NetSig: network-based discovery from cancer genomes

Heiko Horn[1,2,7], Michael S Lawrence[2,3,7], Candace R Chouinard[2], Yashaswi Shrestha[2], Jessica Xin Hu[1,2], Elizabeth Worstell[1,2], Emily Shea[2], Nina Ilic[2,4], Eejung Kim[2,4], Atanas Kamburov[2,3], Alireza Kashani[1,2], William C Hahn[2,4], Joshua D Campbell[2,5], Jesse S Boehm[2,8], Gad Getz[2,3,8] & Kasper Lage[1,2,6,8]

Methods that integrate molecular network information and tumor genome data could complement gene-based statistical tests to identify likely new cancer genes; but such approaches are challenging to validate at scale, and their predictive value remains unclear. We developed a robust statistic (NetSig) that integrates protein interaction networks with data from 4,742 tumor exomes. NetSig can accurately classify known driver genes in 60% of tested tumor types and predicts 62 new driver candidates. Using a quantitative experimental framework to determine *in vivo* tumorigenic potential in mice, we found that NetSig candidates induce tumors at rates that are comparable to those of known oncogenes and are ten-fold higher than those of random genes. By reanalyzing nine tumor-inducing NetSig candidates in 242 patients with oncogene-negative lung adenocarcinomas, we find that two (*AKT2* and *TFDP2*) are significantly amplified. Our study presents a scalable integrated computational and experimental workflow to expand discovery from cancer genomes.

Cancers arise when somatic mutations, copy number alterations, or genomic fusion events of specific genes confer a selective advantage to the corresponding cell, thus promoting tumorigenesis. (Hereafter, we will refer to these genes as driver or cancer genes.) Identifying driver genes in tumors of individual patients provides key mechanistic, diagnostic, and therapeutic insights. Therefore, a central aim of oncology is to provide a complete catalogue of genes underlying human cancers[1–6].

Cancer genes can be identified in an unbiased manner from somatic mutations or copy number changes in genomic sequence data by using gene-based statistical tests such as MutSig, Oncodrive, GISTIC, and RAE[7–10]. These methods have identified many genes that are mutated or amplified at high frequencies (>20%) in tens of tumor types[11]. However, for many tumor types, insufficient sample numbers, compounded by high background mutation and copy number rates render it challenging to confidently pinpoint driver genes at intermediate (2–20%) or low (<2%) frequencies[11]. For this reason, a large number of biologically or clinically relevant driver genes do not meet established statistical cutoffs and remain to be discovered.

Many alternative methods highlight network modules (where genes are connected based on, for example, correlations in gene expression or protein interactions) that are significantly mutated in tumors[12–18]. These analyses have been valuable for illuminating the biological processes and pathways involved in cancers (reviewed in Creixall *et al.*[19]). However, evidence from network-based approaches comes from aggregating weak genetic signals in a set of connected genes and not from mutation signal in any individual gene. This means that no strong direct link can be made between specific genes in a mutated module and the cancer in question. Additionally, most network-based methods are evaluated retrospectively through benchmarks, whereas experimental follow-up is limited to a few novel gene candidates. 'Knowledge contamination' of well-studied genes is thus a major problem that puts the effectiveness of network methods in question; it is impossible to determine how much circularity and bias favors results that agree with more established or classic cancer networks.

We aimed to quantify the real predictive value of network-based approaches and to maximize their benefit for driver prediction by developing a statistic (NetSig) that combines cancer mutation data and molecular network information. NetSig addresses the effects of knowledge contamination and is designed to be independent of gene-based statistical tests, so that it can complement these approaches in any tumor genome analysis pipeline. To test the predictive power of NetSig, we developed a large-scale, *in vivo*, quantitative experimental framework that enabled us to compare the tumorigenic potential of 23 genes with a significant NetSig score (FDR ≤ 0.1) to those of 25 known cancer genes and 79 random genes in mouse experiments. Based on the network

[1]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Pathology and MGH Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [5]Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA. [6]Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark. [7]These authors contributed equally to this work. [8]These authors jointly directed this work. Correspondence should be addressed to J.S.B. (boehm@broadinstiute.org), G.G. (gadgetz@broadinstitute.org), or K.L. (lage.kasper@mgh.harvard.edu).

RECEIVED 29 MARCH; ACCEPTED 19 OCTOBER; PUBLISHED ONLINE 4 DECEMBER; CORRECTED ONLINE 19 DECEMBER 2017 (DETAILS ONLINE); DOI:10.1038/NMETH.4514

analysis and *in vivo* experiments, nine candidates were found to be particularly relevant to lung adenocarcinoma. We reanalyzed copy number data derived from 660 patients with lung adenocarcinoma to discover higher rates of amplification of *TFDP2* and *AKT2* in patients without established genomic driver events compared to patients with mutations and amplifications in known oncogenes. NetSig code is freely available in the **Supplementary Software** and at http://www.lagelab.org/resources, and the algorithm has been implemented in FireCloud (https://software.broadinstitute.org/firecloud/).

Results

Design and properties of the NetSig statistic

NetSig combines data from 4,742 tumor genomes spanning 21 tumor types and InWeb_InBioMap (a human protein–protein interaction network)[20,21] to calculate the mutation signal in a genes' functional protein–protein interaction network. Since we specifically wanted to test the predictive power of mutations in a gene's network, we excluded mutation information on the gene itself in the calculation of the NetSig statistic (see Online Methods).

To benchmark NetSig and to understand the effect of knowledge contamination on the statistic, we defined a set of very well established 'Cosmic classic' cancer genes from the Cosmic database (http://cancer.sanger.ac.uk/cosmic) and a nonoverlapping set of 'recently emerging cancer genes' from recent sequencing studies (see Online Methods; **Supplementary Table 1**). To test for cryptic confounders, we also defined a set of random genes (see Online Methods; **Supplementary Table 1**). We confirmed that the Cosmic classic and recently emerging sets can be classified based on their NetSig score with area under the receiver operating characteristics curves (AUCs) of 0.86 and 0.75, respectively (**Fig. 1a**; adjusted $P < 0.05$ for each of these AUCs using permuted networks; **Supplementary Fig. 1**). As expected, the random control genes fit the null hypothesis and could not be distinguished from other genes represented in InWeb_InBioMap (**Fig. 1a**; AUC 0.49, $P = 0.8$). We further show that NetSig can accurately classify cancer genes in ~60% of the tumor types for which we have data (see Online Methods; **Supplementary Note 1** and **Supplementary Fig. 2**), which illustrates the potential of our statistic to inform many different individual tumor types and also tumor types with relatively few samples.

The majority of genes scored by NetSig fit the null hypothesis and lie on the diagonal in a quantile–quantile plot, but there is an overall genomic inflation ($\lambda = 1.29$) of the significances assigned to genes (**Supplementary Fig. 3**). This could be due to knowledge contamination, the inherent polygenic nature of cancers, or a combination of these two factors. To dissect this phenomenon, we removed the effect of well-studied cancer genes from our analysis (see Online Methods; **Supplementary Note 2**). The ability to predict Cosmic classic cancer genes is reduced from an AUC of 0.86 to 0.79, indicating some knowledge contamination of this set, but the effect on 'recently emerging' cancer genes is much less pronounced (from an AUC of 0.75 to 0.73) (**Fig. 1b**). Consistent with these observations, removing the effect of the Cosmic classic gene set reduces λ from 1.29 to 1.09 in the quantile–quantile plot, and it only changes slightly to 1.07 when the impact of the recently emerging set is also removed (**Supplementary Fig. 3**). Furthermore, running NetSig on random networks results in a noninflated quantile–quantile plot, as expected (**Supplementary Fig. 3**). We also show that NetSig adequately normalizes for the number of interactions a gene has at the protein level (**Supplementary Fig. 4**).

Together, these analyses reveal some knowledge contamination in the protein–protein interaction data of the Cosmic classic set, which leads to a considerably inflated AUC in the benchmark if it is not taken into consideration. Conversely, there is almost no knowledge contamination of genes from recent sequencing studies. This means that when predicting new cancer genes from existing cancer genomes, knowledge contamination should not confound the NetSig method when applied to protein–protein interaction data from InWeb_InBioMap.

Predicting NetSig candidates from tumor genomes

To test whether NetSig can predict new driver genes from existing cancer genome data, we calculated NetSig scores of all genes that had at least one high-confidence protein interaction in InWeb_InBioMap. We calculated NetSig scores both using the pan-cancer cohort of 4,742 tumors and mutation data from each of the individual 21 tumor types represented in ref. 11 (see Online Methods). We declared genes significant at a false discovery rate (FDR) of $Q \leq 0.1$ using the pan-cancer data (**Fig. 1c**) and at $Q \leq 0.1$ in each of the individual 21 tumor types.

The pooled set (named NetSig5000; **Supplementary Table 2**) contains all unique genes that were significant in the pan-cancer analysis or in at least one of the 21 tumor types. NetSig5000 comprises 62 genes, which we divided into groups based on their known connection to cancer. Groups 1 ($n = 12$) and 2 ($n = 9$) contain genes already known to be involved in cancers based on point mutations or gene fusion events, respectively. These groups serve as a positive control that NetSig can identify known cancer genes. Groups 3 ($n = 24$) and 4 ($n = 13$) contain genes that have been speculated to be causal in cancers based on evidence from model systems or from gene expression analyses. Group 5 ($n = 4$) genes have never been linked to cancer (see **Supplementary Table 3** and **Supplementary Note 3** for more information about genes in the NetSig5000 set and **Supplementary Fig. 5** for examples of NetSig networks). All results can be accessed and visualized at http://www.lagelab.org/resources/.

Tumorigenic potential of NetSig candidates

To validate NetSig performance, we tested the tumorigenic potential of 23 genes from the NetSig5000 set (**Supplementary Table 4**; see Online Methods for selection criteria), 79 different patient-derived mutations of 25 known driver genes (positive control; **Supplementary Table 5**), and 79 random genes (random control; **Supplementary Table 6**) using a massively parallel *in vivo* tumorigenesis assay (**Fig. 2a**; see Online Methods). The assay transduces and overexpresses barcoded cDNA constructs of candidate genes (and alleles representing patient-derived mutations) into activated small-airway epithelial cells (SALE-Y cells[22]) or activated immortalized kidney epithelial cells (HA1E-M cells[22–25]). For each cell model (SALE-Y or HA1E-M), all genes (or alleles) to be tested are pooled, grown, and injected subcutaneously into immunocompromised animals at three injection sites per animal. In animals that develop tumors, driver genes can be identified by homogenizing tumors in the animals and by sequencing the barcodes found in the tumor cells (see Online Methods).

To compare the tumorigenic potential of the three gene sets across multiple cell models, we developed a quantitative analytical

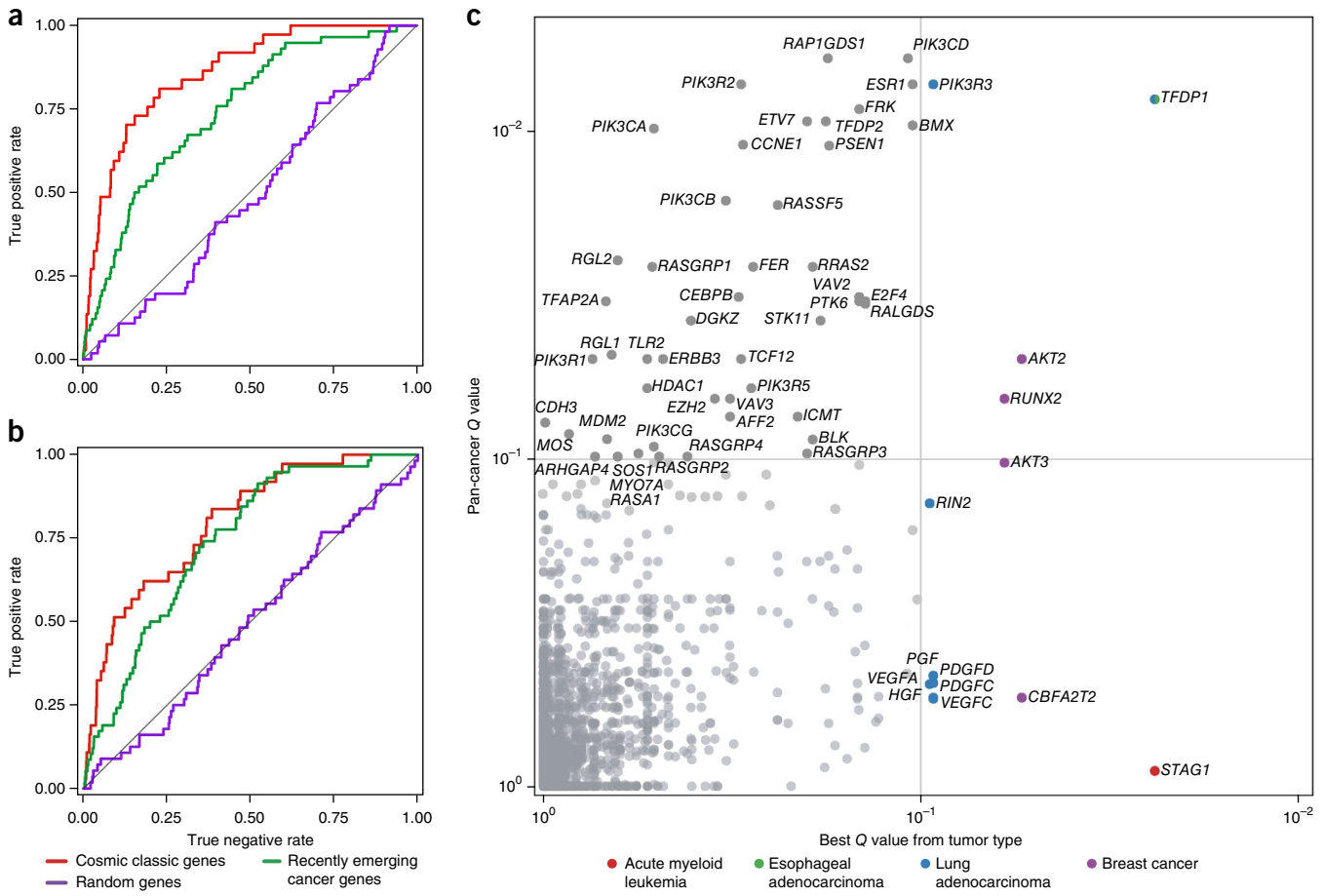

Figure 1 | NetSig predicts true cancer genes. (**a**) Receiver operating characteristic (ROC) curves for genes in the 'Cosmic classic', 'recently emerging', and random sets (AUC 0.86, 0.75, and 0.49, respectively). (**b**) ROC curves after removing the effect of very well established cancer genes (AUC 0.79, 0.73, and 0.5, for Cosmic classic, recently emerging, and random sets, respectively). (**c**) Visualization of the NetSig500 set. Genes are represented as individual dots and plotted along the *x*-axis by the most significant NetSig *Q* value from each of 21 tumor types and on the *y*-axis by the NetSig *Q* value when 4,724 tumors are analyzed as a combined pan-cancer cohort. Gray lines indicate significance at FDR $Q \leq 0.1$.

framework that defines a gene as tumorigenic based on both *in vivo* proliferation rate and the significance of relative tumor growth (see Online Methods; **Supplementary Software**). Our analysis showed that many of the tested NetSig5000 genes (11/23, or 48%) are indeed capable of driving tumorigenesis (**Fig. 2b**,**c** and **Supplementary Table 7**). Specifically, pooled screening supports the tumorigenic potential of *AKT2*, *BLK*, *BMX*, *FER*, *FRK*, *MOS*, *PIK3CG*, *PTK6*, *RASGRP1*, *RASGRP3*, and *TFDP2* (for a comprehensive literature review of these genes, see **Supplementary Note 4**). In comparison, the proportion of known driver genes from the positive control set that induced tumors was 9/25, or 36%, providing an estimate of assay sensitivity (see Online Methods), and the proportion of random genes that induced tumors was 4/79, or 5%, providing an estimate of the false positive rate (**Fig. 2d**). We note that the two random genes that induced tumors are *NTRK1*, which encodes a tyrosine kinase with established tumorigenic properties, and *STRADA*, an interactor of the *STK11* tumor suppressor at the protein level, which suggests that these could be real driver genes that remain to be discovered.

Reanalysis of lung adenocarcinomas

Nine NetSig5000 genes (*AKT2*, *FER*, *FRK*, *MOS*, *PIK3CG*, *PTK6*, *RASGRP1*, *RASGRP3*, and *TFDP2*) validated with high confidence in a cell model (SALE-Y) that is particularly relevant for exploring genes that can induce lung adenocarcinomas[22]. We hypothesized that a subset of these nine genes may be responsible for driving lung adenocarcinomas in oncogene-negative patients (meaning patients that do not have a known oncogenic driver event in the RAS/RAF/receptor tyrosine kinase (RTK) pathway as previously described[26]).

To test this hypothesis, we used a data set of 660 lung adenocarcinomas from TCGA and related studies[26–28]. We first tested for copy number differences between the oncogene-negative (n = 242) and oncogene-positive patients (n = 418), and we showed that the nine genes as a set have a significantly higher copy number in the former group ($P = 7.0 \times 10^{-3}$, Fisher's exact test; **Fig. 3a**). Individually, *TFDP2* and *AKT2* are also found at higher copy numbers in the oncogene-negative group (FDR < 0.1 for each gene, **Fig. 3a**).

Through an in-depth analysis of the surrounding genomic regions, we determined that adjacent potential oncogenes do not underlie the signal (**Fig. 3b**,**c** and **Supplementary Note 5**) and confirmed that there is no overall difference in copy numbers between the two patient groups (**Supplementary Fig. 6**). The genomic events observed for *AKT2* and *TFDP2* are not high-level amplifications. Rather, 3% and 4% of the oncogene-negative

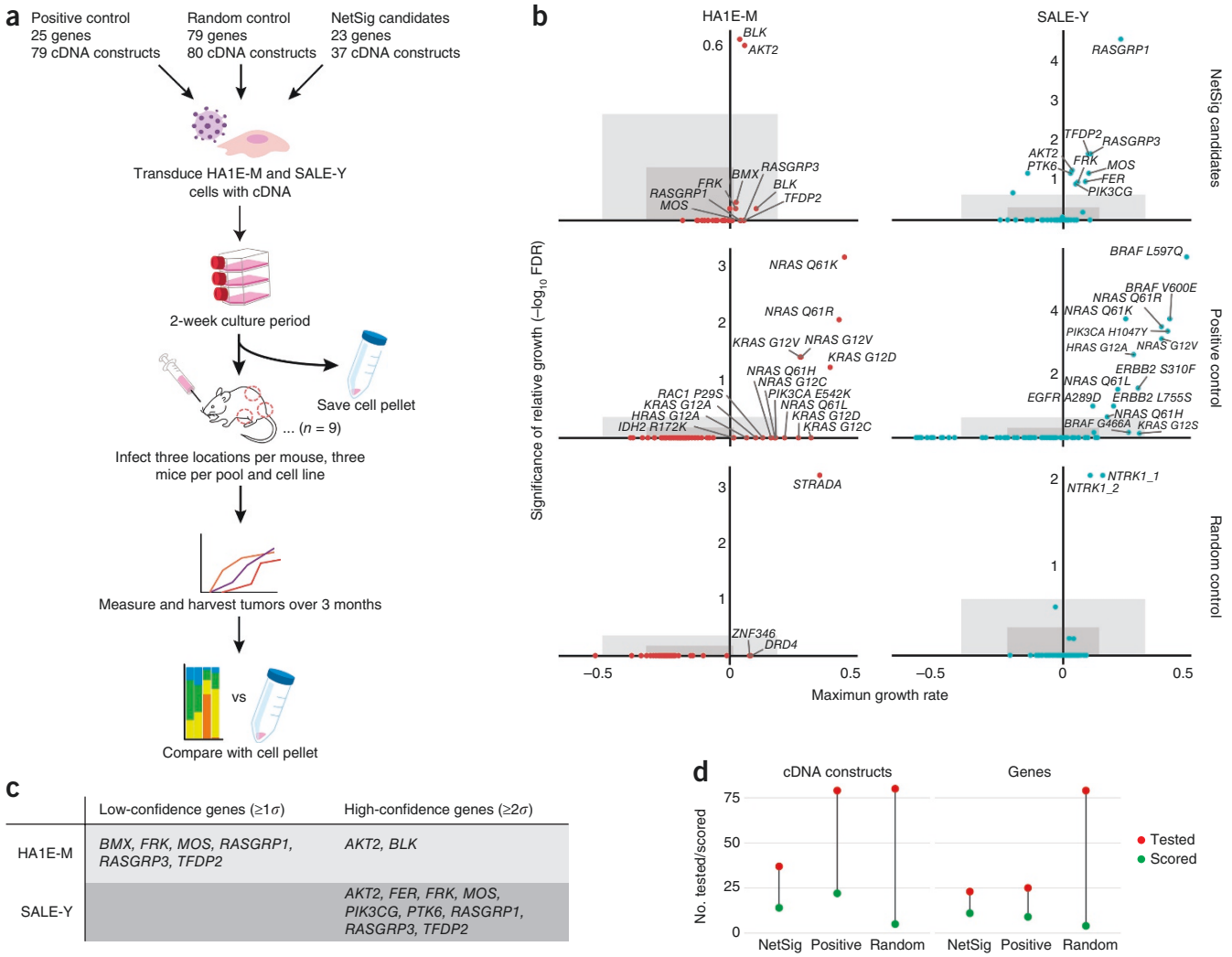

Figure 2 | *In vivo* tumor formation of NetSig5000 and control sets. (**a**) Experimental design. (**b**) Tumorigenic potential of 23 NetSig5000 genes (NetSig candidates), 25 known oncogenes (positive controls), and 79 random genes (random controls) in *in vivo* mouse tumorigenesis experiments. Maximum significance of enrichment in tumors relative to preinjection samples is plotted against maximum proliferation rate. Dark gray boxes indicate one standard deviation from the median (lower confidence), and light grey boxes indicate two standard deviations from the mean (higher confidence). (**c**) Candidates that induce tumors at the higher and lower confidence threshold stratified by cell model. (**d**) Proportion of the NetSig5000 candidates, positive control set, and random set, respectively, that induced tumors in mice. Left panel indicates the results at the level of cDNA constructs. Right panel indicates results at the gene level.

patients have two extra copies of *AKT2* and *TFDP2*, respectively; and 4% and 14% have one extra copy of *AKT2* and *TFDP2*, respectively (**Fig. 3d**,**e**). We found no evidence for increased rates of gain-of-function somatic single-nucleotide variants (SSNVs) or insertions or deletions (indels) for the nine genes in the oncogene-negative versus the oncogene-positive group.

Given the dominating effect of the RAS/RAF/RTK pathway in lung adenocarcinoma, a more straightforward approach to gene discovery would be to make a targeted analysis of mutations or copy number gains in genes in the extended RAS/RAF/RTK pathway (defined here as genes that have at least one protein interaction with a RAS/RAF/RTK pathway member in InWeb_InBioMap). We compared the degree of copy number gains, and activating SSNVs/indels, in our set of nine genes to 100 matched sets of nine RAS-affiliated genes, which showed that the set of nine genes identified through our approach is significantly more enriched for oncogenic copy number gains ($P = 0.04$, using permutation tests; **Supplementary Fig. 7**). This analysis confirms that combining NetSig with tumorigenicity experiments is a better approach to identifying driver genes and events in lung adenocarcinomas than naively choosing genes in the extended RAS/RAF/RTK pathway.

To allow further exploration of pathway relationships relevant to lung adenocarcinomas, the NetSig networks of *AKT2* and *TFDP2* are plotted in **Figure 3f**,**g**.

Discussion

Our integrated computational and experimental analyses firmly establish that network-based approaches can contribute to expanding gene discovery from existing cancer genomes. Not only does the NetSig5000 set identify new genes in well-established oncogenic pathways (e.g., *AKT2*, *PIK3CB*, *PIK3CG*, *RASGRP1*, *RASGRP3*; **Supplementary Note 6** and **Supplementary Fig. 8**), but also our results point to new potential cancer pathways (e.g.,

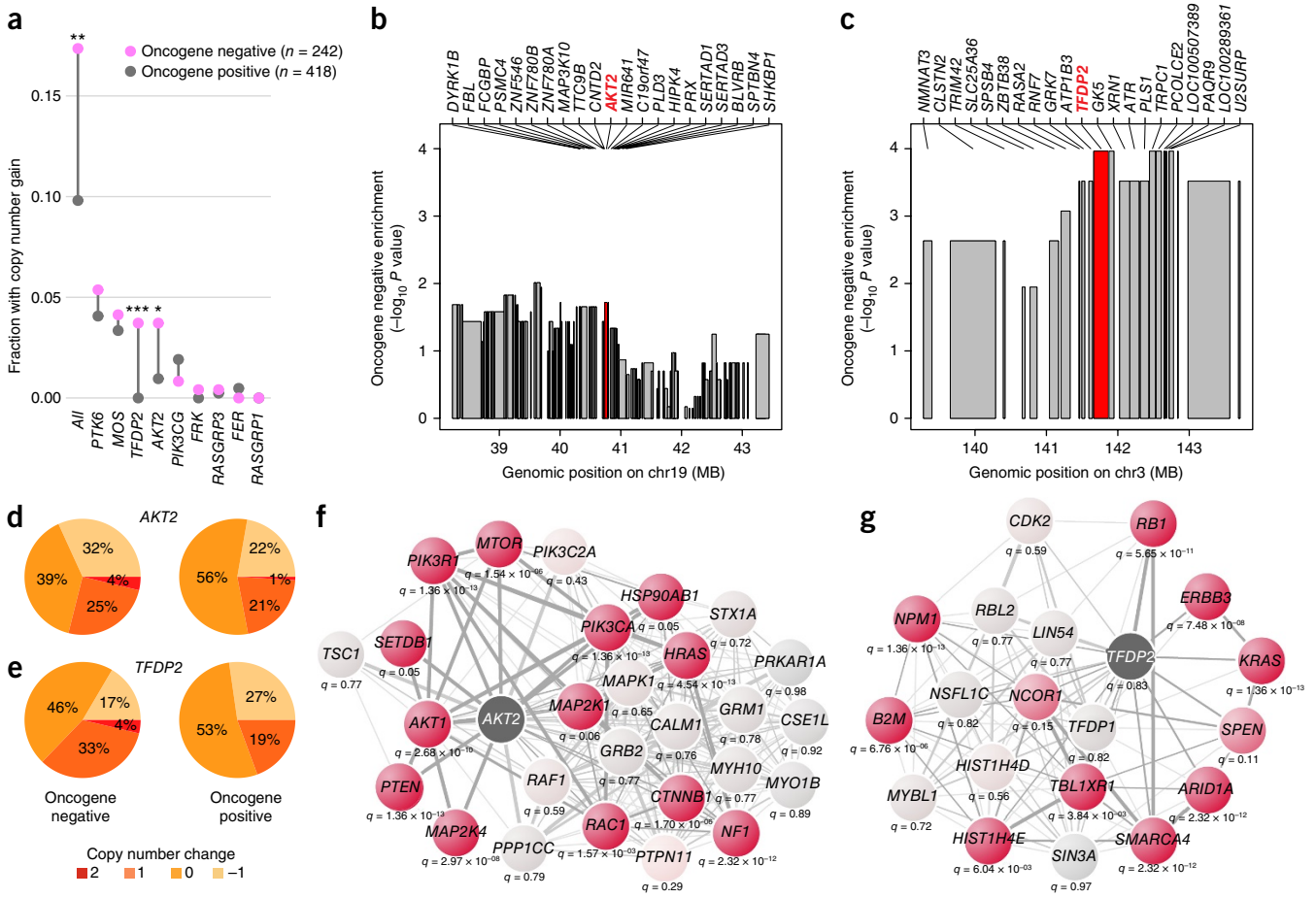

Figure 3 | Targeted reanalysis of oncogene-negative lung adenocarcinoma patients. (**a**) Amplification of the nine genes that induce tumors in the lung-adenocarcinoma-relevant cell model. As a group, the genes are significantly amplified ($P = 7.0 \times 10^{-3}$), and *AKT2* and *TFDP2* are individually significantly amplified (FDR $Q < 0.1$). *, FDR < 0.07; **, FDR < 0.04; ***, FDR < 0.002. (**b**,**c**) In-depth view of the amplified regions surrounding *AKT2* and *TFDP2*. *AKT2* and *TFDP2* are indicated by red and other genes by gray. (**d**,**e**) The proportion of oncogene-positive or oncogene-negative patients with –1, 0, 1, or 2 copy number changes of *AKT2* or *TFDP2*. (**f**,**g**) NetSig networks of *AKT2* and *TFDP2*. Nodes other than *AKT2* and *TFDP2* are colored by the significance of the pan-cancer *Q* value of the corresponding gene, where light gray represents *Q* close to 1 and red Q << 1, with darker red representing more significant *Q* values as indicated below the relevant node.

TFDP2 and *MYO7A*; **Supplementary Note 7**, **Supplementary Figs. 9** and **10**, **Supplementary Tables 8** and **9**).

NetSig has a number of differences from other network-based methods (**Supplementary Note 8**, **Supplementary Fig. 11**, and **Supplementary Table 10**). An important feature is that it is explicitly designed to disregard any mutation information on the gene being tested, so that the signal comes from the gene's network alone. This ensures that NetSig *P* values are fully independent of those from existing gene-based statistical tests such as MutSig, Oncodrive, GISTIC, and RAE. In fact, MutSig and NetSig *P* values for the same genes are only modestly correlated (Pearson correlation coefficient = 0.05, data not shown). This design choice means that NetSig can be seamlessly combined with gene-based statistical tests in any computational cancer genome analysis workflow (**Supplementary Note 9** and **Supplementary Fig. 12**).

NetSig is flexible and can work with many different types of functional genomics network data (**Supplementary Note 9** and **Supplementary Fig. 13**). The average genomic inflation when NetSig is run on two different sets of transcriptional networks[29] (i.e., based on data that cannot be affected by knowledge contamination) is 1.14 and 1.11 (**Supplementary Figs. 14** and **15**). This is comparable to λ in the protein–protein interaction data when the effect of Cosmic classic genes is removed (1.09), which suggests that our approach to removing knowledge contamination is efficient in canceling out that effect. This strongly suggests that the remaining inflation is due to polygenicity of cancers and not to any bias or confounders of the NetSig statistic or network data.

The study's limitations, an estimate of how well the NetSig statistic predicts real cancer genes, and information on the benefit of including several cell models and genetic backgrounds in the validation workflow are discussed further in **Supplementary Note 10**.

While we did not observe evidence for gain-of-function SSNVs or indels across *TFDP2* and *AKT2*, we expect that given more samples, these genes will be enriched for such events. This is consistent with the observation that NetSig5000 is enriched for genes with lower MutSig *P* values in Lawrence *et al.*[11] ($P = 0.04$, two-sample Kolmogorov–Smirnov test). Together, our results strongly suggest that many genes in NetSig5000 are likely real intermediate- or low-frequency driver genes that will reach significance in gene-based statistical tests with more tumor genomes in the future.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

J.D.C. is supported by the LUNGevity Career Development Award (CDA). H.H. was supported by a Fund for Medical Discovery Award from the Executive Committee On Research at Massachusetts General Hospital. H.H. and K.L. are supported by the MGH IRG American Cancer Society. K.L. is supported by a grant from the Stanley Center at the Broad Institute, a Broadnext10 grant from the Broad Institute, 1R01MH109903, a Large Thematic Project Grant from the Lundbeck Foundation (R223-2016-721), and a Research Award from the Simons Foundation (SFARI).

AUTHOR CONTRIBUTIONS

H.H. developed, benchmarked, and implemented the NetSig algorithm with input from M.S.L. and supervision from G.G. and K.L. C.R.C., Y.S., E.S., N.I., and E.K. executed the *in vivo* tumorigenesis experiments with input from H.H. and K.L. and supervision from J.S.B. H.H. developed and implemented the quantitative analytical framework of *in vivo* tumorigenesis data with input from C.R.C., Y.S., and E.S. as well as supervision from J.S.B. and K.L. J.D.C. reanalyzed lung adenocarcinoma data with input from H.H., J.S.B., G.G., and K.L. All authors analyzed data and discussed the results. H.H., W.C.H., J.D.C., J.S.B., G.G., and K.L. wrote the manuscript with input from all authors. J.S.B., G.G., and K.L. designed and directed the work. K.L. initiated and led the study.

COMPETING FINANCIAL INTERESTS

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

1. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
2. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Frampton, G.M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
4. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* **3**, 111ra121 (2011).
5. Van Allen, E.M. *et al.* Somatic *ERCC2* mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. *Cancer Discov.* **4**, 1140–1153 (2014).
6. Wagle, N. *et al.* High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82–93 (2012).
7. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
8. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
9. Taylor, B.S. *et al.* Functional copy-number alterations in cancer. *PLoS One* **3**, e3179 (2008).
10. Lohr, J.G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 3879–3884 (2012).
11. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
12. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
13. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
14. Leiserson, M.D.M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
15. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
16. Babur, Ö. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* **16**, 45 (2015).
17. Miller, C.A., Settle, S.H., Sulman, E.P., Aldape, K.D. & Milosavljevic, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* **4**, 34 (2011).
18. Yeang, C.-H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* **22**, 2605–2622 (2008).
19. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
20. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
21. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2016).
22. Berger, A.H. *et al.* High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* **30**, 214–228 (2016).
23. Boehm, J.S. *et al.* Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell* **129**, 1065–1079 (2007).
24. Dunn, G.P. *et al.* In vivo multiplexed interrogation of amplified genes identifies GAB2 as an ovarian cancer oncogene. *Proc. Natl. Acad. Sci. USA* **111**, 1102–1107 (2014).
25. Kim, E. *et al.* Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. *Cancer Discov.* **6**, 714–726 (2016).
26. Campbell, J.D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
27. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
28. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
29. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13**, 366–370 (2016).

ONLINE METHODS

Calculating the network mutation burden. For a given index gene, NetSig statistic is formalized into a probabilistic score that reflects the index-gene-specific composite mutation burden (i.e., the aggregate of single-gene MutSig suite *Q* values from Lawrence *et al.*[11]) across its first-order biological network and is calculated via a three-step process. First, we identify all genes it interacts with directly at the level of proteins, only including high-confidence, quality-controlled data from the functional human network InWeb_InBioMap[20,21,30] (where the vast majority of connections stem from direct physical interaction experiments at the level of proteins). Second, the composite mutation burden across members of the resulting network is quantified by aggregating single-gene MutSig suite *Q* values from Lawrence *et al.*[11] into one value ϕ using an approach inspired by Fisher's method for combining *P* values:

$$\phi \approx -2\sum_{i=0}^{k} \ln(q_i)$$

Where q_i is the MutSig suite *Q* value for gene *i*, and *k* is the amount of genes in the first-order network of the index gene (i.e., the index gene's degree). Third, by permuting the InWeb network using a node permutation scheme, we compare the aggregated burden of mutations ϕ to a random expectation. In this step, the degree of the index gene, as well as the degrees of all genes in the index gene's network, is taken into careful consideration. The final NetSig score of an index gene is therefore an empirical P value that reflects the probability of observing a particular composite mutation burden across its first-order physical interaction partners (at the level of proteins) normalized for the degree of the index gene as well as the degrees of all of its first-order interaction partners. Because we are interested in estimating the mutation burden independent of the index gene (so that the NetSig results are fully independent of gene-based statistical tests such as MutSig, Oncodrive, GISTIC, and RAE), this gene is not included in the analysis, and it does not affect the NetSig calculation. This also means that for any given gene, MutSig suite significances are independent of NetSig significances (i.e., the Cancer5000 gene set and the NetSig5000 gene set are independently predicted).

Classifying cancer genes. For each gene represented in InWeb_InBioMap (12,507 or 67% of the estimated genes in the genome), we used the gene-specific NetSig probability to classify it as a cancer candidate gene or not. True-positive genes were a set of 'Cosmic classic' genes and a set of 'recently emerging cancer genes'. Specifically, the Cosmic classic set consists of 38 established (or classic) cancer genes from the Catalogue of Somatic Mutations in Cancer (Cosmic, http://cancer.sanger.ac.uk/cosmic; e.g., *TP53*, *BRCA1*, and *BRAF*; **Supplementary Table 1**). The 'recently emerging cancer genes' contains 61 genes that have been recently identified as cancer genes from the Sanger Gene Census dataset (http://cancer.sanger.ac.uk/census/; e.g., *MLL2*, *CDK12*, and *GATA2*; **Supplementary Table 1**). The gene set, for the purposes of the benchmarking analysis, is a set of 87 random genes (**Supplementary Table 1**). True negatives were defined as all genes in InWeb that were not in these three sets; this is likely conservative, as many of these might be yet undetected cancer genes. We used the NetSig probability as the classifier and calculated the AUC for each gene set. For estimating AUC significances, we generated AUCs for 100 random networks (from Rossin *et al.*[31]) and calculated the empirical *P* value.

Using NetSig to classify driver genes across 21 tumor types. For each tumor type we calculated tumor-type-specific NetSig scores and classified the corresponding tumor-specific driver genes. For example, we assembled a set of driver genes from breast tumors (BRCA) by identifying genes significantly (FDR ≤ 0.1) mutated in this tumor type in Lawrence *et al.*[11]. We used mutation data from this tumor type to derive $\text{NetSig}_{\text{BRCA}}$ scores and measured their classification performance on the BRCA driver genes, which they could accurately distinguish with an AUC = 0.76. We compared this result to the results using NetSig scores derived using the pan-cancer data set. The pan-cancer NetSig score increased our ability to accurately classify BRCA driver genes slightly to an AUC of 0.77 (for more information see **Supplementary Note 1** and **Supplementary Fig. 2**).

Testing the robustness of the NetSig approach. To test the robustness of the NetSig approach, we tried several alternative permutation methods and calculated the composite mutation burdens of gene networks using both *Q* and *P* values from Lawrence *et al.*[11]. Specifically, to generate the null distribution of network mutation burdens used to assess the significance of observations in the actual data, we used both a node permutation scheme and a full network permutation scheme. Where the node permutation scheme permutes nodes that have similar degree has the advantage of being much faster than the network permutation scheme (explained in detail in Rossin *et al.*[31]), the architecture of the original network is more precisely mirrored in the random networks using the latter method. We ran the full analysis using both approaches and compared the quantile–quantile plots (data not shown) and classification of genes in Tiers 1–5. This analysis confirmed that the choice of permutation scheme does not have a major influence on the overall results (**Supplementary Fig. 1**). In addition to using *Q* values from Lawrence *et al.*[11] for step 2 in the NetSig calculation (above), we also tried using unadjusted *P* values. For this latter approach the quantile–quantile plots (data not shown) as well as the classification of 'Classic' and 'Recently emerging' cancer genes are similar to the results we report in the main text (**Supplementary Fig. 1**).

Generating the NetSig5000 set. We used a node permutation scheme to create 10^6 permuted networks. NetSig probabilities were determined for every gene in InWeb that was covered by interaction data. The FDR *Q* values were calculated as described by Benjamini and Hochberg[32] based on the nominal *P* values controlled for 12,507 hypotheses. We performed NetSig analyses with the pan-cancer Q values as well as *Q* values from each of the 21 tumor types for which they were available. As it is a technical limitation of the NetSig approach that it is currently not possible to make 5.5×10^6 network permutations, we could not create a data set where we correct for all $12{,}500 \times 22$ hypotheses tested in the NetSig5000 set. For that reason our work does not have the equivalent of the Cancer5000-S (the stringent) set from Lawrence *et al.*[11], where the authors control for all hypotheses is carried out simultaneously.

A multiplexed *in vivo* tumor formation screen in mice. We used the SALE-Y cell model previously described in Berger *et al.*[22]

and the HA1E-M cell model previously described in Kim *et al.*[25]. Specifically, our earlier work revealed that immortalized small-airway epithelial cells harboring an activating *YAP1* variant are rendered tumorigenic via activation of the EGFR/MAPK pathways (SALE-Y cells[22]), and immortalized kidney epithelial cells harboring an activating *MAPK1* variant are rendered tumorigenic via activation of the PI3K/YAP/NFKB pathways (HA1E-M cells[22–25]). Briefly, we inserted each gene into barcoded cDNA clones, and these clones were transduced into SALE-Y and HA1E-M cells in 96 well plates in arrayed format. The cells were selected with puromycin, expanded, and pooled. 2 million cells per pool per site were injected subcutaneously into immunocompromised mice in three sites (interscapular area and left and right flanks) per mouse and tumor formation monitored. This protocol was approved under Broad Institutional Animal Care and Use Committee Protocol ID: 0012-08-14-1. The experimental endpoint was reached when any tumor length exceeded 1 cm. Tumors were homogenized, and genomic DNA was extracted and sequenced to determine the relative proportion of each inserted DNA barcode. The relative proportions of each barcode serve as a proxy indicating the gene driving a tumor. To deal with data from deploying multiple cell models in parallel on a large set of positive controls, random controls, and NetSig candidates, we developed a new quantitative analytical framework (below).

Quantitative comparison of tumorigenicity. We measured the reproducibility and magnitude of the oncogenic signal of the individual gene sets by developing and calculating two complementary metrics: maximum *in vivo* proliferation rate and significance of relative growth.

Calculating maximum proliferation rates. To determine a metric for growth doubling time of cells injected with NetSig5000 genes in the *in vivo* tumors, we calculated the proportion of reads in a tumor normalized to tumor volume and compared to the proportion of reads in the preinjection cell pool, where volume for all pooled cells was set to 1 mm^3 (which roughly corresponds to 2 million cells). This was done for all tumors, and for each tumor we divided the growth rate with the day the tumor was harvested to normalize for tumor age. This led to an estimate of the doubling time of the *in vivo* tumor growth for cells driven by overexpression of a particular NetSig candidate. We call this metric max proliferation rate per gene, which is plotted on the *x*-axis of **Figure 2b**.

Calculating the significance of relative growth. To calculate the significance of relative growth of cells in each cell type (SALE-Y and HA1E-M) transduced with a particular cDNA clone, we plotted the distribution of relative reads in the tumors and compared to the preinjection value. Significances were calculated using a one sided *t*-test and reported as false discovery rates. We call this metric significance of proliferation rate and plotted the maximum significance (after iterative removal of dominant effects; see below) on the *y*-axis of **Figure 2b**.

Computational detection of dominant and subjugated oncogenic clones in tumors. When many oncogenic clones are pooled and injected into mice, a single clone often outcompetes other oncogenic clones to dominate the tumor through a highly stochastic process. We refer to outcompeted, but real, oncogenic clones in the tumors as 'subjugated oncogenic clones'. It is possible to detect subjugated oncogenic clones by iteratively removing dominant clones from the cell pools and repeating the experiments. However, this is very labor intensive. We developed a computational approach where we iteratively removed genes that accounted for more than 50% of the reads in a tumor and repeated the significance of relative growth analysis described above. In **Figure 2b** we report the best FDR after zero, one, or two iterations. We confirmed that the subjugated oncogenic clones detected computationally were indeed driver clones by comparing the results from our computational approach to results from the iterative experimental removal of dominant clones and repetition of the injection of experimentally reduced cell pools into mice from Berger *et al.*[22]. This analysis showed that genes determined to be significant through our computational iterations also became dominant clones when other dominant clones were first removed from the experimental assay.

Sensitivity and specificity of tumorigenesis assay. *Sensitivity.* We determined how many of the 25 positive control genes were correctly classified as tumor inducing at *z*-scores of 1 and 2, in both the HA1E-M and SALE-Y model. In the HA1E-M model, six genes were classified as tumor inducing at a *z*-score of 1 and two genes at a *z*-score of 2 (see **Supplementary Table 7** for details). As we tested a total of 25 genes, this gives a sensitivity of 6/25 = 0.24 and 2/25 = 0.08, respectively (**Fig. 2b**). In the SALE-Y model, seven genes were classified as tumor inducing at a *z*-score of one and six at a *z*-score of two giving a sensitivity of 7/25 = 0.28 and 6/25 = 0.24, respectively (**Fig. 2b**). When combining the two assays together, the sensitivity increases to 9/25 = 0.36, which is likely because we are testing the tumorigenic potential of genes across several genetic backgrounds. Analogous calculations can be seen for constructs in **Supplementary Table 7**.

Specificity. We determined how many of the random gene constructs were correctly classified as non-tumor-inducing at *z*-scores of one or two (see above) in both the HA1E-M and SALE-Y models. In the HA1E-M model, three genes (*STRADA*, *ZNF346*, and *DRD4*) were classified as tumor inducing at a *z*-score of 1, and one gene (*STRADA*) was classified as tumor inducing at a *z*-score of 2. As we tested a total of 79 genes, this gives a specificity of 76/79 = 0.96 and 78/79 = 0.99, respectively (**Fig. 2b**). In the HA1E-M model, one gene (*NTRK1*) was classified as tumor inducing at a *z*-score of 1 and *z*-score of 2. As we tested a total of 79 genes, this gives a specificity of 78/79 = 0.99% at both thresholds (**Fig. 2b**). Analogous calculations can be seen for constructs in **Supplementary Table 7**.

Choosing 25 genes for the validation experiment. We selected the genes based on a number of biological (not being known cancer genes) and technical (available high-quality reagents) criteria. First, we selected a set of genes that were either in group 3, 4, or 5 of our literature curation groups (meaning they have not already been shown to be cancer genes in humans). Second, we chose the subset of genes for which there were already reagents (meaning open reading frame (ORF) constructs) available from the Genetic Perturbation Platform at the Broad Institute. Third, we chose the set of genes where the ORF constructs had been sequenced and (i) did not have any mutations (i.e., the sequence of the cDNA corresponded to the wild type and (ii) where the sequence of the ORF passed a high-quality-sequence cutoff to avoid testing ORFs where the sequence of the clone was ambiguous and could have

doi:10.1038/nmeth.4514

unknown mutations. Fourth, the cell models are optimized for perturbations in certain pathways (i.e., the SALE-Y cells are rendered tumorigenic via activation of the EGFR/MAPK pathways, and the HA1E-M cells are rendered tumorigenic via activation of the PI3K/YAP/NFKB pathways). We hypothesized that choosing a set of genes that linked to the pathways activated in each cell model would likely increase the chance of inducing tumors in these models. We tested this hypothesis by choosing the 25 genes so, when possible, they interacted directly with members of the pathway activated in the HA1E-M model, but not in the SALE-Y model. However, we see similar validation rates in the two models, so it does not seem to have an effect that we are 'fitting' the candidates specifically to the HA1E-M model (**Supplementary Note 10**). It is likely that the higher validation rates observed for Netsig candidates (138% of the theoretical expectation; see "Discussion" in main text) when using both cell models in parallel are due to a combination of these selection criteria (available reagents and connection to known cancer pathways in the HA1E-M model) and underestimates of the sensitivity of the assay because there is an upper limit to how many true positive oncogenes in a cell pool can induce tumors based on the issues with subjugated clones mentioned above.

Analysis of oncogene negative lung adenocarcinoma patients. Segmentation was performed using the Circular Binary Segmentation algorithm followed by Ziggurat Deconstruction to infer the length and amplitude of each segment. Recurrent peaks for focal somatic copy number alteration were identified using GISTIC 2.0 (ref. 8). A peak was considered to be focally amplified or deleted within a tumor if the GISTIC-2.0-estimated focal copy number ratio was greater than 0.1 or less than −0.1, respectively. Purity and ploidy were estimated using ABSOLUTE[33]. Two peaks were considered the same across tumor types if (i) the known target gene of each peak was the same, or (ii) the genomic location of the peaks overlapped after adding 1 Mb to the start and end locations of each gene. For the second criterion, only peaks that contained fewer than 25 genes and were smaller than 10 Mb were considered (for more details see Campbell *et al.*[26]). Because we are executing a case-control analysis of the copy numbers of genes that induce tumors in the SALE-Y model relevant for lung adenocarcinoma, our analysis normalizes out any potential effects of, for example, gene size, amount of protein–protein interactions a gene has, and so forth.

Life Sciences Reporting Summary. Further information on experimental design and reagents is available in the **Life Sciences Reporting Summary**.

Data availability. NetSig code, results, and visualizations are available from http://www.lagelab.org/resources. The protein network data (InWeb version 3.0) are available from http://www.lagelab.org/resources. Tumor genome data are publicly available from Lawrence *et al.*[11]. Lung cancer data sets are available from Campbell *et al.*[26]. Further data that support the findings of this study are available from the corresponding author upon request. NetSig is implemented in FireCloud (https://software.broadinstitute.org/firecloud/).

30. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. USA* **105**, 20870–20875 (2008).
31. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
33. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

Corrigendum: NetSig: network-based discovery from cancer genomes

Heiko Horn, Michael S Lawrence, Candace R Chouinard, Yashaswi Shrestha, Jessica Xin Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Eejung Kim, Atanas Kamburov, Alireza Kashani, William C Hahn, Joshua D Campbell, Jesse S Boehm, Gad Getz & Kasper Lage
***Nat. Methods*; doi:10.1038/nmeth.4514; corrected online 19 December 2017**

In the version of this article initially published online, the color labels for oncogene-positive and oncogene-negative lung adenocarcinomas were swapped in the **Figure 3a** legend. The error has been corrected in the print, PDF and HTML versions of this article.

nature research

Corresponding author(s): Kasper Lage

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

2. Data exclusions

Describe any data exclusions.

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

3. Replication

Describe whether the experimental findings were reliably reproduced.

For each experiment, note whether any attempts at replication failed OR state that all attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Describe how samples were allocated to groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Methods: doi:10.1038/nmeth.4514

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	See methods and SupplementaryMethodsAndData file

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	*Describe any restrictions on availability of unique materials used in the study OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources) OR state that no unique materials were used.*

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	*For all antibodies, as applicable, provide supplier name, catalog number, clone name, and lot number. Also describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript OR state that no antibodies were used.*

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	*Provide information on cell line source(s) OR state that no eukaryotic cell lines were used.*
b. Describe the method of cell line authentication used.	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used have been authenticated OR state that no eukaryotic cell lines were used.*
c. Report whether the cell lines were tested for mycoplasma contamination.	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination OR state that no eukaryotic cell lines were used.*
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	*Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.*

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	*For laboratory animals, report species, strain, sex and age OR for animals observed in or captured from the field, report species, sex and age where possible OR state that no animals were used.*

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	*Provide all relevant information on human research participants, such as age, gender, genotypic information, past and current diagnosis and treatment categories, etc. OR state that the study did not involve human research participants.*

Nature Methods: doi:10.1038/nmeth.4514