A global transcriptional network connecting noncoding mutations to changes in tumor gene expression

Wei Zhang1,8*, Ana Bojorquez-Gomez1,8, Daniel Ortiz Velez2, Guorong Xu3, Kyle S. Sanchez1, John Paul Shen4, Kevin Chen2, Katherine Licon1, Collin Melton6, Katrina M. Olson1,5, Michael Ku Yu1, Justin K. Huang1,6, Hannah Carter5, Emma K. Farley1,5, Michael Snyder4, Stephanie I. Fraley2, Jason F. Kreisberg1 and Trey Ideker1,2,6,7*

Although cancer genomes are replete with noncoding mutations, the effects of these mutations remain poorly characterized. Here we perform an integrative analysis of 930 tumor whole genomes and matched transcriptomes, identifying a network of 193 noncoding loci in which mutations disrupt target gene expression. These ‘somatic eQTLs’ (expression quantitative trait loci) are frequently mutated in specific cancer tissues, and the majority can be validated in an independent cohort of 3,382 tumors. Among these, we find that the effects of noncoding mutations on DAAM1, MTG2 and HYI transcription are recapitulated in multiple cancer cell lines and that increasing DAAM1 expression leads to invasive cell migration. Collectively, the noncoding loci converge on a set of core pathways, permitting a classification of tumors into pathway-based subtypes. The somatic eQTL network is disrupted in 88% of tumors, suggesting widespread impact of noncoding mutations in cancer.

Results

Genome-wide identification of somatic eQTLs in cancer. To identify noncoding mutations associated with functional effects, we performed a systematic analysis of 930 tumors integrating whole-genome sequences, matched mRNA expression profiles and reference transcriptional interaction maps. Using WGS of paired normal and tumor tissues in 930 patients across 22 types of cancer from TCGA (Fig. 1a), we identified $3.5 \times 10^7$ sites with somatic single-nucleotide variations (SNVs). We called these SNVs uniformly across all genomes using the MuTect suite according to GATK best-practice recommendations and those of Melton et al. (Fig. 1b). Clusters of noncoding SNVs located within 50 bp of one another were grouped, defining recurrently mutated loci (Fig. 1c, Methods and Supplementary Fig. 1).

We then tested each locus for its association with changes in mRNA expression of target genes (Fig. 1d). This task made use of two additional datasets. First, enhancer–gene mappings in GeneHancer were used along with promoter-proximal regions, defined as sequences within 1 kb of each transcription start site (TSS), to link recurrently mutated loci to putative target genes considered to be under direct transcriptional control (Methods). Second, for the vast majority of patients with tumor genome sequences, tumor mRNA expression profiles were also available (Fig. 1a). From these data, we developed a multivariate linear regression model of the expression change of each target gene, as a function of the mutation status of its linked loci and covariates, including the presence of copy number alterations (CNAs), DNA methylation status, tissue,
ancestry and sex (Fig. 1d and Methods). Conceptually, this procedure is similar to identifying eQTLs, in which inherited nucleotide variants are mapped to downstream functional changes. Here, however, the variants are somatically acquired rather than inherited. Such ‘somatic eQTL analysis’ simplifies the complexity and scope of eQTL mapping to a relatively small number of unlinked genetic variants: on average, 2.6 loci were tested per gene, with an s.d. of 3.1.

Somatic eQTL analysis

Validation of recurrent mutations

Integration with coding mutations in pathways

Survival analysis

| Identification of recurrently mutated loci | Noncoding mutations (WGS) | Coding mutations (WES) | Copy number alteration (SNP array 6.0) | DNA methylation (HM450) | Gene expression (RNA-seq) | ICWG noncoding mutations | Clinical data | Number patients |
|------------------------------------------|--------------------------|------------------------|----------------------------------------|-------------------------|--------------------------|---------------------------|----------------|----------------|
| Somatic eQTL analysis                    | ✓                        | ✓                      | ✓                                      | ✓                       | ✓                        | ✓                         | ✓              | 930            |
| Validation of recurrent mutations        | ✓                        | ✓                      | ✓                                      | ✓                       | ✓                        | ✓                         | ✓              | 783            |
| Integration with coding mutations in pathways | ✓                      | ✓                      | ✓                                      | ✓                       | ✓                        | ✓                         | ✓              | 3,382          |
| Survival analysis                         | ✓                        | ✓                      | ✓                                      | ✓                       | ✓                        | ✓                         | ✓              | 810            |

Many of the identified somatic eQTLs were frequently mutated in specific cancer tissues (Fig. 2c and Supplementary Table 3). Beyond the promoter of TERT, which is highly mutated in several tissues as previously noted (Supplementary Fig. 3a,b), we found recurrently mutated loci associated with expression of DHX34 (mutated in 43% of diffuse large B cell lymphoma, TUBBP5 (29% of lymphomas and 17% of liver cancers), HYI (21% of melanoma), and PCDH1 (19% of acute myeloid leukemia), among others. While most of the somatic eQTLs were mutated in multiple tissues, 12 of the somatic eQTLs were mutated almost exclusively in melanoma (80% or more of the mutations occurred in melanoma). Such enrichment for a single tissue was not seen for any other tissue type.

Somatic eQTLs are recurrently mutated in a second cohort. To systematically validate this network, we examined an independent pan-cancer cohort from ICGC consisting of genome-wide somatic mutation calls for 3,382 patients. Notably, we found that the majority of the somatic eQTLs identified in the original TCGA discovery set were recurrently mutated in the ICGC validation cohort (107 of the 193 at FDR < 20%; Fig. 2d). These included 10 of the 12 melanoma eQTLs, which again were frequently and almost exclusively mutated in the melanoma samples in ICGC (Fisher’s exact test P = 1.6 × 10−12; Supplementary Table 4). For example, a somatic eQTL associated with increased HYI mRNA expression level was mutated in 21% of US melanomas (TCGA) and 18% of Australian melanomas (ICGC).
Increasing DAAM1 expression leads to cell invasion. We next sought to examine in more detail the somatic eQTL located 191 bp upstream of DAAM1 (Fig. 2a and Methods), which is recurrently mutated in patients with melanoma who have metastatic disease in both cohorts (Fig. 2c,d). The DAAM1 protein forms a complex with Dishevelled and RhoA to recruit the actin cytoskeleton, which is thought to increase the motility and invasiveness of cancer cells in response to Wnt signaling22–24. Mutations at this somatic eQTL are associated with increased mRNA expression levels potentially owing to the loss of an E2F motif and the gain of an enhancer element in all four cell lines tested (Fig. 2d and Supplementary Fig. 4). Furthermore, the GFP-expressing cells had significantly higher levels of GFP expression with the mutant rather than the wild-type DAAM1 element in all four cell lines tested (Fig. 3d and Supplementary Fig. 4,e,g).

We also explored the functional relationship between increased DAAM1 expression and cell motility, using an established 3D collagen hydrogel matrix model25. Genome-wide mRNA sequencing was performed on cells grown within low- or high-density collagen, mimicking the stiffness of normal or tumor tissues and eliciting less and more invasive phenotypes, respectively26, 27 (Methods). In these experiments, DAAM1 was one of the most upregulated transcripts under invasive conditions25 (Supplementary Fig. 5). To test whether invasion was functionally dependent on DAAM1, we quantified cell migration behavior after DAAM1 expression was increased artificially by exogenous overexpression (Fig. 3e, Methods and Supplementary Fig. 6). When cells overexpressing DAAM1 were embedded in the 3D collagen hydrogel, they migrated with significantly greater persistence than did wild-type cells (P = 0.008, two-sided Mann–Whitney U test; Supplementary Fig. 6a). Cells overexpressing DAAM1 also invaded for longer distances than wild-type cells (P = 0.01, two-sided Mann–Whitney U test; Fig. 3f–h), while retaining the same velocities.
as wild-type cells (Supplementary Fig. 6b,c). This invasive phenotype was observed in the absence or presence of additional Wnt5a signaling (Supplementary Fig. 6d). These results suggest that increased DAAM1 expression levels allow cells to more efficiently invade the local microenvironment, thereby linking this noncoding mutation to DAAM1 overexpression and cell invasion.

Noncoding mutations dysregulating MTG2 and HYI. Beyond DAAM1, we examined two additional somatic eQTLs, one in the promoter of MTG2 (+19 to +33) and another in the enhancer of HYI (+95,097 to +95,132) (Methods). The first eQTL was associated with decreased MTG2 mRNA expression levels, likely owing to the disruption of a HIF-1α binding motif by the G-to-A mutation 19 bp downstream of the TSS (Fig. 4a; NC_000001.10:g.43824529G>A). This somatic eQTL was present in several types of cancer, including lung adenocarcinoma and sarcoma. Using another GFP-based reporter assay of promoter activity, we found that this G-to-A mutation greatly decreased reporter gene expression in both A549 lung epithelial carcinoma cells and U2OS bone osteosarcoma cells (Fig. 4b). The second eQTL was present in 21% of melanomas (Fig. 2c) and was associated with increased HYI mRNA expression levels, likely owing to G-to-A or GG-to-AA substitutions altering an Ets family binding motif (Fig. 4c; NC_000001.10:g.43824528G>A, NC_000001.10:g.43824529G>A, or NC_000001.10:g.43824528_43824529GG>GAA). As this somatic eQTL was present in an enhancer region, we used a luciferase-based reporter assay where regulatory elements were cloned upstream of a mini-promoter and luciferase. We found that two of the three HYI enhancer variants led to increased expression levels relative to the wild-type sequence in both A375 melanoma cells and MDA-MB-231 breast cancer cells (Fig. 4d).
Noncoding and coding mutations converge on pathways. Next, we investigated the relationship between the 196 genes transcriptionally regulated by somatic eQTLs and the 138 genes previously documented to have recurrent coding mutations in cancer. This combined set of genes was analyzed by Network-Based Stratification (NBS) (Supplementary Fig. 9), which uses a reference molecular network to implicate network regions associated with the genetic alterations in a tumor and groups tumors into subtypes on the basis of similarity of these implicated regions. As a reference molecular network, we used ReactomeFIA, documenting 229,300 molecular interactions among 12,177 human gene products pertaining to previously reported protein–protein, transcriptional and metabolic interactions.

This approach identified a collection of network regions (henceforth called ‘pathways’ for simplicity) that stratified tumors into a hierarchy of increasingly specific subtypes (Fig. 5b). At a resolution of ten subtypes, each subtype was enriched in 2–5 tumor tissues and tumors of each tissue could be subdivided into 1–3 subtypes (Supplementary Fig. 7). Nonetheless, these subtypes differed significantly in their implications for disease-free survival, beyond the baseline survival for each tissue ($P = 3.3 \times 10^{-4}$, log likelihood ratio test controlling for the tissue types as covariates; Fig. 5c and Supplementary Fig. 8).

Subtypes aggregating noncoding and coding mutations. Among the ten subtypes, four were of particular interest as they contained a large proportion of patients with noncoding mutations (Fig. 5d). The ‘CDKN2A–EGFR–TERT subtype’ (Fig. 5e,f) was defined by disruption of the CDKN2A coding sequence, sometimes in combination with noncoding mutations to the TERT promoter, EGFR activation, or BRAF activation. CDKN2A encodes p14ARF, which can form a complex with HIF-1α and inhibit HIF-1-mediated transcription of TERT. These loss-of-function mutations in CDKN2A may release a key brake on the activity of hTERT. Separately, gain-of-function mutations in EGFR may lead to increased levels of mTOR phosphorylation and activation, which can upregulate telomerase activity by forming a complex with hTERT. The synergy between BRAF and TERT mutations has been previously noted and attributed to modulation of TERT transcription through BRAF–RAS–ERK signaling. This pathway was also linked to DAAM1 promoter mutations (Fig. 5d), validated previously, as DAAM1 forms a complex with Disheveled (DVL3), which indirectly regulates transcription of CDKN2A and EGRF through inhibition of Notch1. This subtype was the most aggressive, with median disease-free survival time at 13 months (Fig. 5c).

A second subtype of interest, the ‘TERT–BRAF–IDH1 subtype’ (Supplementary Fig. 9) was characterized by tumors with TERT noncoding mutations or amplifications, combined in some patients with coding alterations to functionally related genes such as BRAF and SKP2. Beyond the synergy between BRAF and TERT mutations as described above, SKP2 is essential for ubiquitination and degradation of p27kip1 (encoded by CDKN1B), which inhibits the activity of hTERT. Amplification of SKP2 in this pathway may thus increase the activity of hTERT.

A third subtype, ‘PIK3CA–PEX6–GATA3’ (Fig. 5g,h), integrated coding alterations activating PIK3CA and inactivating GATA3 with noncoding alterations downregulating PEX6. In this pathway, members of the peroxisomal biogenesis factor family (PEX6 and PEX6) appear to indirectly interact with PIK3CA and GATA3 through the binding of SMAD family members (SMAD3 and SMAD7).

Finally, the fourth subtype, ‘APOBEC2–ARID1A–CTNNB1’, was characterized by the co-occurrence of noncoding mutations within an enhancer of APOBEC2 and coding alterations in ARID1A and CTNNB1. APOBEC2 encodes a nucleic-acid-editing enzyme with well-known mutagenic effects in cancer. Although ARID1A and CTNNB1 are also known cancer drivers, the connections to APOBEC are unanticipated and create a compelling opportunity for further study.
Fig. 5 | Identification of molecular networks and associated tumor subtypes incorporating noncoding mutations. a, Workflow of NBS. b, Resulting hierarchy of subtypes, at increasing resolution from 2–10 subtypes. c, Disease-free survival probabilities (y axis) are plotted against time after diagnosis (x axis), ordered by subtype (colors). Patients with censored survival data are indicated by a plus sign at the censoring time (last follow-up). d, Signature genes are shown for each subtype with a large proportion of patients with noncoding mutations (y axis) or coding changes (x axis). e, Pathways characterizing CDKN2A–EGFR–TERT (a) or PIK3CA–PEX26–GATA3 (g) subtypes, defined as subnetwork regions extracted from ReactomeFI by NBS. f, h, Mutation matrix of the CDKN2A–EGFR–TERT (f) or PIK3CA–PEX26–GATA3 (h) pathway subtypes showing individual tumors (columns; ordered by cancer tissues) with indicated types of mutations on signature genes for that subtype (rows).

Discussion
Relative to coding changes, interpretation of noncoding mutations poses particular challenges owing to the very large number of events and a limited understanding of their functional consequences. Dealing with these challenges requires strategies to boost signal to noise, which we have pursued here by integrating mutations with
key structural and functional data on transcriptional networks. Structurally, maps of enhancer- and promoter–gene interactions amplify signal by selecting noncoding mutations within defined regulatory regions of specific target genes. These mutations are then characterized functionally as somatic eQTLs by requiring their presence to be significantly associated with expression changes in tumors. The result is a global network of transcriptional regulatory interactions in cancer supported by multiple lines of evidence. Given that most tumors we analyzed had noncoding mutations affecting some part of this network, such mutations appear to represent a widespread feature of cancer biology.

Of the approximately 200 noncoding mutations that have previously been identified as recurrent in cancer\(^1\sim^3\), one-third were also identified here as recurrently mutated loci (Fig. 1c), including well-known mutations in the promoters of PLEKHS1 and DPH3. Notably, though, with the exception of TERT, these mutations did not associate significantly with mRNA expression level changes. This suggests that the effects of these mutations are through mechanisms outside of transcriptional regulation or that the effects on mRNA expression are weaker than could be detected given our statistical power (Supplementary Fig. 2c). On the other hand, hundreds of somatic eQTLs were identified, all of which were unanticipated other than those in the promoter of TERT. Many of the affected genes are not yet widely appreciated as cancer drivers, motivating further studies on the mechanistic basis of noncoding mutations in cancer.

Given an association between gene expression changes and a somatic mutation, it is important to consider whether this association reflects a causal relationship. Although it is tempting to assume that the occurrence of a mutation drives gene expression changes, the opposite could be true, where the change in gene expression levels drives the appearance of the mutation (for example, by increased opening and exposure of chromatin). It is also possible that both effects could be due to a third causal factor. However, the three examples we tested experimentally do support a causal link from mutation to expression changes. These results include transcriptional alterations of DAAM1, impacting cell migration (Fig. 3 and Supplementary Fig. 4); MEG2, which encodes a GTPase that regulates mitochondrial ribosomes\(^8\) (Fig. 4a,b); and HYI, which encodes a putative hydroxyprolyl isomerase and may be involved in carbohydrate transport and metabolism\(^9\) (Fig. 4c,d).

Finally, the somatic eQTL analysis introduced here contrasts with germline eQTL studies in several key aspects. First, in GWAS and germline eQTL studies, testing of multiple SNPs is complicated by the strong codependencies among neighboring SNPs at a genomic locus—so-called linkage disequilibrium\(^10\). In contrast, somatic mutations near to one another in the genome are not in linkage disequilibrium as these alterations, by definition, arise independently in each tumor. Second, population stratification caused by ancestry differences has been a major confounder in the analysis of germline variants\(^10\). It is less of a concern for somatic variants, as these are derived from comparisons between tumor and normal genomes from the same individual, eliminating many, if not all, effects due to ancestry. Nonetheless, we controlled for ancestry diversity and found that the impact on somatic eQTL discovery was minimal. Given these aspects, somatic eQTL analysis may have future interest alongside classical eQTLs as a general mode of mapping transcriptional regulatory architecture.

**Methods**

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0091-2.

Received: 28 June 2017; Accepted: 16 February 2018; Published online: 2 April 2018

**References**

1. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).
2. International Cancer Genome Consortium. International network of cancer genome projects. Nature 464, 993–998 (2010).
3. Hoeffe, M. et al. Challenges in identifying cancer genes by analysis of exome sequencing data. Nat. Commun. 7, 12996 (2016).
4. Venvor, F. et al. A landscape of pharmacogenomic interactions in cancer. Cell 166, 740–754 (2016).
5. Khurana, E. et al. Role of non-coding sequence variants in cancer. Nat. Rev. Genet. 17, 93–108 (2016).
6. Melton, C., Reuter, J. A., Spaceck, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat. Genet. 47, 1258–1263 (2015).
7. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 1160–1165 (2014).
8. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations in breast cancer reveals random alterations across 14 tumor types. Nat. Genet. 46, 1258–1263 (2014).
9. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. Science 339, 957–959 (2013).
10. Hofman, M. M. et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841 (2013).
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
12. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
13. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. 33, 364–376 (2015).
14. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
15. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
16. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2017).
17. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and heterochromatin across cancer types. Nat. Biotechnol. 35, 213–219 (2017).
18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015).
19. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152, 633–641 (2013).
20. Potireal, P. A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).
21. Vögelstein, B. et al. Cancer genome landscapes. Science 339, 1546–1558 (2013).
22. Habas, R., Kato, Y. & He, X. Wnt/Frizzled activation of Rho regulates vertebrate gastrulation and requires a novel Formin homology protein Daam1. Cell 107, 843–854 (2001).
23. Liu, W. et al. Mechanism of activation of the Formin protein Daam1. Proc. Natl Acad. Sci. USA 105, 210–215 (2008).
24. Zhu, Y. et al. Dvl2-dependent activation of Daam1 and Robo regulates Wnt/β-catenin-induced breast cancer cell migration. PLoS One 7, e37823 (2012).
25. Frealey, S. I. et al. A distinctive role for focal adhesion proteins in three-dimensional cell motility. Nat. Cell Biol. 12, 598–604 (2010).
26. Frealey, S. I. et al. Three-dimensional matrix fiber alignment modulates cell migration and MT1-MMP utility by spatially and temporally directing protrusions. Sci. Rep. 5, 14580 (2015).
27. Kumar, S. & Weaver, V. M. Mechanics, malignancy, and metastasis: the force journey of a tumor cell. Cancer Metastasis Rev. 28, 113–127 (2009).
28. Velez, D. O. et al. 3D collagen architecture induces a conserved migratory and transcriptional response linked to vasculogenic mimicry. Nat. Commun. 8, 1651 (2017).
29. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. Nat. Methods 10, 1108–1115 (2013).
30. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell 159, 676–690 (2014).

**URLs**

TCGA Research Network, http://cancergenome.nih.gov/; Firehose, https://confluence.broadinstitute.org/display/GDAC/Home; TCGA RNA-seq data description, https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2; poibin Python package, https://github.com/tsakim/poibin; HOMER, http://homer.ucsd.edu/homer/index.html; somatic mutations of the 930 tumors, http://ideker.ucsd.edu/papers/wzhang2017; GitHub site for custom code, https://github.com/wzhang1984/Noncoding-tumor-mutation-paper.
31. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**, R53 (2010).

32. Fatyol, K. & Szalay, A. A. The p14ARF tumor suppressor protein facilitates nucleolar sequestration of hypoxia-inducible factor-1α (HIF-1α) and inhibits HIF-1-mediated transcription. *J. Biol. Chem.* **276**, 28421–28429 (2001).

33. Nishi, H. et al. Hypoxia-inducible factor 1 mediates upregulation of telomerase (hTERT). *Mol. Cell. Biol.* **24**, 6076–6083 (2004).

34. Fan, Q.-W. et al. EGFR signals to mTOR through PKC and independently of Akt in glioma. *Sci. Signal.* **2**, ra4 (2009).

35. Kawauchi, K., Iijima, K. & Yamada, O. IL-2 increases human telomerase reverse transcriptase activity transcriptionally and posttranscriptionally through phosphatidylinositol 3′-kinase/Akt, heat shock protein 90, and mammalian target of rapamycin in transformed NK cells. *J. Immunol.* **174**, 5261–5269 (2005).

36. Li, Y., Cheng, H. S., Chng, W. J. & Tergaonkar, V. Activation of mutant TERT promoter by RAS–ERK signaling is a key step in malignant progression of BRAF-mutant human melanomas. *Proc. Natl Acad. Sci. USA* **113**, 14402–14407 (2016).

37. Cooper, M. T. & Bray, S. J. Frizzled regulation of Notch signalling polarizes cell fate in the Drosophila eye. *Nature* **397**, 526–530 (1999).

38. Spruck, C. et al. A CDK-independent function of mammalian Cks1: targeting of SCFSkp2 to the CDK inhibitor p27Kip1. *Mol. Cell* **7**, 639–650 (2001).

39. Lee, S.-H. et al. IFN-γ/IRF-1-induced p27Kip1 down-regulates telomerase activity and human telomerase reverse transcriptase expression in human cervical cancer. *FEBS Lett.* **579**, 1027–1033 (2005).

40. Warner, D. R., Roberts, E. A., Greene, R. M. & Pisano, M. M. Identification of novel Smad binding proteins. *Biochem. Biophys. Res. Commun.* **312**, 1185–1190 (2003).

41. Okuyama, S. et al. Excessive activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide 2 (APOBEC2) contributes to liver and lung tumorigenesis. *Int. J. Cancer* **130**, 1294–1301 (2012).

42. Hirano, Y., Ohnwi, R. L., Wada, C., Yoshimura, S. H. & Takeyasu, K. Human small G proteins, ObgH1, and ObgH2, participate in the maintenance of mitochondria and nuclear architectures. *Genes Cells* **11**, 1295–1304 (2006).

43. Ashuchi, M. & Misono, H. Biochemical evidence that Escherichia coli hyi (orfβ0508, gip) gene encodes hydroxypropruvate isomerase. *Biochim. Biophys. Acta* **1435**, 153–159 (1999).

44. Bush, W. S. & Moore, J. H. Chapter 11: genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).

45. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

50. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).

Acknowledgements

The results published here are in whole or part based upon data generated by the TCGA Research Network (see URLs). We would also like to acknowledge the clinical contributors and the data producers from the ICGC who have generated the particular datasets and made them available for public analysis. This work was supported by NIH grants to T.I. (U24CA184427, U54CA209891, P50GM085764, P41GM103504 and R01HG009979) and H.C. (DP5OD017937). G.X. is supported by a UCSD CTRI grant (UL1TR001442). J.K.H. and D.O.V. are supported by a Burroughs Wellcome Fund Career Award at the Scientific Interface (1012027), an NSF CAREER Award (1651855), and UCSD CTRI and FISP pilot grants. We would like to thank members of the Ideker laboratory for valuable comments and critical reading of the manuscript. Finally, we wish to thank the patients and their families for their contributions of valuable data without which this project would not have been possible.

Author contributions

W.Z. and T.I. conceived the study. W.Z. designed and performed most of the analyses. G.X. performed mutation calling of 358 tumors. C.M. and M.S. provided mutation calling of 572 tumors. A.B.-G., K.S.S., J.P.S., K.M.O. and E.K.F. performed the somatic eQTL reporter assays. A.B.-G. and J.F.K. analyzed the flow cytometry and luciferase assay data. A.B.-G., J.P.S. and K.L. performed protein electropherogram analysis. D.O.V., K.C. and S.I.F. performed 3D cell culture assays. M.K.Y. and H.C. helped W.Z. in designing the somatic eQTL analysis. J.K.H. helped W.Z. in network analysis. T.I., J.F.K. and W.Z. wrote the manuscript and formulated all figures.

Competing interests

T.I. is cofounder of Data4Cure, Inc., and has an equity interest. T.I. has an equity interest in Idexia BioSciences, Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies. No potential conflicts of interest were disclosed by the other authors.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0091-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.Z. or J.F.K. or T.I.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Calling and clustering of somatic noncoding mutations. Somatic noncoding mutations from 930 tumors were called as described in the main text. Clusters of noncoding mutations within \(d = 50\) bp of each other were merged using BEDTools\(^{11}\) until no such cluster was located within \(d\) bp from any other. Loci with mutations in \(k \leq 5\) tumors were removed from further analyses. The above parameters \(d\) and \(k\) were chosen to aggregate mutations within a short distance with a modest requirement of recurrence. We achieved very similar results when \(d\) was within the range of 20 to 60 bp (inclusive). Whenever a subset of 930 tumors was used in subsequent analyses (Fig. 1A), this set was again filtered to remove those altered in fewer than \(k\) tumors within the subset. We also calculated a ‘concentration score’ to penalize loci where mutations were spread over a large region rather than concentrated at a single base pair, as might be expected for sites affecting gene transcription. Within each locus, we selected the mutated position present in the largest number of patients. The proportion of patients affected at that position (out of all patients affected by mutations at that locus) was defined as the concentration score. Loci scoring <35% were removed from further study. It is worth noting that the threshold for the concentration score is somewhat arbitrary and could lead to certain loci with multiple closely located somatic mutations being missed. It should also be noted that, by clustering noncoding mutations into loci, we assume that all SNVs in a locus act in a similar way. This assumption is consistent with the previously identified SNVs in the TERT promoter. Our analysis does not attempt to detect loci in which different SNVs alter gene expression in opposite directions.

RNA-seq, CNA and DNA methylation data processing. RNA-seq, CNA (SNP 6.0) and DNA methylation (Illumina HM450) data for TCGA tumors were downloaded from Firehose (see URLs). The data were processed as follows. First, for RNA-seq, the RSEM count for a gene (RNA-seq by expectation maximization)\(^{47}\) was normalized by dividing by the 75th percentile of RSEM values within the tumor sample and multiplying by 1000, according to TCGA practice (see URLS). Genes were retained if the normalized RSEM was \(>1\) in \(>50\%\) of tumors, resulting in 16,413 expressed genes. Normalized RSEM values were log, transformed and \(z\) score standardized for subsequent analyses. Second, for CNA, we used the output of GISTIC2, which indicates gene-level CNAs for all samples. The CNAs are in units of (copy number — 2), so that normal copy number (no amplification or deletion) has a value of 0, whereas genes with amplifications and deletions have positive and negative values, respectively. A gene is assigned the highest amplification or deletion as the lowest deletion value among the markers it covers. Among the 783 patients with both mRNA expression and genome sequence data, 761 also had copy number data available. The remaining patients were assigned 0 for all CNAs. Third, methylation data for the remaining patients were retained if the normalized RSEM was \(>1\) in \(>50\%\) of tumors, resulting in 16,413 expressed genes. Normalized RSEM values were log, transformed and \(z\) score standardized for subsequent analyses. Among the 783 TCGA patients with both mRNA expression and genome sequences data, 605 had methylation data available. Methylation data for the remaining patients were imputed using mean values for the DNA methylation of each gene.

Linking recurrently mutated loci to transcriptional target genes. Our recurrently mutated loci were extended by \(100\) bp on each side when mapping to promoters or enhancers. Transcriptional regulatory interactions from recurrently mutated loci to target genes were defined whenever a locus had \(>50\%\) of its sequence overlap with either the promoter region of a gene (\(\pm 1\) kb from the TSS) or a gene enhancer region defined by GeneHancer\(^{17}\). In the case where an enhancer was concentrated at a single base pair, as might be expected for sites affecting gene transcription, we used the Storey approach\(^{50}\) with a threshold of FDR \(\leq 20\%\). Finally, a somatic eQTL analysis with five mutations if \(W > 1.2\) or with ten mutations if \(W > 0.9\).

Independent validation of recurrence. To validate the recurrence of mutations in the identified somatic eQTLs, we downloaded simple somatic mutations (substitutions) called from the WGS of \(n = 3382\) publicly available non-US donors from the ICGC.\(^{33}\) For each eQTL, the number of mutated patients \(k\) was used as the test statistic. To determine whether \(k\) was greater than expected owing to the background mutation rate (BMR), we developed an approach for estimating BMR that was conceptually similar to MutSigCV\(^{18}\). First, a large pool of 2000 candidate background sequences was created by randomly reassigning (without replacement) the location of the eQTL to the same type of noncoding genomic regions (promoters or putative enhancers\(^{19}\)) while retaining the eQTL length. Each of these 20,000 sequences was placed in a 3D feature space taking into account nucleotide content, DNA replication timing and gene expression. Nucleotide content was represented as the percentages of all possible mononucleotides (A/T versus C/G), dinucleotides (e.g., AA, AC and AG) and trinucleotides (for example, AAA, AAC and AAG), encoded as a 44-dimensional vector. This information was then compressed into a single feature representing nucleotide content, using the Pearson’s correlation between the vector of the candidate sequence and the vector of the original eQTL. DNA replication timing was obtained from ENCODE via the UCSC Genome Browser\(^{10}\). To create a single replication timing feature, we used the average waveform—smoothed signal from the following 14 cell lines: BJ, GM06990, GM12801, GM12812, GM12813, GM12878, HeLa-S3, HepG2, HUVEC, IMR-90, K562, MCF-7, NHEK and SK-N-SH, according to the method of Melton and colleagues.\(^{21}\) For gene expression, the median expression value of the nearest gene (log, transformed RNA-seq data, 783 TCGA patients) was used as a feature. The above three features were \(z\) score standardized. Within this feature space, the top 5% (1,000 of 20,000) background sequences with the smallest Euclidean distance to the eQTL of interest were selected. For each patient, a patient-specific BMR was estimated as the number of mutations with at least one mutation in that patient out of the 1,000 selected sequences. Finally, we estimated the probability of having observed \(k\) or more mutations in \(n\) patients in the eQTL of interest using a Poisson binomial model

\[
P(k \geq k) = \sum_{i=k}^{n} \sum_{i=0}^{\min(A_i, k)} \sum_{i=0}^{\min(B_i, k)} \sum_{i=0}^{\min(C_i, k)} \sum_{i=0}^{\min(D_i, k)} (1 - P_f) i \prod_{i=A, B, C, D} (1 - P_f)
\]

where \(P_f\) is the probability for each of \(k\) or more integers that can be selected from \(\{1, 2, \ldots, n\}\), \(P_f\) or \(P_i\) is the probability that patient \(i\) or patient \(j\) is mutated, \(A\) is a set of \(k\) integers that can be selected from \(\{1, 2, \ldots, n\}\) and \(A'\) is the complement of \(A\). In practice, we used an approximation for the Poisson binomial in the poibin Python package (see URLs).

Power analysis. Statistical power depends on various parameters, including the number of samples, the eQTL effect size, the noise, and the significance threshold. Instead of a simulation based on a model of noise, we evaluated statistical power using the actual data. All locus–gene pairs were plotted in Supplementary Fig. 2c, evaluated by the number of patients with mutations (\(x\)-axis) versus the change in gene expression given the mutation (\(y\)-axis; defined by \(W = \frac{\delta_{\text{mut}} - \delta_{\text{wild}}}{\lambda\delta_{\text{wild}} + \delta_{\text{mut}}}\), one unit of \(W\) represents 1 sd of change in the final gene expression). Power was defined as 1 – (Type II error) at a significance level of \(P\) (type I error) = 0.0085, which is approximately at 20% FDR. We calculated power using the pwr.f2.test function in R, where the \(f^2\) effect size was calculated on the basis of the proportion of variance explained by two nested models (\(f^2 = \frac{\delta_{\text{mut}} - \delta_{\text{wild}}}{\lambda\delta_{\text{wild}} + \delta_{\text{mut}}}\)). Our somatic eQTL analysis has 50% power to detect a somatic eQTL with five mutations if \(W > 1.2\) or with ten mutations if \(W > 0.9\).
Transcription factor binding motif analysis. Each reference and somatically altered nucleotide site, along with ±7 bp of flanking sequence, was analyzed using HOMER3 (see URLs). HOMER searches for matches within a library of 319 vertebrate motifs (position weight matrices). Specifically, we ran the findMotifs.pl program with default parameters to find motifs from FASTA files. The reference and altered sequences were used as the background for each other to control the nucleotide context. The command line is findMotifs.pl seqlist_mappable_alt.fa fasta logo/-fastaBg seqlist_mappable_ref.

The list of somatic eQTLs that disrupt or create transcription factor binding motifs in four or more patients is reported in Supplementary Table 2.

Prioritizing somatic eQTLs for subsequent functional validation. The three somatic eQTLs selected for functional studies (DAAM1, HYI1 and MTG2) were chosen based on the specific biological interest of the authors and several rules of thumb:

1. The somatic eQTL alters a known transcription factor binding motif in many patients;
2. The somatic eQTL falls in open chromatin in previously mapped cell lines and conditions (for example, in regions with markers such as H3K27ac and H3K4me1)45;
3. The affected target gene has high endogenous mRNA expression levels in cell lines46 that match where the somatic eQTL was detected;
4. The somatic eQTL is not present in a region with repetitive DNA.

Note that none of this information was used to filter loci before somatic eQTL analysis, as it is not complete, conclusive or cancer specific.

Generation of reporter plasmids. To examine the effect of the DAAM1 somatic eQTL on reporter gene expression, the wild-type and mutant regulatory regions from −233 bp to +148 bp relative to the TSS, the somatic eQTL at −202 to −191 bp, were synthesized and cloned upstream of GFP (Fig. 3b). For MTG2, the cloned region spanned −200 bp to +200 bp relative to the TSS, the somatic eQTL at +19 to +33 bp.

For the somatic eQTL located in the HYI enhancer, the region corresponding to +94,931 to +95,332 bp relative to the TSS, including the somatic eQTL at +95,097 to +95,132 bp, was cloned into the firefly luciferase reporter plasmid pGL4.23 (Promega). Mutations were generated using the Q5 Site-Directed Mutagenesis kit (New England Biolabs). All inserts for the GFP and luciferase reporter plasmids were confirmed to match the human reference genome hg19 by Sanger sequencing.

Promoter and enhancer activity assays. Cell lines used to evaluate promoter activity were plated in six-well dishes at 300,000 cells per well, with three replicates per group. The next day, plasmid DNA (1 μg) was transfected using Lipofectamine 3000 (Thermo Fisher). Forty-eight hours after transfection, cells were harvested and suspended in ice-cold PBS with 1% FBS. GFP expression was measured by flow cytometry on a FACSCalibur or FACSCan (BD Biosciences). Flow cytometry data were analyzed with FlowJo v10 (BD Biosciences). Cells with typical forward (size) and side (granularity) scatter properties were further analyzed for GFP expression. As a negative control, cells were transfected with the empty lentivector Puro plasmid (Addgene) for the DAAM1 experiments (Fig. 3d) and Supplementary Fig. 4) or a promoterless GFP plasmid (pMT-GFP; Origene) for the MTG2 experiments (Fig. 4b). As a positive control for all GFP experiments, we used a plasmid with the cytomegalovirus promoter upstream of GFP. All flow cytometry experiments were performed at least three times. Early pilot experiments were often performed on single or duplicate samples with the final triplicate version often performed at least twice.

To evaluate the activity of the enhancer region of HYI, A375 and MDA-MB-231 cells were plated in white, opaque, 96-well plates at 10,000 cells per well, with four replicates per group. Cells were transfected 24 h later using Lipofectamine 3000 with 33 ng of total DNA: 27.5 ng of the firefly pGL4.23 constructs and 5.5 ng of control Renilla pGL4.75 (Promega) plasmid. Firefly and Renilla luciferase activities were measured 48 h after transfection using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer’s instructions. Luciferase values were collected on a BioTek Synergy HT, and data were collected via Gen5 2.0.11 software. To calculate relative luciferase values, background signal was first subtracted from each channel. Then, firefly luminescence was divided by Renilla luminescence. The average value for the wild-type enhancer was set to 1, and the mutated samples were evaluated in comparison to this control. Experiments in both cell lines were performed three times, with each experiment consisting of samples in quadruplicate.

DAAM1 overexpression. Wild-type MDA-MB-231 breast cancer cells were transfected with a plasmid encoding the full DAAM1 cDNA (Origene, RC217675). Cells were then selected using G418 (500 μg/ml) for 7 d to ensure stable expression of the DAAM1 construct. DAAM1 overexpression was verified by extracting total protein and quantifying using the Wes electropherogram (Proteinsimple) with anti-DAAM1 antibody (clone W2-3, sc-100942, lot B1815, Santa Cruz, 1:250 dilution) and anti-tubulin antibody (clone YL1/2, MAB1864, lot 2886723, Millipore, 1:250 dilution). DAAM1 expression was 5.3-fold greater in cells with the DAAM1 overexpression construct relative to wild-type cells (Supplementary Fig. 6e).

3D collagen cell migration assays. Collagen matrices were prepared by mixing cells suspended in culture medium and 10× reconstitution buffer, one-to-one with soluble rat tail type I collagen in acetic acid (Corning). Sodium hydrosulfite was used to neutralize pH to 7.4 in each time point in each cell line used for imaging and using image recognition software (Metamorph/Metave, Molecular Devices). Tracking data were processed to calculate cell speed using an extension of previously published scripts47. Cell migration assays (Fig. 3f–h) were performed two times, and both attempts showed the same trend.

RNA sequencing from cells in 3D culture. In Supplementary Fig. 5, cell migration assays were performed using wild-type MDA-MB-231 breast cancer cells and HT-1080 fibrosarcoma cells. 3D collagen gels were seeded in three independent experiments and harvested after 24 h of culture for RNA extraction and directly homogenized in TRIzol (Thermo Fisher). Total RNA was purified using the High Pure RNA Isolation kit (Roche), and the integrity of the samples was verified using RNA Analysis ScreenTape (Agilent Technologies). Total RNA samples were sequenced using the TrueSeq Stranded mRNA Sample Prep kit (Illumina) and the Illumina MiSeq platform at a depth of >25 million reads per sample. Paired-end reads were aligned to the hg19 UCSC human genome reference using Bowtie248 and streamed to the Express tool for transcript abundance quantification.

Tumor genetic profiles integrating noncoding and coding alterations. Integrated genetic alteration profiles were constructed for the 810 tumors with WGS, WES and CNA data (Fig. 1a) as follows. Known oncogenes or tumor suppressors49 were combined with the set of target genes of eQTLs identified by the somatic eQTL analysis (see above); each of these genes was then classified as wild type (0) or altered (1) in each tumor, constituting its tumor genetic profile. In this profile, an alteration was defined as follows. Most oncogenes (for example, EGRF) were considered altered (activated) if impacted by a missense mutation, in-frame indel or copy number amplification. For oncogenes typically altered only by amplification (CCND1, MDM2, MDM4, MYC, MYCL, MYCN, NCOA3 and SPT2), only copy number amplifications were considered as alterations and not SNVs or indels. Tumor suppressors (for example, CDKN2A) were considered altered (inactivated) if there was any type of non-silent mutation or a copy number deletion. For each target gene, we defined a dominant direction of regulation Δd = {1, −1} as the sign of the coefficient (q, in equation (1) of its most significantly associated eQTL. Noncoding mutations in eQTLs that led to a transcriptional change in the dominant direction were considered alterations of such genes. For TERT, copy number amplifications in the coding region were also considered as alterations, as both promoter mutations and gene amplifications have been associated with growth advantage of tumor cells and poor prognosis of patients50,51,52.

Network-based stratification to identify tumor subtypes. Network propagation53 was used to compute the pairwise similarities among tumor genetic alteration profiles (see above) within the Reactome functional interaction network (ReactomeF1)54. Each tumor genetic profile was propagated across this network on the basis of a random walk model (equivalent to heat diffusion) with a restart probability of 0.5. After convergence, the score of each gene (temperature) represents its network proximity to genetic alterations. The top 70 principal components of these scores, representing the tumor’s network-transformed profile (Fig. 5a), were analyzed using the sklearn.cluster.SpectralClustering package (https://scikit-learn.org/0.18/modules/generated/sklearn.cluster.SpectralClustering.html) to divide tumors into subtypes with the shortest Euclidean distance. We chose k = 170, which ensures that the similarity graph is connected, as previously recommended55. Next, this graph is analyzed to partition tumors into subtypes at different resolutions (number of subtypes n = {2, …, 10}). Following spectral clustering, each set of n (parent) subtypes was divided into the n + 1 (child) subtypes to track the similarity of tumor assignments (Fig. 5b). An arrow was drawn from a parent to child subtype if they shared ≥18 tumors.

Characterizing tumor subtypes with signature genes and subnetworks. For each subtype, we defined a set of ‘signature genes’ as those that had higher network-transformed scores in that subtype than others (t test, Benjamin–Hochberg FDR <0.1) and, among these, were more frequently altered in that subtype (Fisher exact test, FDR <0.05; Fig. 5b–e). To identify subnetworks, this set was expanded to include ‘intermediate genes’ with relatively high network-transformed scores (t test, FDR <0.05) that lay on the shortest paths between each pair of signature.
Survival analysis. We used the coxph package in R statistical software to fit Cox proportional-hazard models. \( P \) values were calculated by log likelihood ratio test. To evaluate whether the subtype classifications provided additional prognostic power beyond the baseline survival expectancy due to cancer tissue, we compared the likelihood for the complete model, including NBS-derived molecular subtypes \( s \) and cancer tissues \( c \) as covariates, against that of a null model that included cancer tissues only.

\[
\begin{align*}
\text{Complete model: } & \lambda(t \mid s, c) = \lambda_0(t) \exp(\beta_s + \beta_c c) \\
\text{Null model: } & \lambda(t|c) = \lambda_0(t)\exp(\beta_c c)
\end{align*}
\]  

where \( \lambda_0(t) \) is the baseline hazard function. Then, a log likelihood ratio statistic was defined as

\[ D = -2 \ln \left( \frac{\text{likelihood for null model}}{\text{likelihood for complete model}} \right) \]
Experimental design

1. Sample size

Describe how sample size was determined.

All of the flow cytometry experiments presented (Fig 3, Fig 4, Sup Fig 4) were performed in three independent cell culture replicates (Sup Table 5). Each triplicate consists of 50,000 events counted through the flow counter. Although the sample size was not pre-determined, it proved sufficient to observe a very significant difference in % GFP+ and GFP intensity (two-tailed t-test p = 5.63E-07 to 5.74E-04). Experiments with this sample size are in accordance to conventions in this field.

For the luciferase assays (Fig 4d), each experiment was performed in four independent cell culture replicates (Sup Table 6). Although the sample size was not pre-determined, it proved sufficient to observe a very significant difference in % GFP+ and GFP intensity (two-tailed t-test p = 4.60E-08 to 6.21E-03). Experiments with this sample size are in accordance to conventions in this field.

In the cell migration assay, 74 and 83 cells were imaged in the two groups in Fig. 4h and Sup Figs 6a-c, and 63 and 15 cells were images in Fig 6d. The sample size is sufficient to observe a significant difference in invasion distance, persistence, and invasion distance with additional Wnt5a signaling (two-tailed Mann–Whitney U test p = 0.01, 0.008, and 0.0002, respectively).

In somatic eQTL analysis (Fig 2a), all 783 TCGA tumors with both genome sequence and mRNA expression data were used for the study. The sample size is sufficient to identify 193 somatic eQTLs at a FDR of 20% (F-test).

In the validation of somatic eQTL recurrence (Fig 2d), all 3,382 publicly available non-US ICGC tumors with whole genome sequence data were used for the study. The sample size is sufficient to validate that the majority of the somatic eQTLs identified in the original TCGA discovery set were recurrently mutated in the ICGC validation cohort (107 of the 193 at an empirical FDR of 20%).

In the RNA-seq analysis (Sup Fig 5b), each experiment was performed in three independent cell cultures. Although the sample size was not pre-determined, it proved sufficient to observe a significant difference in DAAM1 expression (two-tailed t-test p = 0.056 and 0.016). Experiments with this sample size are in accordance to conventions in this field.

2. Data exclusions

Describe any data exclusions.

No data were excluded from the analysis.

Note that in our analysis, although loci were selected in a series of consecutive steps (Figs 1c, d), each locus was tested in the full set of tumors (n = 783) for which whole genome sequence and mRNA expression are both available.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All flow cytometry experiments were performed at least three times. Early pilot experiments were often performed on single or duplicate samples with then the
final triplicate version often also performed at least twice. The figure presented always represented the majority of experimental findings.

All luciferase assays experiments were performed three times, with each experiment consisting of samples in quadruplicate. The figure presented always represented the majority of experimental findings.

Cell migration assays (Fig. 4f-h) were performed two times and both attempts showed the same trend.

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

No method of randomization was used.

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

The investigators were not blinded to group allocation during data collection and/or analysis.

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.

Software
Policy information about availability of computer code

7. Software
Describe the software used to analyze the data in this study.

Custom code is publicly available at https://github.com/wzhang1984/Noncoding-tumor-mutation-paper. Otherwise we used bedtools v2.23.0 to cluster mutations, GeneHancer for enhancer-gene mappings, R v3.2.5 for statistical learning, probabilistic estimation of expression residuals (PEER) v1.3 to identify hidden factors, UCSC Genome Browser to obtain DNA replication timing, poibin Python package (https://github.com/tsakim/poibin) for the Poisson binomial model, HOMER v4.8.2 for motif analysis, FlowJo v10.2-4 for flow cytometry data analysis, Metamorph/Metavue v7.8 for tracking cells, Bowtie2 v2.2.6 and eXpress v1.5.1 for RNA-seq analysis, and Cytoscape v3.5.1 for network visualization.

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.

There are no restrictions on the availability of the materials used for this project except for the GFP and Luciferase reporter constructs, which will be available upon publication.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

DAAM1 overexpression was verified by extracting total protein and quantitating it using the Wes electropherogram (ProteinSimple) with an anti-DAAM1 antibody (clone WW-3, cat# sc-100942, lot# B1815, Santa Cruz, 1:250 dilution, mouse) and an anti-tubulin antibody (clone YL1/2, cat# MAB1864, lot# 2886723, Millipore, 1:250 dilution, rat).

Both antibodies were raised against human antigens and have been tested by their manufacturers for use with human samples.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

The following cell lines were acquired directly from ATCC: A375, RPMI-7951, U2OS, A549 and HT1080. MDA-MB-231 cells were acquired from the PSOC network.

b. Describe the method of cell line authentication used.

U2OS, MDA-MB-231 and HT-1080 cell genomic DNA was submitted for STR characterization to IDEXX BioResearch. The others were not, as they were recently obtained specifically for this project.

c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines are tested for mycoplasma upon receipt, or 48-72 hours post-thawing from cryostorage. All tests gave negative results.

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.

None of the cell lines used are listed in ICLAC’s v8 records.

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.

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.

The study does not involve human research participants