ONLINE METHODS

Sequence and domain analysis. We extracted 3,068,965 mRNA sequences from GenBank and mapped them to the human genome by BLAT (Kent, 2002)[21]. Sequences that aligned to exon boundaries of two different genes were considered fusion chimeras and compared to the Mitelman database of known fusions to identify deposited fusion sequences (http://cgap.nci.nih.gov/Chromosomes/Mitelman). The fusion proteins were delineated based on the exon recombination sites and the open reading frames of both partners. The conserved domains in each fusion protein were delineated based on the protein-domain mapping data extracted from the Entrez Gene database (http://www.ncbi.nih.gov/gene).

Interrogation of the gene-fusion network with the molecular-interaction network. The molecular interactions for human genes were extracted from the HPRD database[7], a resource that contains expert-curated reference protein-protein interactions. The gene fusion network was constructed using established fusions from the Mitelman database. We applied hypergeometric probabilities to detect the enrichment of gene fusion partners in the molecular interactions sets. Suppose an interaction gene set for gene j, consisting of N interacting genes, and a fusion partner set for gene i, consisting of x partners; the intersection of these two sets is calculated as k_{ij}. Then, taking the complete set of all human genes (size n), the probability that k_{ij} is a more significant overlap than expected by chance is calculated using the hypergeometric distribution (Fig. 1a). Using these statistics, the gene fusion network was interrogated with the molecular interaction network. To evaluate the top fusion-interaction network hub candidates, we resolved the fusion-interaction network for shared interacting genes with $P < 10^{-7}$ ($\geq$3 connectivities with a fusion partner group). The fusion-interaction network was visualized by VisANT[9] and then processed by the spring embedded relax function. The fusion partner groups that fall into the six major clusters were exhibited together with their shared interacting genes on Fig. 1d. The hubs were nominated based on the significance from the above statistical test within each subset of connected fusions, and ablating drugs were identified by mapping the hubs to the DrugBank database (http://www.drugbank.org/) as of August 8, 2008 (ref. 23).

Enrichment analysis of cancer genes in the compendium of molecular concepts and calculation of the ConSig score. We compiled 28,963 molecular concepts from the Gene Ontology database (http://www.geneontology.org/), the Reactome database (http://www.reactome.com/), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), Biocarta (http://cgap.nci.nih.gov/Pathways), the HPRD database[7] (http://www.hprd.org/) and the Entrez Gene conserved domain database (http://www.ncbi.nlm.nih.gov/gene) (Table 1). In the processing of gene ontologies, the genes that appeared in the child ontologies were subtracted from the parents to avoid duplicate representation. Next, we mapped and analyzed the enrichment of established fusion or point mutation genes against all concepts and calculated the fusion and mutation ConSig score for all known human genes based on their participation in signature concepts. The point mutation genes were compiled from the Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/). Computationally, let k be the number of concepts associated with a specified gene. Let n_i represent the number of total genes and x_i represent the number of fusion or mutation genes participating in a given concept i, $i = 1,\ldots,k$. The ConSig score then integrates a signal measure of fusion or mutation genes participating in concept i ($x_i/n_i^{0.5}$) over all possible i, with the incorporation of normalization factor for k

using the formula:

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\log_{10}\left(1+\frac{x_i}{\sqrt{n_i}}\right)$$

With this computation, if a gene has high probability to be involved in gene fusions or mutations, the fusion/mutation ConSig score will be high; thus the radius in the two-dimensional ConSig-score plot for fusions and mutations will correlate with the role of tested genes in cancer. To eliminate the bias from the gene itself in the overlap, the seeding genes were subtracted from the signature concepts during the calculation of their own ConSig score. A step-by-step protocol for ConSig analysis is available at: http://s333404265.onlinehome.us/consig_protocol.htm

Kolmogorov-Smirnov analysis for ConSig score. The established cancer genes from the Mitelman (http://cgap.nci.nih.gov/Chromosomes/Mitelman) and cancer gene consensus databases (http://www.sanger.ac.uk/genetics/CGP/Census/) were used as a prototype, and compiled into ordered gene lists by descending *r*ConSig score. The enrichment of these established cancer genes in top scored genes was measured using the Kolmogorov-Smirnov rank statistic (K-S, $P = 1.39e^{-114}$). Let X be the number of known cancer genes in the ordered gene list ($X = 470$). Set $Y = n/X\text{-}1$ where n represents the total number of human genes interrogated and construct a vector V where V(i) is the component corresponding to gene i. Let V(i) = Y if i is in the target gene set and V(i) = $-X$ if not. Thus, our K-S statistical score is the maximum value of the running sum of consecutive values of V(i).

Random gene set statistics. Randomization tests were performed to evaluate the statistical significance of our observations. First, to test whether the fusion partner groups are significantly more linked by mutual interacting genes than by chance, randomized gene sets were generated with the same gene sizes and an equal amount of interacting genes as the fusion partner groups. Fusion genes that have fewer than 58 interacting genes will be substituted by genes with the same number of interactions; the others will be substituted randomly by genes having ≥58 interactions. Then the number of statistically significant links generated by the HPRD database were calculated ($P < 0.01$). This process was permutated for 1,000 times; none of the random gene family sets generated more significant links than fusion partner groups ($P < 0.001$). Second, to test the significance of ConSig score in isolating known cancer genes, randomized gene sets were generated corresponding to the sizes of the fusion and mutation gene lists. Then ConSig scores were calculated as if these random genes were actual cancer genes. As above, the K-S score was calculated and recorded. This process was repeated ten times for each cancer gene list size, resulting in nonsignificant K-S statistical scores, thus validating the K-S score as unbiased and providing a null distribution of ConSig score under the null hypothesis of no functional signal in the input gene list.

Meta-analysis of public array CGH/SNP data sets for multiple human cancers. Public array CGH/SNP data sets were compiled from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). A total of seven data sets were included in this study (GSE4659, GSE8918, GSE7255, GSE9611, GSE9113, GSE3930 and GSE8398), covering six cancer types (leukemia, lymphoma, sarcoma, salivary adenoma, brain and prostate tumors). The samples from each data set were manually curated and classified according to pathological associations. For Affymetrix SNP arrays, model-based expression was per-

formed to summarize signal intensities for each probe set using the perfect-match/mismatch (PM/MM) model. For copy number inference, raw copy numbers were calculated for each tumor sample by comparing the signal intensity of each SNP probe set against a diploid reference set of samples. In two-channel array CGH data sets, the differential ratio between the processed testing channel signal and processed reference channel signal was calculated. All resulting relative DNA copy number data were log2 transformed, which reflects the DNA copy number difference between the testing and reference channels. For normalization, log ratios were transformed into a normal distribution with a mean of 0 under the null model assumption. The data were then segmented by the circular binary segmentation (CBS) algorithm[23]. Cutoffs of 0.3 and −0.4 were used to call amplifications and deletions, respectively. To explore the evidence of fusion breakpoint pattern at the *NFE2* loci in lung cancer, we compiled the SNP array data of lung cancer tissues and cell lines from publication[24] and array express (E-MTAB-38) respectively. The relative copy number data were inferred and segmented as discussed above to reveal the DNA breakpoint patterns.

Analysis of paired-end transcriptome sequencing data. Mate pair transcriptome reads were mapped to the human genome (hg18) and Refseq transcripts, allowing up to two mismatches, using Efficient Alignment of Nucleotide Databases (ELAND) program within the Illumina Genome Analyzer Pipeline. Using a Perl script, we parsed the Illumina export output files to identify chimerical mate pairs with the following criteria: (a) putative chimeras must be supported by at least one mate pair that is the best unique match across genome; and at least three mate pairs in total; (b) the distances between the 5′ and 3′ partners of the intrachromosome chimeras must be more than 1 Mb. The resultant candidate chimeras were aligned by an *r*ConSig score of 3′ partner genes to reveal functionally important gene fusions in lung cancer cell lines.

RT-PCR and sequencing. RNAs from lung cancer cell lines, obtained from the American Type Culture Collection, were extracted and reverse transcribed with superscript III (Invitrogen) and random primers. Polymerase chain reaction was performed with Platinum Taq High Fidelity and fusion or NFE2 specific primers for 35 cycles. The primers used in this study are listed in Supplementary Table 3. Products were resolved by electrophoresis on 1.5% agarose gels, and TOPO TA cloned into pCR 4-TOPO. Purified plasmid DNA from at least four colonies was sequenced bidirectionally using M13 Reverse and M13 Forward primers on an ABI Model 3730 automated sequencer at the University of Michigan DNA Sequencing Core. Quantitative PCR (qPCR) was performed using the Step One Real Time PCR system (Applied Biosystems). The amount of each target gene relative to the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH) for each sample was determined using the comparative threshold cycle (Ct) method (Applied Biosystems User Bulletin #2, http://docs.appliedbiosystems.com/pebiodocs/04303859.pdf). For the experiments presented in Figure 4b, the relative amount of the target gene was calibrated to the relative amount from a lung cancer cell line with the latest Ct value.

Gene expression data analysis. To determine the expression of *R3HDM2* and *NFE2* in lung cancer cell lines and normal tissues, we interrogated the gene expression study of 73 lung cancer cell lines[25], and the 40 normal tissue data set[26], using the Oncomine database (http://www.oncomine.org/)[27]. Descriptions of tissue types from both data sets are provided in Supplementary Table 12.

Fluorescence in situ hybridization (FISH). To detect possible translocations on lung cancer

cell lines involving *R3HDM2* and *NFE2* loci, we used break-apart and colocalizing probe FISH strategies, with two probes spanning the *R3HDM2* locus (digoxin-dUTP labeled BAC clone RP11-258J5 (5′ *R3HDM2*) and biotin-14-dCTP labeled BAC clone RP11-799O6 (3′ *R3HDM2*)) and *NFE2* locus (digoxin-dUTP labeled BAC clone RP11-753H16 (5′ *NFE2*) and biotin-14-dCTP labeled BAC clone RP11-621J12 (3′ *NFE2*)). All BAC clones were obtained from the Children's Hospital of Oakland Research Institute (CHORI). Prior to FISH analysis, the integrity and purity of all probes were verified by hybridization to metaphase spreads of normal peripheral lymphocytes. For interphase FISH on lung cancer cell lines, interphase spreads were prepared using standard cytogenetic techniques. For interphase FISH on a lung cancer tissue microarray, tissue hybridization, washing and color detection were performed as described[28,29]. The total evaluable cases include 76 lung adenocarcinoma cases. For evaluation of the interphase FISH on the tissue microarray, an average of 50–100 cells per case were evaluated for assessment of the *NFE2* rearrangement. In addition, formalin-fixed paraffin-embedded (FFPE) tissue sections from a positive case were used to confirm the tissue microarray results.

Small RNA interference, cell proliferation and invasion assays. The NFE2-fusion-positive H1792 cell line and an H460 cell line with low NFE2 expression were plated into 10-cm dishes and transfected with siRNA against NFE2 or nontargeting controls. Transfection was performed with oligofectamine following manufacturer's suggestion (Invitrogen). Forty-eight hours post-transfection, cells were trypsinized and counted. For each treatment, equal amounts of cells were plated into 24-well plates for cell counting, 96-well plates for WST-1 assay and Boyden invasion chambers for invasion assay. The rest of the cells were harvested for qPCR analysis. The knockdown study on H1792 cell lines was performed twice.

Cell-counting analysis was performed by Coulter counter (Beckman Coulter) at the indicated time points in triplicate. WST-1 proliferation assay was performed using manufacturer's protocol (https://www.roche-applied-science.com/pack-insert/1644807a.pdf). Invasion assay was performed as described previously[30].