**Supplementary Methods**

*Data resources*

TCGA DNA and RNA sequencing data were downloaded from Cancer Genomics Hub (CGHub, <https://cghub.ucsc.edu>). Copy number segmentation data and gene expression data were downloaded from Firehose (<https://gdac.broadinstitute.org/>). Somatic mutation data were downloaded from UCSC Xena repository (<https://xenabrowser.net/datapages/?cohort=TCGA%20Pan-Cancer%20(PANCAN)>). Genomic Variants database was retrieved from <http://dgv.tcag.ca/dgv/app/home>. All data used in this study were summarized in **Table S1**. We excluded 41 samples from the 689 normal samples because they clustered with tumor samples in unsupervised hierarchical clustering. The clustering was done within each cancer type using expression of all genes and ward’s method. The resulting panel of normal samples (n=648) were subjected to the same fusion detection pipeline and were used as controls to filter out potential germline fusion events and artifacts.

*Identification of fusion transcripts*

We applied PRADA (1) to all RNAseq samples for data preprocessing and fusion calling. In brief, RNA sequencing reads were aligned to a composite reference consisting of both genome (hg19) and transcriptome (Ensembl 64), followed by a remapping step that aligns transcriptome coordinates to the reference genome(2). GATK best practices were implemented in the pipeline, including marking duplication and base quality recalibration. More information about PRADA can be found at <http://bioinformatics.mdanderson.org/main/PRADA:Overview>.

PRADA detects fusion transcripts based on discordant read pairs (reads mapping to different protein-coding genes) and junction spanning reads (reads mapping to the exon–exon junctions). We required at least two discordant read pairs and one junction spanning read to call a fusion candidate. All fusion candidates were collected and were subject to additional filtering. The filters were described as follows: (1) candidates observed in normal controls were removed; (2) candidates with highly similar partners in sequence (blastn e-value≤0.001) were removed; (3) candidates with low transcriptional  allelic fraction were removed (TAF, minimum 0.01 for both partner genes); (4) candidates with very promiscuous partner genes were removed (the Partner Gene Variety filter, see below); (5) Candidates with identical junctions in more than 15 samples were removed; (6) candidates with supporting reads mapped disproportionately to sense and antisense strands were removed. Transcriptional allelic fraction (TAF) was calculated as the ratio of fusion supporting junction spanning reads to the total number of reads spanning the junction involved in the fusion. Partner Gene Variety (PGV) was defined as the number of unique chromosomal arms where the partner genes were found. A higher PGV suggests a gene was found to fuse with more partner genes in a cancer lineage. For genes with PGV greater than 10, we used permutations (n=100,000) to model the background distribution of the random chances of obtaining the observed PGV (empirical p value). We removed fusions with empirical p value less than 0.001%. For filter (6), we hypothesized that ratio of sense and antisense strand mapping reads was proportional to the distance from the start of the fusing transcript to the junctions of the two partner genes. Since lower coverage and short distance may confound this ratio, we limited our filtering to fusions with more than 100 spanning reads and such distance more than 500 base pairs. We removed fusions that had this ratio greater than 100. The effectiveness of each filter was summarized in Supplementary table 3.

To establish a positive control fusion list, we integrated three resources including Mitelman(3), ChimerPub(4), and Cosmic fusions(5). Fusions reported in all three independent references were curated as a list of known fusions (n=321). Of these 321, 38 fusions were detected 359 times in our data set. The relatively limited overlap between the control fusion list and our uniformly detected fusions may the differences between the cancer types included in TCGA (n=33) and the much larger set of cancers informing the literature, sometimes with highly specific fusions (i.e. DNAJB1-PRKACA transcripts in fibrolamellar hepatocellular carcinoma).

*Validation of fusion transcripts through integrating structure variants and copy number changes*

For cases where both copy number profile and gene fusion were available, we aligned fusion junctions with copy number breakpoints. We allowed a 100 Kb window to the expected orientation for both partner genes when searching array based copy number data.

We detected structural variants (SVs) from whole genome sequencing (WGS) data using Speedseq with default parameters (6). Speedseq categorizes SVs into deletion (DEL), inversion (INV), duplication (DUP), and breakend (BND). Compared with other three categories, BND is less specific and often used to represent translocations and complex rearrangements. We filtered SVs requiring more than 3 supporting reads, i.e. at least one split read and one discordant read pair. For fold-back inversions (BND on the same chromosome) we required more than 9 supporting reads. We removed SVs with breakpoints falling in low-complexity regions (e.g. repeat region DNA), or stacking across different tumor types. We further removed SVs where the flanking 100 bp of the two breakpoints share high sequence similarity (blastn E-value > 0.0001). Germline events were filtered out by comparing with matched normal samples.

We scanned the intersection between the edge of confident interval from the supported structure variants including large fragment duplication, deletion, insertion and inversion and truncated intron region flanking the junction upon fusion events. We assigned two partner genes into three groups based on their relevant location of break points to adjacent break point of structure variants. High confidence group was defined when a break point of structure variants fell into the immediate intron of the fusing exon for both partner genes; low confidence group was defined when a break point of structure variants fell between the fusion junction and the start or end of the partner gene depending on the fusion orientation, or fell into the 100K window from the corresponding gene boundary; Intermediate confidence group was defined when one partner gene met criteria of high confidence group and the other met that of the low confidence group. For those fusion pairs with only one junction points supported by structure variants, we assigned as one-sided. For DUP, INV, and DEL we assigned fusion associations to each group. However, for BND we considered the scenarios of fusion-SV alignment. We counted a SV as a translocation only if the fusion junction of gene A located at downstream of SV break point of geneA, and fusion junction of gene B located at upstream of SV break point of gene B, and if the distance between the fusion junction / boundary of the gene and SV break point fall into 100kb window.

*Exons and transcription expression analysis of fusions*

Exons and transcripts expression of fusion partners are retrieved from normalized RSEM value of level3 RNA-seq from Firehose (<https://gdac.broadinstitute.org/>). We performed Z score transformed expression level across all samples in each cancer type to plot exon expression heatmaps.

*Fusion centrality analysis*

Fusion transcript centrality score was calculated based on domain-based fusion model using default parameters (https://bmsr.usc.edu/software/targetgene/), to predict the oncogenic driver in which partner genes act as hubs in a cancer pathway network (7). Fusions with centrality score > 0.37 were considered as potential drivers.

*Kinase analysis*

Genes were classified as kinases based on the UniProt/Swissprot database (https://www.ebi.ac.uk/uniprot) but this resource does not include information on the position of the kinase domain. We retrieved kinase domain location from the Human Protein Reference Database ((http://hprd.org/) which provided the relevant annotation on approximately half the kinase genes (1,052 of 2,126).

**References**

1. Torres-Garcia, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G. and Verhaak, R.G. (2014) PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*, **30**, 2224-2226.

2. Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res*, **20**, 413-427.

3. Mitelman, F., Johansson, B. and Mertens, F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*, **7**, 233-245.

4. Lee, M., Lee, K., Yu, N., Jang, I., Choi, I., Kim, P., Jang, Y.E., Kim, B., Kim, S., Lee, B. *et al.* (2017) ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res*, **45**, D784-D789.

5. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**, D805-811.

6. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R. and Hall, I.M. (2015) SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*, **12**, 966-968.

7. Wu, C.C., Kannan, K., Lin, S., Yen, L. and Milosavljevic, A. (2013) Identification of cancer fusion drivers using network fusion centrality. *Bioinformatics*, **29**, 1174-1181.