CORRECTION NOTICE

Nat. Biotechnol. doi:10.1038/nbt.3080; 8 December 2014

A comprehensive transcriptional portrait of human cancer cell lines

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

In the version of the Supplementary Data 1, 2 and 4 originally posted online, a column containing necessary identifiers was omitted in each. The error has been corrected in this file as of 26 January 2015.

A COMPREHENSIVE TRANSCRIPTIONAL PORTRAIT OF HUMAN CANCER CELL LINES

Supplemental Information

Christiaan Klijn[1], Steffen Durinck[2], Eric 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 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,*]

1. Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA 94080, USA

2. Department of Molecular Biology, Genentech Inc., South San Francisco, CA 94080, USA

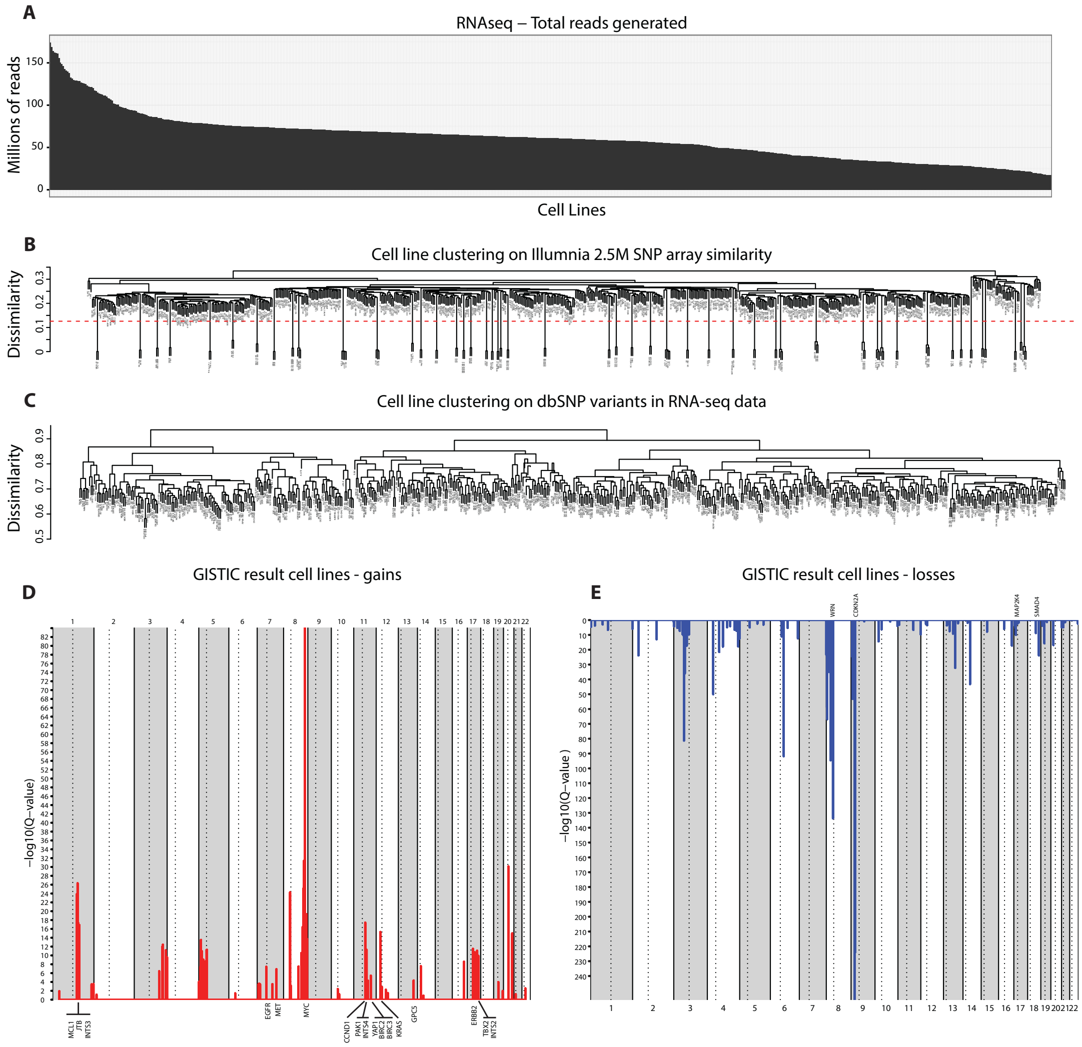
3. Department of Discovery Oncology, Genentech Inc., South San Francisco, CA 94080, USA

Nature Biotechnology: doi:10.1038/nbt.3080

Table of Contents

Supplemental Figures

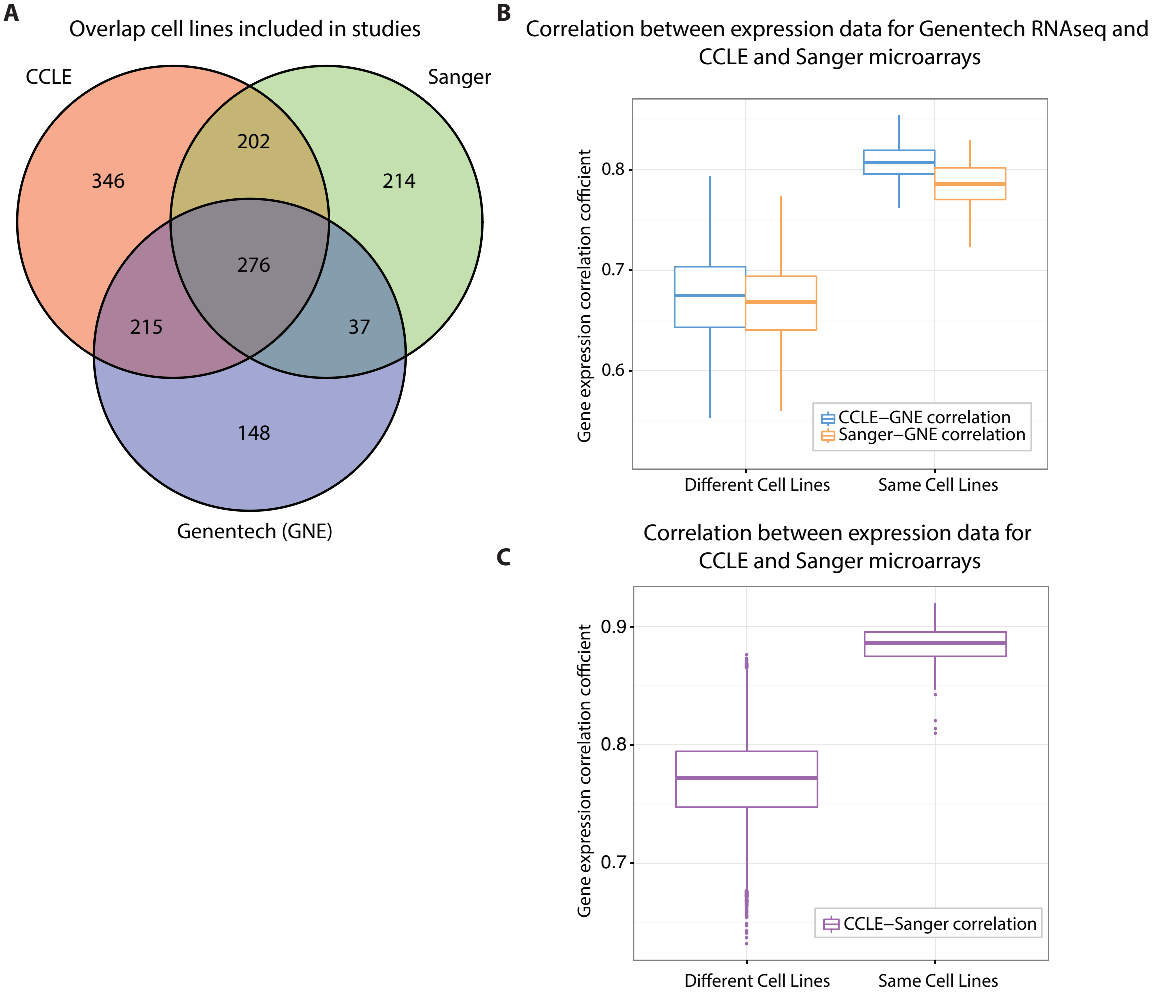
Supplementary Figure 1 – Data generation and SNP-array GISTIC analysis



A. Histogram of total RNA-Seq reads generated per cell line. **B.** Hierarchical clustering of 676 cancer cell lines based on SNP concordance, as in Figure 1. **C.** Concordance of 610 cell lines remaining after SNP array filtering, using variants detected from RNA-Seq found in the dbSNP database version 132. GISTIC results for either copy number gains (**D.**) or copy number losses (**E.**). The y-axis shows the GISTIC q value and the x-axis the genomic location. Candidate driver genes are annotated.

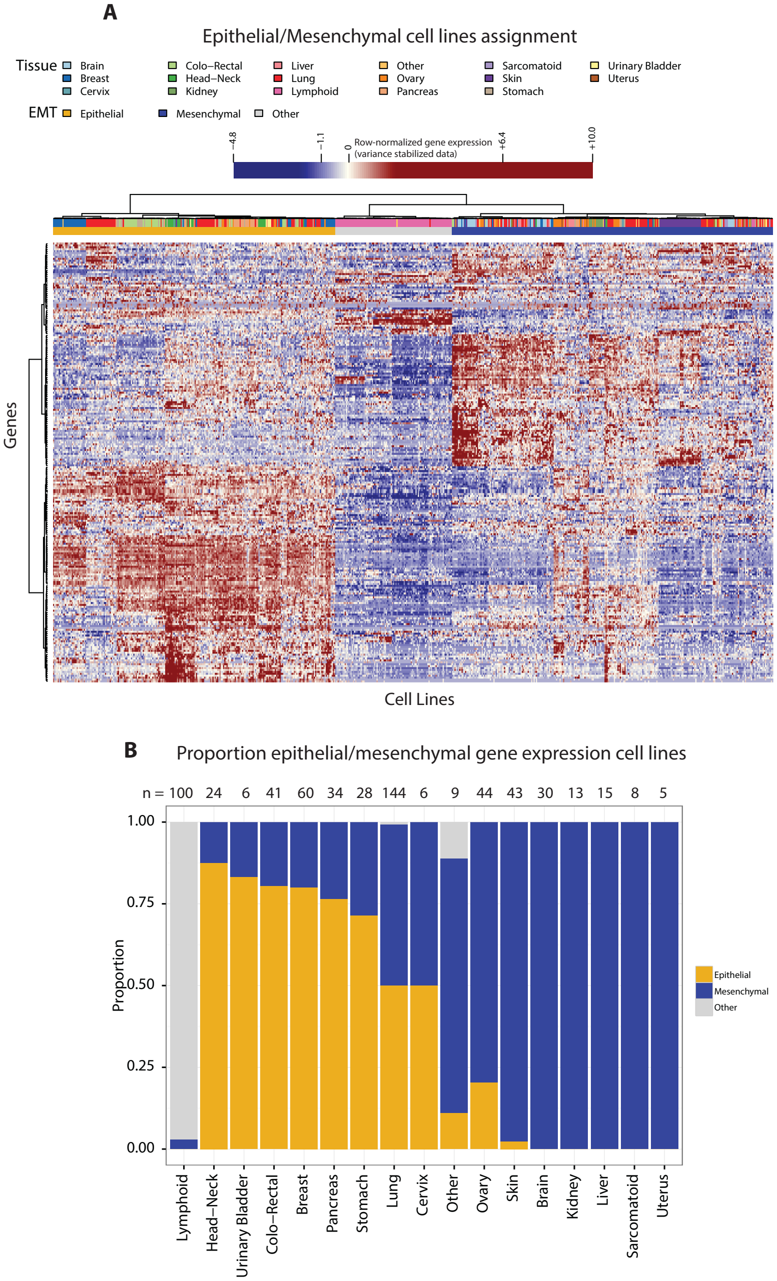
Nature Biotechnology: doi:10.1038/nbt.3080

Supplementary Figure 2 – Cell line and gene expression overlap with previous studies



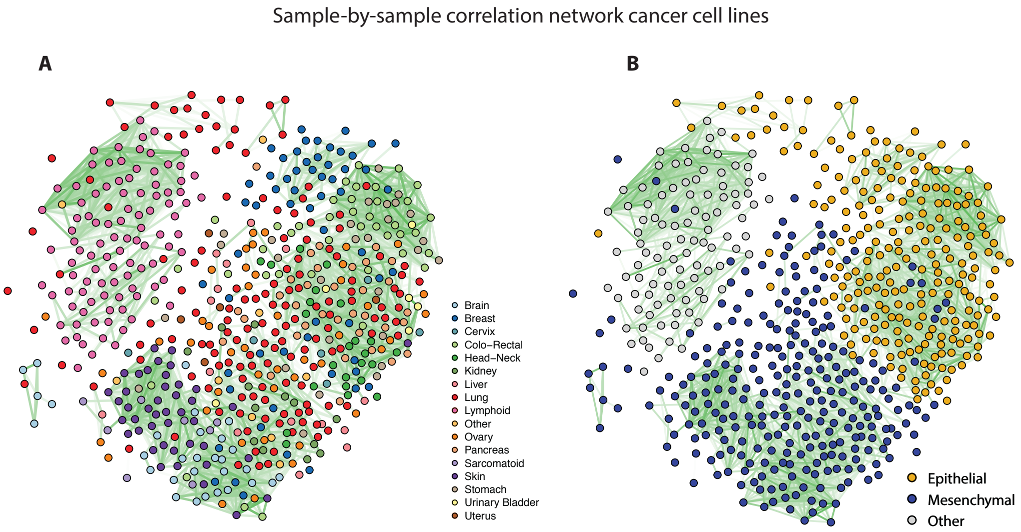
A. Venn diagram showing the cell line inclusion in our study (GNE) and two previously published studies. These overlaps were based on the analysis of Haibe-Kains (Nature, 2013), with 5 additional overlapping lines between CCLE and Sanger that were missed by Haibe-Kains et al. **B.** Boxplots showing gene expression correlation coefficients for the 276 overlapping cell lines between the CCLE and GNE (blue) and the Sanger lines and GNE (yellow). **C.** Boxplots showing gene expression correlation coefficients for 478 overlapping cell lines between the CCLE and Sanger lines.

Supplementary Figure 3 – Gene expression analysis of cancer cell lines



A. 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 1000 most variable genes as determined by inter-quartile range. Clustering was done using Euclidean distance and Ward linkage. Colors represent the cell line tissue of origin. **B.** Proportional bar graph showing the distribution over cell line tissues of epithelial, mesenchymal and non-scoring cell lines as assigned according to an epithelial-to-mesenchymal gene expression signature

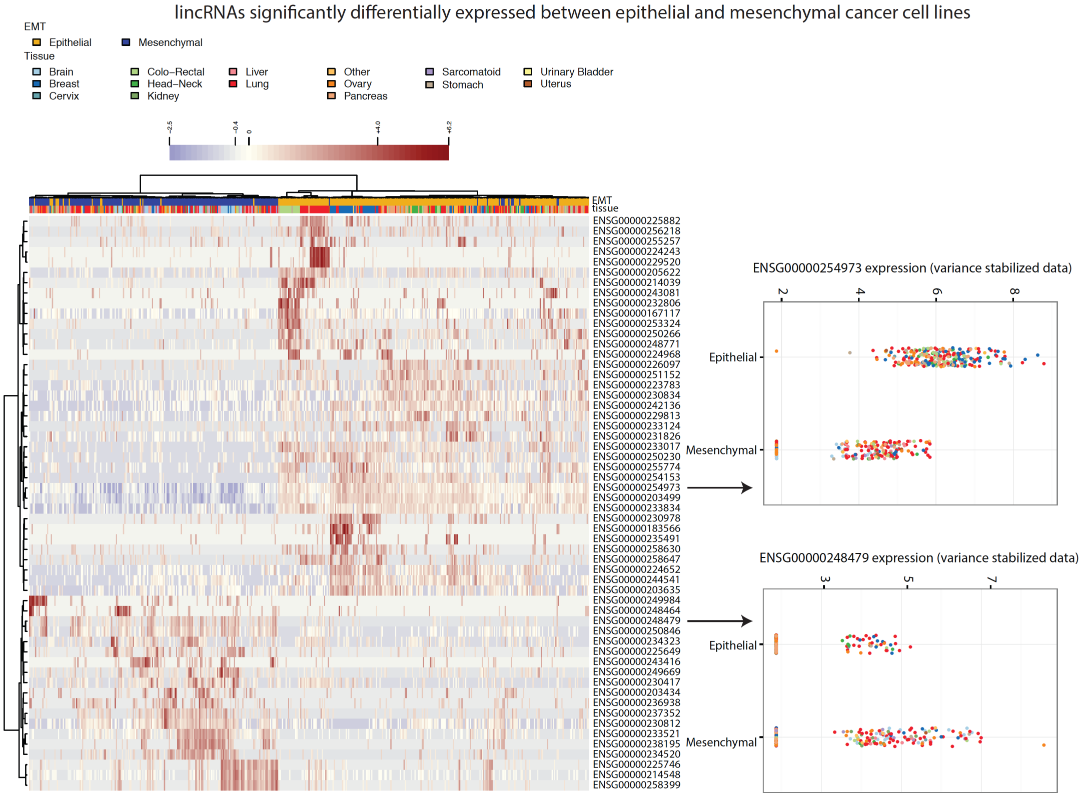
Supplementary Figure 4 – Sample-by-sample gene expression correlation networks



Networks derived from the correlation matrix of the 1000 most variable genes (as determined by inter-quartile range). Edges are shown for correlation values > .75 and higher edge intensity signifies higher correlation. The Fruchterman-Reingold algorithm was used to determine the network layout. Colors of the nodes denotes either tissue of origin (**A**) or epithelial (yellow) or mesenchymal (blue) gene expression state as assigned according to an epithelial-to-mesenchymal gene expression signature (**B**).

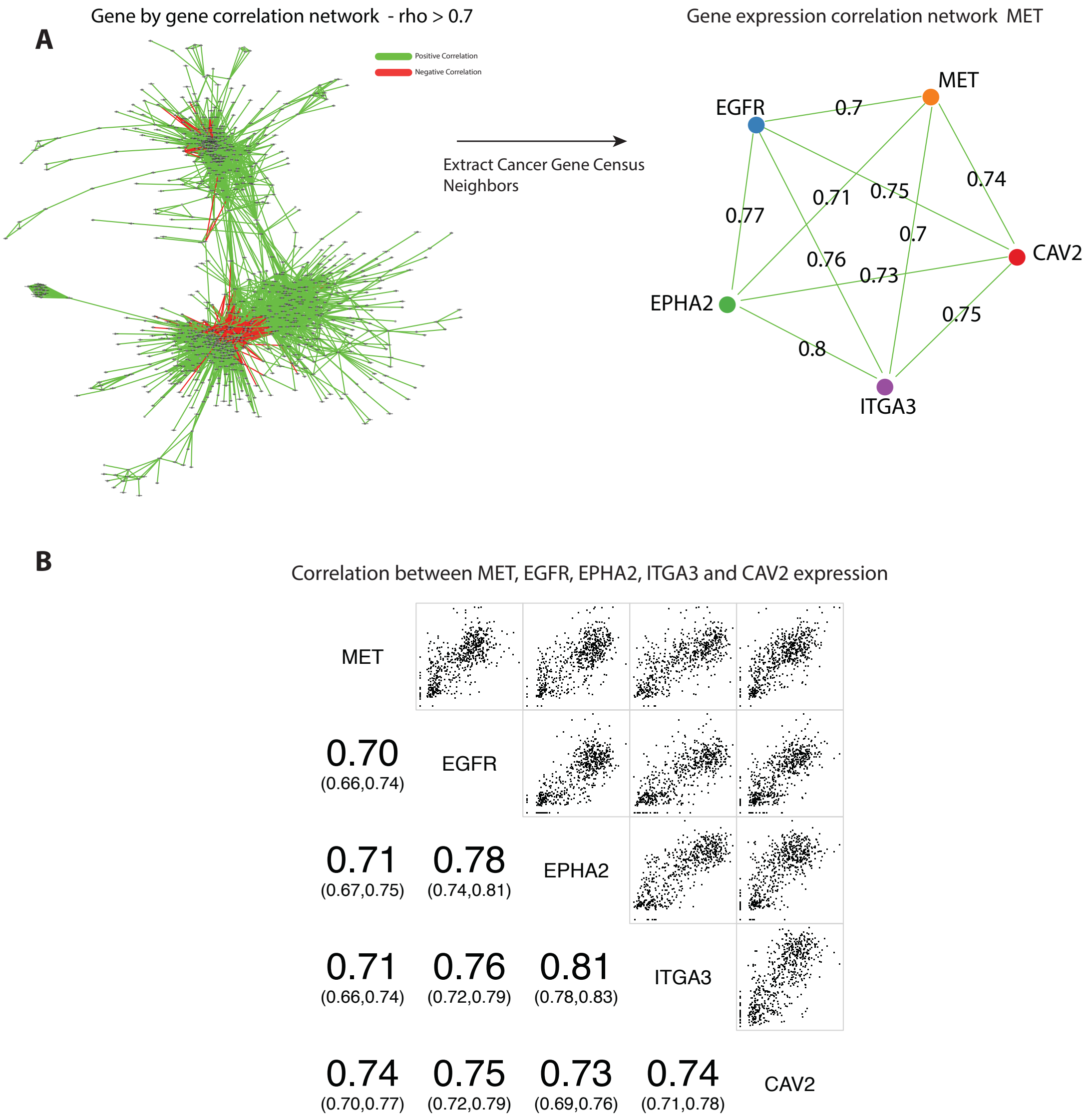
Nature Biotechnology: doi:10.1038/nbt.3080

Supplementary Figure 5 – Hierarchical clustering of lincRNA genes associated with EMT.



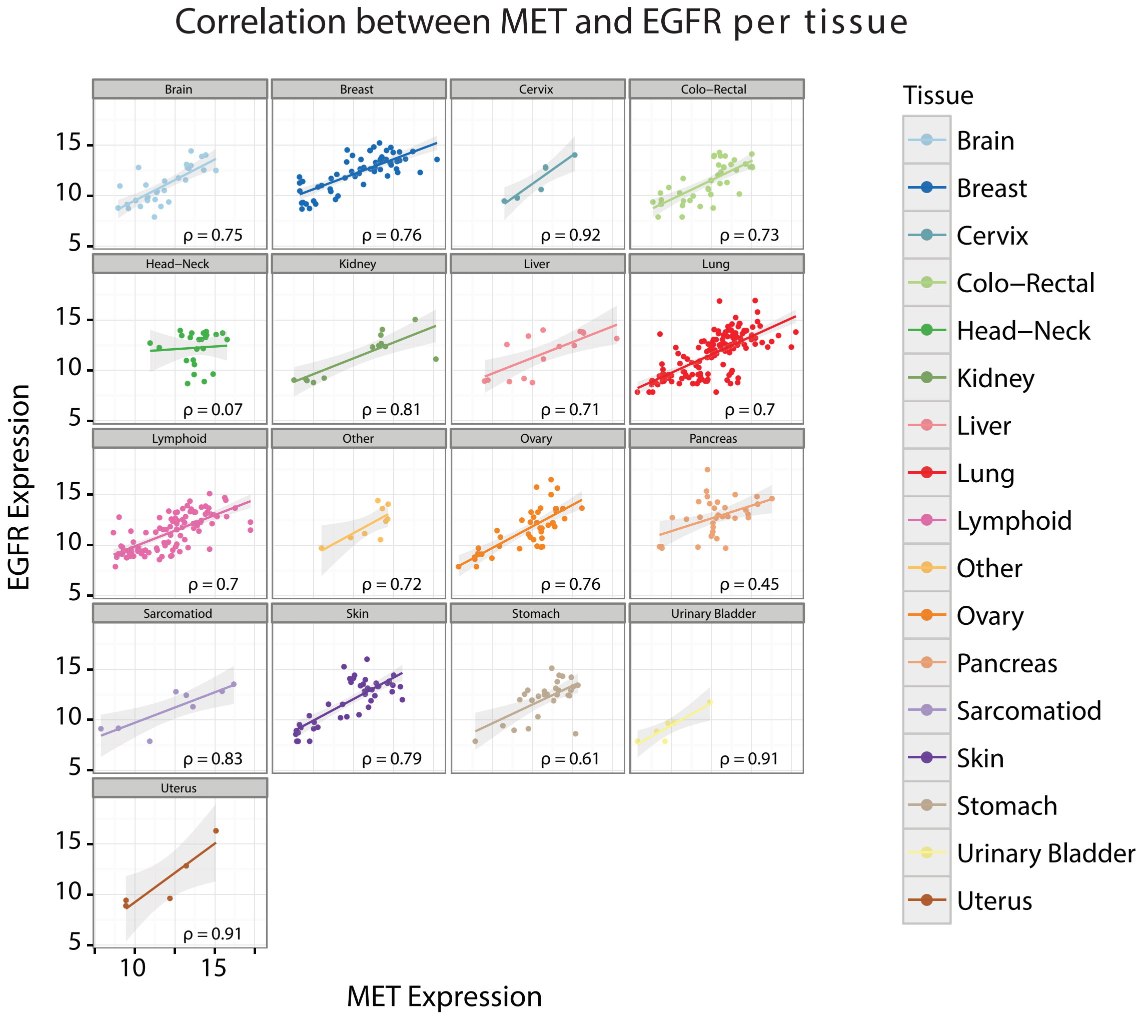
LincRNA genes found differentially expressed between epithelial-type cell lines and mesenchymal-type cell lines are used for hierarchical clustering. Cell lines that could not be categorized as either epithelial or mesenchymal were excluded from this representation. Two examples of lincRNAs are highlighted with either expression predominantly in mesenchymal cells (ENSG00000248479) or in epithelial cells (ENSG00000254973) of various tissues of origin. Both genes encode lincRNAs for which there is no functional information.

Supplementary Figure 6 – Gene expression correlation reveals co-regulation of MET, EGFR, ITGA3 and EPHA2 expression



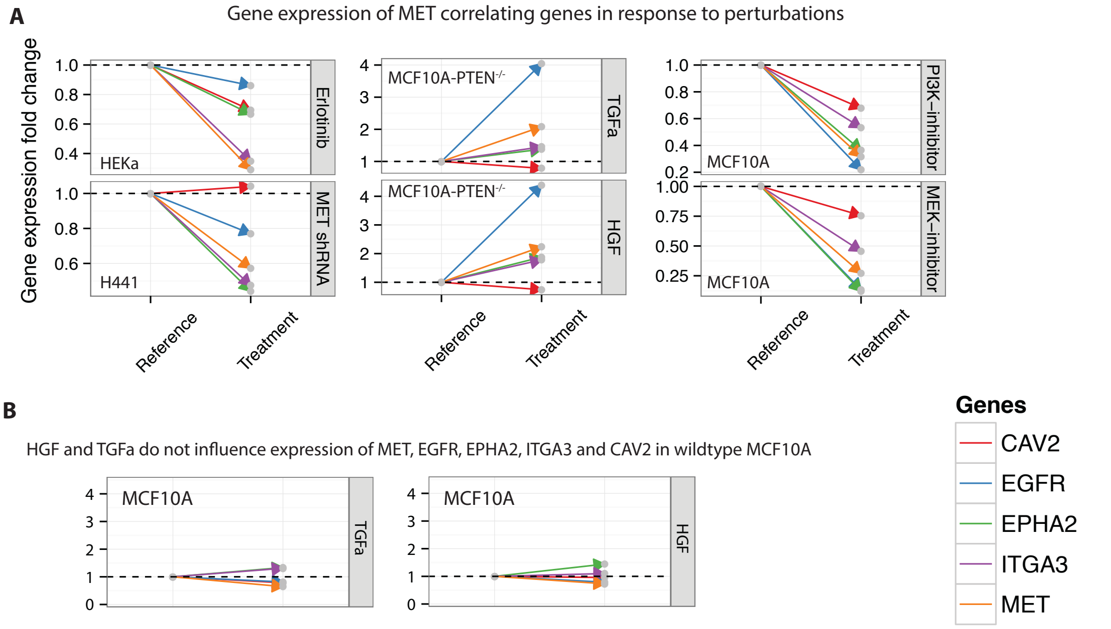
All gene expression values are variance stabilized data values as determined by the *DESeq* R package. **A.** The largest connected network of the gene expression correlation. Only edges with a correlation of > 0.7 or < -0.7 are shown, with green edges indicating positive correlation and red edges indicating negative correlation. The smaller network shows the direct neighbors of the *MET* gene. Correlation coefficients are plotted onto the edges. **A.** Correlation matrix showing the pair-wise gene expression for *MET, EGFR, EPHA2, ITGA3* and *CAV2* over all cell lines. Numbers in the lower left panels indicate the Spearman correlation coefficients and their 95% confidence interval as determined by the *cor.test* function in *R*.

Supplementary Figure 7 – MET and EGFR gene expression correlation is conserved across cell line tissues



Scatterplots, separated per tissue, showing the expression of *MET* on the x axis and the expression of *EGFR* on the y axis. Gene expression values are variance stabilized data values as determined by the *DESeq* R package.

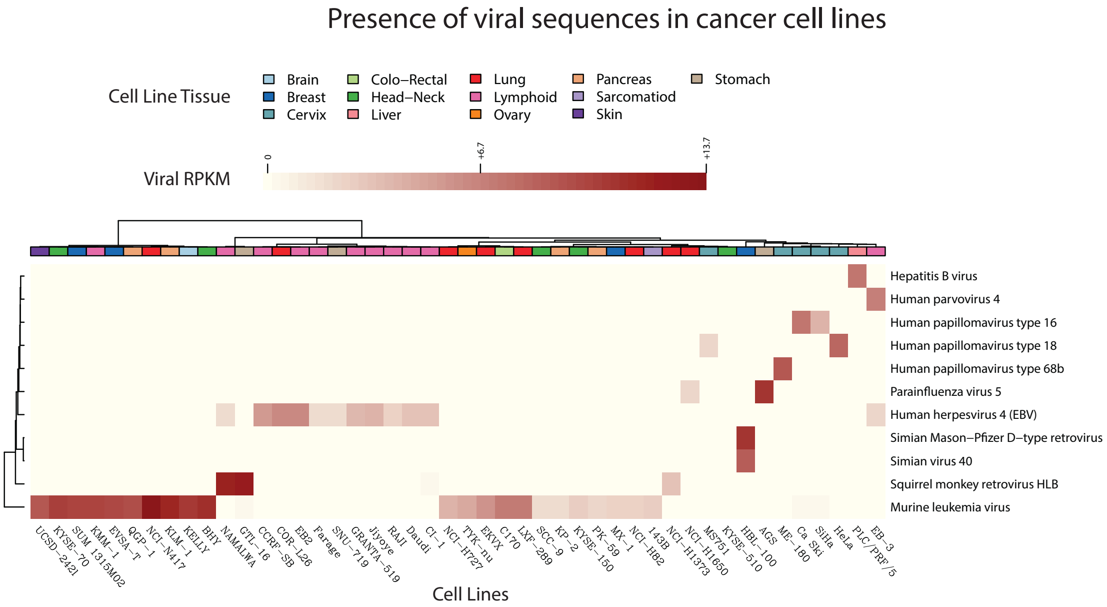
Supplementary Figure 8 – EGFR, MET, EPHA2 and ITGA3 show gene expression regulation by MET and EGFR signaling.



A. Changes in transcript expression of MET, EGFR, ITGA3, EPHA2 and CAV2 by perturbation with the indicated compounds and shRNAs. The cell line in which the experiment was performed is annotated in the graph area. See also Supplementary Figure 3. **B.** Arrow charts representing the fold change of gene expression values of *MET*, *EGFR*, *ITGA3*, *EPHA2* and *CAV2* before and after perturbation with either an EGFR ligand (TGFa) or a MET ligand (HGF) in MCF10A cell lines carrying a wild-type *PTEN* gene.

Nature Biotechnology: doi:10.1038/nbt.3080

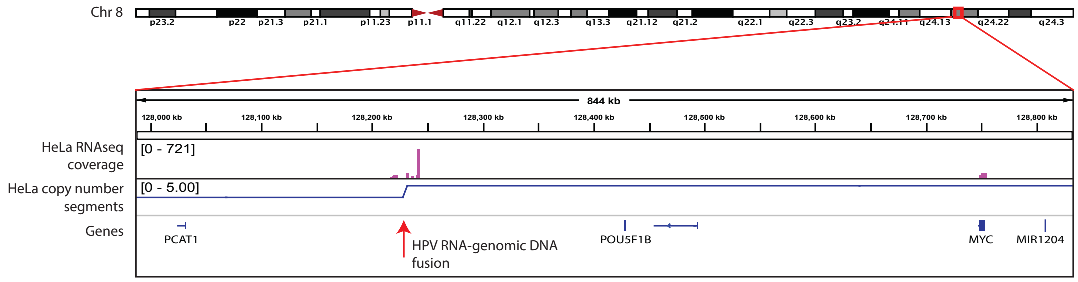
Supplementary Figure 9 - Viral sequences in cell lines



Heatmap showing viral RPKM (Reads Per Kilobase per Million mapped reads) within human cell lines. Rows and columns are clustered using Euclidean distance and Ward's linkage. Tissue of origin groups are color-coded above the columns.

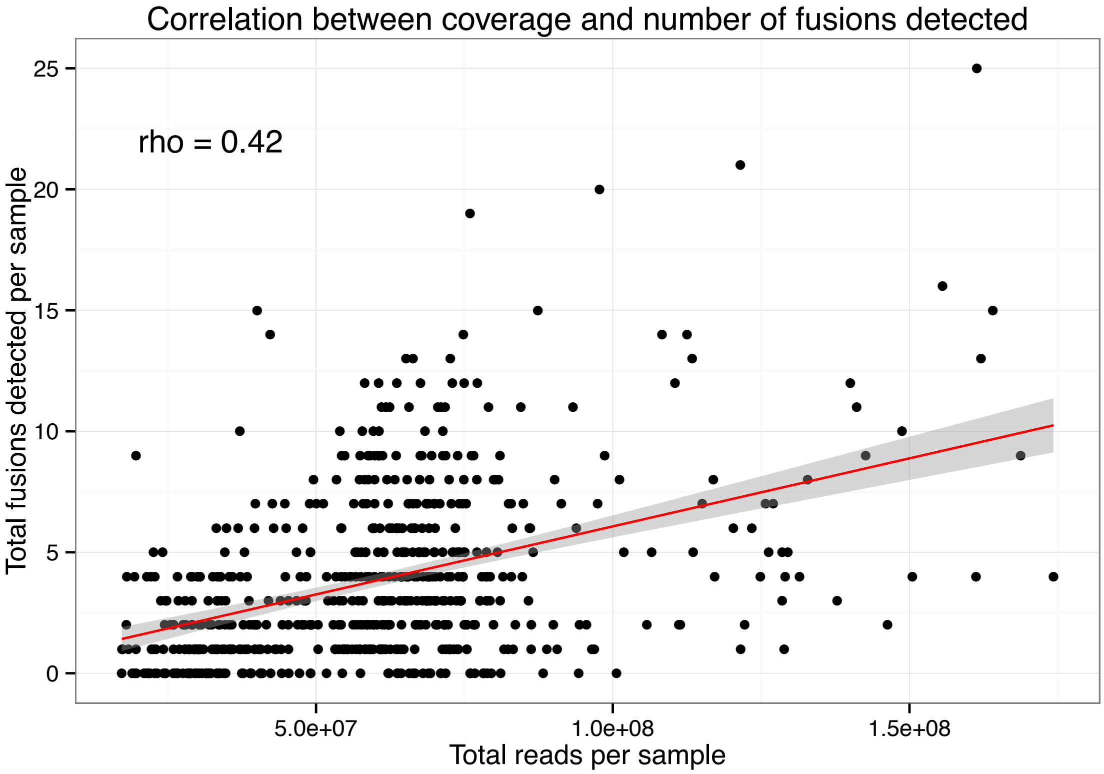
Supplementary Figure 10 – Integration of human papillomavirus near the MYC gene in the HeLa cell line co-localizes with copy number breakpoint.

HeLa HPV integration at DNA copy number breakpoint



Integration of HPV at the *MYC* genomic locus in the HeLa cell line. The first row shows the coverage of mapped RNA-Seq reads, the second row shows absolute copy number segments called using the PICNIC from Illumina 2.5 SNP array data. The arrow indicates the location of the detection of human papillomavirus type 18 RNA fused to transcribed human genomic DNA.

Supplementary Figure 11 – Sequence depth and number of fusion candidates found are correlated



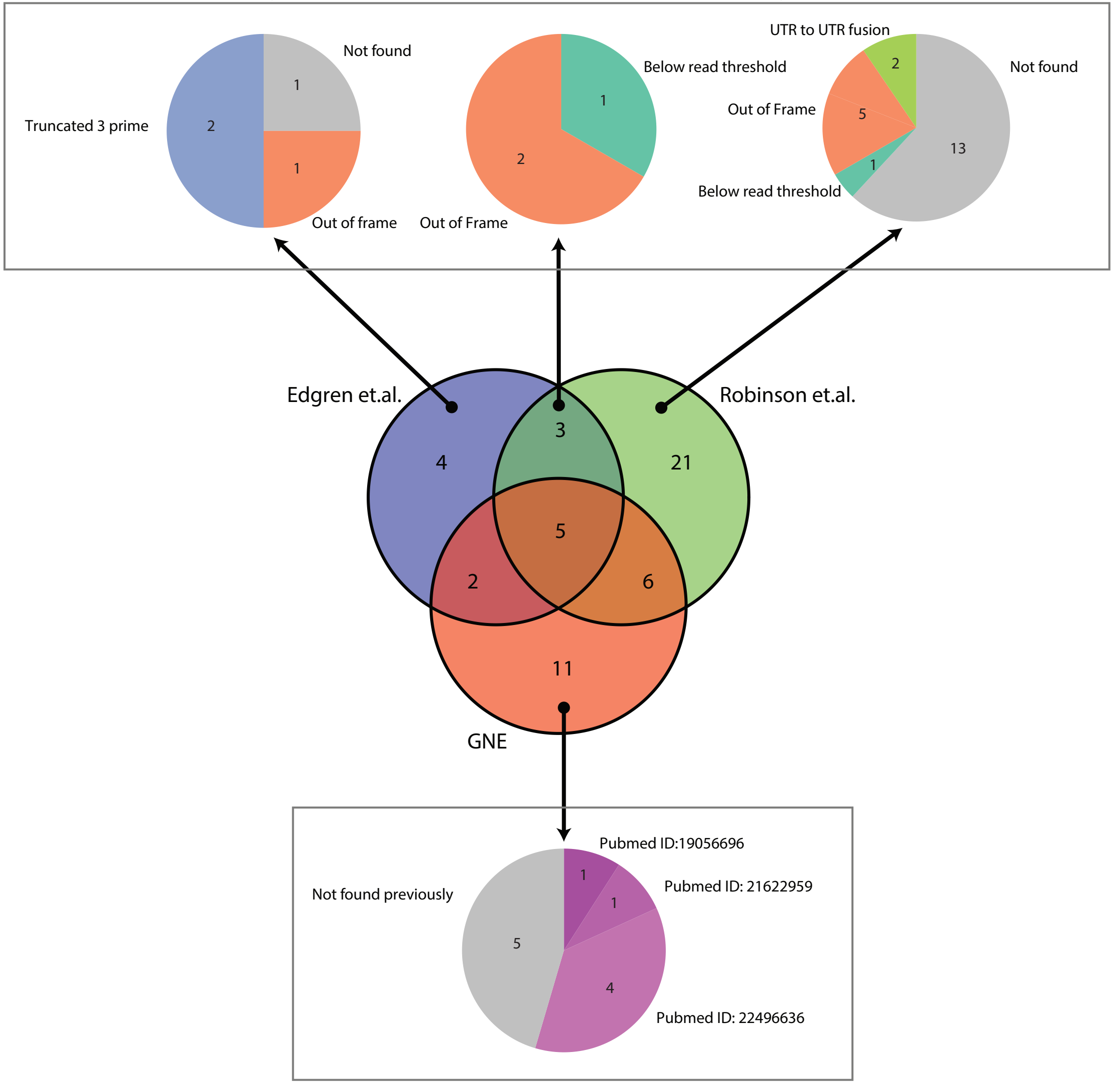
Plotted are the sequence depth (x-axis) and the number of fusion candidates found (y-axis). The red line indicates a linear regression, and the gray shaded area indicates the 95% confidence interval for this regression. Linear regression was applied as implemented in the ggplot2 package for the R programming language.

Nature Biotechnology: doi:10.1038/nbt.3080

Supplementary Figure 12 – Comparison of GNE fusions detected in cell lines MCF-7 and BT-474 with two published studies

Fusions found in cell lines MCF-7 and BT-474 in three studies

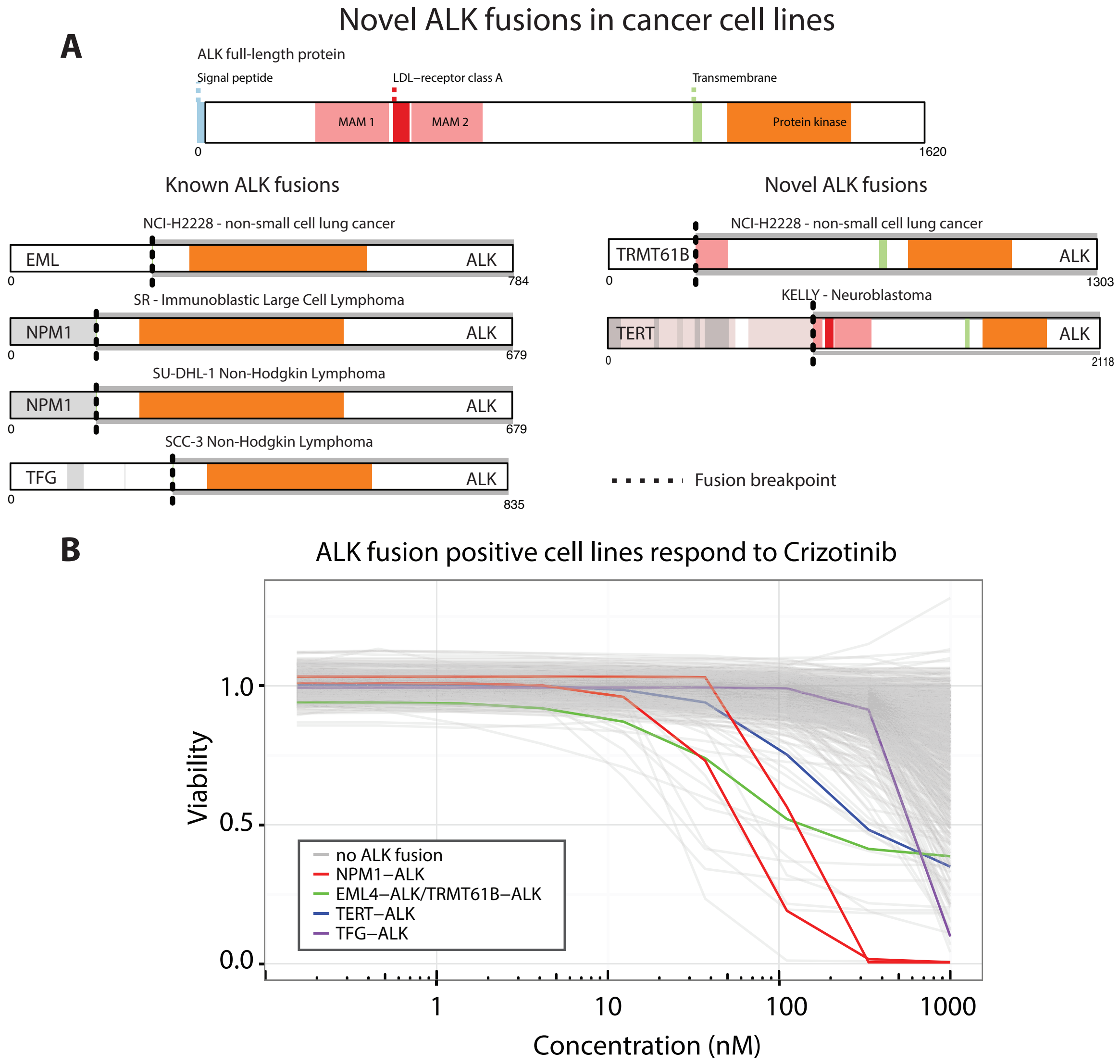
Edgren and Robinson results compared to pre-filtering GNE results



GNE-specific fusions compared to additional studies on MCF-7 and BT-474

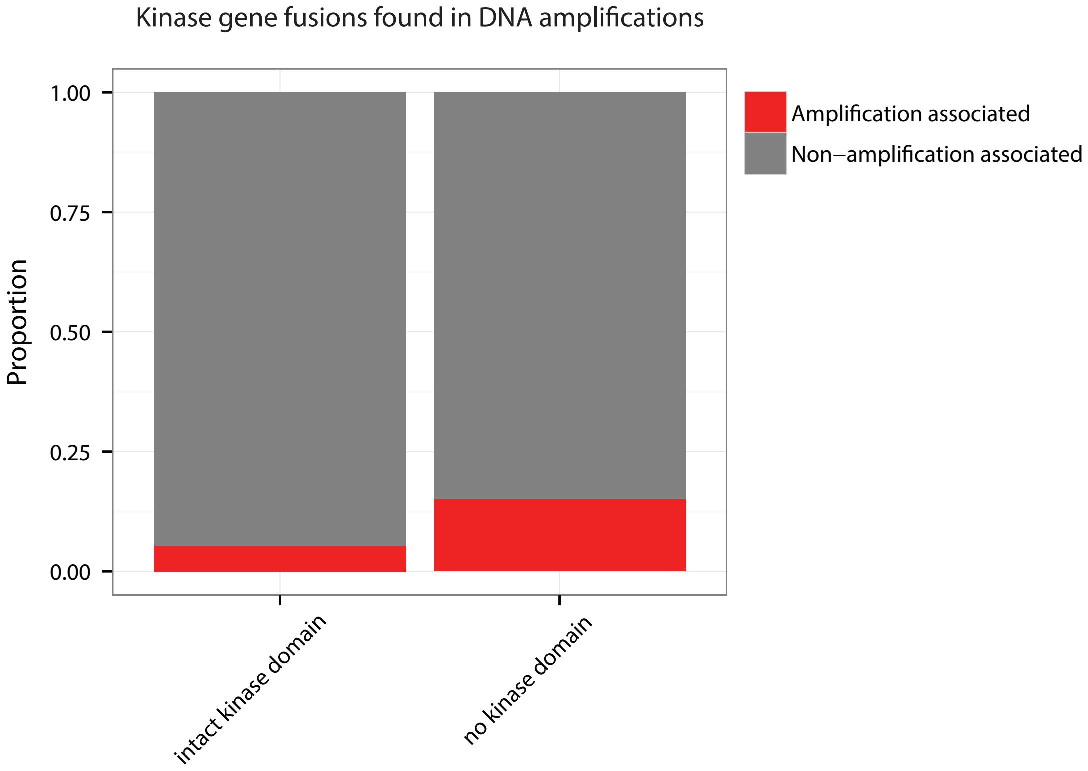
The Venn diagram in the middle of the figure depicts the overlap of fusions found in our study, after the filtering pipeline, compared with published fusion results from two previously published studies: Edgren *et al*. 2011, *Genome Biology* and Robinson *et al*. 2011, *Nature Medicine*. For the Venn subsets that contain fusions specific to the Edgren and Robinson studies, pie graphs are shown in the top panel detailing whether these fusions were detected in our study, but removed by filtering (colored slices), or not detected (gray slice). The bottom panel shows a pie graph for the fusions specific to our study. The colored slices indicate that this fusion was found in additional published studies (for which the Pubmed identification number is shown) or was not reported previously (gray slice).

Supplementary Figure 13 – Fusions involving Anaplastic Lymphoma Kinase (ALK) found in cancer cell lines



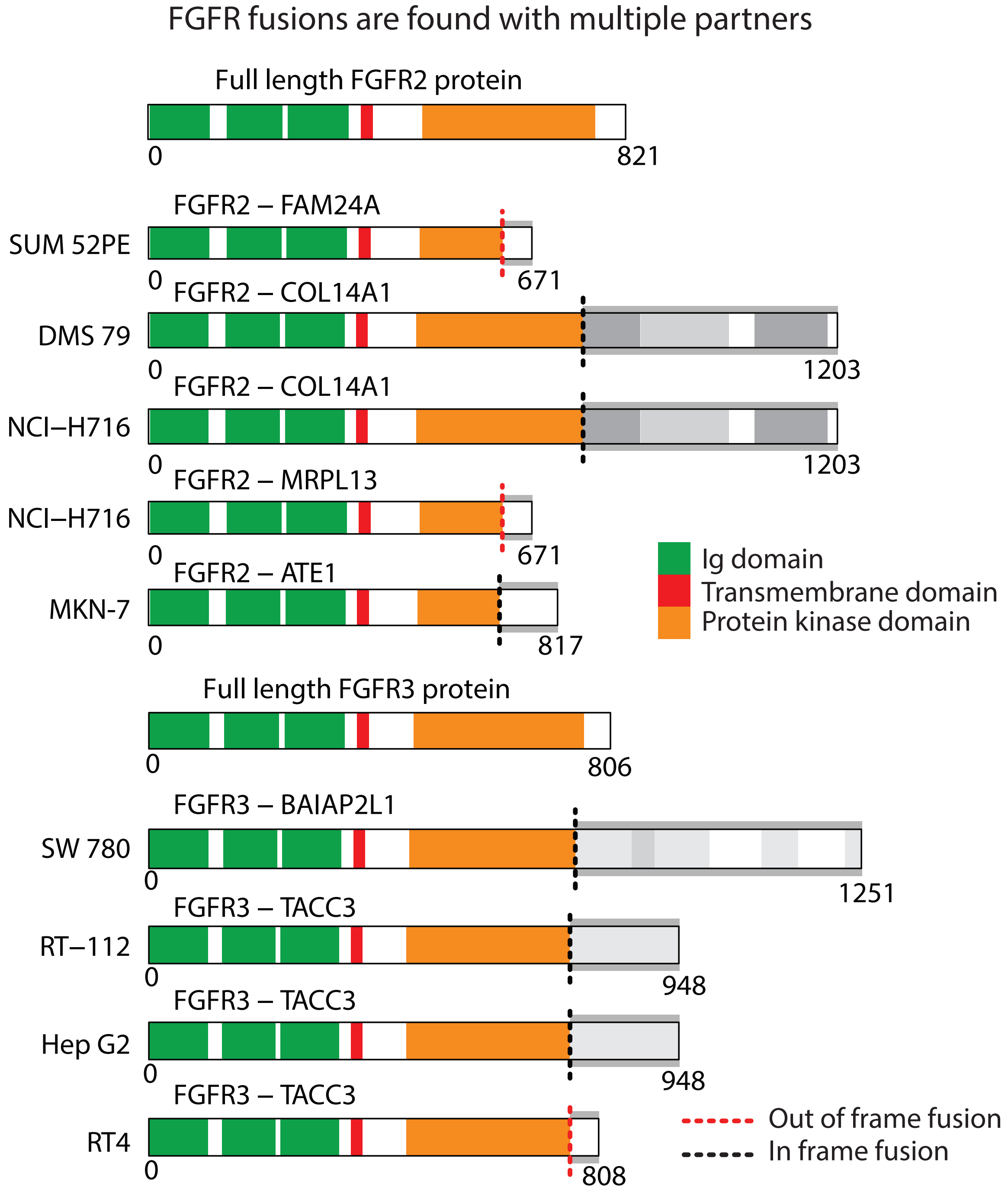
A. Schematic representation of predicted fusion proteins involving ALK. All fusion proteins are predicted to be in-frame and contain a fully intact protein kinase domain. B. Cell line drug response curves for crizotinib treatment. ALK fusion-positive cell lines are indicated in color.

Supplementary Figure 14 – Kinase gene fusions in amplified regions are more likely to have a non-complete kinase domain



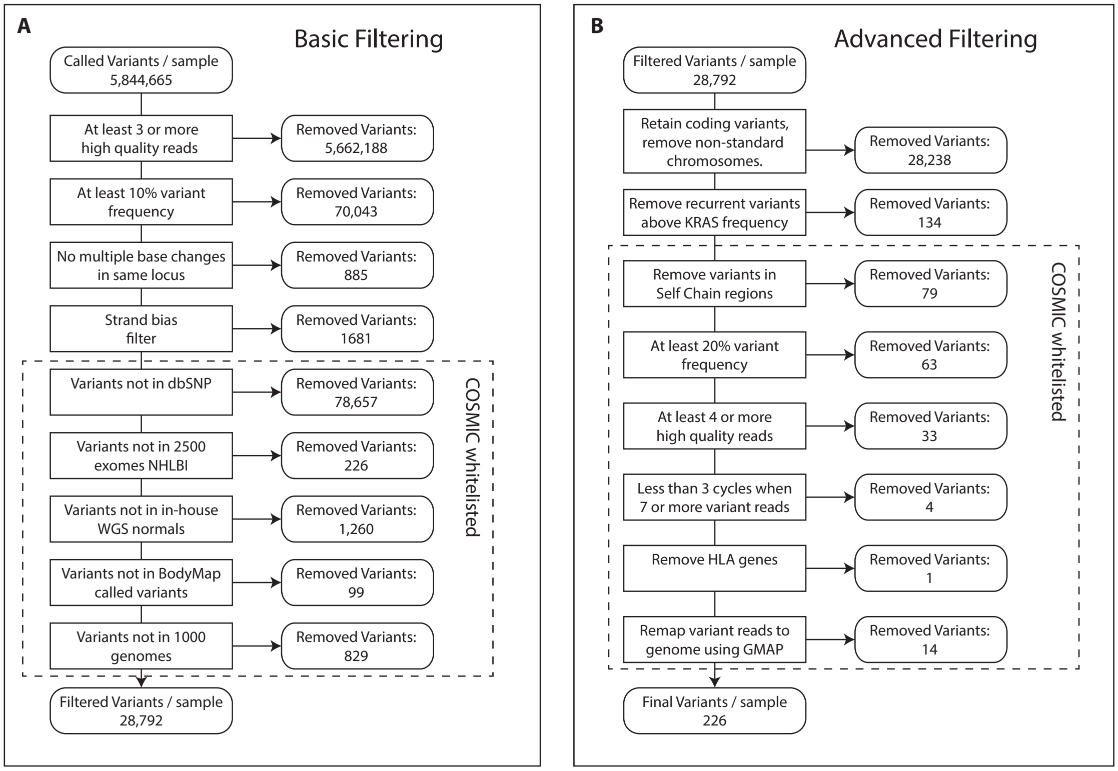
Proportional bar chart showing the proportion of amplification-associated fusions for kinases with either a functional kinase domain, or a truncated or absent kinase domain.

Supplementary Figure 15 – Overview of FGFR2 and FGFR3 fusion genes found in cell lines with complete or partial kinase domain.



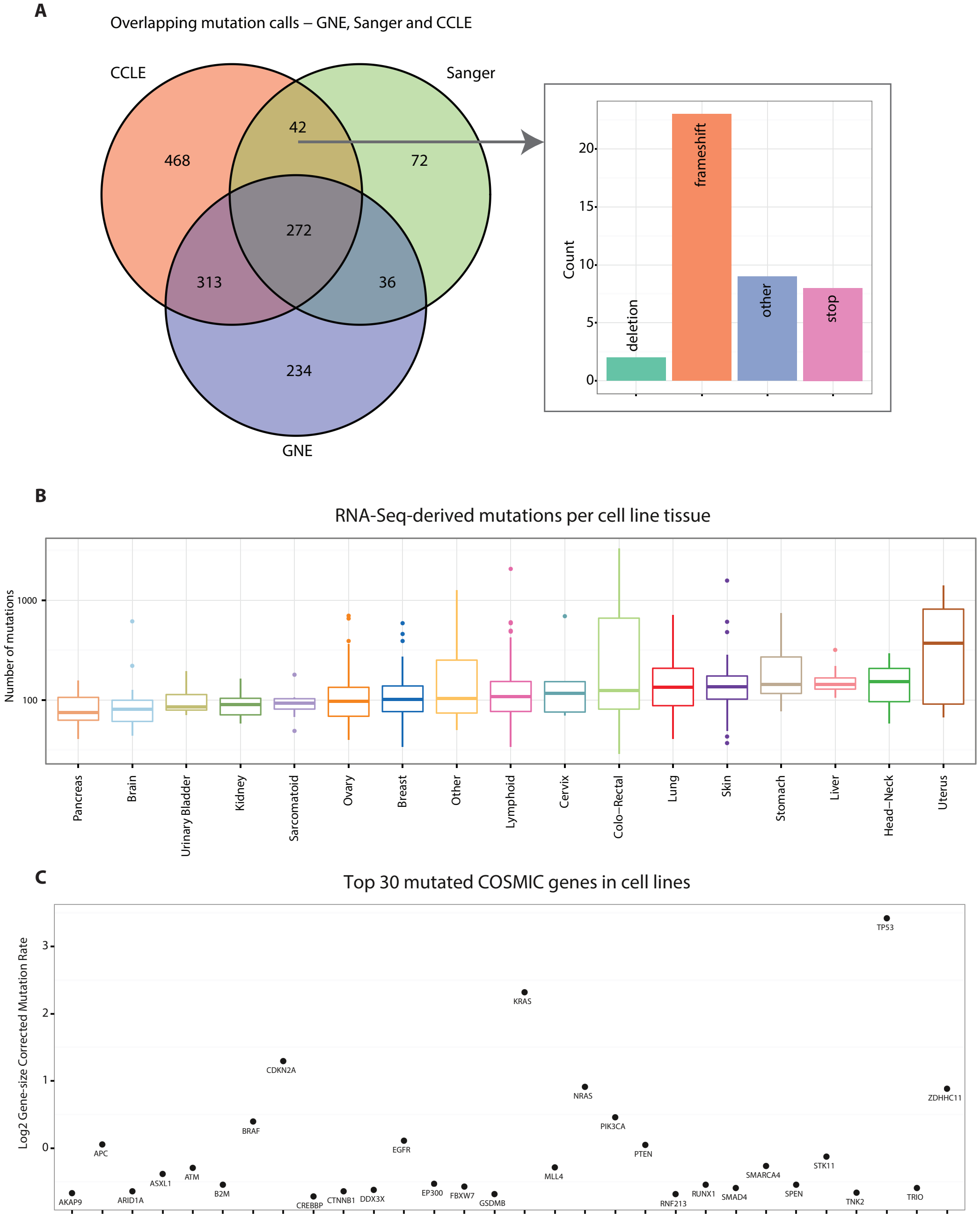
Schematic overview of fusions found involving FGFR2 and FGFR3 and containing at least part of the kinase domain (indicated in orange). The dotted lines indicate either an in-frame fusion (black dotted line) or an out-of-frame fusion (red dotted line).

Supplementary Figure 16 – Overview of RNA-seq variant filtering pipeline


A
Basic Filtering
Called Variants / sample
5,844,665
At least 3 or more high quality reads
Removed Variants: 5,662,188
At least 10% variant frequency
Removed Variants: 70,043
No multiple base changes in same locus
Removed Variants: 885
Strand bias filter
Removed Variants: 1681
Variants not in dbSNP
Removed Variants: 78,657
Variants not in 2500 exomes NHLBI
Removed Variants: 226
Variants not in in-house WGS normals
Removed Variants: 1,260
Variants not in BodyMap called variants
Removed Variants: 99
Variants not in 1000 genomes
Removed Variants: 829
COSMIC whitelisted
Filtered Variants / sample
28,792
B
Advanced Filtering
Filtered Variants / sample
28,792
Retain coding variants, remove non-standard chromosomes.
Removed Variants: 28,238
Remove recurrent variants above KRAS frequency
Removed Variants: 134
Remove variants in Self Chain regions
Removed Variants: 79
At least 20% variant frequency
Removed Variants: 63
At least 4 or more high quality reads
Removed Variants: 33
Less than 3 cycles when 7 or more variant reads
Removed Variants: 4
Remove HLA genes
Removed Variants: 1
Remap variant reads to genome using GMAP
Removed Variants: 14
COSMIC whitelisted
Final Variants / sample
226

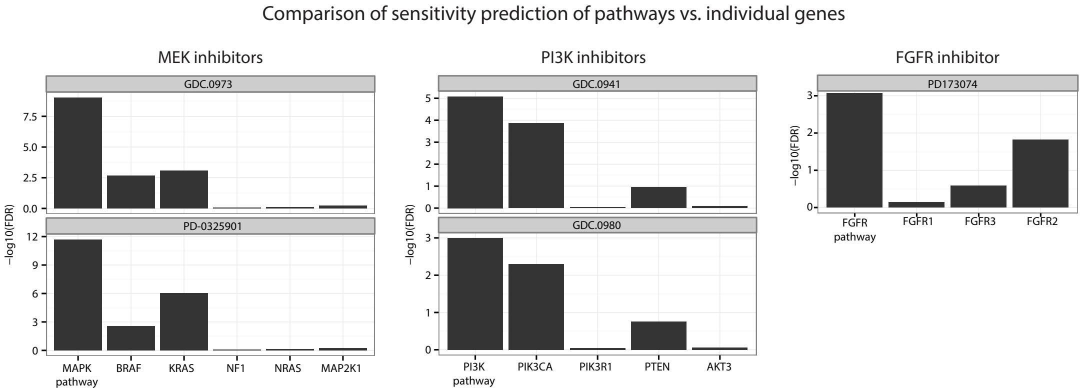
This Supplementary Figure hows our variant filtering pipeline for variants found in RNA seq data. We use a two step approach. We use basic filtering to exclude most common germline variants and sequencing error-derived variants. The advanced filtering results in a more stringent selection and results in the list finally used for the paper.

Supplementary Figure 17 – RNA-seq derived high confidence gene mutations



A. Venn diagram showing the overlap of mutations found by the CCLE, Sanger and Genentech mutation detection. Only genes for which all studies had data were included and overlap was performed on amino acid position and change. For the mutations missed by Genentech, but found by both CCLE and Sanger we show the nature and occurrence (inset graph). As can be seen, most missed mutations cover indels. **B.** Boxplot showing the distribution of high confidence mutations found in cell lines over tissue groups. **C.** Plot showing the gene-size corrected mutation rate for the top 30 most frequently mutated genes determined from RNA sequencing data in cancer cell lines. Only genes that have recorded mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) are shown.

Supplementary Figure 18 – Pathway-based response prediction outperforms single gene predictors.



Barplots showing the predictive value of the aggregated aberration profiles on the sensitivity (in IC50) for MEK inhibitors GDC-973 and PD901, PI3K inhibitors GDC-941 and GDC-980, and the FGFR inhibitor PD173074. Y values denote the –log10 False Discovery Rate for the Wilcoxon Rank Sum Test (corrected by Benjamini/Hochberg correction, n=12 for MEK inhibitors, n=10 for PI3K inhibitors and n=4 for the FGFR inhibitor) of IC50s between aberrant and wildtype cell lines for the pathways or single genes shown on the x-axis.

Supplementary Tables

Supplementary Table 3 – Resolution of genomic concordance among cell lines

Cell line 1	Cell line 2	Concordance Score	Description	How To Resolve	Cell Line Kept	Previously described	Category
105KC	JJ012	0.997	Both are chondrosarcoma derived cell lines, but should be derived from different patients. This was an internal sample handling error.	Highest read count	JJ012	No	mixup
501A	624 mel	0.998	No evidence that they are from the same individual	Highest read count	624 mel	No	new
C170	HCT-15	0.998	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
C170	DLD-1	0.998	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
C170	HCT-8	0.993	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
C32	C32TG	0.993	C32TG is a 6-thioguanine resistant clone of C32.	C32	C32	Reported by Sanger	known
Caco-2	C2BBe1	0.992	C2BBe1 is cloned from Caco-2 in 1988	Highest read count	C2BBe1	Reported ATCC	known
CHL-1	COLO 699	0.966	Probably derived from the same patient	Highest read count	COLO 699	Reported by Sanger, ECACC	known
COLO 201	COLO 205	0.990	All three established from a pleural effusion in the same colon cancer patient	Highest read count	COLO 206F	Reported by Sanger	known
COLO 201	COLO 206F	0.985	All three established from a pleural effusion in the same colon cancer patient	Highest read count	COLO 206F	Reported by Sanger	known
COLO 205	COLO 206F	0.993	All three established from a pleural effusion in the same colon cancer patient	Highest read count	COLO 206F	Reported by Sanger	known
COLO 800	COLO-818	0.991	COLO 800 and 794 are both from a 14 YO male, 818 is annotated to be from a 42 YO female. Discard 818, keep either 794 or 800	Highest read count	COLO 794	Reported by DSMZ, ECACC	known
COLO 800	COLO 794	0.991	COLO 800 and 794 are both from a 14 YO male, 818 is annotated to be from a 42 YO female. Discard 818, keep either 794 or 800	Highest read count	COLO 794	Reported by DSMZ, ECACC	known
COLO 829	COLO 857	0.965	Extracted from the same patient, different sites, name mismatches with vendor-quoted paper (PMID: 8402545) discard all	None	None	Reported by Sanger	known
COLO 829	COLO 853	0.961	Extracted from the same patient, different sites, name mismatches with vendor-quoted paper (PMID: 8402545) discard all	None	None	Reported by Sanger	known
COLO 853	COLO 857	0.995	Extracted from the same patient, different sites, name mismatches with vendor-quoted paper (PMID: 8402545) discard all	None	None	Reported by Sanger	known
COLO-818	COLO 794	1.000	COLO 800 and 794 are both from a 14 YO male, 818 is annotated to be from a 42 YO female. Discard 818, keep either 794 or 800	Highest read count	COLO 794	Reported by DSMZ, ECACC	known
COV413B	COV413A	0.986	Extracted from the same tumor	Highest	COV413B	Derived from	known

				read count		same patient	
CX-1	HT-29	0.987	HT-29 and CX-1 seem to be from the same patient (annotated as colon carcinoma from a 44 YO female). WiDr should be from a 78 YO female, but SRT also shows as a derived from HT-29. Discard WiDr.	Highest read count	HT-29	Reported by Sanger, ATCC	known
CX-1	WiDr	0.984	HT-29 and CX-1 seem to be from the same patient (annotated as colon carcinoma from a 44 YO female). WiDr should be from a 78 YO female, but SRT also shows as a derived from HT-29	Highest read count	HT-29	Reported by Sanger, ATCC, ICLAC	known
DLD-1	HCT-15	0.998	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
DLD-1	HCT-8	0.993	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
EB1	EB2	0.991	Not known if officially from the same individual, isolated around the same time, both Burkitt's lymphoma	Highest read count	EB2	Reported by Sanger	known
EFM-192A	EFM-192B	0.983	All three established from a pleural effusion in the same breast cancer patient	Highest read count	EFM-192A	Reported by DSMZ	known
EFM-192A	EFM-192C	0.973	All three established from a pleural effusion in the same breast cancer patient	Highest read count	EFM-192A	Reported by DSMZ	known
EFM-192B	EFM-192C	0.984	All three established from a pleural effusion in the same breast cancer patient	Highest read count	EFM-192A	Reported by DSMZ	known
G112	G122	0.986	Both were estabished at a different time. Possible contamination	Highest read count	G122	No	new
G118	G142	0.997	Established in the same lab, so either the same patient or contamination	Highest read count	G118	No	new
G118	G141	0.996	Established in the same lab, so either the same patient or contamination	Highest read count	G118	No	new
G142	G141	0.993	Established in the same lab, so either the same patient or contamination	Highest read count	G118	No	new
G44	G96	0.995	Derived in the same lab, but should be from two different patients. Possible contamination	Highest read count	G44	No	new
GP2d	GP5d	0.999	Take from same primary tumor.	Highest read count	GP2d	No	new
GTL-16	MKN-45	1.000	GTL-16 is a subclone of MKN-45 (PMID: 1486568)	Highest read count	GTL-16	Published	known
H322T	NCI-H322T	0.992	Same cell line, naming issue	Highest read count	NCI-H322T	ATCC knows about this- no longer available	known
HCC2157	HeLa	0.999	likely HeLa contamination	HeLa	HeLa	Reported	known
HCC2157	HEp-2	0.980	likely HeLa contamination	Highest read count	HeLa	Reported by Sanger	known

HCT 116	ATRFLOX	1.000	ATRFLOX is a derivative of ATCC Catalog No. CCL-247 (HCT 116) in which one copy of the ATR (ataxia telangiectasia related) gene has been disrupted, and the other allele has been fixed with lox sites flanking exon 2 making it susceptible to Cre deletion.	HCT 116	HCT 116	Reported by ATCC	known
HCT-15	HCT-8	0.993	Known to be a part of a group of cross-contaminated cell lines (PMID: 9809040)	Highest read count	C170	Reported by Sanger, ATCC	known
HeLa	HEp-2	0.979	likely HeLa contamination	HeLa	HeLa	Reported by Sanger	known
Hep G2	C3A	0.992	C3A is a derivative of Hep G2, selected for strong contact inhibition of growth, high albumin production, high production of alpha fetoprotein (AFP) and ability to grow in glucose deficient medium.	Hep G2	Hep G2	Reported by ATCC	known
HM7	LS 174T	0.999	HM-7 and LS 174T are known derivatives of colon cancer line LS 180. MT-3 is supposed to be a breast carcinoma line.	LS 180	LS 180	Published	known
HM7	LS 180	0.999	HM-7 and LS 174T are known derivatives of colon cancer line LS 180. MT-3 is supposed to be a breast carcinoma line.	LS 180	LS 180	Published	known
HPAC	KCI-MOH1	0.983	DSMZ "DNA fingerprinting and cytogenetic analysis showed cross-contamination with cell line STR profile matches 100% with DSMZ KCI-MOH1 and ATCC HPAC. 02/27/12 SS. HPAC which was established in 1985 from the pancreas adenocarcinoma of a 64-year-old Caucasian woman". 06/22/11 SS.	Highest read count	KCI-MOH1	Reported by DSMZ	known
Hs 69ST	Hs 695T	0.993	Hs 69ST is a typo, Hs 695T is the correct version	Hs 695T	Hs 695T	No	mixup
HS-Sultan	Jiyoye	0.988	HS-Sultan deposited with the ATCC as a plasmacytoma cell line, DNA fingerprinting has shown this line to be a derivative of Jiyoye (ATCC CCL-87), a Burkitt's lymphoma cell line.	Jiyoye	Jiyoye	Reported by ATCC, ICLAC	known
HT-29	WiDr	1.000	HT-29 and CX-1 seem to be from the same patient (annotated as colon carcinoma from a 44 YO female). WiDr should be from a 78 YO female, but SRT also shows as a derived from HT-29	Highest read count	HT-29	Reported by Sanger, ATCC, ICLAC	known
IM-95	IM-95m	0.993	IM-95m is a subclone of IM-95. Keep parental strain	IM-95	IM-95	Reported JCRB	known
KMS-12-BM	KMS-12-PE	0.968	Extracted from the same patient, different sites	Highest read count	KMS-12-BM	Reported by DSMZ	known
KMS-28BM	KMS-28PE	0.986	Extracted from the same patient, different sites	Highest read count	KMS-28BM	CLIMA	known
KPL-1	MCF-7	0.996	KPL-1 shown to be cross-contaminated with MCF-7. Keep MCF-7	MCF-7	MCF-7	Reported by DSMZ, Sanger	known
LS 180	LS 174T	1.000	HM-7 and LS 174T are known derivatives of colon cancer line LS	LS 180	LS 180	Reported by Sanger	known

			180. MT-3 is supposed to be a breast carcinoma line, but is not in gCell.				
M059J	M059K	0.984	Isolated from the same patient	Highest read count	M059K	Reported by ATCC	known
MCF 10A	MCF10DCIS.com	0.991	DCIS.com is propagated through xenograft. Keep parental	MCF 10A	MCF 10A	Published	known
MT-3	LS 174T	0.998	HM-7 and LS 174T are known derivatives of colon cancer line LS 180. MT-3 is supposed to be a breast carcinoma line.	LS 180	LS 180	Reported by Sanger	known
MT-3	LS 180	0.998	HM-7 and LS 174T are known derivatives of colon cancer line LS 180. MT-3 is supposed to be a breast carcinoma line.	LS 180	LS 180	Reported by Sanger	known
MT-3	HM7	0.998	HM-7 and LS 174T are known derivatives of colon cancer line LS 180. MT-3 is supposed to be a breast carcinoma line.	LS 180	LS 180	Reported by Sanger	known
NCI-H1155	HCC12	0.992	HCC12 is supposedly a liver cell line and H1155 a lung cancer cell line	None	None	No	new
NCI-H2106	NCI-H1770	0.981	NCI-H1770 is annotated to be from a 57 YO individual and NCI-2106 from a 58 YO individual, possible clerical error in history?	Highest read count	NCI-H1770	Reported by Sanger	known
NCI-H2198	NCI-H2196	0.993	Same patient characteristics in ATCC	Highest read count	NCI-H2198	Reported by Sanger	known
NCI-H23	HCC60	0.962	HCC60 is supposedly Ovarian and NCI-H23 is annotated to be lung cancer	None	None	No	new
OvCA 429	OVCAR433	0.978	OVCA433 and OvCA 429 have identical fingerprints. Communication with MDAnderson investigators indicates that there was likely a mix up at source. These lines should be regarded as of common ancestry.	Highest read count	OVCAR433	No	new
PA-TU-8988T	PA-TU-8988S	0.986	Extracted from the same patient, both from liver met (PMID: 1348891)	Highest read count	PA-TU-8988T	Extracted from the same patient, both from liver met (PMID: 1348891)	known
PK-45H	PK-45P	0.953	Extracted from the same patient (10.1159/000015091)	Highest read count	PK-45P	Reported by Riken	known
PSN1	GR-M	0.983	PSN1 is annotated as pancreatic, GR-M as melanoma. Possible mixup	None	None	Reported by Sanger	known
RKO	RKO-E6	0.993	RKO-E6 is a transfected derivative of RKO. Keep the untransfected one.	RKO	RKO	Reported by ATCC	known
SCLC-21H	SCLC-22H	0.986	Extracted from the same patient (10.1007/BF00389964), both effusions	Highest read count	SCLC-22H	Extracted from the same patient (10.1007/BF00389964), both effusions	known
SR-786	SR	0.994	Annotated as different types of lymphoma, and SR-786 is not in gCell, only in Biospecimen	SR	SR	Reported by DSMZ, Sanger, not ATCC	known

SUM 149PT	SUM 229PE	1.000	Supposedly isolated from different patients in the same lab. (PMID: 10604729). This was a vendor mix up. Their early stocks were mixed up and they have since corrected this.	SUM 149PT	SUM 149PT	No	mixup
SW 480	SW 527	0.972	SW480 is primary colon, SW620 a linked metastasis, SW527 is annotated a a breast cancer cell line. Keep parental tumor	SW 480	SW 480	Reported by Sanger, ICLAC, not ATCC	known
SW 620	SW 527	0.971	SW480 is primary colon, SW620 a linked metastasis, SW527 is annotated a a breast cancer cell line. Keep parental tumor	SW 480	SW 480	Reported by Sanger, ICLAC, not ATCC	known
SW 620	SW 480	0.959	SW480 is primary colon, SW620 a linked metastasis, SW527 is annotated a a breast cancer cell line. Keep parental tumor	SW 480	SW 480	Reported by Sanger, ATCC	known
TYK-nu	TYK-nu.CP-r	0.999	CP-r is a cisplatin resistant clone of TYK-nu. Keep non resistant.	TYK-nu	TYK-nu	Reported JCRB	known
WM-115	WM-266-4	1.000	WM-115 is from the primary tumor, WC-266-4 from the metastasis. Keep the primary	WM-115	WM-115	Reported by ATCC	known
YMB-1	YMB-1-E	1.000	ZR-75-1 has been described in the literature before YMB-1 and its (possibly related) cell line YMB-1E. The estabishment of those two is published in a Japanese journal only.	Highest read count	YMB-1	Published	known
ZR-75-1	YMB-1	0.993	ZR-75-1 has been described in the literature before YMB-1 and its (possibly related) cell line YMB-1E. The estabishment of those two is published in a Japanese journal only.	Highest read count	YMB-1	Reported by Sanger	known
ZR-75-1	YMB-1-E	0.993	ZR-75-1 has been described in the literature before YMB-1 and its (possibly related) cell line YMB-1E. The estabishment of those two is published in a Japanese journal only.	Highest read count	YMB-1	Reported by Sanger	known

Supplementary Table 5 – Significantly differentially expressed lincRNAs between Epithelial and Mesenchymal cell lines

lincrna	mean expression mesenchymal	mean expression epithelial	fold change	adjusted p-value
ENSG00000229813	1.69	28.82	17.04	0.01
ENSG00000230812	42.47	3.28	0.08	0.01
ENSG00000237352	6.09	0.47	0.08	0.01
ENSG00000224968	0.03	1.26	41.75	0.01
ENSG00000203635	3.58	45.61	12.73	0
ENSG00000242136	2.29	34.73	15.14	0
ENSG00000225649	12.24	0.68	0.06	0
ENSG00000231826	4.04	88.54	21.92	0
ENSG00000235491	0.03	1.84	68.35	0
ENSG00000243081	0.22	3.46	16.1	0.01
ENSG00000244541	1.38	16.5	11.97	0.01
ENSG00000223783	1.03	16.31	15.77	0
ENSG00000224652	2.68	50.45	18.84	0
ENSG00000251152	0.91	17.4	19.1	0.01
ENSG00000250846	176	4.39	0.02	0
ENSG00000248479	15.25	0.51	0.03	0
ENSG00000248771	4.77	106.28	22.3	0
ENSG00000250266	0.4	26.82	67.71	0
ENSG00000248464	9.77	0.63	0.06	0.01
ENSG00000249984	4.96	0.13	0.03	0
ENSG00000249669	26.76	1.43	0.05	0.01
ENSG00000233834	9.77	150.24	15.37	0
ENSG00000226097	0.48	6.14	12.89	0.01
ENSG00000253324	0.65	12.28	18.86	0.01
ENSG00000236938	4.19	0.21	0.05	0
ENSG00000234520	17.42	1.14	0.07	0
ENSG00000243416	4.99	0.41	0.08	0.01
ENSG00000254153	0.37	16.38	44.85	0
ENSG00000203499	155.98	3330.25	21.35	0
ENSG00000254973	5.44	72.96	13.4	0
ENSG00000234323	31.37	1.7	0.05	0
ENSG00000230417	23.32	1.14	0.05	0
ENSG00000203434	9.19	0.77	0.08	0
ENSG00000230834	2.26	30.98	13.7	0.01
ENSG00000255257	0.95	12.01	12.65	0
ENSG00000250230	0.95	13.84	14.61	0.01
ENSG00000255774	9.8	134.28	13.71	0.01
ENSG00000256218	0.32	4.91	15.27	0
ENSG00000214039	2.76	135.8	49.2	0
ENSG00000233124	0.22	3.42	15.72	0
ENSG00000229520	1.14	23.24	20.34	0
ENSG00000224243	1.82	37.84	20.82	0
ENSG00000258630	0.1	1.62	16.69	0.01
ENSG00000214548	3599.86	212.96	0.06	0
ENSG00000258399	36.63	3.51	0.1	0
ENSG00000225746	35.9	4.64	0.13	0.01
ENSG00000258647	0.46	10.19	21.92	0
ENSG00000167117	1.44	84.54	58.52	0
ENSG00000183566	0.2	12.88	65.7	0
ENSG00000233017	0.24	7.18	30.21	0
ENSG00000230978	0.15	3.71	24.06	0
ENSG00000205622	0.75	8.9	11.89	0.01
ENSG00000232806	0.03	1.69	53.71	0.01
ENSG00000238195	3.95	0.1	0.03	0
ENSG00000233521	116.31	4.82	0.04	0
ENSG00000225882	1.42	18.73	13.19	0

Supplementary Table 9 – TCGA samples analyzed for fusions

TCGA indication	Samples Analysed	Samples with a fusion	% sample with fusion
ACC	79	32	40.5
BLCA	211	133	63
BRCA	991	632	63.8
CESC	144	72	50
COAD	419	162	38.7
DLBC	27	11	40.7
GBM	169	97	57.4
HNSC	398	227	57
KICH	66	14	21.2
KIRP	649	138	21.3
LAML	123	56	45.5
LGG	297	123	41.4
LIHC	147	78	53.1
LUAD	504	298	59.1
LUSC	457	277	60.6
OV	306	284	92.8
PAAD	55	22	40
PRAD	247	167	67.6
READ	157	71	45.2
SARC	105	75	71.4
SKCM	355	194	54.6
STAD	191	166	86.9
THCA	499	152	30.5
UCEC	544	227	41.7

Supplementary Table 12 – Cancer-related pathways with member genes that are known to be functionally altered in cancer.

Pathway Name	Genes In Pathway
PI3K	PIK3CA, PTEN, PIK3R1, AKT1, AKT3
MAPK	KRAS, NRAS, BRAF, MAP2K1, NF1
WNT	APC, CTNNB1, AXIN1, AXIN2
RTK	EGFR, ERBB2, MET, ALK, JAK2, RET, ROS1, FGFR1, FGFR2, PDGFR, CRKL
FGFR	FGFR1, FGFR2, FGFR3
P53	TP53, MDM2
NOTCH	NOTCH1, NOTCH2, NOTCH3
TGFB	SMAD2, SMAD4, TGFBR2
Cell cycle	CDKN2A, CDKN2B, CCND1, CCNE1, CDKN1B
TOR	STK11, TSC1, TSC2
SWI/SNF	SMARCA1, SMARCA4, ARID1A, ARID2, ARID1B, PBRM1
Lineage transcription factors	MITF, NKX2-1, SOX2, ERG, ETV1, CDX2
MYC	MYC
Apoptosis	BCL2A1, BCL2L1, MCL1
Chromatin histone acetyltransferases	CREBBP, EP300
Chromatin histone methyltranferases	MLL, MLL2, MLL3, EZH2, NSD1, WHSC1L1
Chromatin histone demethylases	KDM6A, KDM5A, KDM5C
Protein metabolism	SPOP, FBXW7, WWP1, FAM46C, XBP1
Splicing	SF3B1,U2AF1,SFRS1,SFRS7,SF3A1,ZRSR2,SRSF2,U2AF2,PRPF40B
Metabolism	IDH1, IDH2
DNA repair	MSH2, MSH3, MSH6, ATM, MLH1

Supplementary Notes

Supplementary note 1: PICNIC algorithm adjustments

PICNIC was designed to work only with the Affymetrix SNP 6.0 array and thus required modification to work with the Illumina arrays. The segment initialization component was replaced with the CBS algorithm[66]. The prior distribution of "alpha", the raw copy number signal expected when zero copies of a SNP are present, was altered to have a mean of 0.7 with a standard deviation of 0.05. The prior distribution for overall ploidy was expanded in its right tail to allow for ploidy values > 3.5. Population allele frequencies were calculated from 159 normal samples [31], and in this work, were supplemented with 10 pseudocounts for each allele to avoid "impossible" HMM states, which otherwise would lead to spurious single-SNP segments. Adapting PICNIC to Illumina arrays also required a new method to transform the raw allele-specific signal intensities into normalized values appropriate for PICNIC's HMM strategy. The new per-SNP normalization method is as follows:

Final estimation strategy

The hidden Markov model used for copy number inference requires that two-dimensional probe intensity data for each SNP be transformed so that the normal AA, AB, and BB genotype centroids are separated by one unit in both the horizontal and vertical directions. To achieve this, we first used a Bayesian model to estimate cluster centroids for each SNP using 159 normal samples, and then applied a smooth, non-linear transform to send each centroid to the appropriate location.

Bayesian model

For SNP *k* and genotype *g*, observed data in normal samples were modeled as following a bivariate Gaussian distribution with mean $\mu_{\mathrm{AA}}^{(\mathrm{k})}$ and covariance $\Sigma_{\mathrm{g}}^{(\mathrm{k})}$. The $\mu_{\mathrm{AA}}^{(\mathrm{k})}$, $\mu_{\mathrm{AB}}^{(\mathrm{k})}$, and$\mu_{\mathrm{BB}}^{(\mathrm{k})}$ were observed to be strongly positively correlated. (If one cluster was further from the origin than usual for a given SNP, other clusters for that SNP were highly likely to be further from the origin as well.) To take advantage of this, cluster centers were modeled jointly by a 6-dimensional Gaussian distribution, with mean μ and covariance Σ. μ was treated as a hyperparameter; Σ was modeled as inverse Wishart with two hyperparameters: scale matrix V_μ and degrees of freedom d_μ. The $\Sigma_{\mathrm{g}}^{(\mathrm{k})}$ were modeled as scaled inverse Wishart, each with a corresponding V_g scale matrix, but with common degrees of freedom *d*.

Hyperparameter specification

Hyperparameters were set empirically using a training set of 156 normal samples that included the matched normal from our cancer patients. μ was set to the global average in the training data, with weighting to adjust for differing numbers of observations from SNP to SNP for a given genotype. The V_μ, V_{AA}, V_{AB}, and V_{BB} were constructed similarly. (To estimate V_μ, only SNPs with at least three observations in each genotype cluster were used.) *d* and d_μ were manually tuned to provide satisfactory results for a wide range of probe behavior and minor allele frequencies.

Non-linear transformation of centroids

To account for the fact that in the training data, probe response was observably non-linear with respect to true underlying copy number, signal for SNP *k* (for the A and B alleles separately) was transformed with a non-linear function: $y = \alpha_k x^{\gamma_k} + \beta_k$. For each SNP, the α_k, β_k and γ_k were selected based on the posterior distributions computed above, so that 0, 1, or 2 true underlying copies of a given allele mapped to 1, 2, or 3, as required by the hidden Markov model.

Per-gene copy number

We used per-gene averaging to determine the total copy number per gene in each cell line. We calculated the ploidy-corrected copy number (CN_{pc}) per gene as follows:

$$\mathrm{CN_{pc}} = \frac{\mathrm{CN_{total}}}{\mathrm{ploidy}} - 1$$

We called amplified genes as > 1 CN_{pc} and deleted genes as < -0.75 CN_{pc}. As an example for a tetraploid (4n) cell line the ploidy corrected copy number of 1 or higher corresponds to 8 or more DNA copies and a corrected copy number of -.75 or less corresponds to less than 1 copy.