

An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer

Xiao-Song Wang¹⁻³, John R Prensner^{1,3,8}, Guoan Chen^{4,8}, Qi Cao^{1,3}, Bo Han^{1,3}, Saravana M Dhanasekaran^{1,3}, Rakesh Ponnala¹, Xuhong Cao^{1,3}, Sooryanarayana Varambally^{1,3,5}, Dafydd G Thomas³, Thomas J Giordano³, David G Beer⁴, Nallasivam Palanisamy^{1,3}, Maureen A Sartor², Gilbert S Omenn² & Arul M Chinnaiyan^{1-3,5-7}

Cancer genomes contain many aberrant gene fusions—a few that drive disease and many more that are nonspecific passengers. We developed an algorithm (the concept signature or ‘ConSig’ score) that nominates biologically important fusions from high-throughput data by assessing their association with ‘molecular concepts’ characteristic of cancer genes, including molecular interactions, pathways and functional annotations. Copy number data supported candidate fusions and suggested a breakpoint principle for intragenic copy number aberrations in fusion partners. By analyzing lung cancer transcriptome sequencing and genomic data, we identified a novel *R3HDM2-NFE2* fusion in the H1792 cell line. Lung tissue microarrays revealed 2 of 76 lung cancer patients with genomic rearrangement at the *NFE2* locus, suggesting recurrence. Knockdown of *NFE2* decreased proliferation and invasion of H1792 cells. Together, these results present a systematic analysis of gene fusions in cancer and describe key characteristics that assist in new fusion discovery.

Gene fusions resulting from chromosomal rearrangements often define molecular subtypes of cancers and appear as initial events in oncogenesis¹. The discovery of recurrent fusions in common epithelial cancers²⁻³ has stimulated a widespread search for novel gene fusions. Yet, new fusion discovery and molecular targeting of known fusions is complicated by the complex biological behavior displayed by fusion genes. First, most genes involved in fusions recombine with many different partners, forming interrelated gene fusion networks⁴. Second, recurrent gene fusions in carcinomas are often found in the background of many nonspecific gene fusions, which illustrates the karyotypic complexity of solid tumor evolution. Distinguishing

nonspecific (passenger) fusions from recurrent (driver) fusions is a formidable task. In this study, we sought to investigate the functional and genetic landscape of fusion genes and characterize fundamental principles to help facilitate new gene fusion discovery from large-scale genomic data and next-generation sequencing data.

RESULTS

Understanding the recombination of fusion partners

To determine common characteristics of fusion gene recombinations, we explored the hypothesis that fusion genes sharing a common partner might share common domain architectures. Using GenBank, we extracted core nucleotide sequences of chimeras representing known fusions. Open reading frames and their domain architectures were determined using the Entrez Gene conserved domain database. The resulting unique domain architectures were clustered by domain similarities, enabling the global analysis of domain recombination in gene fusions (**Supplementary Results**). Interestingly, the domain architectures of fusion proteins are very diverse, especially for 5' partners. In addition, clustering gene fusions according to their domain architectures resulted in few pathologically related clusters; the majority of the clusters did not show tumor-entity specificity. This suggested the possible existence of other major factors influencing fusion gene recombinations, such as preferential selection for shared pathways or gene ontologies.

We compiled pathway data from Reactome⁵, Kyoto Encyclopedia of Genes and Genomes (KEGG)⁶ and Biocarta, and analyzed the shared pathways within fusion partner groups. However, most fusion genes with a mutual partner are involved in distinct cell signaling pathways (data not shown).

Yet, because canonical pathways may not encompass the complexities of cell biology, we interrogated a molecular interaction database to generate a comprehensive view of cancer signaling. We derived 90 fusion partner groups from the Mitelman database and mapped these to the molecular interaction network extracted from the Human Protein Reference Database (HPRD)⁷. For all human genes in the database, we defined the interaction gene set J to be all genes that interact with gene j . If we denote a given fusion gene and its fusion partners as i and I , respectively, we can then individually test the significance of overlap between every set of fusion partners I with every gene interaction set J using the hypergeometric distribution (**Fig. 1a**). In aggregate, this analysis yielded a total

¹Michigan Center for Translational Pathology, Ann Arbor, Michigan, USA.

²National Center for Integrative Biomedical Informatics, CCMB, Ann Arbor, Michigan, USA. ³Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. ⁴Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA. ⁵Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan, USA. ⁶Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, Michigan, USA. ⁷Department of Urology, University of Michigan, Ann Arbor, Michigan, USA. ⁸These authors contributed equally to this work. Correspondence should be addressed to G.S.O (gomenn@umich.edu) or A.M.C (arul@umich.edu).

Published online 1 November 2009; doi:10.1038/nbt.1584

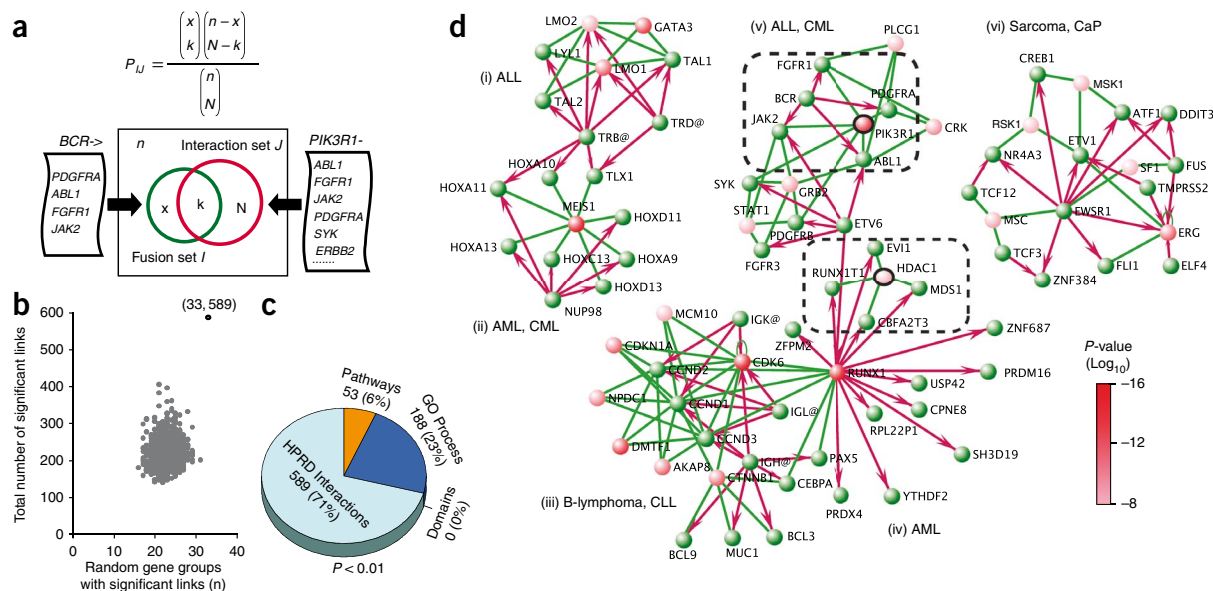


Figure 1 Exploring cancer-related gene fusions in the context of known molecular interaction networks. **(a)** The hypergeometric statistics for assessing whether a group of fusion gene partners (e.g., all *BCR* partners) contains an unexpected number of genes that physically interact with the same gene (e.g., all genes that interact with *PIK3R1*). **(b)** The total number of significant links (589) and the number of fusion partner groups having these links (33) were plotted with the distribution calculated from randomly chosen gene sets with an equal amount of connectivity (1,000 permutations). **(c)** Analysis of the fusion partner groups with a compendium of molecular concepts by hypergeometric statistics. The numbers in the pie chart represent the number of significant concepts in each functional category ($P \leq 0.01$). **(d)** Network visualization of the most significant ($P < 10^{-7}$) instances where many fusion partners also interact with a shared gene. Fusion genes are green nodes; shared interacting genes are red (with color intensity indicating P -value). Red arrows designate gene fusions (from 5' partners to 3' partners); green lines represent molecular interactions. For simplicity, genes and proteins are both given in roman type in the diagram. For each fusion partner set, the shared interacting gene having the most significant P -value was designated as a fusion-interaction hub. Clusters are limited to known fusions joined by established molecular interactions. (i) Acute lymphoblastic lymphoma (ALL) fusions with a hub of *GATA3*. (ii) Acute/chronic myelogenous leukemia (AML and CML) fusions with a hub of *MEIS1*. (iii) B-cell lymphoma and chronic lymphoblastic lymphoma (CLL) fusions through the hubs of *CDK6* and *CTNNB1*. (iv) AML fusions partially focusing on *HDAC1*; *RUNX1* is the hub of immunoglobulin fusions, and also involved in multiple fusions in AML, and thus links clusters iii and iv. (v) ALL and CML fusions through the hub of *PIK3R1*. (vi) Sarcoma and prostate cancer fusions around *ERG* and *MSK1*. The *PIK3R1* and *HDAC1* hubs are within dashed lines.

of 589 genes whose interacting genes were enriched for genes in 33 out of 90 fusion partner groups in the Mitelman database ($P < 0.01$). The top shared interacting genes are supplied in **Supplementary Results**.

To test whether fusion genes are significantly enriched for mutual interacting genes, we randomly chose 90 gene sets with an equivalent level of connectivity as the fusion partner groups (Online Methods), and determined the extent to which they were linked by mutual interacting genes. This process was repeated 1,000 times, and then the total number of significant links and the number of gene groups having these links were plotted (**Fig. 1b**). The number of links generated is significantly greater for fusion genes, validating our observation ($P < 0.001$).

To systematically evaluate the importance of shared interacting genes in fusion gene recombinations, we applied these statistics to the pooled domains, pathways, Gene Ontology (GO) database biological process and HPRD interactions data. We refer to these data sets collectively as 'molecular concepts' (**Table 1**). We benchmarked each of these data sets by statistically assessing the number of molecular concepts shared by fusion partner groups (**Fig. 1c**).

We next focused on the network of the most significant fusion-interaction (FI) links. We visualized fusion-interaction networks using the VisANT program⁸ and found six major clusters of interactions that connected gene fusions from similar tumor entities (**Fig. 1d**). The shared interacting genes with the greatest statistical significance in each subset of connected

fusions were designated as 'fusion-interaction hubs' in each cluster. For example, *BCR* has four 3' partners (*ABL1*, *FGFR1*, *JAK2* and *PDGFR*), all of which interact with *PIK3R1*, one of the fifteen subunits encoding PI3K ($P = 9.54 \times 10^{-11}$). This finding suggested that *BCR* fusion partners interact with and presumably activate *PIK3R1* as part of leukemogenesis, which we confirmed by mining the literature^{9–14}. These results show the utility of the fusion-interaction networks in elucidating fusion biology by distinguishing key genes that serve as network hubs with functional importance in mediating fusion signaling (**Supplementary Results**).

Quantification of concept signatures

The fact that cancer-related fusion partner groups tend to cluster around shared interacting genes or share common gene ontologies prompted us to generalize this finding to develop a method that could filter out nonspecific gene fusions. We hypothesized that such 'signatures' of molecular concepts frequently found in fusion genes may be used to define biologically meaningful gene fusions underlying cancer, similar to signature genes defining certain phenotypes. This requires a systematic characterization of all fusion genes as a coherent group from multiple functional perspectives.

To benchmark the functional characteristics of fusion genes, we compared fusion genes to point mutation genes in cancer. We used Fisher's exact test to identify molecular concepts enriched for fusion genes and concepts enriched for point mutations, generating two sets of minimally overlapping concepts (**Fig. 2a**). Fusion genes were

Table 1 The compendia of molecular concepts for integrative functional analysis of fusion genes

Class	Source	Web link	Type	Concepts (n)	Connectivity (n)
Annotation	Gene Ontology	http://www.geneontology.org/	Biologic process	3,920	46,530
			Cellular component	732	42,463
			Molecular function	2,561	47,026
Pathways	Biocarta	http://cgap.nci.nih.gov/Pathways	Signaling pathways	263	4,459
	KEGG	http://www.genome.jp/keg	Metabolic pathways	112	2,985
	Reactome	http://www.reactome.com/	Signaling pathways	2,456	52,238
			Biochemical reactions	5,450	44,347
Interactions	HPRD	http://www.hprd.org/	Protein interaction sets	7,819	37,206
Domains	Entrez Gene	http://www.ncbi.nlm.nih.gov/gene	Conserved domains	5,650	5,693

Four classes of molecular concepts were compiled from six sources. Connectivity represents the total number of concept to gene connections in each concept type.

enriched for molecular concepts related to signal transduction and transcription activation; in contrast, mutation genes were enriched for molecular concepts related to DNA repair and cell cycle checkpoints. Thus, we defined these two sets as ‘concept signatures’—a fusion concept signature and a mutation concept signature.

Using these two concept signatures, we hypothesized that genes involved in fusions or point mutations could be distinguished from each other and from the remaining human genes. We designed an algorithm, termed concept signature score (ConSig score), to quantitatively rank genes underlying cancer by the strength of their association with the two concept signatures (Fig. 2b). The algorithm first determines the ‘relevance’ of each concept in a signature, where relevance is defined as \log_{10} of the number of fusion (or point mutation) genes that are associated with that concept divided by the square root of the total number of genes in the concept. Then, the ‘fusion ConSig score’ of a gene is calculated by summing the relevances of the fusion signature concepts associated with the gene, normalized for the total number of assigned concepts k . The ‘mutation ConSig score’ is similarly calculated except using the concepts in the mutation signature.

An important step in this analysis was to remove the redundant information from the calculation of the ConSig score. First, to avoid redundant representation in the GO database, we subtracted the genes that appeared in the child ontologies from the parents. Second, to eliminate the bias from the gene itself in the overlap, we subtracted the seeding genes from the signature concepts during the calculation of their own ConSig score. Finally, to minimize the redundant information in the interactome and pathway databases, we removed the pathways significantly overlapping with the molecular interactions (Fisher’s exact test, $P < 0.01$) in the calculation of the ConSig score. However, this adjusted ConSig score did not show an advantage over the unadjusted score (Supplementary Results).

We calculated fusion and mutation ConSig scores for all known human genes. Plotting the fusion and mutation ConSig scores separated known fusion genes from mutation genes (Fig. 2c). The distinction line (D-line), $y = 1.67x$, was determined by testing optimal separation capacity, which separates 85% of mutation genes from 80% of fusion genes (Supplementary Results). In this setting, the radius to the zero point is defined as the radial ConSig score of a gene (r ConSig score), which indicates the strength of association with signature concepts of both fusion and mutation genes, thus implies the functional relevance of candidate genes in cancer. The distance vector from the node to the D-line, which illustrates a distinction between fusion and mutation genes, is defined as the distinction ConSig score (d ConSig score). Rating all human genes by the r ConSig score produced enrichment of established cancer genes

in top-scoring genes, with the majority of fusion or mutation genes matching the prediction from the d ConSig score (Fig. 2d). Replacing the fusion or mutation gene sets with random gene sets produced no enrichment of the randomly selected genes. Although the ConSig algorithm is able to segregate fusion genes and mutation genes, we propose that its main utility is in the identification of biologically important gene fusions from next-generation sequencing data, where a large number of candidate gene fusions hinders a quick discussion evaluation of their functional importance (Supplementary Discussion).

Genetic characteristics of unbalanced fusion genes

Having evaluated fusion genes by functional traits, we next used high-throughput copy number data to explore the genomic imbalance pattern that could inform unidentified gene fusions. Using leukemia as a genetic model, we studied the recurrent fusion genes in a high-resolution single nucleotide polymorphism (SNP) microarray data set with 304 leukemia samples^{15,16}. A total of 157 samples are annotated with seven gene fusions in this data set (Supplementary Results). The percentage of unbalanced fusions ranged from 21.2–94.1% for different fusions, with most *TCF3-PBX1* fusions identifiable by unbalanced breakpoints (Fig. 3a). The physical lengths of amplifications or deletions associated with fusion genes were 0.08–84.21 Mb (averaging 19.7 Mb). We observed a surprising heterogeneity in the genomic aberrations generating gene fusions. Often two fusion partners were found to possess different degrees of copy number gain or loss; elsewhere one fusion partner harbors a balanced translocation whereas the other partner has an unbalanced translocation.

Despite this diversity, an association analysis of unbalanced breakpoints with fusion gene placements revealed a consistent genetic pattern: copy number increases generally affect the 5’ region of 5’ partners and the 3’ region of 3’ partners, whereas deletions generally remove the 3’ region of 5’ fusion partners and the 5’ region of 3’ partners. Of 56 samples with 7 unbalanced fusions in this data set, 55 samples follow this pattern (Fig. 3b and Supplementary Results). We further analyzed the data for 36 leukemia cell lines¹⁵ and associated gene fusions from published sources¹⁷; 11 of 12 unbalanced fusions from these cell lines were found to follow this pattern (Fig. 3c and Supplementary Results). We termed this pattern the ‘fusion breakpoint principle’. Based on this reasoning, we can deduce an inferred principle for the unbalanced gene fusions within the same chromosome (Supplementary Results). For gene fusions having two partners on the same DNA strand, we define the fusion as ‘consistent’ if the genomic location of the two partners parallels their positions within the fusion transcript (that is, 5’ partner at the 5’ side of the 3’ partner), or ‘inconsistent’ if the two partners display the opposite genomic positioning (that is, 5’ partner at the 3’ side of the 3’ partner). Then, consistent fusions cannot be generated by a copy number increase, whereas inconsistent fusions cannot be generated by a deletion. For gene fusions having two partners on different strands (inversion), the fusion cannot be generated by simple interstitial deletions or copy number increases.

Although the fusion breakpoint principle can be inferred based on conventional cytogenetics analysis, it should be stressed that unlike G-banding and fluorescence *in situ* hybridization (FISH), array-based high-throughput genomic data loses balanced genomic translocation information, and may misrepresent individual cases of complex genomic rearrangements (Supplementary Results on

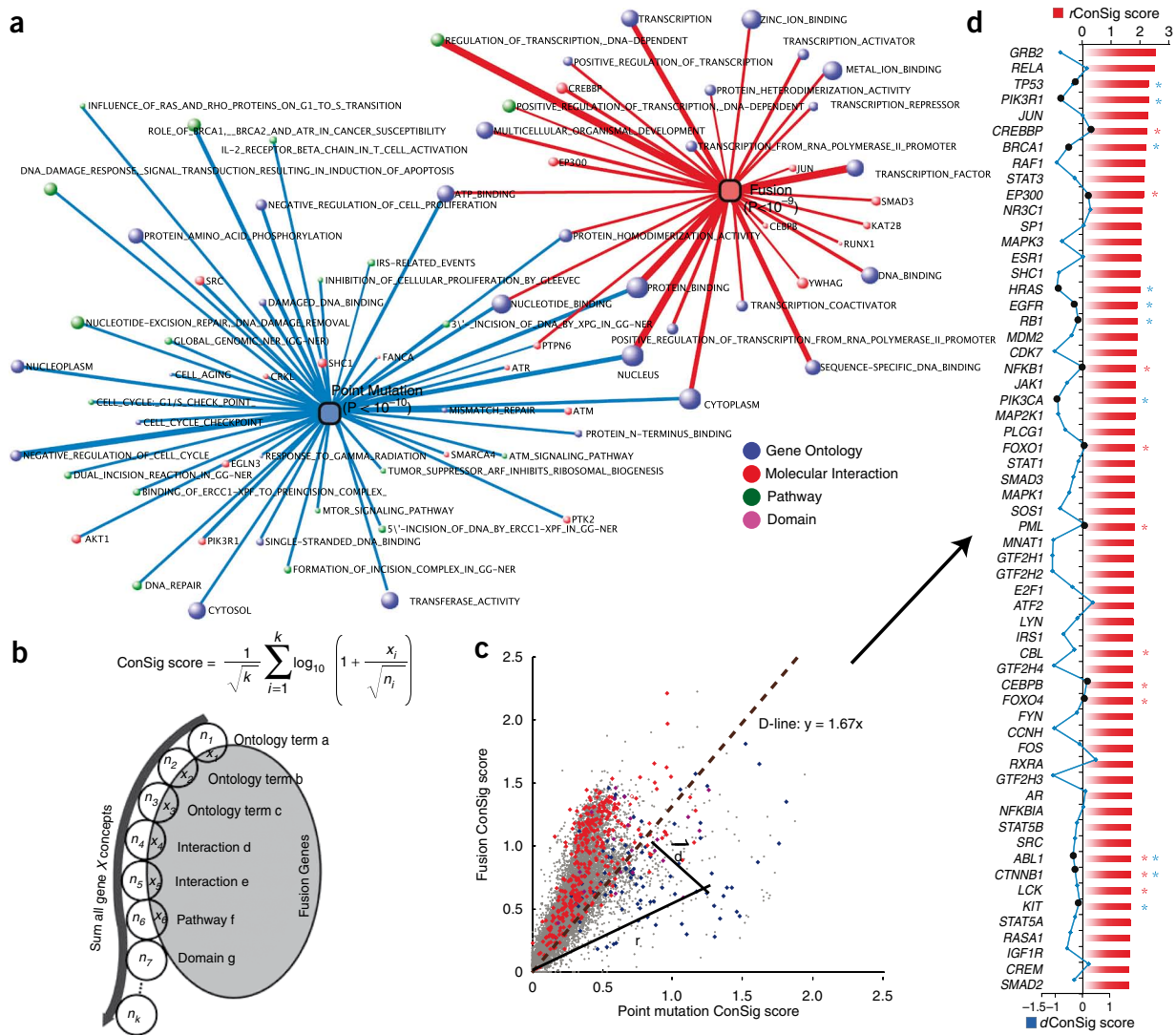


Figure 2 Distinguishing biological features of gene fusions and point mutations in cancer. **(a)** Enrichment analysis with a compendium of molecular concepts generates two sets of minimally overlapping signature concepts for fusion and point mutation genes. Molecular concepts are depicted as nodes with the size of each node corresponding to the number of genes in each concept. The thickness of the lines correlates with the significance of overlap tested by hypergeometric statistics. **(b)** The ConSig algorithm. Circles represent concepts associated with gene *X*. *k* is the total number of concepts associated with gene *X*, *x_i* = number of fusion genes in concept *i*, *n_i* = total genes in concept *i*. A corresponding figure for mutation genes is not shown. **(c)** Fusion and mutation ConSig scores for known fusion genes (red dots), cancer point mutation genes (blue dots) and all other human genes (gray dots). Genes known to be both fusion and mutation genes are purple. *r*, *r*ConSig score; *d*, *d*ConSig score, D-line, distinction line. **(d)** Identifying the top 60 genes rated by *r*ConSig score (red column chart) produced a list highly enriched for established cancer genes (known fusion genes labeled with red stars; mutation genes blue stars). The *d*ConSig scores are depicted by a blue line chart, with positive and negative values indicating the dots above or below the D-line. Known mutation or fusion genes matching the prediction by *d*ConSig scores are marked by black dots in the line chart.

the *MLL-AF9* fusion). For this reason, extensive evidence from large numbers of malignancies is required to confirm the applicability of this principle to high-throughput genomic data.

To confirm the breakpoint principle, we performed a large-scale meta-analysis of recurrent gene fusions based on high-resolution array comparative genomic hybridization (array-CGH) and SNP array data sets annotated with gene fusions, as well as literature curation (**Supplementary Results**). In total, 276 tumor samples were identified as having unbalanced fusions, including 85 leukemia, 15 lymphoma, 23 sarcoma and 153 epithelial tumor samples. Although diverse breakpoint patterns were observed on these samples (**Supplementary Results**), the unbalanced fusions from 273 samples conformed to the principle (98.9%). Furthermore, we

also confirm the inferred principle by analyzing the reports for all unbalanced intrachromosome fusions from the Mitelman database (**Supplementary Results**).

An integrative approach to new fusion discovery

To demonstrate the application of those principles to new fusion discovery, we analyzed next-generation sequencing data and large-scale genomic data from lung cancer. First, we used the ConSig score to nominate biologically important fusion candidates from paired-end transcriptome data from lung cancer cell lines run in a single lane on an Illumina Genome Analyzer II flow cell. We extracted the chimeric paired reads from the paired-end libraries, and then ranked the 3' partners by *r*ConSig score. Second, the DNA breakpoints at

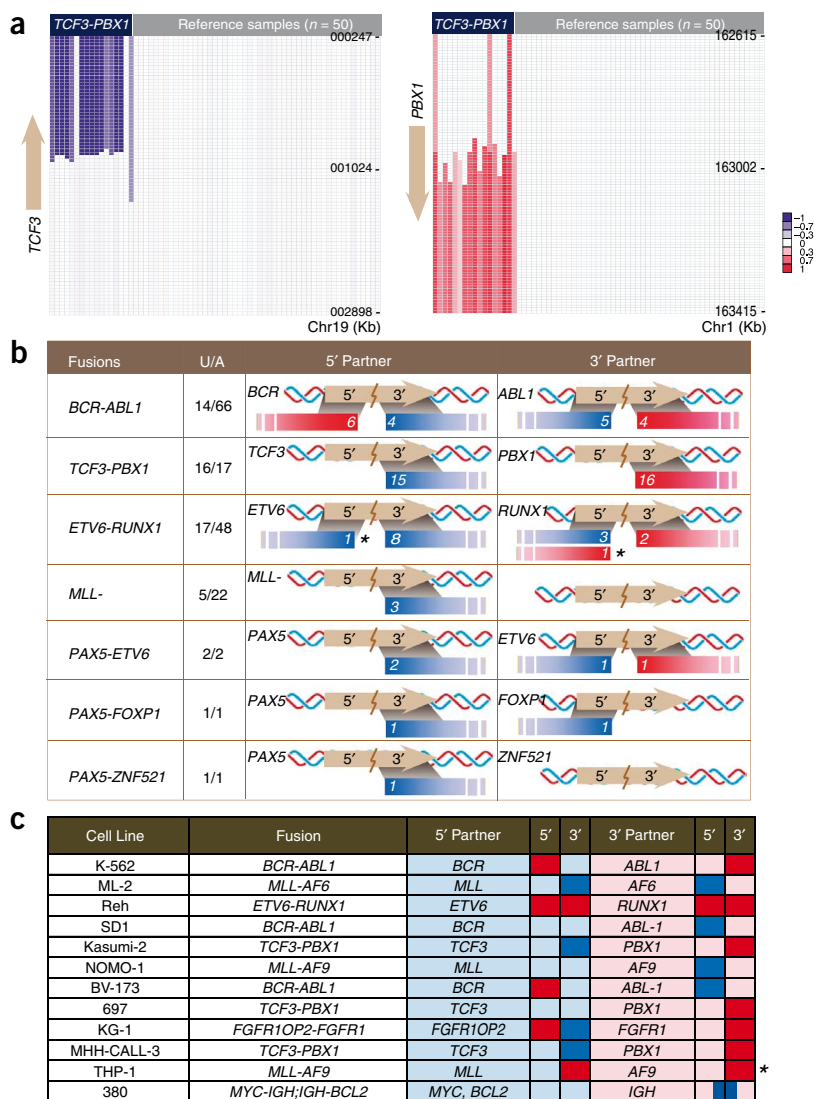


Figure 3 Characterizing the genomic imbalances of recurrent gene fusions in acute lymphocytic leukemia. An SNP array data set was used to evaluate genomic aberrations associated with gene fusions in acute lymphocytic leukemia (ALL). **(a)** The recurrent *TCF3-PBX1* fusion ($n = 17$) was associated with deletion of the 3' region of *TCF3* and duplication of the 3' region of *PBX1*. Color scales indicate \log_2 transformed relative copy number. **(b)** Of the 56 samples with unbalanced gene fusions in this data set, 55 samples conformed to the fusion breakpoint principle. Fusion genes on their corresponding chromosomes were aligned with the unbalanced breakpoints. "U/A" represents the number of samples with unbalanced fusions out of the total samples with gene fusions. Colored bars in the 5' partner and 3' partner columns indicate copy number increases (red) or deletions (blue), and given in white is the number of samples having the DNA aberration. For example, the first row shows that out of 66 cases of ALL patients with *BCR-ABL1* fusions, 14 cases are unbalanced, of which 6 have 5' *BCR* duplication and 4 have 3' *BCR* deletion. *ABL1* can be interpreted in a similar fashion. **(c)** Unbalanced fusions in the 12 leukemia cancer cell lines follow the fusion breakpoint principle. The exceptions to the principle are marked by asterisks.

several additional lung adenocarcinoma cell lines (**Fig. 4b**); however, no rearrangements were detected in these samples by FISH, suggesting other mechanisms activating *NFE2* expression (**Supplementary Results**).

The *R3HDM2-NFE2* fusion was predicted to encode the full-length open reading frame of *NFE2*, with only untranslated promoter sequences contributed from *R3HDM2* (**Fig. 4d**), and exon-walking qRT-PCR demonstrated the specific overexpression of the *NFE2* coding exons 2–3 under the regula-

tion of the *R3HDM2* promoter (**Supplementary Results**). In H1792, knockdown of *NFE2*, which encodes a transcription factor normally expressed during erythropoiesis, resulted in a marked decrease in cell proliferation and to a lesser extent cell invasion (**Fig. 4e**), whereas no effect was seen in H460, which has low levels of endogenous *NFE2* (**Supplementary Results**).

Analysis of SNP array data for 139 lung adenocarcinoma tissues revealed copy number gain consistent with the fusion breakpoint principle at the 3' *NFE2* locus in two people with lung cancer (**Fig. 4f**), suggesting possible recurrent aberrations involving the *NFE2* locus in this cancer. We therefore performed FISH analysis on a lung cancer tissue microarray comprised of a cohort of 76 lung adenocarcinoma samples, which confirmed recurrent *NFE2* rearrangements in two individuals (**Fig. 4g**).

DISCUSSION

The complex biological events contributing to tumorigenesis are frequently driven by chromosomal rearrangements. Although previous studies have observed the generation of recurrent fusions at the edges of genomic imbalances¹⁸, efforts to identify cancer-promoting fusions from unbalanced breakpoints have been met with limited success, often discovering nonfunctional aberrations that appear to

the genomic loci of top candidate fusion genes were evaluated on the basis of the fusion breakpoint principle using publicly available lung cancer SNP array data encompassing a large number of tumor samples to search for recurrent rearrangements.

We first tested the ConSig approach on the H2228 cell line known to harbor the recurrent *EML4-ALK* fusion. Rating 3' partners of paired-end chimeras by *r*ConSig score revealed *EML4-ALK* as the top-ranked candidate on the H2228 cell line, which was supported by six mate pairs (**Fig. 4a**, left). This showed the effectiveness of the *r*ConSig score in preferentially nominating driver gene fusions from numerous paired-end chimeras.

We then applied this method to reveal driver gene fusions from the transcriptome sequencing data of 12 lung cancer cell lines. Although there were 530 gene fusions in total supported by more than two paired reads, the 3' *r*ConSig score prioritized *R3HDM2-NFE2* as the lead in the H1792 lung cancer cell line (supported by three paired reads, **Fig. 4a**, right), and this fusion was confirmed by quantitative RT-PCR (qRT-PCR) (**Fig. 4b**), conventional capillary sequencing and interphase FISH, the latter of which showed high copy number gain of *R3HDM2-NFE2* in H1792 (**Fig. 4c**). Consistent with previous microarray data on lung cancer cell lines (**Supplementary Results**), qRT-PCR also revealed marked overexpression of *NFE2* on H1792 and

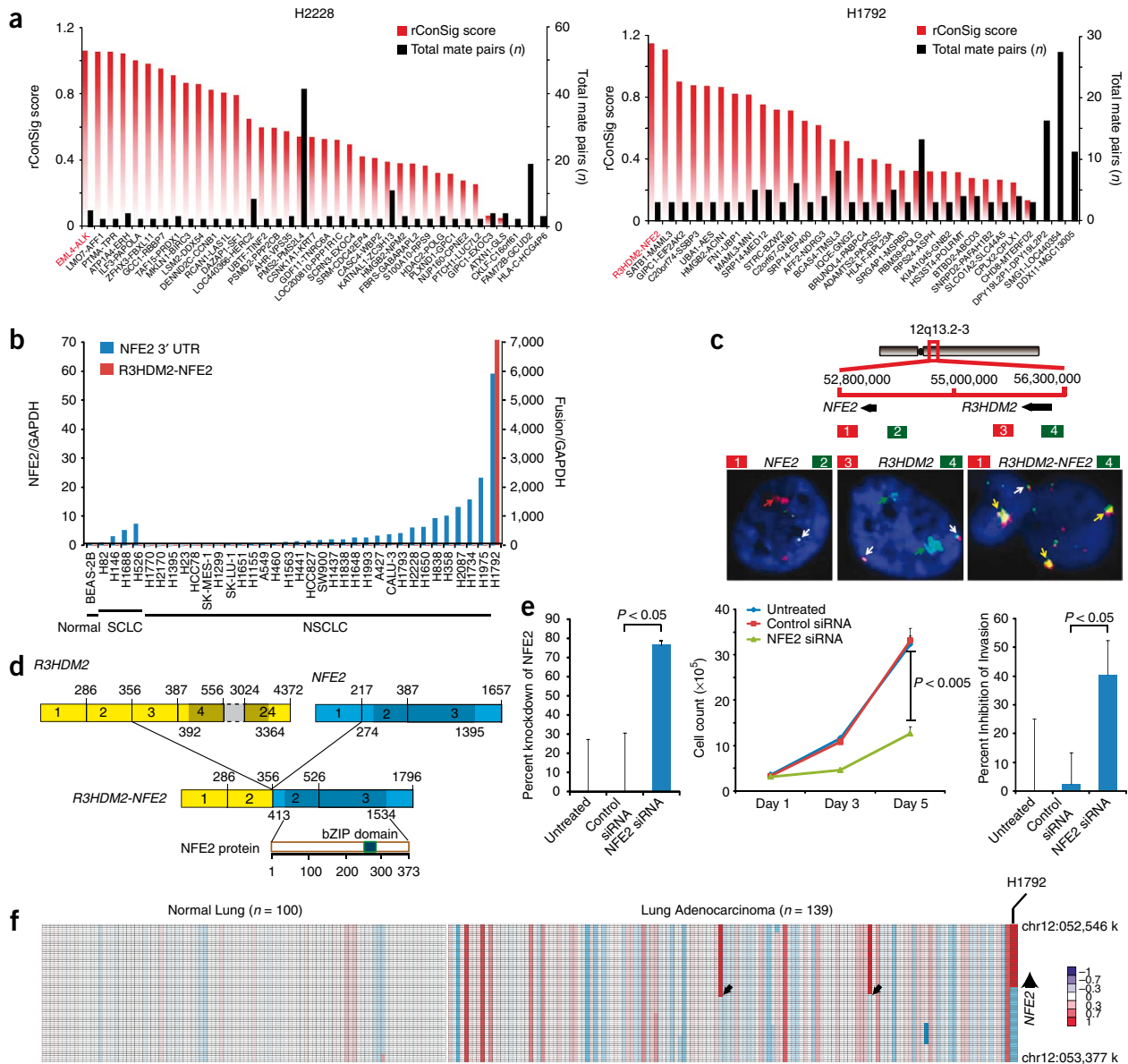


Figure 4 Discovery and validation of the *R3HDM2-NFE2* fusion using the ConSig algorithm and the fusion breakpoint principle. (a) Pair-end transcriptome sequencing of 12 lung cancer cell lines, followed by prioritizing the 3' partners of paired-end chimeras (≥ 3 paired reads) by *rConSig* score (red bars). Left, *EML4-ALK* is a candidate fusion, and is supported by six paired reads (black bars), in the H2228 lung cancer cell line (known to harbor this fusion). Right, ConSig analysis nominates the *R3HDM2-NFE2* fusion as the top candidate in the H1792 lung cancer cell line. (b) The *R3HDM2-NFE2* fusion was confirmed by RT-PCR and sequencing of the PCR product. qRT-PCR of wild-type *NFE2* revealed overexpression of *NFE2* in a subset of lung cancer cell lines; only H1792 cells express the chimeric *R2HDM2-NFE2*. (c) Top, schematic of the genomic organization of *R3HDM2-NFE2* fusion, with red and green bars indicating the location of BAC clones. This fusion was generated by an intrachromosomal translocation. Bottom, interphase FISH analysis showing amplification signals of 3' *NFE2* and 5' *R3HDM2* (left, middle) and *R3HDM2-NFE2* fusion (right) on H1792 cell line. Normal signals are indicated by white arrows; aberrant colocalizing signals by yellow arrows; aberrant split signals by green or red signals. (d) Schematic of the *R3HDM2-NFE2* fusion mRNA and protein. Structures for the *R3HDM2* and *NFE2* genes are derived from GenBank reference sequences. The numbers above the exons indicate the last base of each exon. Open reading frames are shown in darker shades. The exons of *R3HDM2-NFE2* fusion are numbered from the original reference sequence. The lower schematic shows wild-type *NFE2* protein and its domain architecture. (e) siRNA knockdown of *NFE2* in H1792 cells leads to decreased cell proliferation (middle graph) and invasion (right graph). Percent knockdown of the fusion transcript revealed by qRT-PCR is shown in the left graph. (f) Analysis of SNP array data from 139 lung adenocarcinoma tissues revealed recurrent copy number aberrations in two patients at the 3' *NFE2* locus, as well as the focal amplification of *R3HDM2-NFE2* fusion on H1792. (g) As in c, except the data are from three lung adenocarcinoma patients. L41 is a negative case with two colocalizing signals; L83 has split and high copy number gain at *NFE2* locus; L18 showed one additional 3' *NFE2* signal (red).

be biological by-products. This also holds true for next-generation sequencing transcriptome data, which routinely generates a large number of putative chimeras, most of which are nonfunctional^{19,20}. Here, we describe a methodology to nominate biologically important fusions from an integrative analysis of next-generation transcriptome data and high-throughput genomic data.

By undertaking a comprehensive analysis of the biological associations of all genes contributing to gene fusions, we demonstrate that, although analysis of domain architectures and shared pathways was less informative, cancer-related fusion genes tend to engage distinct interaction networks or share common gene ontologies. Using such information, we generalized this finding to a genomic scale and developed an algorithm, ConSig score, to assay the probability that any given gene may contribute to a driving gene fusion based on the strength of that gene's association with biological concepts characteristic of cancer genes. Although ConSig analysis can nominate putative cancer genes, the association of a gene with a specific tumor type requires additional evidence compiled from other biological data sets. To integrate use of high-throughput genomic data, we characterized the chromosomal imbalances associated with gene fusions, finding that recurrent gene fusions exhibit distinctive patterns of copy number alteration corresponding to differential portions of fusion partners. To our knowledge, this is the first evidence that integrative bioinformatics may be able to predict which genes are preferentially subject to chromosomal rearrangements important in tumorigenesis.

We applied the ConSig score to next-generation sequencing transcriptome data to benchmark fusion candidates, which were then assessed for chromosomal aberrations complying with the fusion breakpoint principle by integrating high-quality copy number data. We found that the ConSig score was able to identify the known *EML4-ALK* fusion as the top-ranked candidate in the H2228 lung cancer cell line, and in addition, we found further evidence of a *R3HDM2-NFE2* fusion in H1792 cell line. We show that the *R3HDM2-NFE2* fusion, which results in overexpression of wild-type *NFE2*, promotes cell proliferation and invasion. Moreover, through analysis of SNP arrays and lung tissue microarrays, we find that chromosomal rearrangements at the *NFE2* locus are recurrent in a small subset of patient tumors, suggesting that *NFE2* may contribute to a new class of lung cancer molecular biology. These data suggest that such approaches may have broad applicability to the analysis of multidimensional cancer genomic data.

The methodology described here can filter the large number of fusion candidates generated by paired-end next-generation sequencing data and preferentially identify driver gene fusions in cancer. The ConSig technology suggests the functional importance of putative fusions in cancer, whereas the breakpoint principle helps interpret large-scale cancer genomic data sets to explore potential recurrence. Although we have not applied this methodology to the discovery of novel mutations, we hypothesize that a similar computational schematic may yield insights in this area as well. Ultimately, we hope that this integrative methodology will elucidate key aspects of tumor biology as well as facilitate the development of targeted therapy of human cancers.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturebiotechnology/>.

Accession code. GenBank: GU068583 for nucleotide sequence of the *R3HDM2-NFE2* fusion from H1792 cell line.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank F. Mitelman for offering the fusion genes list from the Mitelman database; T.W. Glover for the important comments for improving the manuscript; J. Granger for help with editing the manuscript; S. Qin for useful discussions about biostatistics; NCIBI colleagues L. Ke and A. Ade for helping in the implementation of tools and technologies; C. Yang for the guidance in drug informatics; Z. Hu from Boston University for help with network visualization. This work was supported by the National Institutes of Health (NIH; U54 DA021519) and a National Institutes of Health Cancer Biology Training Grant (CA009676-18 to J.R.P.). J.R.P. is a Fellow of the University of Michigan Medical Scientist Training Program. A.M.C. is supported by NIH early detection network U01 CA111275, DOD W81XWH-09-2-0014, the Doris Duke Foundation and the American Cancer Society.

AUTHOR CONTRIBUTIONS

X.-S.W., G.S.O. and A.M.C. designed the study. X.-S.W., J.R.P. and M.A.S. performed bioinformatics analyses. X.-S.W., J.R.P., G.C., Q.C., S.M.D., R.P., X.C. and S.V. performed experimental studies. B.H. and N.P. performed FISH analysis. D.G.T., T.J.G. and D.G.B. coordinated the clinical and pathology components. X.-S.W., J.R.P., G.S.O. and A.M.C. wrote the manuscript.

Published online at <http://www.nature.com/naturebiotechnology/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
- Tomlins, S.A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
- Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).
- Kumar-Sinha, C., Tomlins, S.A. & Chinnaiyan, A.M. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* **8**, 497–511 (2008).
- Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39 (2007).
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–280 (2004).
- Prasad, T.S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–72 (2009).
- Hu, Z. *et al.* VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.* **35**, W625–632 (2007).
- Chen, C. *et al.* Leptin induces proliferation and anti-apoptosis in human hepatocarcinoma cells by up-regulating cyclin D1 and down-regulating Bax via a Janus kinase 2-linked pathway. *Endocr. Relat. Cancer* **14**, 513–529 (2007).
- Chen, G.J., Weylie, B., Hu, C., Zhu, J. & Forough, R. FGFR1/PI3K/AKT signaling pathway is a novel target for antiangiogenic effects of the cancer drug fumagillin (TNP-470). *J. Cell. Biochem.* **101**, 1492–1504 (2007).
- Vantier, M. *et al.* PI3-kinase/Akt-dependent antiapoptotic signaling by the PDGF alpha receptor is negatively regulated by Src family kinases. *FEBS Lett.* **580**, 6769–6776 (2006).
- Walz, C., Cross, N.C., Van Etten, R.A. & Reiter, A. Comparison of mutated ABL1 and JAK2 as oncogenes and drug targets in myeloproliferative disorders. *Leukemia* **22**, 1320–1334 (2008).
- Fuhrer, D.K. & Yang, Y.C. Complex formation of JAK2 with PP2A, PI3K, and Yes in response to the hematopoietic cytokine interleukin-11. *Biochem. Biophys. Res. Commun.* **224**, 289–296 (1996).
- Kharas, M.G. *et al.* Ablation of PI3K blocks BCR-ABL leukemogenesis in mice, and a dual PI3K/mTOR inhibitor prevents expansion of human BCR-ABL+ leukemia cells. *J. Clin. Invest.* **118**, 3038–3050 (2008).
- Mullighan, C.G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110–114 (2008).
- Mullighan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
- Drexler, H.G. *The Leukemia-Lymphoma Cell Line Factsbook* (Academic Press, San Diego, 2000).
- Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet.* **15** Spec No 417–474 (1997).
- Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
- Bashir, A., Volik, S., Collins, C., Bafna, V. & Raphael, B.J. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.* **4**, e1000051 (2008).

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 database⁷, 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 $P < 10^{-7}$ (≥ 3 connectivities with a fusion partner group). The fusion-interaction network was visualized by VisANT⁹ 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 database⁷ (<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 , $i = 1, \dots, k$. 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.

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 $Y = n/X - 1$ 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 permuted 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 performed 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 \log_2 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) algorithm²³. 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 publication²⁴ 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 rConSig 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 Results**. 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/pebiiodocs/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 lines²⁵, and the 40 normal tissue data set²⁶, using the Oncomine database (<http://www.oncomine.org/>)²⁷. Descriptions of tissue types from both data sets are provided in **Supplementary Results**.

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 previously³⁰.

21. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
22. Wishart, D.S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
23. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
24. Weir, B.A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007).
25. Richard, W. Overall experiment characteristics. *National Cancer Institute* (<https://array.nci.nih.gov/caarray/project/woost-00041>) (2009).
26. Roth, R.B. *et al.* Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**, 67–80 (2006).
27. Rhodes, D.R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–180 (2007).
28. Rubin, M.A. *et al.* Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res.* **64**, 3814–3822 (2004).
29. Garraway, L.A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
30. Cao, Q. *et al.* Repression of E-cadherin by the polycomb group protein EZH2 in cancer. *Oncogene* **27**, 7274–7284 (2008).