107**, 15449–15454 (2010).
19. Pádua Alves, C. *et al.* Brief Report: The lincRNA Hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. *Stem Cells* **31**, 2827–2832 (2013).
20. Folkes, A.J. *et al.* The identification of 2-(1H-Indazol-4-yl)-6-(4-methanesulfonyl-piperazin-1-ylmethyl)-4-morpholin-4-yl-thieno[3,2-d]pyrimidine (GDC-0941) as a potent, selective, orally bioavailable inhibitor of class I PI3 kinase for the treatment of cancer. *J. Med. Chem.* **51**, 5522–5532 (2008).
21. Hoeflich, K.P. *et al.* Intermittent administration of MEK inhibitor GDC-0973 plus PI3K inhibitor GDC-0941 triggers robust apoptosis and tumor growth inhibition. *Cancer Res.* **72**, 210–219 (2012).
22. Lai, A.Z., Abella, J.V. & Park, M. Crosstalk in Met receptor oncogenesis. *Trends Cell Biol.* **19**, 542–551 (2009).
23. Acunzo, M. *et al.* Cross-talk between MET and EGFR in non-small cell lung cancer involves miR-27a and Sprouty2. *Proc. Natl. Acad. Sci. USA* **110**, 8573–8578 (2013).
24. Lin, Z. *et al.* Detection of murine leukemia virus in the Epstein-Barr virus-positive human B-cell line JY, using a computational RNA-seq-based exogenous agent detection pipeline, PARSES. *J. Virol.* **86**, 2970–2977 (2012).
25. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22**, 593–601 (2012).
26. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660–664 (2012).
27. Singh, D. *et al.* Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science* **337**, 1231–1235 (2012).
28. Druker, B.J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).
29. McDermott, U. *et al.* Genomic alterations of anaplastic lymphoma kinase may sensitize tumors to anaplastic lymphoma kinase inhibitors. *Cancer Res.* **68**, 3389–3395 (2008).
30. Robinson, D.R. *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.* **17**, 1646–1651 (2011).
31. Edgren, H. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* **12**, R6 (2011).
32. Mitelman, F., Johansson, B. & Mertens, M. (eds.) *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer* http://cgap.nci.nih.gov/Chromosomes/Mitelman (2013).
33. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7**, e1001138 (2011).
34. Berger, M.F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).
35. Shah, N. *et al.* Exploration of the gene fusion landscape of glioblastoma using transcriptome sequencing and copy number data. *BMC Genomics* **14**, 818 (2013).
36. Wu, Y.-M. *et al.* Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discov.* **3**, 636–647 (2013).
37. Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. *Nat. Rev. Cancer* **10**, 116–129 (2010).
38. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
39. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
40. Barrett, S.D. *et al.* The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. *Bioorg. Med. Chem. Lett.* **18**, 6501–6504 (2008).
41. Sutherlin, D.P. *et al.* Discovery of a potent, selective, and orally available class I phosphatidylinositol 3-kinase (PI3K)/mammalian target of rapamycin (mTOR) kinase inhibitor (GDC-0980) for the treatment of cancer. *J. Med. Chem.* **54**, 7579–7587 (2011).
42. Mohammadi, M. *et al.* Crystal structure of an angiogenesis inhibitor bound to the FGF receptor tyrosine kinase domain. *EMBO J.* **17**, 5896–5904 (1998).

ONLINE METHODS

Collection of cell lines. For each cell line, we catalogued the primary tissue of origin, the cancer subtype and the site of extraction, as well as the sex, age and ethnicity of the donor where available. We grouped tissues into classically defined tissue types (such as rectal, colon and ileum were grouped into colorectal). For each cell line we established a canonical alphanumeric name. Cell line growth conditions are detailed in **Supplementary Table 2**. We used the Qiagen AllPrep DNA/RNA kit was used to prepare RNA and DNA, according to the manufacturer's protocol. RNA integrity was assessed using the Bioanalyzer (Agilent).

RNA-seq data. RNA libraries were made with the TruSeq RNA Sample Preparation kit (Illumina) according to the manufacturer's protocol. The libraries were sequenced on an Illumina HiSeq 2000, using one to four lanes per cell line. We generated a median of 61 million reads per sample, of which we were able to map a median of 49 million reads uniquely to the human genome and concordant with established gene models. Reads were trimmed to 75 bp and filtered for quality and rRNA contamination. Genomic alignment was performed using GSNAP version 2011-12-28 (ref. 43). GSNAP parameters used were default, with the exception of the following flags: –i 1 –M 2 –n 10 –N 1 –E 1. Further read trimming was not performed, reads for which 30% of all bases were of Phred quality 23 or lower were discarded. Multimapping reads were discarded. RNA variant calls were made using the *VariantTools* package available from *Bioconductor*[44] for the R programming language.

SNP array data. Illumina HumanOmni2.5_4v1 arrays were used to assay 668 cell lines for genotype, DNA copy and LOH at ~2.5 million SNP positions according to manufacturer's protocols. A subset of 2,295,239 high-quality SNPs was selected for all analyses. These SNPs were concordant in >17/20 cell lines assayed by both Illumina HumanOmni2.5_4v1 array and Complete Genomics full-genome sequencing[26]. We applied a modified version of the PICNIC[45] algorithm to estimate total copy number and allele-specific copy number and LOH. The modifications to the PICNIC algorithm are described in **Supplementary Note 1**. We corrected absolute copy number calls for the ploidy of the cell lines, as detailed below.

Data availability. The raw data (SNP array and RNA sequencing) has been submitted to the European Genome-phenome Archive (EGA) under accession number EGAS00001000610. Quantified gene expression information from the RNA-seq data—as described in the 'gene expression' heading—is available in the ArrayExpress database under accession number E-MTAB-2706. Additional intermediate data (RNA-seq counts, per-gene copy numbers and all single-nucleotide mutations) can be found at the site accompanying this publication (http://research-pub.gene.com/KlijnEtAl2014/), as well as in **Supplementary Data 1–4**.

Concordance between cell lines. We first used the SNP array data to calculate SNP concordance between the 668 cell lines. Concordance was defined as the percentage of identical SNP calls between two samples. Hierarchical clustering was applied using 1 – concordance as distance measure. Any cell lines with >85% concordance were manually examined to determine how to resolve the similar pairs. In case of a known relationship between the cell lines, we prioritized the cell lines that were closest to the primary tumor. We discarded all known contaminated cell lines, cell lines that were made resistant to specific treatments *in vitro*, and cell lines derived from metastases as opposed to primary tumors. The pairs of cell lines that were highly similar, but annotated with a different tissue were discarded. For cell lines for which we could not find a known relationship and that originated from the same samples, we kept the line with the highest coverage of mapped RNA-seq reads. Second, we used called variants from RNA-seq data (see the mapping and calling section of this document) that overlapped with dbSNP version 132 and for which we had at least five high-quality reads. Concordance between two cell lines was defined as the percentage of shared variants of the union of dbSNP variants found in the two cell lines. Hierarchical clustering was again applied using 1 – concordance as distance measure.

External data comparison. A comparison between the CCLE and Sanger data was published previously[14]. We used those results as the basis for our comparison. Through manual curation we found an additional five cell lines that overlapped between the CCLE and Sanger. For mutation comparisons we restricted ourselves to the genes tested in the Sanger data set. We treated missing values as wild-type calls.

GISTIC analysis. Genomic regions with recurrent DNA copy gain and loss were identified using GISTIC, version 2.0.12 (ref. 15). Segmented integer total copy number values obtained from PICNIC, c, were converted to $\log_2$ ratio values, y, as $y = \log_2(c + 0.1) - 1$. Cutoffs of ±0.3 were used to categorize $\log_2$ ratio values as gain or loss, respectively. A minimum segment length of 20 SNPs and a $\log_2$ ratio "cap" value of 3 were used. Chromosomes X and Y were not included in analysis to avoid spurious deletion calls. Genes were matched to GISTIC results using the locations of Entrez Genes on the hg19 genome build. We excluded GISTIC peaks with more than ten known germline copy number variations (CNVs) from the analysis.

Gene expression analysis. Per-gene RNA counts were retrieved from GSNAP. For all subsequent analyses we used variant stabilized data as produced by the *DESeq R*-package[46] with *method* = *'blind'* and *fitType* = *'local'*. For all gene expression analyses, we used the DESeq size factor correction to account for differences in sequencing depth between the samples. Hierarchical clustering used Euclidean distance and Ward's linkage. Correlation networks were built with the *qgraph* R-package. Only edges of correlation coefficient >0.7 are shown, and the network layout is automatically determined using the Fruchterman-Reingold algorithm. We classified cell lines as epithelial or mesenchymal by the unsupervised hierarchical clustering using the gene expression signature as described by Taube *et al.*[18]. We were able to match 220 genes from the signature to genes in our expression data. Overlap with the gene signature was calculated using all expressed genes with appreciable variance (gene IQR > 0.25). Three clusters were chosen by manual inspection, with labels of the cluster determined by overexpression of either known epithelial marker genes, mesenchymal marker genes or absence of either.

LincRNA analysis. We took per gene read count values for all transcripts with biotype 'lincRNA' from the Ensembl database version 67 that did not overlap a coding gene. We used DESeq to calculate differential gene expression between epithelial and mesenchymal cell lines for all tissues except lymphoid and skin lines. We kept all lincRNAs that were significantly expressed at multiple testing corrected P-value < 0.01.

Correlation analysis. We calculated the Pearson's correlation matrix between genes, then extracted genes correlating with MET at rho > 0.7. For each gene we set the expression in the nontreated sample to 1 and calculated the corresponding fold change in the treated sample.

Perturbation experiments. H441 and EBC1 cell lines were engineered to express an inducible MET-targeting short-hairpin (sh)RNA through retroviral-mediated gene delivery. An oligonucleotide (5′-GATCCCC**GAACAGAATC ACTGACATA**TTCAAGAGA**TATGTCAGTGATTCTGTTC**TTTTTTGGAA A-3′; bold text signifies the target hybridizing sequence) encoding an shRNA sequence against MET was cloned into BglII/HindIII sites of the pShuttle-H1 vector downstream of the H1 promoter. This plasmid was recombined with the retroviral pHUSH-GW vector using Clonase II enzyme (Invitrogen, Carlsbad, California), generating constructs in which shRNA expression was controlled by an inducible promoter. GP-293 packaging cells were co-transfected using FuGene 6 Transfection Reagent (Promega Corporation, Madison, Wisconsin) and CalPhos Mammalian Transfection Kits with the pVSV-G retroviral vector (both Clontech Laboratories, Mountain View, California) and the above recombinant retroviral constructs. Medium containing the recombinant virus was then added to EBC1 and H441 cells, and recombinants were selected in puromycin. H441 cells were seeded in 50:50 F:12/DMEM 10% FBS. Plates were incubated overnight and media was replaced with fresh media containing either 10% FBS or 0.1% FBS ± 200 ng/ml doxycycline. Cells were then incubated for an additional 4 d. Cells were then trypsinized, spun down, and pellets frozen at −80 °C. Pellets were processed for RNA using a Qiagen RNeasy Kit.

Erlotinib was synthesized at Genentech. Normal adult human epidermal keratinocytes (HEKa) were obtained from Invitrogen. Cells of low passage

number (<4) were seeded onto 15-cm tissue culture dishes with complete keratinocyte media (Invitrogen) at a density of 1 million per dish. This density was previously determined to yield sufficient, yet subconfluent, numbers of cells for the length of the experiment. Cells were incubated at 37 °C and 5% CO_2 and allowed to adhere for 2 d before proceeding. After 2 d, fresh media containing 2 μM of erlotinib was added to the plates in three replicates. Cells were incubated with the inhibitor for 4 d. After 4 d, cells were lysed and RNA was extracted using the Qiagen RNA-easy kit.

Both the H441 and HEK experiments were analyzed on hgu133plus2 microarrays (Affymetrix) according to manufacturer's specifications.

MCF10A and MCF10A-PTEN$^{-/-}$ cells were grown in DMEM/F12 medium with 2% charcoal dextran–treated FBS, 20 ng/ml EGF, 0.01 mg/ml insulin and 500 ng/ml Hydrocortisone under adherent conditions. Cells were treated for 6 h with either 3 μM TGFa (R&D Systems), 50 ng/ml HGF (R&D Systems), 1 μM GDC-0941 or 1 μM GDC-0973 (both synthesized in-house). RNA was extracted using the Qiagen RNA-easy kit. The RNA was sequenced as described above.

Viral sequence detection. The raw paired-end reads from transcriptome sequencing were mapped to the hg19 human reference genome build using GSNAP[43]. The unmapped reads were re-aligned by GSNAP to a comprehensive viral reference database containing 3,279 viral genomes obtained from the NCBI Genome Resources (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239). The normalized viral reads coverage was summarized as RPKM (reads per kilobase of viral genome per million reads). For viral reference genomes that share substantial sequence similarity (>100 BLAST bit score), we only kept one viral genome with the highest read coverage as the representative reference. We eliminated viral genomes with <5 RPKM to retain only confident viral calls. Finally, we only represented samples that have at least one viral genome detected at >2 RPKM. To identify viral-human fusions, we searched for chimeric read pairs where one read mapped to a viral sequence and the partner read mapped uniquely to a human genomic location. Adjacent chimeric reads were clustered to obtain nonredundant fusion events. Clusters with over ten chimeric reads were considered as putative human-viral fusion events. We removed sequence for phage Phi X174 as it is often used as a control lane in sequencing runs. For this analysis we did not choose a representative strain among sequence-similar viral genomes. Instead, for each unique human integration coordinate we kept the viral genome with the highest normalized chimeric read count.

Gene fusions. We analyzed the RNA-seq data to find chimeric transcripts with reads spanning the breakpoint between the two distinct transcripts as previously described[26]. In brief: paired-end reads were aligned using our alignment program GSNAP[43]. GSNAP has the ability to detect splices representing translocations, inversions and other distant fusions within a single read end. These distant splices provided one set of candidate fusions for the subsequent testing stage. The other set of candidate fusions was derived from unpaired unique alignments, in which each end of the paired-end read aligned uniquely to a different chromosome. We also derived candidates from paired, but discordant unique alignments, in which each end aligned uniquely to the same chromosome. For these we required an apparent genomic distance that exceeded 200,000 bp or genomic orientations that suggested an inversion or scrambling event.

Candidate fusions were then filtered against known transcripts from RefSeq and aligned to the genome using GMAP[47]. We required that both fragments flanking a distant splice, or both ends of an unpaired or discordant paired-end alignment, map to known exon regions. We further eliminated candidate inversions and deletions that suggested rearrangements of the same gene, as well as apparent read-through fusion events involving adjacent genes in the genome.

For the final analysis we retained only chimeric transcripts for which we had two or more breakpoint-spanning reads. We excluded any chimeric transcripts that we also detected in a set of in-house normal samples from gastric, colon and lung tissue. We kept only gene pairs that were predicted to be in-frame. We used fusions reported in the Mitelman database (http://cgap.nci.nih.gov/Chromosomes/Mitelman, version current on March 13, 2014) as a source for previously identified fusions.

We processed the fusions reported in two studies that previously reported fusions found in MCF-7 and BT-474 (refs. 30,31). We overlapped our fusion results with the previously published fusion on a gene-pair basis. For fusion detected in the previous studies we cross-referenced our prefiltered list of fusion candidates. We used the Mitelman database as described above to find additional studies that described fusion only found in our study in the three-way comparison.

We downloaded raw FASTQ data for RNA-seq data of 6,730 TCGA tumor samples and 633 normal samples. Gene fusions were detected in the tumor samples as described above with the addition that we removed fusions that we also detected in two or more TCGA normal samples, and we removed fusion pairs that were between genes where the gene symbol was an identical three-letter string, such fusion pairs were found in 40 or more samples. For the final set we only kept TCGA fusions where one of the genes was also found in the cell line data set.

To correct the amount of fusions found per cell line for the overall read coverage in a sample, we divided the number of fusions by the number of sequencing reads generated.

Variant calling from RNA-seq data. We applied a two-step filtering approach. Basic filtering for quality and known SNPs was performed first, and advanced filtering was performed next to generate a high-quality variant set.

We required at least four reads of quality > 23 and at least 20% variant frequency. We did not allow multiple base changes per locus in the same sample. We used a strand bias filter based on the Fishers' Exact test modified from the *VariantTools* implementation. We constructed a 2 × 2 table for reference and variant alleles and plus and minus strands. We automatically passed any variant with >1 plus-strand and >1 minus-strand reads, and we failed any other allele for which the two-tailed Fishers' Exact $P < 0.05$. A panel of sources of normal variant was used to filter known germline variants. We used dbSNP v. 132, variants from the 1000 Genomes Project, 6,515 exomes as published by the NHLBI[48], 69 normal genomes made public by Complete Genomics[49], a set of in-house normal called from gastric, colorectal and lung normal tissue and variants called using our pipeline from the Illumina BodyMap RNA data set (GEO accession GSE30611). During germline SNP filtering we whitelisted any variant that was both present in the Catalogue of Somatic Mutations in Cancer (COSMIC) v. 57 and found in at least two independent samples in COSMIC.

To further reduce the variant count per sample and generate a high-confidence list of variants, we applied several additional filtering steps. First, we removed nongenic and silent variants as well as variants mapped to chromosomes other than the 22 autosomes and the two sex chromosomes. We then eliminated any variant with a higher sample frequency as the KRAS V12 variant, as we are unlikely to find a higher-penetrance mutation and other high-frequency variants are probably germline variation. During subsequent steps we whitelisted any variant that was both present in COSMIC v. 57 and found in at least two independent samples in COSMIC. We trained all thresholds to minimize COSMIC gene filtering while removing obvious artificial variants. We set a threshold on the number of unique mapping positions for reads reporting a variant. With seven or more reads reporting a variant we require at least three unique mapping positions among the reads. We eliminated HLA genes as they tend to accumulate many germline variants. To confirm the mapping for the remaining variant reads we remapped them to the genome using the GMAP algorithm (version 2012-07-20)[47]. We removed any variant that had 25% or more reads that mapped outside the original location. To eliminate variants reported due to minute contaminations with mouse sequence probably caused by sequencer carry-over we remapped variant-bearing reads for all remaining variants to either the human genome (hg19) or the mouse genome (mm9) using the GMAP algorithm (version 2012-07-20)[47]. We removed reads that mapped to mouse with fewer mismatches.

Drug response screening. Cells were maintained in RPMI-1640, 10% FBS (heat-inactivated FBS for suspension lines), and 2 mM glutamine. Before plating, adherent lines were trypsinized and suspension lines were resuspended. Cells were assessed with a Vi-CELL Cell Viability Analyzer; viability of at least 90% was required for screening. Cells were diluted to a previously determined, line-specific level intended to achieve 75% confluency at 96 h. After dilution, a Thermo Multidrop Combi Reagent Dispenser was used for plating cells into

doi:10.1038/nbt.3080

barcoded Falcon 384 black clear-bottom plates. Serial drug dilutions (9 doses per drug) were performed using an Oasis Liquid Handling Robot. After 72 h, 25 µl CellTiter-Glo was added using a Biotek Multiflo Microplate Dispenser. Cell lysis was induced by mixing for 2 min on an orbital shaker, and plates were incubated at room temperature for 10 min to stabilize luminescent signal. Luminescence readout was done with a 2104 EnVision Multilabel Reader.

Three to four independent biological replicates were produced; same-plate technical replication was observed to have negligible effect on outcome and therefore omitted. Per-dose viability was computed relative to median RLU from undrugged wells physically near drugged wells on the plate to reduce noise and the impact of spatial effects. To minimize the impact of isolated aberrant wells, a four-parameter log-logistic model relating viability to log-dose was fit by minimizing the mean of residual absolute value (as an alternative to standard least square). Cases with a high degree of variability across biological replicates were flagged and excluded from further analysis. In some cases, grossly aberrant runs were also dropped, leaving less than three replicates. Fitted viability statistics for retained biological replicates were integrated by maximum likelihood and a hierarchical model, which accounted for both within-run noise and cross-run biological variability. This approach induces weighted averaging of runs, with weight inversely proportional to within-run noise level. Reported IC_{50} is dose at which cross-run estimated inhibition is 50% relative to undrugged wells, that is, absolute IC_{50}.

The crizotinib data were generated in a separate run (**Supplementary Table 10**). The drug was dosed starting from a 1 nM concentration and serially diluted down using threefold serial dilution.

Drug response data integration. We assembled multiple genomic features into pathway aberration models based on previously described pathway genes known to be mutated in cancer[39]. We used all fusions present in our fusion results that involved any of the pathway genes. Amplification and deletion of a gene was defined as > 1 or < -0.75 of the ploidy-corrected copy number, respectively. Gene mutations were only included in the pathway integration if they were present in COSMIC, caused a premature stop codon or were predicted to have a functional deleterious effect according to Condel[50]. Overexpression of a gene was defined as 4 s.d. or higher than the mean expression. We treated activating (amplification, mutation) and inactivating (deletion, mutation) identically when constructing the pathway aberration model. Association of pathway aberration with drug response was calculated using the two-sided Wilcoxon Rank Sum test as implemented in R between the IC_{50} of pathway aberrated lines and pathway neutral cell lines. Multiple testing correction was performed using the Benjamini-Hochberg algorithm.

43. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
44. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
45. Greenman, C.D. *et al.* PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164–175 (2010).
46. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
47. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
48. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
49. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
50. González-Pérez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011).