nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0091-2

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

Wei Zhang[1,8*], Ana Bojorquez-Gomez[1,8], Daniel Ortiz Velez[2], Guorong Xu[3], Kyle S. Sanchez[1], John Paul Shen[1], Kevin Chen[2], Katherine Licon[1], Collin Melton[4], Katrina M. Olson[1,5], Michael Ku Yu[1], Justin K. Huang[1,6], Hannah Carter[1], Emma K. Farley[1,5], Michael Snyder[4], Stephanie I. Fraley[2], Jason F. Kreisberg[1*] and Trey Ideker[1,2,6,7*]

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.

Human cancers are fundamentally heterogeneous, with many distinct subtypes associated with differences in molecular, cellular and clinical characteristics. To gain insight into this complexity, projects such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) have used massively parallel DNA sequencing to construct large catalogs of somatic mutations in many types of tumors[1–3]. Focusing initially on protein-coding regions, several hundred genes were found to be recurrently mutated in cancer, a few of which are targetable therapeutically[4].

As coding regions account for less than 2% of the human genome, attention is now shifting to the greater number of somatic mutations in noncoding regions[5]. Thus far, the clearest role for noncoding mutations in cancer has been in the promoter of the telomerase reverse transcriptase gene (*TERT*)[6–8], with such mutations leading to increases in *TERT* expression levels in many types of tumors[8,9]. Although whole-genome sequencing (WGS) of tumor–normal pairs has found recurrent somatic mutations at several other noncoding loci, assessing the function of these mutations, if any, has been challenging[6–8]. In this respect, the task of functional interpretation is greatly aided by recent efforts of consortia such as ENCODE[10,11] and Roadmap[12,13], which have published extensive reference maps of noncoding regions and their likely transcriptional regulatory connections to genes. Here we show that such networks provide critical information for identifying noncoding mutations with functional impacts among the many others that may be spurious[6].

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[1] (Fig. 1a), we identified 3.5×10^7 sites with somatic single-nucleotide variations (SNVs). We called these SNVs uniformly across all genomes using the MuTect suite[14] according to GATK best-practice recommendations[15, 16] and those of Melton et al.[6] (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[17] 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,

[1]Department of Medicine, University of California, San Diego, La Jolla, CA, USA. [2]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [3]Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [5]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [6]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA, USA. [7]Cancer Cell Map Initiative (CCMI), University of California, La Jolla and San Francisco, CA, USA. [8]These authors contributed equally: Wei Zhang and Ana Bojorquez-Gomez. *e-mail: wez124@ucsd.edu; jkreisberg@ucsd.edu; tideker@ucsd.edu

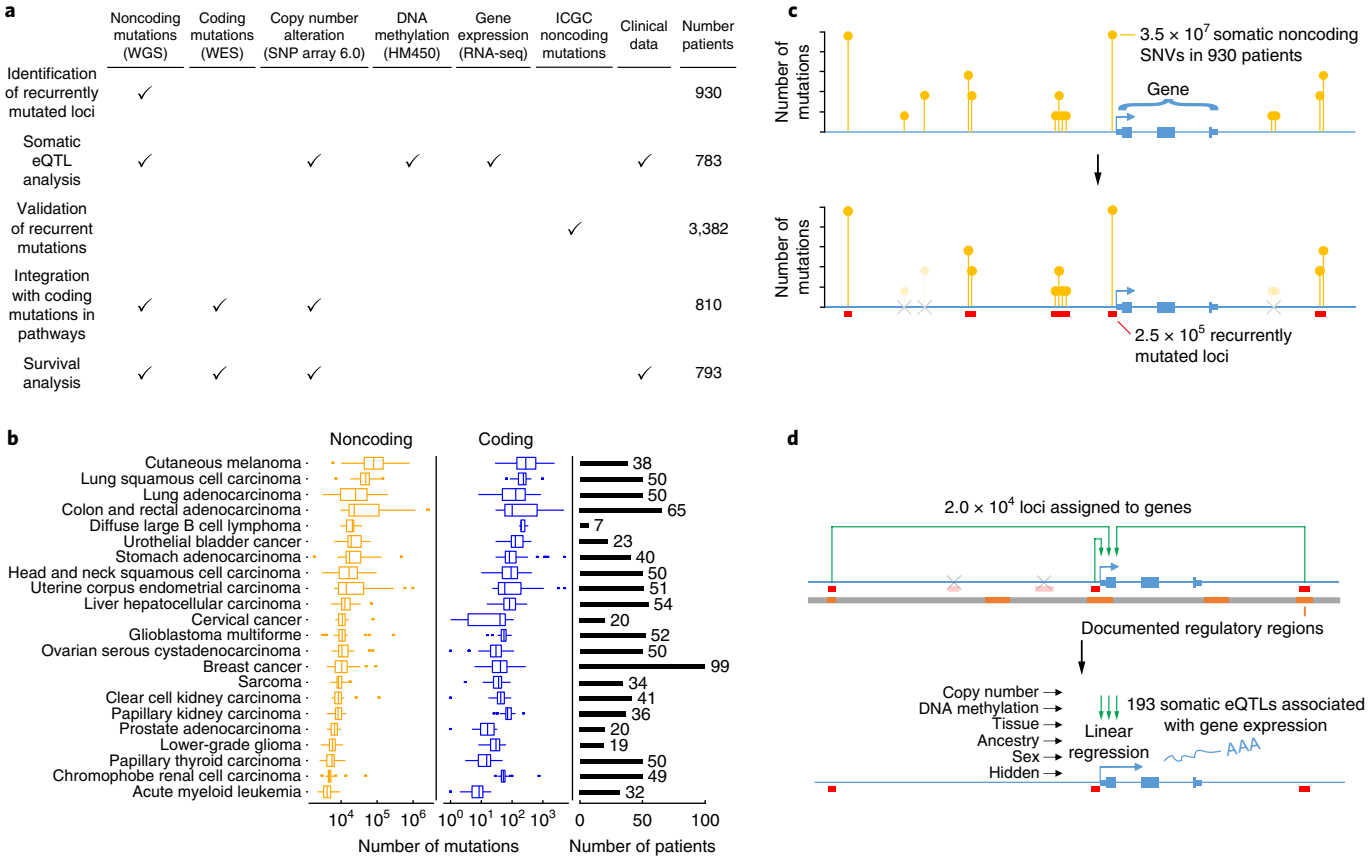

Fig. 1 | Mutation calling and somatic expression quantitative trait locus analysis. a, Types of data and numbers of tumors used in this study. **b**, Number of mutations called per tumor. Box plots show the distribution of this number within tumors of each tissue type (center line, median; upper and lower hinges, first and third quartiles; whiskers, highest and lowest values within 1.5 times the interquartile range outside hinges; dots, outliers beyond 1.5 times the interquartile range). The number of tumors of each type (sample size) is shown in the right panel. **c**, Clustering of somatic noncoding mutations resulting in identification of recurrently mutated loci. **d**, Workflow of somatic eQTL analysis. WGS, whole-genome sequencing; WES, whole-exome sequencing; SNV, single-nucleotide variation.

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[18,19]. 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 and a maximum number of 53 (Supplementary Fig. 1e).

Altogether, this approach identified a cancer transcriptional network of 206 regulatory interactions between 193 somatic eQTLs and 196 gene-expression-level changes, at a false discovery rate (FDR) of 20% (Fig. 2a,b and Supplementary Fig. 2; somatic eQTLs at different FDR thresholds are provided in Supplementary Table 1 and the Supplementary Note). At least one locus in this network was somatically mutated in 88% of cases studied (820 of 930), suggesting that transcriptional dysregulation through noncoding mutations is a general property of most tumors. Somatic eQTLs linked noncoding mutations to the expression levels of 13 known tumor-suppressor genes or oncogenes[4,20,21] (Supplementary Table 1), although, interestingly, known cancer-associated genes were not significantly enriched overall (Fisher's exact test $P=0.3$). We also found that 43% of somatic eQTLs disrupted or created a transcription factor binding motif (83 of 193; Supplementary Fig. 3 and Supplementary Table 2), although this percentage was very similar for recurrently mutated loci not detected as somatic eQTLs (40%; 2409 of 8607).

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[6–8] (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[2]. 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 $P=4.1\times10^{-12}$; 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).

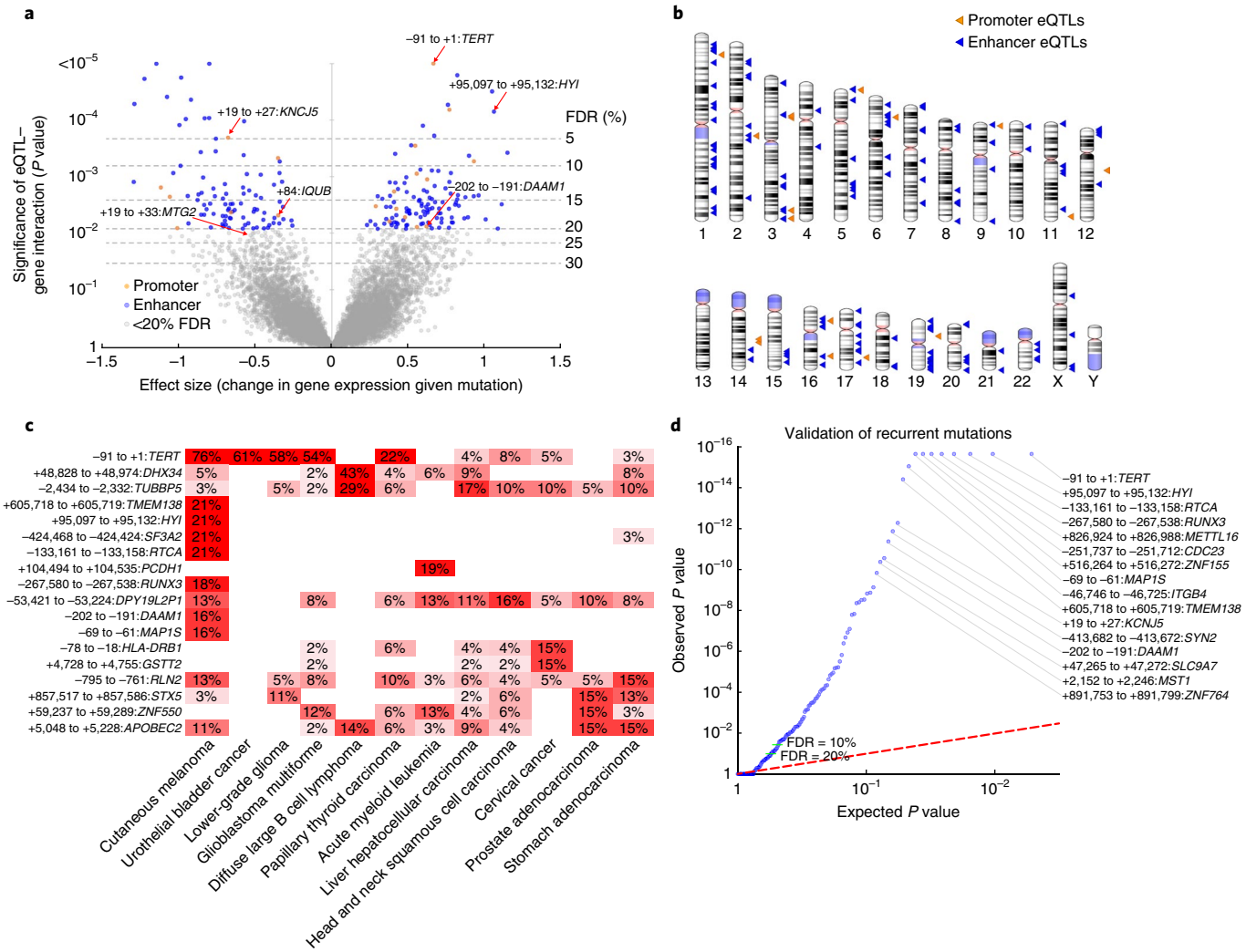

Fig. 2 | Effect size and recurrence of somatic eQTLs. a, Volcano plot of associations between somatic eQTLs and the expression level changes of their target genes, evaluated by significance (*y* axis; *F*-test *P* value, $n = 783$ tumors) versus effect size (*x* axis). One unit on the *x* axis represents 1 s.d. of change in gene expression. FDR was calculated using the Storey approach[50]. Selected somatic eQTLs are labeled by coordinates in base pairs relative to the TSS of the target gene. **b**, Ideogram of the 193 significant somaitc eQTLs at FDR < 20%. **c**, Heat map showing the percentage of patients in various cancer tissues with alterations in each somatic eQTL. Somatic eQTLs and cancer tissues with mutation rates of ≥15% are shown. **d**, Validation of somatic eQTL recurrence in a pan-cancer cohort from ICGC. The quantile–quantile plot shows the observed empirical *P* values of mutation recurrence ($n = 3{,}382$ tumors) compared to the random expectation for the 193 somatic eQTLs. FDR was calculated using the Benjamini–Hochberg approach.

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 signaling[22–24]. Mutations at this somatic eQTL are associated with increased *DAAM1* mRNA expression levels potentially owing to the loss of an E2F motif and the gain of an Ets motif (Fig. 3a; NC_000014.8:g.59655190 G > A). To confirm a causal relationship between the somatic eQTL and gene expression level changes, wild-type and mutant *DAAM1* regulatory elements were inserted upstream of the GFP gene (Fig. 3b). Analysis by flow cytometry showed that the mutated regulatory element led to a significantly higher percentage of cells expressing GFP in melanoma, sarcoma and breast cancer cell lines (Fig. 3c,d 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. 4c,e,g).

We also explored the functional relationship between increased *DAAM1* expression and cell motility, using an established 3D collagen hydrogel matrix model[25]. 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, respectively[26,27] (Methods). In these experiments, *DAAM1* was one of the most upregulated transcripts under invasive conditions[28] (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. 6e). 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

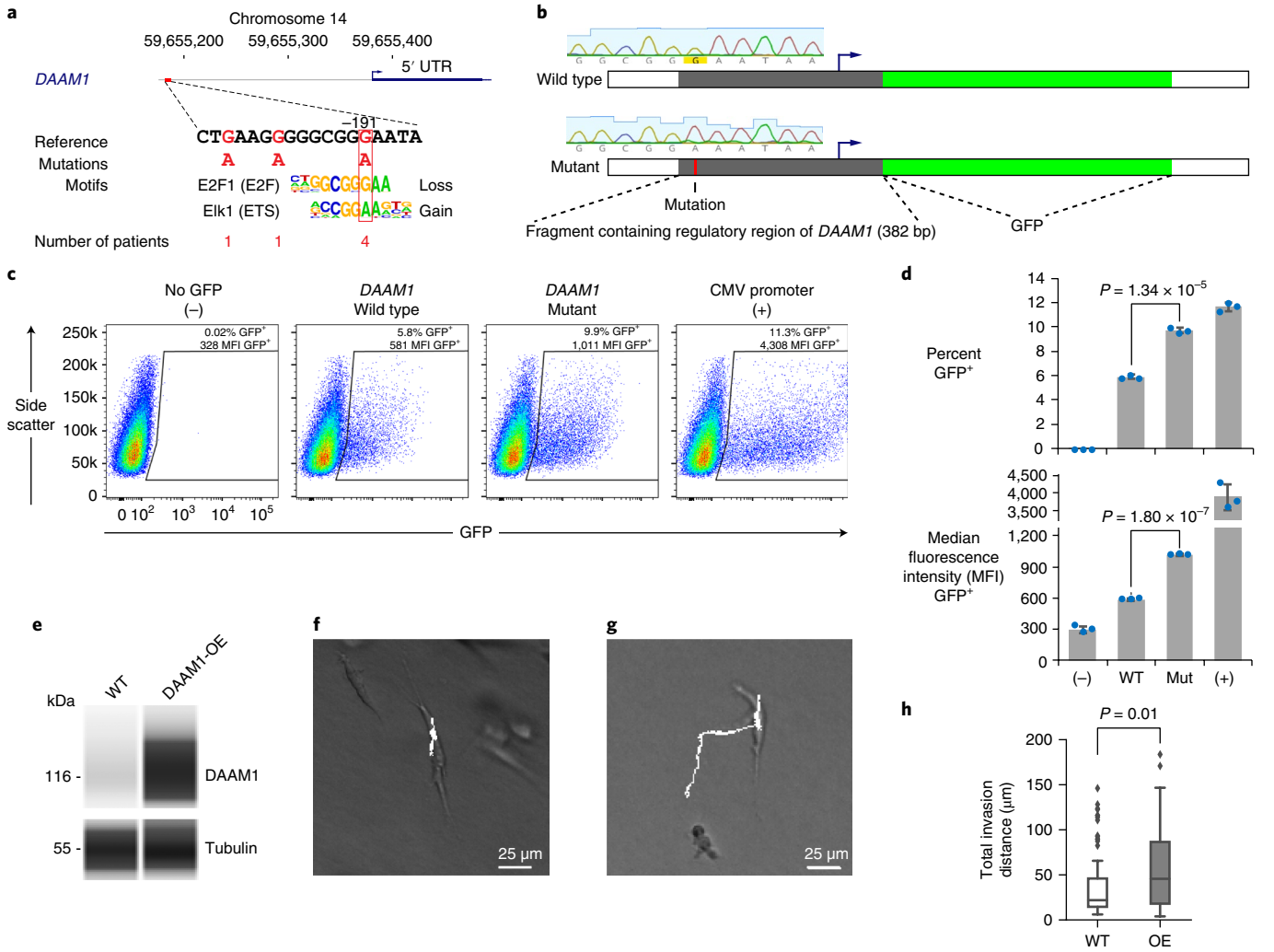

Fig. 3 | Functional validation of the mutated *DAAM1* regulatory element. a, A somatic eQTL in the *DAAM1* promoter region is associated with increased mRNA expression levels. **b**, Schematic of wild-type and mutant GFP reporter constructs along with Sanger sequencing traces confirming the sequence of the key nucleotide. **c**, Flow cytometry analysis of A375 human melanoma cells 48 h after transient transfection. The polygon delineated by black lines shows the gated region used to define GFP⁺ cells. **d**, Bar graphs (average ± s.d. across three cell culture replicates; *P* values from two-tailed *t* tests) showing the percentage of GFP⁺ cells and the median fluorescence intensity of the GFP⁺ cells. Individual data points are in Supplementary Table 5. **e**, Protein electropherogram analysis of wild-type and *DAAM1*-overexpressing MDA-MB-231 cells using antibodies against DAAM1 and tubulin. The complete electropherogram is in Supplementary Fig. 6e. The image is representative of two independent cell culture experiments. **f**,**g**, Sample trajectories of wild-type (**f**) and *DAAM1*-overexpressing (**g**) cells embedded in 2.5 mg/ml 3D collagen hydrogels. **h**, Total invasion distance traveled by individual cells (the *P* value is from a two-tailed Mann–Whitney *U* test; the 95% confidence intervals of the mean were (32.3 μm, 48.2 μm) and (47.6 μm, 67.0 μm) for wild-type and *DAAM1*-overexpressing cells, respectively). Imaging and quantification was performed on 74 and 83 cells in the wild-type and *DAAM1* overexpression groups, respectively. Box plot elements are defined as in Fig. 1b.

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-1β binding motif by the G-to-A mutation 19 bp downstream of the TSS (Fig. 4a; NC_000020.10:g.60758100G>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>AA). 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).

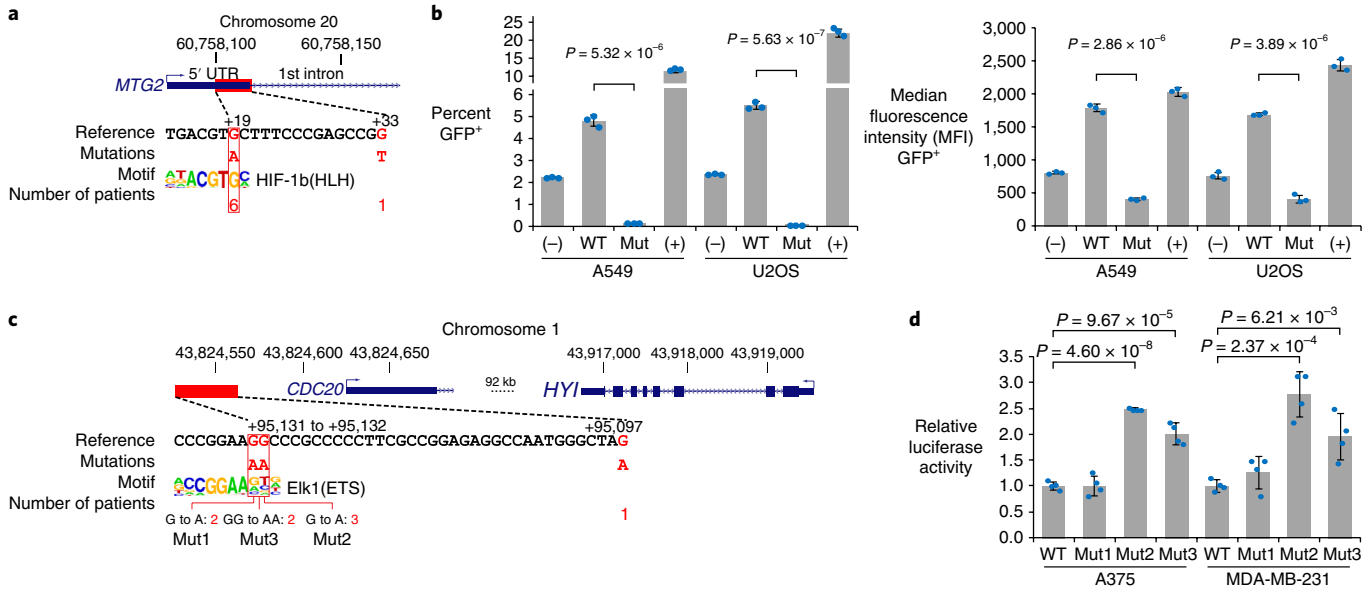

Fig. 4 | Additional case studies. a, The somatic eQTL associated with downregulation of *MTG2* is located in its 5′ UTR (19 bp downstream of the TSS) and frequently alters a potential HIF-1β binding motif. **b**, Flow cytometry analysis of A549 lung epithelial carcinoma cells and U2OS bone osteosarcoma cells 48 h after transient transfection with *MTG2* GFP reporter constructs. Bar graphs (mean ± s.d. across three cell culture replicates; *P* values from two-tailed *t* tests) showing the percentage of GFP+ cells and the median fluorescence intensity of GFP+ events. **c**, The somatic eQTL associated with upregulation of *HYI* is located 95 kb downstream of the TSS and frequently alters a potential Ets family binding motif. **d**, Luciferase assay results (mean ± s.d. across four cell culture replicates; *P* values from two-tailed *t* tests) for the *HYI* somatic eQTLs 48 h after transient transfection in A375 melanoma cells and MDA-MB-231 breast cancer cells. Individual data points are available in Supplementary Tables 5 and 6.

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[21]. This combined set of genes was analyzed by Network-Based Stratification (NBS)[29,30] (Fig. 5a), 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 ReactomeFI[31], documenting 229,300 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 ($P = 3.3 \times 10^{-6}$, 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 $p14^{ARF}$, which can form a complex with HIF-1α and inhibit HIF-1-mediated transcription of *TERT*[32,33]. 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[34], which can upregulate telomerase activity by forming a complex with hTERT[35]. The synergy between *BRAF* and *TERT* mutations has been previously noted and attributed to modulation of *TERT* transcription through BRAF–RAS–ERK signaling[36]. This pathway was also linked to *DAAM1* promoter mutations (Fig. 5d), validated previously, as DAAM1 forms a complex with Dishevelled (DVL3)[22,23], which indirectly regulates transcription of *CDKN2A* and *EGFR* through inhibition of Notch1[37]. 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 $p27^{KIP1}$ (encoded by *CDKN1B*)[38], which inhibits the activity of hTERT[39]. Amplification of *SKP2* in this pathway may thus increase the activity of hTERT.

A third subtype, '*PIK3CA–PEX26–GATA3*' (Fig. 5g,h), integrated coding alterations activating *PIK3CA* and inactivating *GATA3* with noncoding alterations downregulating *PEX26*. In this pathway, members of the peroxisomal biogenesis factor family (PEX26 and PEX6) appear to indirectly interact with PIK3CA and GATA3 through the binding of SMAD family members (SMAD3 and SMAD7)[40].

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[41]. Although *ARID1A* and *CTNNB1* are also known cancer drivers, the connections to *APOBEC* are unanticipated and create a compelling opportunity for further study.

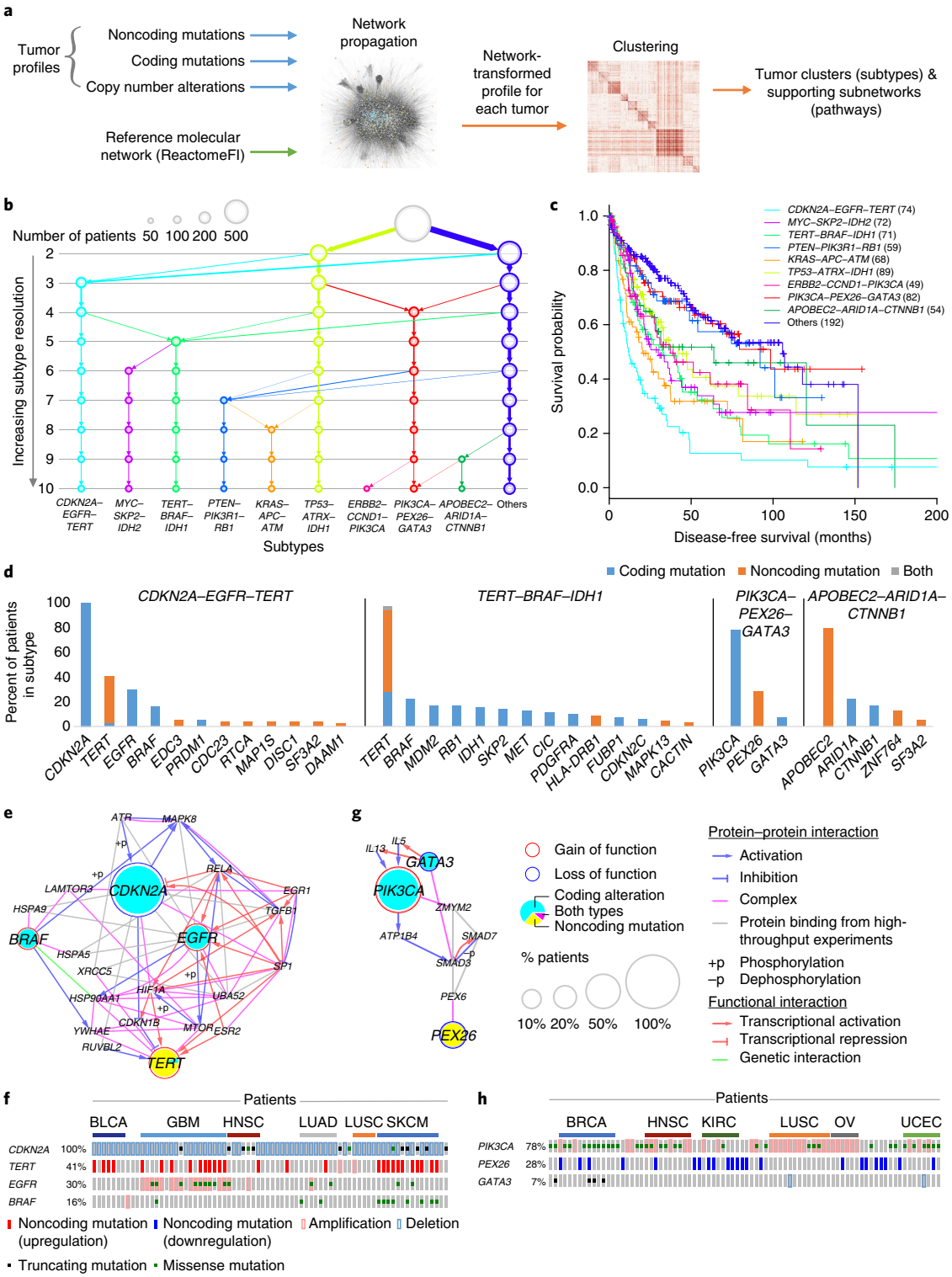

Fig. 5 | Identification of molecular networks and associated tumor subtypes incorporating noncoding mutations. a, Workflow of NBS. **b**, Resulting hierarchy of subtypes, at increasing resolution from 2–10 subtypes. **c**, Disease-free survival probabilities (*y* axis) are plotted against time after diagnosis in months (*x* axis) for each of the identified cancer subtypes (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 (*x* axis), ordered by the percentage of patients with alterations (*y* axis). **e**,**g**, Pathways characterizing *CDKN2A–EGFR–TERT* (**e**) 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[6–8], 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); *MTG2*, which encodes a GTPase that regulates mitochondrial ribosomes[42] (Fig. 4a,b); and *HYI*, which encodes a putative hydroxypyruvate isomerase and may be involved in carbohydrate transport and metabolism[43] (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[44, 45]. 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 diversity has been a major confounder in the analysis of germline variants[44, 45]. 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.

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.

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. Hofree, M. et al. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* **7**, 12096 (2016).
4. Iorio, 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., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (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 and gene expression 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. Hoffman, 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 (2013).
17. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database **2017**, https://doi.org/10.1093/database/bax028 (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. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
21. Vogelstein, 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 RhoA regulates Wnt5a-induced breast cancer cell migration. *PLoS One* **7**, e37823 (2012).
25. Fraley, S. I. et al. A distinctive role for focal adhesion proteins in three-dimensional cell motility. *Nat. Cell Biol.* **12**, 598–604 (2010).
26. Fraley, 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).

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., Ihjima, K. & Yamada, O. IL-2 increases human telomerase reverse transcriptase activity transcriptionally and posttranslationally 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., Ohniwa, R. L., Wada, C., Yoshimura, S. H. & Takeyasu, K. Human small G proteins, ObgH1, and ObgH2, participate in the maintenance of mitochondria and nucleolar architectures. *Genes Cells* **11**, 1295–1304 (2006).
43. Ashiuchi, M. & Misono, H. Biochemical evidence that Escherichia coli hyi (orfb0508, gip) gene encodes hydroxypyruvate 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). S.I.F. 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 Ideaya 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[46] until no such locus was located within d bp from any other. Loci with mutations in $k < 5$ 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 $\log_2$ transformed and z score standardized for subsequent analyses. Second, for CNAs, 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 have positive values and genes with deletions have negative values. A gene is assigned the highest amplification or 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 probes were mapped to the promoter regions of genes (±1 kb from the TSS), and each gene was assigned the mean methylation (beta) values of these probes. 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 (±1 kb from its TSS) or a gene enhancer region defined by GeneHancer[17]. In the case where an enhancer was shorter than a locus, the mapping was performed when 50% of the enhancer sequence overlapped with the locus.

Somatic eQTL analysis using multivariate linear regression. For each gene target linked to recurrently mutated loci, we fit a regression model of the normalized gene expression level e as a function of l, the alteration status of its recurrently mutated loci (1, mutated; 0, wild type), controlling for the impact of CNA status c (0, wild type; positive value, amplification; negative value, deletion), DNA methylation m (mean beta value), 21 tumor tissues t (binary variables), 3 ancestries r (binary variables: Asian; black or African American; white), gender g (1, female; 0, male) and 20 hidden factors h (real values) as covariates

$$e = \beta_0 + \beta_1 l + \beta_2 c + \beta_3 m + \beta_4 t + \beta_5 r + \beta_6 g + \beta_7 h \quad (1)$$

The hidden factors h were identified using probabilistic estimation of expression residuals (PEER)[48,49], while accounting for the effect of known covariates t, r and g. The number of hidden factors was determined by the posterior variance of the factor weights, as previously recommended[49]. The parameters β were estimated from data from 783 tumors with matched RNA-seq and WGS data. Somatic eQTLs were identified as follows. First, for each gene, we selected features by adding an L1-norm to the objective function based on the least-squared error between the true and predicted gene expression levels.

$$(e - \hat{e})^2 + \lambda \|\beta\|_1 \quad (2)$$

The sparsity parameter λ was optimized by cross-validation. For genes in which the L1-norm resulted in $\beta_1 = 0$ for all loci, we decreased λ to include at least one locus. Second, to assess whether the mutation status of any locus contributed significantly to gene expression, the accuracy of the complete model was compared to that of a simple model under the null hypothesis of no genetic associations (i.e., $\beta_1 = 0$ for all loci). The F-test P value between the two nested models was used as the test statistic. Third, having derived an F-test P value for each gene, q values were calculated using the Storey approach[50] with a threshold of FDR ≤ 20%. And finally, for each gene that passed the selection, this threshold was mapped back to the equivalent F-test P value of each locus. Loci with F-test P values below or equal to this threshold were included in the final list and defined as somatic eQTLs.

We elected to perform one test per gene for three reasons. First, in GWAS and typical (germline) eQTL studies, linkage disequilibrium complicates the simultaneous testing of multiple SNPs in a single model because these SNPs are usually codependent. Unlike inherited SNPs, somatic mutations observed in a tumor population are not in linkage disequilibrium no matter how closely they are located. Therefore, a simple F test, which assumes independent influences of multiple factors, is sufficient to simultaneously test whether any loci are associated with gene expression. Second, for each gene, all eQTLs share the same set of covariates along with the associated phenotype of mRNA expression level. If multiple eQTLs are associated with gene expression levels, they can be covariates of one another. It is then convenient to fit them all in a single model and enjoy the benefit of gene-based approaches such as feature selection by L1 regularization. Third, there is precedent in the literature to fit gene-level models in eQTL studies[51,52].

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 $W = \frac{\text{Coefficient}}{\text{Residual s.d.}}$; one unit of W represents 1 s.d. of change in residual gene expression). Power was defined as 1 − P(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 ($f^2 = \frac{R^2_{\text{alternative}} - R^2_{\text{null}}}{1 - R^2_{\text{alternative}}}$). Our somatic eQTL analysis has 50% power to detect a somatic eQTL with five mutations if $W > 1.2$ or with ten mutations if $W > 0.9$.

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[2]. 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[53]. First, a large pool of 20,000 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[17]) while retaining the eQTL's 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[54]. To create a single replication timing feature, we used the average wavelet-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[6]. For gene expression, the median expression value of the nearest gene ($\log_2$-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 sequences 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

$$P(K \geq k) = \sum_{l=k}^{n} \sum_{A \in F_l} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j) \quad (3)$$

where F_l is the set of all subsets of k integers that can be selected from {1, 2, …, n}, p_i or p_j 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,…, n} and A^c is the complement of A. In practice, we used an approximation for the Poisson binomial in the poibin Python package (see URLs).

Transcription factor binding motif analysis. Each reference and somatically altered nucleotide site, along with ±7 bp of flanking sequence, was analyzed using HOMER[55] (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 log/ -fastaBg seqList_mappable_ref.fa -p 16 -find ~/soft/homer/data/knownTFs/vertebrates/known.motifs

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)[11];
3. The affected target gene has high endogenous mRNA expression levels in cell lines[56] 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 gene expression levels, the wild-type and mutant regulatory regions, from −233 bp to +148 bp relative to the TSS, including 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, including 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 FACSCanto (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 an empty lentiGuide-Puro plasmid (Addgene) for the *DAAM1* experiments (Fig. 3c,d and Supplementary Fig. 4) or a promoterless GFP plasmid (pRMT-tGFP, 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 2.01.14 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 WW-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 5.5-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)[25]. Sodium hydroxide was used to normalize pH (pH 7.0, 10–20 µl 1 M NaOH), and the mixture was placed in 48-well culture plates for polymerization at 37 °C. Final gel volumes were approximately 200 µl with the final collagen concentration set to 2.5 mg/ml. The polymerized cell-laden hydrogels were incubated for 24 h under a standard cell culture environment before imaging. Gels were then transferred to a microscope stage-top incubator, and cells were imaged at low magnification (10×) every 2 min for 48 h. The coordinates of cell location in each time frame were determined using image recognition software (Metamorph/Metavue, Molecular Devices). Tracking data were processed to calculate cell speed using an extension of previously published scripts[57]. 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 I gels were seeded in three independent experiments and harvested after 24 h of culture for RNA extraction and directly homogenized in TRIzol reagent (Thermo Fisher). Total RNA was purified using the High Pure RNA Isolation kit (Roche), and the integrity of the sample was verified using RNA Analysis ScreenTape (Agilent Technologies). Total RNA samples were sequenced using the TruSeq 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 Bowtie2[58] and streamed to eXpress[59] 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 suppressors[21] 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, *EGFR*) were considered altered (activated) if impacted by a missense mutation, in-frame indel or copy number amplification. For oncogenes typically altered only by amplification[21] (*CCND1*, *MDM2*, *MDM4*, *MYC*, *MYCL*, *MYCN*, *NCOA3* and *SKP2*), 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 $d \in \{+1, -1\}$ as the sign of the coefficient (β_1 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 patients[60,61].

Network-based stratification to identify tumor subtypes. Network propagation[29] was used to compute the pairwise similarities among tumor genetic alteration profiles (see above) within the Reactome functional interaction network (ReactomeFI)[31]. 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[62] (affinity = *k*-Nearest-Neighbors, assign-labels = discretize, n_clusters = [2…10]). This method first constructs a similarity graph on all pairs of tumors, where each tumor is connected to the *k* others with the shortest Euclidean distance. We chose $k = 170$, which ensures that the similarity graph is connected, as previously recommended[62]. 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 compared to 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, Benjamini–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

genes. The union of the signature and intermediate genes was used to induce a subnetwork within ReactomeFI[31], referenced in the main text as the corresponding 'pathway' impacted in that subtype (Fig. 5d). An additional filter was applied in Fig. 5e and Supplementary Fig. 9a, where we only visualized the signature genes with ten or more mutations and the shortest paths among them with at most one intermediate gene. All networks were visualized in Cytoscape[63].

Survival analysis. We used the coxph package in R statistical software to fit Cox proportional-hazard models[64]. *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 c only

$$Complete\ \ model: \lambda(t \mid s, c) = \lambda_0(t)\exp(\beta_0 + \beta_1 s + \beta_2 c) \quad (4)$$

$$Null\ model: \lambda(t|c) = \lambda_0(t)\exp(\beta_0 + \beta_2 c) \quad (5)$$

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

$$D = -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for complete model}}\right) \quad (6)$$

Finally, a chi-squared test *P* value was calculated on the basis of D with the number of degrees of freedom equal to the number of NBS-derived molecular subtypes.

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

Code availability. Custom codes for annotating mutations, somatic eQTL analysis, validation of recurrence, motif analysis and NBS are available through GitHub (see URLs).

Data availability. The somatic mutations of the 930 tumors are publicly available (see URLs). RNA-seq data are accessible through GEO series accession GSE101209.

References

46. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
47. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
48. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
49. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
51. Michaelson, J. J., Alberts, R., Schughart, K. & Beyer, A. Data-driven assessment of eQTL mapping methods. *BMC Genomics* **11**, 502 (2010).
52. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
53. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
54. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107**, 139–144 (2010).
55. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
56. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
57. Wu, P.-H., Giri, A., Sun, S. X. & Wirtz, D. Three-dimensional cell migration does not follow a random walk. *Proc. Natl Acad. Sci. USA* **111**, 3949–3954 (2014).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).
60. Cao, Y., Bryan, T. M. & Reddel, R. R. Increased copy number of the TERT and TERC telomerase subunit genes in cancer cells. *Cancer Sci.* **99**, 1092–1099 (2008).
61. Xie, H. et al. TERT promoter mutations and gene amplification: promoting TERT expression in Merkel cell carcinoma. *Oncotarget* **5**, 10048–10057 (2014).
62. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
63. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
64. Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982).

nature research

Corresponding author(s): Trey Ideker

☐ 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.

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, U20S, 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