ARTICLES
https://doi.org/10.1038/s41588-019-0572-y
nature genetics

Identification of cancer driver genes based on nucleotide context

Felix Dietlein[1,2,7*], Donate Weghorn[3,4,5,7], Amaro Taylor-Weiner[1,2], André Richters[2,6], Brendan Reardon[1,2], David Liu[1,2], Eric S. Lander[2], Eliezer M. Van Allen[1,2,8*] and Shamil R. Sunyaev[3,4,8*]

Cancer genomes contain large numbers of somatic mutations but few of these mutations drive tumor development. Current approaches either identify driver genes on the basis of mutational recurrence or approximate the functional consequences of nonsynonymous mutations by using bioinformatic scores. Passenger mutations are enriched in characteristic nucleotide contexts, whereas driver mutations occur in functional positions, which are not necessarily surrounded by a particular nucleotide context. We observed that mutations in contexts that deviate from the characteristic contexts around passenger mutations provide a signal in favor of driver genes. We therefore developed a method that combines this feature with the signals traditionally used for driver-gene identification. We applied our method to whole-exome sequencing data from 11,873 tumor–normal pairs and identified 460 driver genes that clustered into 21 cancer-related pathways. Our study provides a resource of driver genes across 28 tumor types with additional driver genes identified according to mutations in unusual nucleotide contexts.

Only a small proportion of the somatic mutations found in tumor cells drive tumor development[1–3], whereas the vast majority are functionally neutral passengers that do not confer selective advantage to cancer cells[4]. A major goal of cancer genomics is to identify these rare driver mutations amid the myriad passengers[5]. A number of highly sophisticated computational methods have been developed to identify driver mutations[6–13]. Applied to thousands of tumor exomes, these methods have contributed greatly to our understanding of which genes are involved in carcinogenesis[5,11,12,14–16].

Current algorithms generally exploit two features of driver mutations: first, they occur in functionally important genomic positions corresponding to amino acids that are critical for the protein function[6–9], and second, they occur in excess over the background mutability of the genome owing to positive selection in the tumor[10–13]. For most positions in the genome, the functional importance is not known[17,18] and is usually proxied by differences between synonymous and nonsynonymous mutations[12], the positional clustering of mutations[7] and bioinformatically predicted scores of functional significance[8,9]. To detect the excess of driver mutations over a carefully modeled background, current methods model the regional variation in the mutation rate with the help of synonymous mutations or epigenomic features[10–13]. Recent approaches further calibrate their background models to the mutability of different nucleotide contexts[9,12,13]. These methods typically aggregate mutation counts over genes or genomic regions and compare them with a context-dependent background expectation[9,12,13]. Current methods further combine different tests to identify driver genes, for example by statistical methods[11] or random forests[19].

Nucleotide contexts around passenger mutations reflect the mutational process active in a given tumor[20–23]. For instance, APOBEC enzymes scan single-stranded DNA for specific nucleotide sequence motifs and deaminate cytidine to uracil within these motifs[24–26]. Similarly, mutant polymerase ε randomly introduces mutations in a non-uniform manner, as its fidelity depends strongly on the local nucleotide context[27–30]. Passenger mutations are thus embedded in nucleotide contexts characteristic of the underlying mutational process[20–23], whereas driver mutations are localized towards functionally relevant positions. To the best of our knowledge, these functionally relevant positions are not surrounded by a particular nucleotide context. This suggests that driver mutations tend to occur more frequently than passengers in 'unusual' nucleotide contexts, deviating from the contexts of the underlying mutational process. Consequently, an excess of mutations in unusual nucleotide contexts gauges the shift of driver mutations from functionally neutral towards functionally important positions.

Nucleotide contexts can therefore inform driver-gene identification in two complementary ways. Some of the recent methods calibrated their background models to the abundance of passengers in highly mutable nucleotide contexts[9,12,13]. Instead, we here examined the other end of the mutability spectrum and assessed the sparsity of passenger mutations in unusual nucleotide contexts. Previous studies have focused on the enrichment of passenger mutations in process-specific nucleotide contexts[20–23], whereas few attempts have been made to quantify the absence of passenger mutations per nucleotide context (that is, scoring its degree of 'unusualness'). However, this is an important endeavor as it helps identify positions in the cancer genome in which passenger mutations are rare, and mutations are thus a strong indicator of the shift of driver mutations towards functionally important positions.

[1]Department of Medical Oncology, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [3]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Centre for Genomic Regulation, Barcelona, Spain. [6]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]These authors contributed equally: Felix Dietlein, Donate Weghorn. [8]These authors jointly supervised this work: Eliezer M. Van Allen, Shamil R. Sunyaev. *e-mail: Felix_Dietlein@dfci.harvard.edu; EliezerM_VanAllen@dfci.harvard.edu; ssunyaev@rics.bwh.harvard.edu

The use of unusual nucleotide contexts does not require previous knowledge of the exact location of the functionally relevant positions. This is essential, as the location of functionally important positions is generally unknown[17,18]. We thus hypothesized that the performance of current methods to detect driver genes could be further improved by using mutations in unusual nucleotide contexts as an indirect proxy of functional importance. We developed a method that searches for genes harboring an excess of mutations in unusual nucleotide contexts and combined this feature with the signals used by existing methods to detect driver genes[6–13]. As such, our method is well suited to identify driver genes in cancer types with both low and high background mutation rates. We demonstrate that our method expands existing catalogs of driver genes in tumor types with high background mutation rates, in which the search for drivers has proven intrinsically challenging[5,11,31,32].

Results

A framework for identifying driver genes in cancer. The main steps of our method are as follows (Fig. 1a). (1) The mutation probability of each genomic position in the human exome is modeled depending on its surrounding nucleotide context[20–23] and the regional background mutation rate[33–35]. (2) Given a gene g with n_g nonsynonymous mutations in positions $\vec{\mathbf{p}}_g$, a Monte Carlo simulation approach[36,37] is used to simulate random 'scenarios' in which n_g or more nonsynonymous mutations are randomly distributed along the same gene g. (3) The number and positions of mutations in each random scenario are compared with the observed mutations in gene g. Based on these comparisons, a P value for gene g is derived (Fig. 1a). (4) This P value is combined with additional statistical components that test for mutational clustering and the abundance of loss-of-function mutations[19], including insertions and deletions.

In steps (2) and (3), we had to evaluate the joint probability of observing n_g nonsynonymous mutations in positions $\vec{\mathbf{p}}_g$ by chance, assuming they were all passengers. This probability can be expressed as a product of the probability of observing n_g nonsynonymous mutations in total and the probability that these mutations occupy specific positions $\vec{\mathbf{p}}_g$ in the gene sequence:

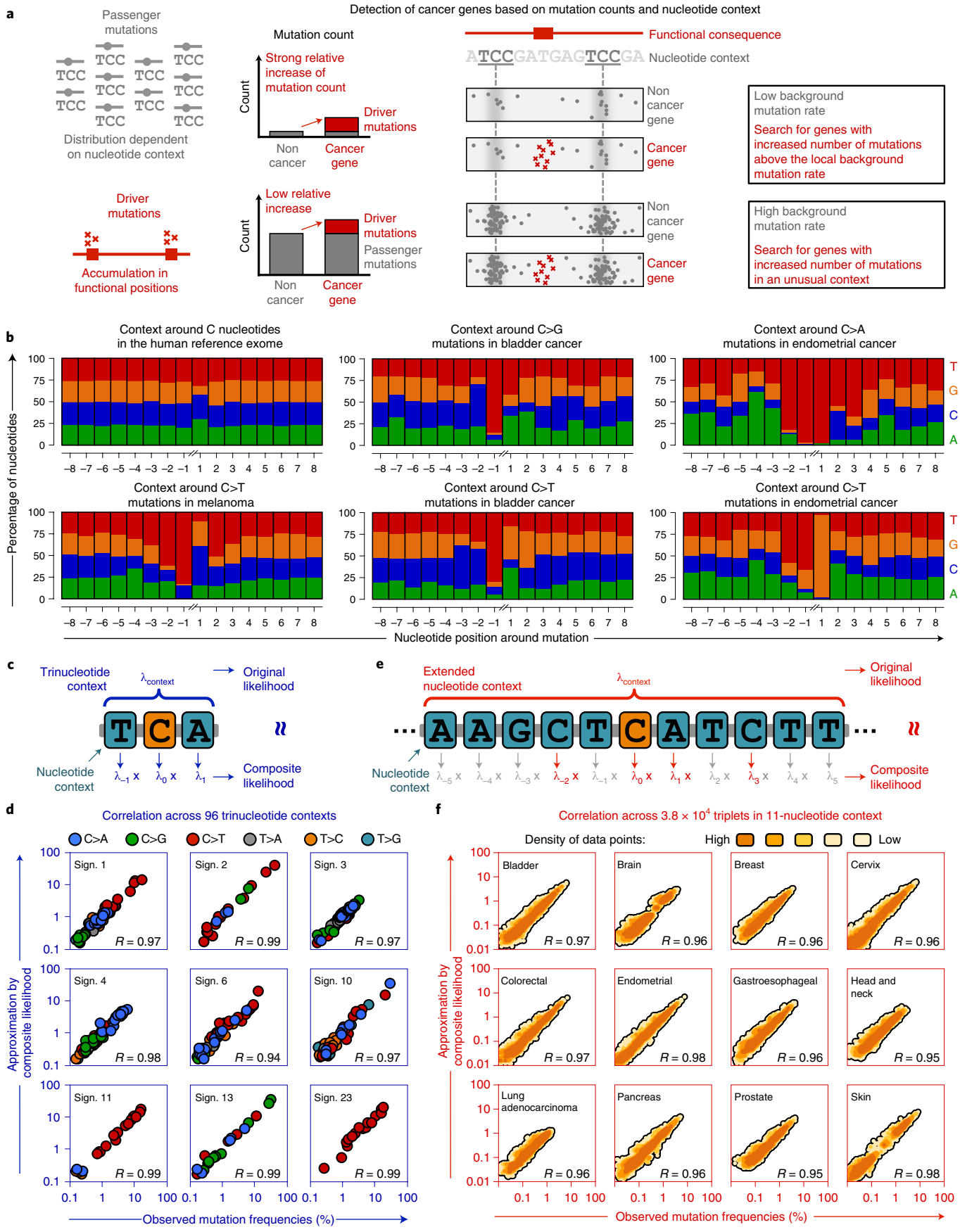
$$P\left(n_g, \vec{\mathbf{p}}_g | s_g; \vec{\boldsymbol{\lambda}}_g\right) = \underbrace{P\left(n_g | s_g\right)}_{\substack{\text{regional} \\ \text{mutation rate}}} \cdot \underbrace{P\left(\vec{\mathbf{p}}_g | n_g; \vec{\boldsymbol{\lambda}}_g\right)}_{\substack{\text{nucleotide} \\ \text{context}}} \quad (1)$$

Here, $P(n_g|s_g)$ is the probability of observing n_g nonsynonymous mutations in gene g, given the number of synonymous mutations s_g. This factor accounts for regional variation in the background mutation rate on the megabase scale[33,34] and is based on a previous study[13]. $P\left(\vec{\mathbf{p}}_g | n_g; \vec{\boldsymbol{\lambda}}_g\right)$ denotes the probability of these n_g nonsynonymous mutations falling in positions $\vec{\mathbf{p}}_g$, conditional on their context-dependent mutability scores $\vec{\boldsymbol{\lambda}}_g$. This factor accounts for context-specific variation in the background mutation rate on the single-base scale[20–23]. We similarly computed the second factor for synonymous mutations and filtered out potential false positives caused by local deviations from the overall context-dependent distribution of passenger mutations[20,22]. Finally, we determined the P value of gene g in step (3) as the fraction of random 'scenarios' that had at least n_g nonsynonymous mutations and had a joint probability lower than that of the observed data (evaluated through equation 1).

Context-dependent mutability of genomic positions. Our method requires the quantification of the mutability of genomic positions depending on their surrounding nucleotide context ($\vec{\boldsymbol{\lambda}}_g$ in the model above). To robustly characterize the mutability signal, our method first performs a Bayesian hierarchical clustering step that groups samples with similar mutational processes together (Supplementary Fig. 1). The 5′ and 3′ nucleotides immediately adjacent to a position have the strongest effect on its mutability[20–23] (Fig. 1b and Supplementary Figs. 2,3). However, as reported previously[38,39], additional upstream and downstream nucleotides flanking a position may also influence its mutability (Fig. 1b and Supplementary Figs. 2,3). The effect of the neighboring nucleotides has traditionally been modeled by determining the mutation probabilities of all possible 96 trinucleotide contexts independently[20–23], thus ignoring the effect of the broader nucleotide context. Here, we employed a composite likelihood model to account for the impact of flanking nucleotides outside the trinucleotide context on local mutation probabilities. In brief, this model returns a mutational likelihood score for each genomic position and incorporates the effect of each flanking nucleotide as a multiplicative factor (Fig. 1c–f, Extended Data Fig. 1, Supplementary Figs. 4,5 and Supplementary Note). In particular, the composite likelihood does not model the mutability of each possible nucleotide context separately (Supplementary Figs. 6 and 7), which is crucial for the use of broad nucleotide contexts in the background model (sparsity of mutation counts per possible nucleotide context, prevention of overfitting of mutational hotspots in the context-dependent background signal). When applied to trinucleotide contexts, this model closely matched the mutation probabilities of the 30 widely used COSMIC mutation signatures[20–23] (Fig. 1c,d and Extended Data Fig. 1a). The composite likelihood model robustly generalized to broader nucleotide contexts for the 28 cancer types examined in this study despite signal sparsity (Fig. 1e,f, Extended Data Figs. 1–4 and Supplementary Figs. 6–11).

Considering the effects of the flanking nucleotides outside the trinucleotide context contributed to the accuracy of the composite likelihood model. For instance, considering heptanucleotide instead of trinucleotide contexts increased the correlation between the observed and predicted mutation probabilities of C>T mutations in melanoma from 0.76 to 0.91, thus refining the approximation of the local mutation probabilities (Extended Data Fig. 1c and Supplementary Fig. 11). Furthermore, we estimated the residual variance between the predicted and observed mutability scores across nucleotide contexts as a function of the number of nucleotides included in the composite likelihood model (Extended Data Fig. 4). Accounting for extended nucleotide contexts beyond

Fig. 1 | Dependency of mutations on extended nucleotide contexts. a, We searched for mutations in unusual nucleotide contexts that deviate from the context around passenger mutations to identify driver genes. We combined this feature with other signals for driver-gene identification. **b**, Frequency at which each nucleotide occurs around recurrent mutations in bladder cancer (middle; $n=317$), endometrial cancer (right; $n=327$) and melanoma (bottom left; $n=582$). **c**,**d**, We applied the composite likelihood model to the mutation frequency vectors of nine COSMIC mutation signatures[20–23]. For each trinucleotide context, we plotted the original frequency against the mutation frequency obtained from the composite likelihood model. Sign., signature. **e**,**f**, We tested whether the composite likelihood model generalized to broader nucleotide contexts in 12 cancer types (bladder, $n=317$; brain, $n=760$; breast, $n=1{,}443$; cervix, $n=192$; colorectal, $n=223$; endometrial, $n=327$; gastroesophageal, $n=833$; head and neck, $n=425$; lung adenocarcinoma, $n=446$; pancreas, $n=729$; prostate, $n=880$ and skin, $n=582$). For any three nucleotides in the 11-nucleotide context, we counted how many mutations were surrounded by the nucleotide triplet ($n=38{,}400$ triplets, not necessarily adherent, ≥1 nucleotide on the 5′ and 3′ sides). We plotted these counts against the prediction of the composite likelihood model. We compared original and modeled mutation frequencies by using Pearson's correlation coefficient (R). Plots for other mutation signatures and cancer types are provided in the Supplementary Information.


a
Detection of cancer genes based on mutation counts and nucleotide context
Passenger mutations
Distribution dependent on nucleotide context
Driver mutations
Accumulation in functional positions
Mutation count
Strong relative increase of mutation count
Driver mutations
Non cancer
Cancer gene
Low relative increase
Passenger mutations
Functional consequence
Nucleotide context
Non cancer gene
Cancer gene
Low background mutation rate
Search for genes with increased number of mutations above the local background mutation rate
High background mutation rate
Search for genes with increased number of mutations in an unusual context
b
Context around C nucleotides in the human reference exome
Context around C>G mutations in bladder cancer
Context around C>A mutations in endometrial cancer
Context around C>T mutations in melanoma
Context around C>T mutations in bladder cancer
Context around C>T mutations in endometrial cancer
Percentage of nucleotides
Nucleotide position around mutation
c
Trinucleotide context
Original likelihood
Nucleotide context
Composite likelihood
e
Extended nucleotide context
Original likelihood
Nucleotide context
Composite likelihood
d
Correlation across 96 trinucleotide contexts
Approximation by composite likelihood
Observed mutation frequencies (%)
f
Correlation across 3.8 × 10^4 triplets in 11-nucleotide context
Density of data points:
Approximation by composite likelihood
Observed mutation frequencies (%)

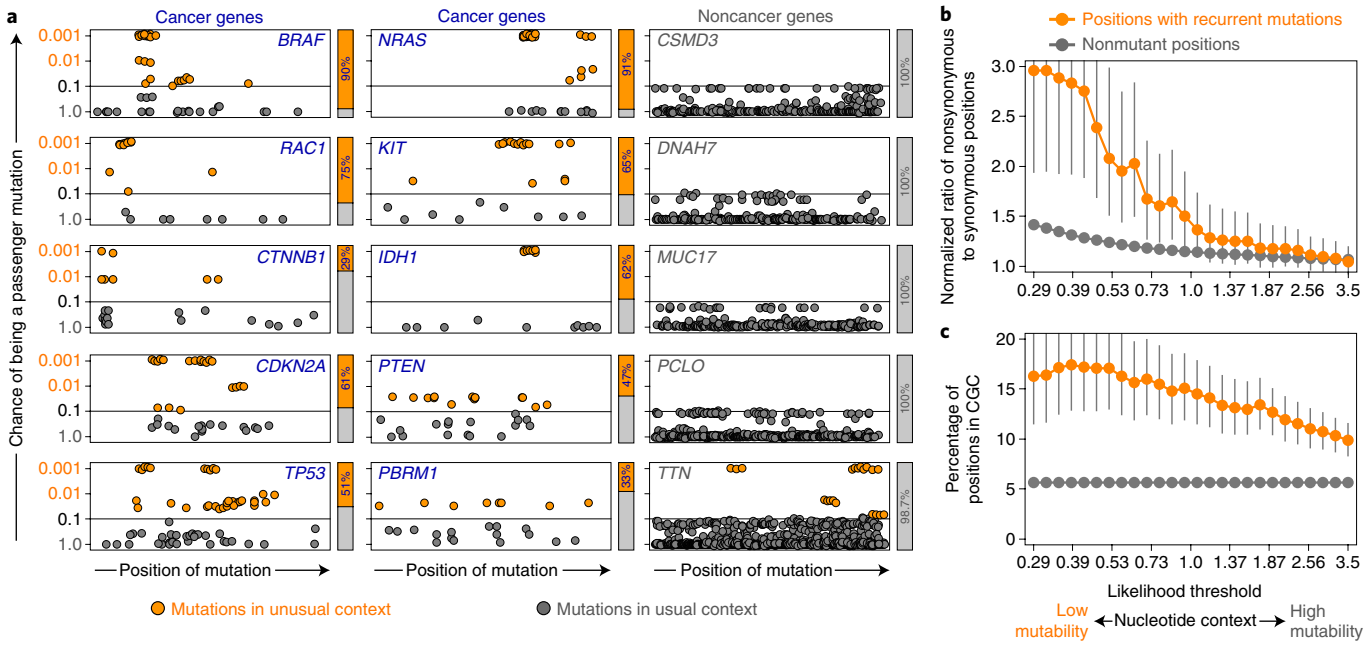


Fig. 2 | Mutations in unusual contexts provide a signal in favor of driver genes. a, On the basis of 582 melanoma samples, we examined the nucleotide contexts around mutations in ten cancer and five noncancer genes. We estimated the mutability of positions by using the composite likelihood. We tested which positions contained more mutations than expected (one-sided test, binomial distribution) and adjusted for multiple testing (FDR). We used an FDR threshold of 0.1 to classify whether the number of mutations per position was usual (gray) or unusual (orange) compared with its surrounding nucleotide context. Each nonsynonymous mutation is visualized as a dot. A small amount of jittering was added to separate mutations at the same position. **b**,**c**, The recurrence of mutations in the same position results from passenger mutations in highly mutable contexts or driver mutations at functionally important sites. On the basis of 582 melanoma samples, we examined whether the nucleotide contexts could distinguish between these two possibilities. We gradually modulated the mutational likelihood cutoff (*x* axis) from weakly mutable to highly mutable nucleotide contexts. We computed the ratio of nonsynonymous to synonymous positions (**b**) and the fraction of positions in established cancer genes listed in the CGC[42,43] (**c**) for each cutoff. The error bars depict the 95% confidence intervals based on the beta distribution and the dots indicate the distribution mean. We determined the same measures for positions without mutations as a negative control. Recurrence is a better indicator of selection for sites with low mutational likelihood than for sites with high mutational likelihood.

the trinucleotide context substantially reduced the residual variance for six tumor types (bladder, breast, cervix, colorectal, endometrium and melanoma; Extended Data Fig. 4). For other tumor types, the residual variance remained largely the same when nucleotides beyond the trinucleotide context were added to the composite likelihood model. Therefore, accounting for extended nucleotide contexts in the composite likelihood model helps with the identification of nucleotide contexts at both ends of the mutability spectrum, which is important to account for the abundance of passenger mutations in usual nucleotide contexts and the relative sparsity of passenger mutations in unusual nucleotide contexts.

Unusual contexts provide a signal for driver mutations. We next tested whether driver mutations occurred more frequently in unusual nucleotide contexts than passenger mutations, which is the biological rationale underlying our method. We first examined the nucleotide contexts around mutations in ten known melanoma genes and five genes unrelated to cancer (previously reported as false positives in cancer gene discovery studies[10]). Most mutations in the genes unrelated to cancer were surrounded by the characteristic nucleotide contexts of passenger mutations, whereas several mutations in the cancer genes occurred in unusual nucleotide contexts (Fig. 2a).

We next analogously analyzed the nucleotide contexts around recurrent mutations (Fig. 2b,c). Recurrent mutations in the same position result from either driver mutations in functionally important sites[40,41] or passenger mutations accumulating in highly mutable contexts[20–23]. We calculated the ratio of nonsynonymous to synonymous positions (Fig. 2b) and the fraction of positions falling into established cancer genes (Fig. 2c; Cancer Gene Census (CGC)[42,43]) to examine whether the nucleotide contexts could help distinguish between these two possibilities. Both measures suggested that positions with recurrent mutations in weakly mutable nucleotide contexts contain higher fractions of driver mutations than positions with recurrent mutations in highly mutable contexts (Fig. 2b,c). In particular, the ratio of nonsynonymous to synonymous positions differed significantly from the baseline expectation for the positions surrounded by unusual nucleotide contexts ($P = 1.47 \times 10^{-4}$ for likelihood < 0.5 based on a beta-binomial distribution). In contrast, the ratios did not differ significantly from the baseline for the usual contexts (Fig. 2b; $P = 0.74$ for likelihood < 3.5). Similarly, the positions with recurrent mutations in unusual nucleotide contexts fell into established cancer genes more frequently compared with the usual contexts (Fig. 2c; 16.7% versus 9.7%, $P = 6.48 \times 10^{-4}$, χ^2 test). Additional analyses are presented in Supplementary Figs. 12–15.

Hence, mutations in unusual nucleotide contexts provide an indirect measure of the shift of driver mutations towards functionally important positions without knowledge of their exact location. They may be particularly useful when the applicability of other proxies of functional excess is limited, owing to a high abundance of functionally neutral nonsynonymous passengers (diluting the statistical power of the difference between nonsynonymous and synonymous mutations[11]) or context-dependent positional clustering of passenger mutations (interfering with the search for driver mutations in mutational hotspots[40]).

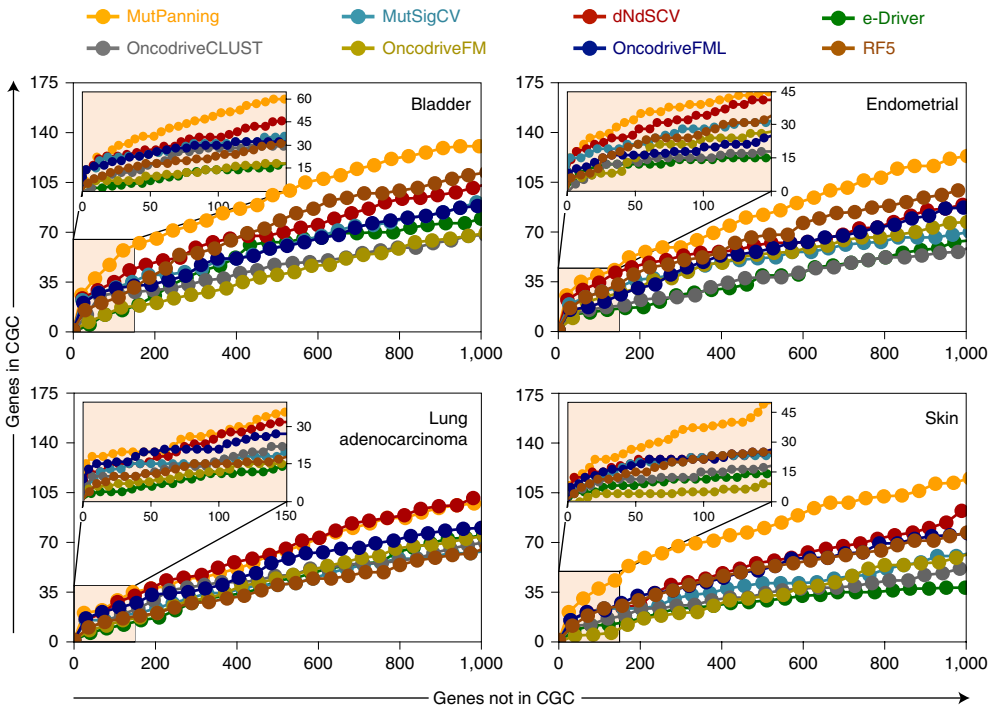


Fig. 3 | Comparison of different methods to identify driver genes. We benchmarked the performance of our method against seven other methods for driver-gene identification. Given that the full set of driver genes per cancer type is unknown, we used the CGC[42,43] for a conservative approximation of the true-positive rate (that is, not every non-CGC gene is necessarily a false positive). On the basis of the top genes returned by each method, we plotted the number of non-CGC genes (*x* axis) against the number of CGC genes (*y* axis) until the list contained 1,000 non-CGC genes. Insets: 150 non-CGC genes. This figure shows this benchmarking analysis for three cancer types with a high context dependency based on the TCGA subcohort (bladder, $n=130$; endometrial, $n=305$ and skin, $n=342$) and one cancer type with a low context dependency based on the TCGA subcohort (lung adenocarcinoma, $n=230$). Similar curves for other cancer types and the full study cohort are provided in Extended Data Figs. 6–9 and the Supplementary Information.

Comparison with other methods for driver-gene detection. We next examined whether the rationale behind our method provided an enhanced ability to identify driver genes. For this purpose, we used whole-exome sequencing data from a collection of 11,873 tumor–normal pairs spanning 28 different tumor types (Extended Data Fig. 5 and Supplementary Table 1). Furthermore, we used two homogeneously processed datasets (The Cancer Genome Atlas (TCGA) and Multi-Center Mutation Calling in Multiple Cancers (MC3); Supplementary Note) to confirm our results. We applied seven current methods for benchmarking, representing major sources for driver-gene detection, including mutational recurrence above a modeled background (MutSigCV[10,11]), difference between synonymous and nonsynonymous mutations (dNdScv[12]), positional clustering into mutational hotspots (OncodriveCLUST[7]), bioinformatically predicted scores of functional impact (e-Driver[6], OncodriveFM[8] and OncodriveFML[9]) and a combination of different sources of mutational significance (RF5 method[19]). We used the CGC[42,43] as a conservative approximation of the true-positive rate (that is, not every non-CGC gene is necessarily a false positive) and plotted a receiver-operating-characteristic curve up to the top 1,000 significant non-CGC genes for each method.

Our method (MutPanning) exhibited the highest performance in two homogeneously processed datasets as well as our study cohort of 11,873 samples (Fig. 3, Extended Data Figs. 6–9 and Supplementary Figs. 16–19). In our study cohort, our method outperformed the seven other methods in 26 of 28 cancer types (Extended Data Figs. 6,7, Supplementary Fig. 16 and Supplementary Table 2), whereas none of the other methods displayed a robust second-best performance across all cancer types (Extended Data Fig. 6). Our method exhibited similarly improved performance relative to other methods when we used the OncoKB[44] instead of the CGC[42,43] database for comparison (Extended Data Figs. 6–8, Supplementary Fig. 17 and Supplementary Table 2). We obtained analogous results when using the precision at 5% recall[45] (Extended Data Figs. 6–8, Supplementary Fig. 18 and Supplementary Table 2) and in additional analyses (Supplementary Figs. 20–23).

We performed two power analyses to examine whether the nucleotide contexts contributed to the performance of our approach (Supplementary Fig. 24). The impact of the nucleotide contexts on the performance of MutPanning was most prominent in cancers with highly context-specific distributions of passenger mutations (Supplementary Fig. 24). In these cancer types, extended nucleotide contexts enhanced the fit of the composite likelihood model (Extended Data Fig. 4). These analyses further suggest that mutational recurrence and unusual nucleotide contexts define complementary signals, both of which are important for the performance of MutPanning (Fig. 3, Extended Data Figs. 4,6–9 and Supplementary Figs. 16–19,24,25). The mutational recurrence was highly informative in cancer types with low background mutation rates, such as thyroid cancer (Fig. 3, Extended Data Figs. 4,6–9 and Supplementary Figs. 16–19,24,25). The nucleotide contexts were the dominant criterion used by our method in cancer types with highly context-specific distributions of passenger mutations, such as melanoma (Fig. 3, Extended Data Figs. 4,6–9 and Supplementary Figs. 16–19,24,25). Two cancer types (lung adenocarcinoma and squamous-cell lung cancer) with high mutation rates and context-independent distributions of passenger mutations may represent a potential challenge for MutPanning and the other methods in our benchmarking panel (Supplementary Figs. 24 and 25).

Stratification of driver genes based on literature support. We combined the findings identified by our method (Figs. 4–6, Extended

Data Fig. 10, Supplementary Figs. 26–64 and Supplementary Tables 3,4) into a driver gene catalog of 460 genes and 827 gene–tumor pairs (pairs of significantly mutated genes and their associated tumor type). The number of gene–tumor pairs varied between tumor types (for example, 42 pairs for cutaneous melanoma versus four pairs for uveal melanoma), depending on the cohort size[11] ($R=0.66$; Fig. 4 and Supplementary Fig. 26a) and the background mutation rate[46] ($R=0.24$; Fig. 4 and Supplementary Fig. 26b). Furthermore, some cancer types exhibited overlaps in driver genes (Supplementary Fig. 27). Most findings could be similarly identified in the MC3 (refs. [5,47]) and TCGA datasets (Supplementary Fig. 28). We compared our results with both the CGC[42,43] and a systematic literature search for experimental or clinical support of our findings (Fig. 5a). On the basis of these comparisons, we stratified our findings into four levels according to supporting evidence in the literature (Fig. 5a): level A includes gene–tumor pairs involving canonical cancer genes in the CGC (523 of 827, 63%); level B contains gene–tumor pairs with experimental literature support in the same tumor type as was identified by our method (106 of 827, 13%); and level C consists of gene–tumor pairs with experimental literature support in a different tumor type (115 of 827, 14%). The fraction of gene–tumor pairs with no literature support (level D) varied in accordance with the false-discovery rate (FDR) thresholds used for cancer gene identification: 4% for FDR < 0.01, 6% for FDR < 0.05, 8% for FDR < 0.1 and 10% for FDR < 0.25.

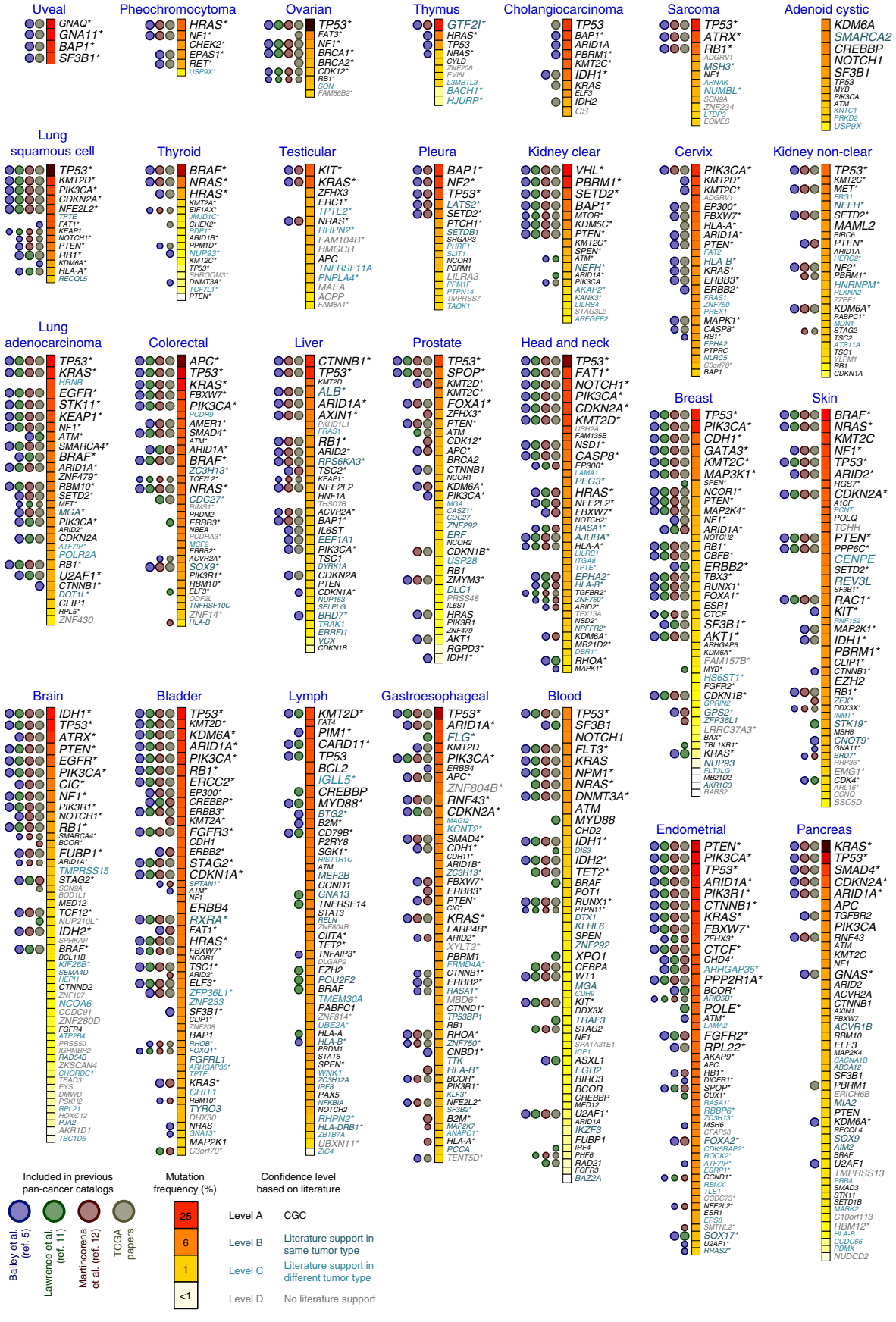
We next examined the overlap between our catalog and results reported as significant in previous pan-cancer studies for driver-gene discovery (Fig. 5b–d and Supplementary Figs. 29–32). Lawrence et al. used the MutSigCV suite to detect driver genes across 4,742 tumors[11]. Martincorena et al. applied the dNdScv algorithm to 7,664 tumors[12]. Most marker papers from TCGA employ MutSigCV[10,11] or MuSiC[48] to discover cancer genes[14–16]. Bailey et al. recently combined 26 different computational tools to search for driver genes in 9,423 tumors[5]. We identified 85% of the CGC gene–tumor pairs reported in at least two of these studies. Hence, our findings are consistent with previously reported results (Fig. 5b,c). Moreover, our catalog contained 169 additional gene–tumor pairs that were part of the CGC but were missing from all previous driver-gene catalogs (Figs. 4, 5b,d and Supplementary Tables 3,4). This number was larger than the corresponding numbers identified in previous studies (Lawrence[11], 25; Martincorena[12], 12; TCGA[14–16], 11 and Bailey[5], 51). Both the robust performance of our method (Fig. 3, Extended Data Figs. 6–9 and Supplementary Figs. 16–19) and the marginally larger size of the sequencing dataset underlying our study (11,837 tumors in this study versus up to 9,423 tumors in previous studies) may have contributed to the larger size of our driver-gene collection. Even after removing all gene–tumor pairs identified in at least two studies, 47%, 50% and 84% of our findings involved canonical cancer genes in the CGC[42,43], the OncoKB genes[44] or had experimental support in the literature, respectively (Fig. 5a). Analogous numbers were 40%, 42% and 82% for genes in our catalog that were not part of any of the other driver-gene catalogs. These rates are considerably higher than those obtained for random gene–tumor pairs (3.8%, 5.3% and 17%, respectively). Moreover, several of the additional driver genes are differentially expressed between mutated and wild-type samples, a pattern that is common for known cancer genes (Supplementary Fig. 33a,b). The additional driver genes in our catalog, which were not included in any of the previous catalogs, were 5.4-fold enriched for this pattern compared with random controls ($P=4.90\times10^{-37}$, χ^2 test). This adds an additional layer of support for their driver candidacy (Supplementary Fig. 33c,d). Furthermore, the protein products of the following additional genes in our catalog have known functional roles in tumor development: *NOTCH2*, *MAML2*, *FGFR4*, *ERRFI1*, *FGFRL1*, *IKZF3*, *ERF*, *ETV6*, *HNF1A*, *CTNND2*, *TCF7L1*, *ANAPC1*, *BTG1*, *CCNQ*, *ROCK2*, *AIM2*, *STAT3*, *BIRC3*, *BIRC6*, *SF3B2*, *ESRP1*, *KLHL6*, *UBE2A*, *UBR5*, *POLR2A*, *REV3L*, *RECQL4*, *RECQL5*, *JMJD1C*, *SMARCA2* and *SMAD3* (see Supplementary Table 5 for literature references and Extended Data Fig. 10, Supplementary Figs. 34,35 and Supplementary Note for a discussion of their functional roles). Although they had been reported individually and in separate publications focusing on a certain cancer subtype or gene, they had not been identified together in a systematic pan-cancer analysis and were missing from all previous pan-cancer studies[5,11,12,14–16]. Our full driver-gene catalog is available as an online resource (www.cancergenes.org).

Clustering of driver genes based on physical interactions. We examined whether the additional driver genes in our catalog revealed insights into tumor signaling when analyzed in combination with established driver genes. Using a large-scale protein–protein interaction dataset[49–52], we studied the physical interactions between the protein products of established (that is, CGC genes) and less well-established driver genes (that is, non-CGC genes) in our catalog. We noticed that several CGC/non-CGC interactions in our catalog had well-defined functional roles in tumor signaling (Fig. 6a). For instance, the protein product of the non-CGC gene *TCF7L1* directly mediates the Wnt signaling activity of CTNNB1 (ref. [53,54]), which is listed in the CGC; the non-CGC gene *ERRFI1* encodes a protein that inhibits the activation of EGFR[55] (listed in the CGC); and the transcriptional activity of POLR2A (not in the CGC) is mediated by MED12, which is part of the transcriptional mediator complex[56,57] and the CGC (Fig. 6a). Thus, the physical interactions between the protein products of CGC and non-CGC genes informed the characterization of less well-established driver genes in our catalog.

Driver genes clustered into 21 pathways on the basis of their physical interactions (Fig. 6a). These 21 pathways include major cancer hallmark pathways[58,59] (for example, MAPK signaling, mTOR–PI3K signaling, cell-cycle regulation, DNA repair and chromatin modification) as well as additional pathways involved in carcinogenesis (for example, RNA binding[60,61], ribosome function[62,63], Rho GTPases[64,65] and immune signaling[66,67]). Some pathways were mutated across most of the 28 cancer types examined (for example,

Fig. 4 | A catalog of driver genes in human cancer. We derived a catalog of driver genes across 28 cancer types based on the whole-exome sequencing data from 11,873 tumor–normal pairs. Extended Data Fig. 5 lists the exact number of samples per cancer type. The *P* values were derived using our approach (MutPanning) and then adjusted for multiple testing. The most significant gene–tumor pairs (FDR < 0.25) for each cancer type are listed in decreasing order of their mutation frequencies (indicated by the color of the square next to the gene name). A maximum of 50 gene–tumor pairs are shown per cancer type. The full catalog can be found in Supplementary Table 3. The font size of the gene name reflects its significance, with highly significant genes in large font and less significant genes in small font. We compared our driver-gene catalog with four catalogs from previous pan-cancer studies. The colored dots indicate which gene–tumor pairs were listed as significant in previous catalogs. The font colors reflect which gene–tumor pairs had been reported in the literature (confidence levels A–D). Heterogeneity in variant calling, tissue collection protocols and mutation reports (synonymous mutations were not reported for 6.1% of the samples; studies marked in Supplementary Table 1) may represent a potential limitation for driver-gene identification. We therefore ran MutPanning on two uniformly processed datasets (TCGA, $n=7{,}060$ samples and MC3, $n=9{,}079$ samples) that did not have these limitations. We marked the gene–tumor pairs that also reached statistical significance in these smaller datasets (TCGA or MC3) with asterisks. The TCGA and MC3 datasets did not include adenoid cystic carcinoma.

apoptosis regulation and chromatin modification), whereas other pathways were more specific to tumor types (for example, G proteins, metabolism, TGFβ signaling and Wnt signaling; Fig. 6b). Moreover, several pathways exhibited either positive (for example, chromatin/apoptosis regulation, Wnt/TGFβ signaling and RTK/MAPK signaling) or negative (for example, PI3K/MAPK



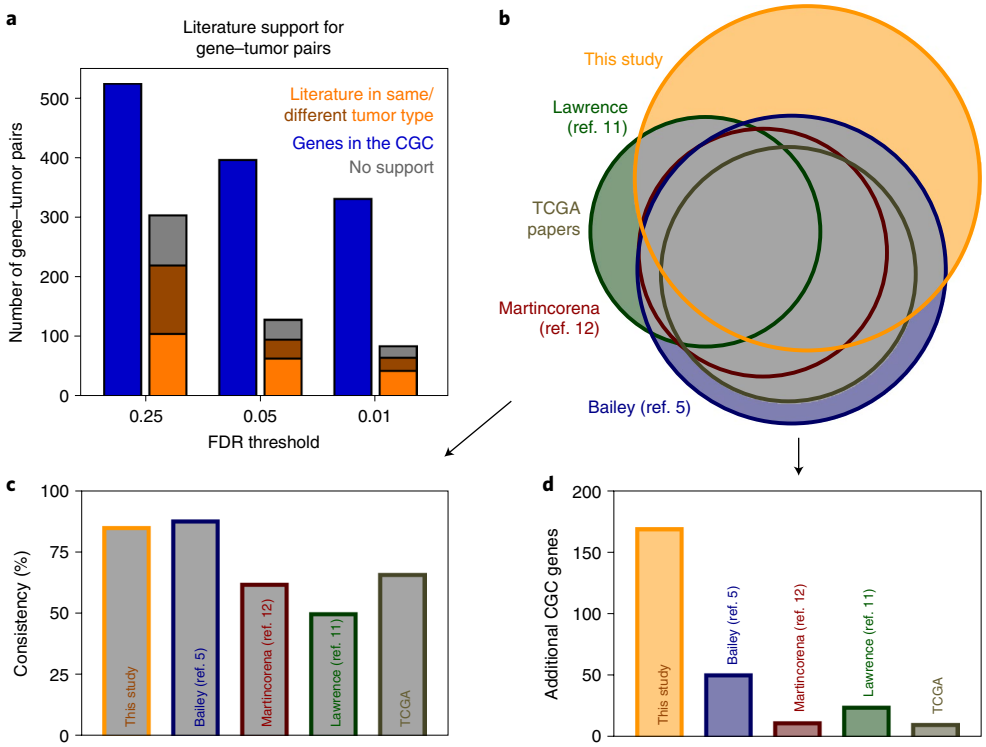


Fig. 5 | Stratification of driver genes based on literature support. a, We stratified 827 gene–tumor pairs (based on 11,873 samples; the significance values were derived using MutPanning and adjusted for multiple testing) on the basis of their literature support. Gene–tumor pairs involving canonical cancer genes in the CGC[42,43] are shown in blue, gene–tumor pairs reported by experimental studies for the same/different tumor type as those identified by our method are shown in orange/brown and gene–tumor pairs with no literature support are shown in gray. **b**, Area-proportional Venn diagrams show the overlap in CGC genes between our catalog and significant results from previous studies. The gray area reflects the CGC gene–tumor pairs that were reported for the same tumor type in at least two independent catalogs. **c**, As a measure of consistency, we counted how many CGC genes from previous studies were also identified by our study (*y* axis, fraction of CGC gene–tumor pairs in at least two independent catalogs). **d**, We counted the number of CGC gene–tumor pairs in our catalog that were not part of previous studies. This measure reflects whether our catalog expanded existing catalogs with additional candidate driver genes. Our catalog (orange) recapitulated 85% of the CGC gene–tumor pairs from at least two previous studies (**c**) and contained 169 additional CGC gene–tumor pairs that were not part of previous pan-cancer catalogs (**d**).

signaling, RTK/Wnt signaling and ubiquitination/transcription factors) associations with one another (Fig. 6b). In eight pathways, more than 60% of the mutational signal was concentrated in ≤2 genes (for example, mTOR–PI3K signaling, apoptosis regulation, Wnt signaling and Notch signaling). In the other 13 pathways, the signal was widely spread across rare driver genes and <60% of the mutations occurred in the two most frequently mutated genes (for example, chromatin modification, DNA repair and immune signaling; Supplementary Fig. 36).

Discussion

We developed a method for driver-gene identification that utilizes mutations in unusual nucleotide contexts in combination with established sources for driver-gene discovery (Fig. 1)[6–13]. Passenger mutations are enriched in characteristic nucleotide contexts, depending on the tumor type and mutational process[20–23], whereas driver mutations are localized towards functionally important positions[40,41,68,69] that do not follow any particular context-specific distribution patterns. As a result, we expect that functionally important mutations occur, on average, more frequently in unusual nucleotide contexts relative to passenger mutations. Hence, a shift in mutations from usual to unusual nucleotide contexts mimics the shift from functionally neutral to functionally important positions (Figs. 1 and 2). Our method compares the nucleotide context around each genomic position in the human exome with the observed number of mutations at that position. Thus, our method weighs each nonsynonymous mutation in the human exome differentially; nonsynonymous mutations in weakly mutable nucleotide contexts have a higher impact on the *P* value of a gene than nonsynonymous mutations in highly mutable nucleotide contexts.

To benchmark our method, we compiled a large-scale whole-exome sequencing dataset of 11,873 samples from TCGA and non-TCGA studies (Extended Data Fig. 5). Although all samples were processed with the same sequencing strategy and a homogeneous variant filter, differences in tissue collection protocols, variant calling pipelines and mutation reports (for example, synonymous mutations were not reported in 6.1% of the samples) may represent a potential source of heterogeneity. Hence, we used two uniformly processed datasets for validation (Extended Data Fig. 9 and Supplementary Fig. 19). Furthermore, although solid tumors in TCGA were largely unaffected by tumor-in-normal contamination[70], tumor-in-normal contamination may have affected variant calling in blood tumors, thereby missing potential driver genes.

Our method enabled us to systematically aggregate large numbers of driver genes that were missing from the catalogs of previous pan-cancer studies (Figs. 4 and 5). For most tumor signaling pathways, mutations are spread across long tails of driver genes[58]. The mutation frequencies of genes at the ends of these tails are below the detection thresholds of current methods used for driver-gene identification[5,11,31,32]. Given that our catalog contained multiple rare driver genes with mutation frequencies as low as 1%, it may represent a valuable resource for aggregating mutations across these tails, thereby enabling driver mutations to be characterized at a pathway level rather than a gene level (Figs. 4–6). Our study further

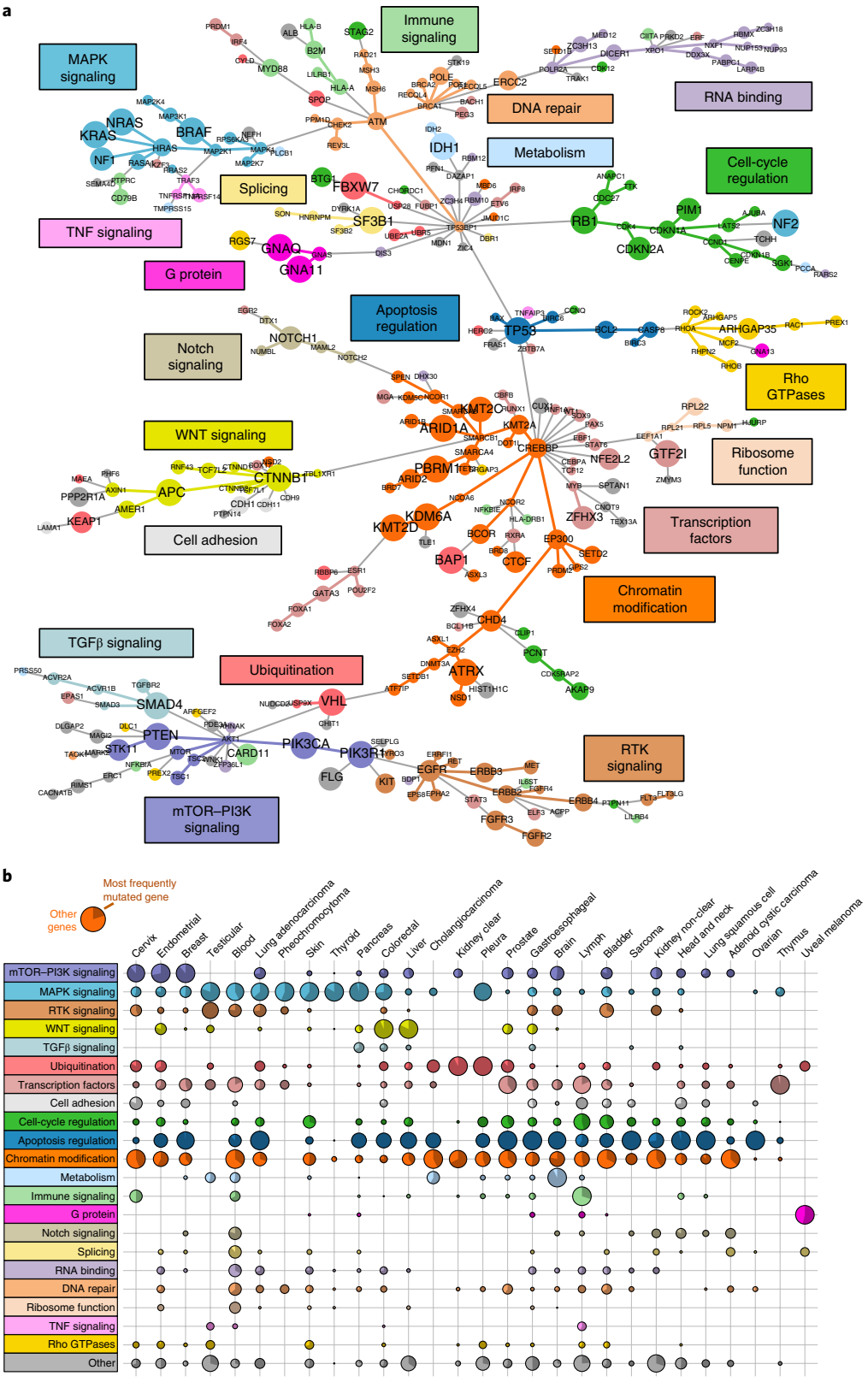


Fig. 6 | Characterization of driver genes based on physical interactions. a, Physical interactions between driver genes (based on 11,873 samples; identified using MutPanning) are visualized as a minimum-spanning tree based on a large-scale protein–protein interaction database[49]. The color of each gene reflects its associated pathway and the dot size indicates its maximum mutation frequency across the 28 cancer types examined in this study. **b**, We aggregated mutations across all driver genes in the same pathway and determined the relative contributions (dot sizes) of different pathways (rows) to the mutational landscape of 28 different cancer types (columns). The contribution of the most frequently mutated gene in each pathway is shown as a dark area in each dot.

demonstrates that the identification of multiple driver genes in the same pathway facilitates the biological interpretation of mutations in less well-established driver genes (Fig. 6). Our catalog may similarly inform the clinical annotation of tumor patients with mutations in less-established driver genes and thereby enhance comparisons of mutation profiles across patients[51,71].

Moving forward, we anticipate that mutations in unusual nucleotide contexts may also be useful in related areas, including the capture of low-frequency mutational hotspots[40,72] and probabilistic annotation of mutations as drivers in the genomes of individual tumor patients[18,73]. Furthermore, our approach may directly inform driver-gene identification in ongoing and future large-scale cancer-genome sequencing efforts such as GENIE[74], MSK-IMPACT[75], PCAWG[76], ICGC[77] and HMF[78]. Our method is available as an interactive software tool called MutPanning (www.cancer-genes.org) and can be run online as a module on the GenePattern platform[79,80] (www.genepattern.org).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0572-y.

Received: 17 May 2019; Accepted: 16 December 2019;
Published online: 3 February 2020

References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
4. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
5. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
6. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
7. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
8. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
9. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
10. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
11. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
12. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
13. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
14. Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
15. The Cancer Genome Atlas Research Network Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
16. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
17. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
18. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
19. Kumar, R. D., Searleman, A. C., Swamidass, S. J., Griffith, O. L. & Bose, R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics* **31**, 3561–3568 (2015).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
21. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
22. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
23. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
24. Ebrahimi, D., Alinejad-Rokny, H. & Davenport, M. P. Insights into the motif preference of APOBEC3 enzymes. *PLoS ONE* **9**, e87679 (2014).
25. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
26. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
27. Church, D. N. et al. DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. *Hum. Mol. Genet.* **22**, 2820–2828 (2013).
28. Shinbrot, E. et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).
29. Goodman, M. F. & Fygenson, K. D. DNA polymerase fidelity: from genetics toward a biochemical understanding. *Genetics* **148**, 1475–1482 (1998).
30. Ganai, R. A. & Johansson, E. DNA replication—a matter of fidelity. *Mol. Cell* **62**, 745–755 (2016).
31. Hofree, M. et al. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* **7**, 12096 (2016).
32. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci. USA* **113**, 14330–14335 (2016).
33. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
34. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
35. Polak, P. et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
36. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical *P* values from Monte Carlo procedures. *Am. J. Hum. Genet.* **71**, 439–441 (2002).
37. Ewens, W. J. On estimating *P* values by the Monte Carlo method. *Am. J. Hum. Genet.* **72**, 496–498 (2003).
38. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.* **11**, e1005657 (2015).
39. Fredriksson, N. J. et al. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS Genet.* **13**, e1006773 (2017).
40. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
41. Chang, M. T. et al. Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* **8**, 174–183 (2018).
42. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–11 (2015).
43. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
44. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00011 (2017).
45. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
46. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl Acad. Sci. USA* **112**, 118–123 (2015).
47. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281 (2018).
48. Dees, N. D. et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
49. Szklarczyk, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).
50. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
51. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
52. Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
53. Murphy, M., Chatterjee, S. S., Jain, S., Katari, M. & DasGupta, R. TCF7L1 modulates colorectal cancer growth by inhibiting expression of the tumor-suppressor gene EPHB3. *Sci. Rep.* **6**, 28299 (2016).

54. Morrison, G., Scognamiglio, R., Trumpp, A. & Smith, A. Convergence of cMyc and β-catenin on Tcf7l1 enables endoderm specification. *EMBO J.* **35**, 356–368 (2016).
55. Cairns, J. et al. Differential roles of ERRFI1 in EGFR and AKT pathway regulation affect cancer proliferation. *EMBO Rep.* **19**, e44767 (2018).
56. Taatjes, D. J. The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends Biochem. Sci.* **35**, 315–322 (2010).
57. Soutourina, J. Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018).
58. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
59. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
60. Pereira, B., Billaud, M. & Almeida, R. RNA-binding proteins in cancer: old players and new actors. *Trends Cancer* **3**, 506–528 (2017).
61. Neelamraju, Y., Gonzalez-Perez, A., Bhat-Nakshatri, P., Nakshatri, H. & Janga, S. C. Mutational landscape of RNA-binding proteins in human cancers. *RNA Biol.* **15**, 115–129 (2018).
62. Pelletier, J., Thomas, G. & Volarevic, S. Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nat. Rev. Cancer* **18**, 51–63 (2018).
63. Sulima, S. O., Hofman, I. J. F., De Keersmaecker, K. & Dinman, J. D. How ribosomes translate cancer. *Cancer Discov.* **7**, 1069–1087 (2017).
64. Wilson, K. F., Erickson, J. W., Antonyak, M. A. & Cerione, R. A. Rho GTPases and their roles in cancer metabolism. *Trends Mol. Med.* **19**, 74–82 (2013).
65. Porter, A. P., Papaioannou, A. & Malliri, A. Deregulation of Rho GTPases in cancer. *Small GTPases* **7**, 123–138 (2016).
66. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 (2018).
67. Disis, M. L. Immune regulation of cancer. *J. Clin. Oncol.* **28**, 4531–4538 (2010).
68. Chakravorty, D. et al. MYCbase: a database of functional sites and biochemical properties of Myc in both normal and cancer cells. *BMC Bioinform.* **18**, 224 (2017).
69. Izarzugaza, J. M., Redfern, O. C., Orengo, C. A. & Valencia, A. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. *Proteins* **77**, 892–903 (2009).
70. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
71. Creixell, P. et al. Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
72. Hess, J. M. et al. Passenger hotspot mutations in cancer. *Cancer Cell* **36**, 288–301 (2019).
73. Carter, H. et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
74. AACR Project GENIE Consortium. AACR project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818–831 (2017).
75. Cheng, D. T. et al. Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med. Genomics* **10**, 33 (2017).
76. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. Preprint at *bioRxiv* https://doi.org/10.1101/237313 (2017).
77. Zhang, J. et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
78. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
79. Reich, M. et al. GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
80. Reich, M. et al. The genepattern notebook environment. *Cell Syst.* **5**, 149–151 (2017).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

Sequencing-data curation and variant filtering. We compiled whole-exome sequencing data from 32 TCGA-related projects (7,091 samples) as well as from 55 TCGA-independent publications (4,856 samples). Mutation annotation files (MAFs) for TCGA-related projects were directly obtained from the TCGA Gene Data Analysis Center data portal hosted by the Broad Institute (gdac.broadinstitute.org; latest data version from 28 January 2016, doi:10.7908/C11G0KM9). The MAFs for the TCGA-independent studies were either downloaded from the cBioPortal platform (www.cbioportal.org)[81,82] or—if not available there—directly from the supplements of the publications. Details on how we selected these studies and samples can be found in the Supplementary Note.

We integrated all mutations into a combined MAF and removed duplicate patients from the combined MAF. We grouped patients into subcohorts according to their cancer type. Most of these tumor types were defined as in the TCGA marker papers (27 of 28 tumor types).

Finally, mutations from this combined MAF were processed through a homogeneous filtering step to minimize sequencing artifacts, mutation calling errors and germline variants that might have slipped through the variant filters applied in each study. We applied the following filters.

Filtering of common germline variants. Each mutation was compared against the Exome Aggregation Consortium database[83], which reports the germline variants of 60,706 individuals. Similarly to a previous study[74], we removed all variants from the MAF that occurred more than ten times in any of the seven Exome Aggregation Consortium subpopulations.

Removal of 8-oxoguanine and strand-bias sequencing artifacts. Oxoguanine artifacts result from excessive oxidation during sequence-library preparation[84], whereas strand-bias artifacts produce disparities between G>T and C>A mutation counts at low variant allele frequencies[47]. We used the annotation of the MC3 dataset[47] to reduce the number of oxoguanine and strand-bias artifacts in our MAF.

Removal of low-quality samples. Samples for which >10% of the somatic mutations were flagged as artifacts or germline variants were removed from the study.

In this way, we arrived at a study cohort of 11,873 tumor samples spanning 28 different cancer types.

Statistical analyses to identify driver genes. The first step of MutPanning is to cluster samples with similar passenger-mutation distributions together and to characterize the context-dependent background signal in each cluster. In brief, we first counted the number of mutations of each base substitution type t (C>A, C>G, C>T, T>A, T>C and T>G) for each sample and summarized these counts into a type count vector $\mathbf{v}^{\text{type}} \in \mathbb{N}^6$. Each element $\mathbf{v}_t^{\text{type}}$ in this vector corresponds to the number of base substitutions of type $t \in \{1, \ldots, 6\}$. We further counted the nucleotides that occurred in a 20-nucleotide window around the mutations identified for each sample to capture the extended nucleotide context around mutations. We summarized these counts into the nucleotide context count vector $\mathbf{v}^{\text{seq}} \in \mathbb{N}^{6\times 20\times 4}$. Each element $\mathbf{v}_{t,p,n}^{\text{seq}}$ in this vector denotes the count of nucleotide $n \in \{\text{A, C, G, T}\}$ in position $p \in [-10; 10]\backslash\{0\}$ around mutations of type $t \in \{1, \ldots, 6\}$. MutPanning then quantified the similarity between two count vectors $\mathbf{v}, \mathbf{w} \in \mathbb{N}^l$ by examining whether updating a distribution prior $\mathbf{x}$ by $\mathbf{w}$ made the observation of $\mathbf{v}$ more likely (Dirichlet multinomial distribution). More details on the choice of the distribution prior $\mathbf{x}$ as well as the metrics to compare count vectors are provided in the Supplementary Note.

In the second step, MutPanning establishes a composite likelihood model for each cluster C of samples. In brief, MutPanning derives likelihood ratios for each cluster C as

$$\lambda_{t,p,n}^{\mathcal{C}} := \frac{\mathbf{v}_{t,p,n}^{\text{seq}}}{\mathbf{v}_t^{\text{type}} \cdot f_{n(t),p,n}^{\text{ref}}}$$

where $f_{n,p,n'}^{\text{ref}}$ denotes the frequency of nucleotide n' around nucleotide n at position p in the human exome, and $n(t)$ denotes the reference nucleotide of base substitution type t (that is, C for types $t=1$, 2 and 3 and T for types $t=4$, 5 and 6). Hence, $\lambda_{t,p,n}^{\mathcal{C}}$ reflects the ratio of the observed number of mutations ($\mathbf{v}_{t,p,n}^{\text{seq}}$) and the expected number of mutations ($\mathbf{v}_t^{\text{type}} \cdot f_{n(t),p,n}^{\text{ref}}$) if all mutations were equally distributed across the human exome.

Similarly, given a substitution type $t \in \{1, \ldots, 6\}$ we define the likelihood ratio as

$$\lambda_t^{\mathcal{C}} := \frac{\mathbf{v}_t^{\text{type}}}{|\mathbf{v}^{\text{type}}|/6}$$

with $|v| := \sum_k |v_k|$. Hence, $\lambda_t^{\mathcal{C}}$ reflects the ratio of the observed number of mutations ($\mathbf{v}_t^{\text{type}}$) of substitution type t and the expected number of mutations ($|\mathbf{v}^{\text{type}}|/6$) if all substitution types occurred at the same frequency.

Given a base substitution type t and a genomic position that is surrounded by nucleotides n_p at position p, we define its composite likelihood as

$$\lambda_{\text{pos}} := \lambda_t^{\mathcal{C}} \cdot \prod_{\substack{-10\le p\le 10\\ p\neq 0}} \lambda_{t,p,n_p}^{\mathcal{C}}$$

for reference nucleotides $n_0 = \text{C}$ and T, and

$$\lambda_{\text{pos}} := \lambda_t^{\mathcal{C}} \cdot \prod_{\substack{-10\le p\le 10\\ p\neq 0}} \lambda_{t,-p,\overline{n_p}}^{\mathcal{C}}$$

for reference nucleotides $n_0 = \text{A}$ and G; $\overline{n_p}$ denotes the complementary nucleotide to n_p.

This likelihood score indicates whether the position is expected to contain more ($\lambda_{\text{pos}} > 1$) or fewer mutations ($\lambda_{\text{pos}} < 1$) compared with a flat mutation distribution. That way, highly mutable nucleotide contexts ($\lambda_{\text{pos}} \gg 1$) and mutations in highly unusual nucleotide contexts ($\lambda_{\text{pos}} \ll 1$) can be identified and weighted differently in the statistical model. More details on the full composite likelihood model can be found in the Supplementary Note.

In the third step, MutPanning examines how likely the number and positions of its nonsynonymous mutations might occur by chance for each gene. Three different base substitutions are possible for each reference nucleotide. Hence, given a gene of length l_g, we defined a count vector $\mathbf{v}^g \in \mathbb{N}^{l_g\times 3}$ that contains the number of mutations at each position and for each substitution type. Similarly, we defined the vector $\boldsymbol{\lambda}^g$ that contains the composite likelihood for each position and substitution type in gene g. We then split these vectors into $\mathbf{v}^g = (\mathbf{v}^{g,\text{s}}, \mathbf{v}^{g,\text{n}})$ and $\boldsymbol{\lambda}^g = (\boldsymbol{\lambda}^{g,\text{s}}, \boldsymbol{\lambda}^{g,\text{n}})$, reflecting synonymous and nonsynonymous positions, respectively.

MutPanning then determines the probability of observing $\mathbf{v}^{g,\text{n}}$ by chance, given the number of synonymous mutations $|\mathbf{v}^{g,\text{s}}|$ and the context-dependent composite likelihood scores $\boldsymbol{\lambda}^{g,\text{n}}$ in the same gene. This probability factorizes into two factors

$$P(|\mathbf{v}^{g,\text{n}}| \,|\, |\mathbf{v}^{g,\text{s}}|) \cdot P(\mathbf{v}^{g,\text{n}} \mid |\mathbf{v}^{g,\text{n}}|; \boldsymbol{\lambda}^{g,\text{n}})$$

The first factor ($P(|\mathbf{v}^{g,\text{n}}| \,|\, |\mathbf{v}^{g,\text{s}}|)$) reflects the chance of observing $|\mathbf{v}^{g,\text{n}}|$ nonsynonymous mutations in a gene with $|\mathbf{v}^{g,\text{s}}|$ synonymous mutations. This factor is modeled by a convoluted Poisson distribution, that is $|\mathbf{v}^{g,\text{n}}| \sim \text{Pois}(\mu)$, where the mutation rate μ is drawn from another distribution, conditional on the number of synonymous mutations $|\mathbf{v}^{g,\text{s}}|$ (see Supplementary Note for more details). This factor accounts for mutational recurrence above the regional background mutation rate. The second factor ($P(\mathbf{v}^{g,\text{n}} \mid |\mathbf{v}^{g,\text{n}}|; \boldsymbol{\lambda}^{g,\text{n}})$) reflects the chance that these $|\mathbf{v}^{g,\text{n}}|$ nonsynonymous mutations occur in their observed positions ($\mathbf{v}^{g,\text{n}}$) conditional on the context-dependent mutational likelihood scores $\boldsymbol{\lambda}^{g,\text{n}}$. This factor is modeled by a multinomial distribution that distributes the $|\mathbf{v}^{g,\text{n}}|$ nonsynonymous mutations across genomic positions proportionally to their composite likelihood scores in $\boldsymbol{\lambda}^{g,\text{n}}$. This factor accounts for an excess of mutations in unusual nucleotide contexts. This enabled us to obtain the probability of observing the number (first factor) and positions (second factor) of nonsynonymous mutations by chance. More details on these distribution models are provided in the Supplementary Note.

In the fourth step, MutPanning examines whether the probability derived in the previous step is small or large compared with a 'random' scenario of $\geq|\mathbf{v}^{g,\text{n}}|$ nonsynonymous mutations in the same gene obtained from Monte Carlo simulations. For each scenario, we randomly drew the total number of nonsynonymous mutations from a randomized Poisson model[13], conditional on the number of synonymous mutations $|\mathbf{v}^{g,\text{s}}|$. We simulated the positions of nonsynonymous mutations across the gene by using a multinomial distribution, conditional on the context-dependent composite likelihood scores $\boldsymbol{\lambda}^{g,\text{n}}$ (see Supplementary Note for more details on both distributions). To derive a P value for each gene, we compared the probability of each scenario with the observed number and positions of nonsynonymous mutations (see Supplementary Note for more details on this comparison). More details on the simulation step and the computation of P values are provided in the Supplementary Note.

In the fifth step, MutPanning computes two additional P values for each gene, which account for destructive mutations (which are an important source to detect tumor suppressors) and for positional clustering (which is an important source to detect mutational hotspots in oncogenes). These P values are then combined with the P value from the previous step using the Brown method. More details on the calculation of these additional P values and their combination to a final P value are provided in the Supplementary Note.

In the last step, MutPanning adjusts its significance values for multiple testing (FDR). Furthermore, it performs additional filtering steps to reduce the number of false positives. For instance, MutPanning determines whether the nucleotide contexts around synonymous mutations deviate from the overall distribution pattern (for example, due to local accumulation of APOBEC-related mutations). If the null hypothesis is locally violated (local deviation from the context-dependent distribution), significant P values do not necessarily reflect positive selection and these genes are filtered. More details on this filtering step as well as the adjustment for multiple testing can be found in the Supplementary Note.

Stratification of driver genes based on literature support. To explore the relevance of our findings, we systematically examined which significantly mutated genes were supported by the literature. In brief, we stratified our findings into four different 'confidence' levels (levels A–D).

Level A. The gene was listed as a canonical cancer gene in the CGC[42,43].

Level B. The gene had not been reported in the CGC, but there were experimental data implicating the gene in the tumor type in which we discovered it as significantly mutated.

Level C. The gene had not been reported in the CGC and there were no experimental data to support the gene in the tumor type in which we discovered it. However, there were experimental data that the gene has a functional role in cancer.

Level D. The gene had not been reported in the CGC and there was no experimental evidence that this gene plays a role in cancer.

To characterize the functional roles of significant findings that were not part of the CGC[42,43] (level A), we systematically searched for publications with experimental and clinical data that implicated our findings in cancer. In brief, our literature search proceeded in two main stages. The first stage entailed searching for experimental evidence in the same tumor type in which we had detected the gene as significantly mutated (steps 1a–4a). In the second stage, we examined whether genes for which we had not found any functional data in the same tumor type had been reported as functionally relevant in any cancer type (steps 1b–4b).

Both of these stages contained a fully automated part (steps 1–3) and a manual review part (step 4). In steps 1–3, we automatically retrieved the abstracts from the PubMed database of publications supporting our findings, pre-filtered them and sorted them by relevance. In step 4, we determined whether the publications contained any experimental data to support our findings.

Step 1a. For each gene–tumor pair, we searched for the gene name plus the cancer type through the Esearch tool (NCBI Entrez Programming Utilities, E-utilities). The Esearch tool provided automated access to the PubMed database. For the gene name, we used the officially approved symbol from the NCBI Reference Sequence Database (RefSeq). For the name of the cancer type, we used all names that commonly appear in the literature (Supplementary Note). If more than one name existed, we searched for all names separately and combined the search results. In this manner, we obtained a list of PubMed IDs (PMIDs) for each gene–tumor pair. If we retrieved more than 100 IDs, we added the search term 'mutation' to narrow our results.

Step 1b. We proceeded in parallel to step 1a. We used the search terms 'cancer', 'tumor', 'tumour' and 'carcinoma' instead of the cancer type.

Steps 2a/b. For each PMID from steps 1a and 1b, we obtained the abstracts and meta-information from the PubMed database using the Efetch tool (NCBI E-utilities). We pre-filtered our results on the basis of this information to guarantee that an abstract was available in English and that the PMID referred to original work. Reviews and case reports were excluded if annotated in the meta-data.

Step 3a/b. For several gene–tumor pairs, we obtained more abstracts than we could manually review. Hence, we retained a maximum of 15 publications per gene–tumor pair for manual review. To retain the most relevant publications, we prioritized abstracts according to the relevance score (Supplementary Information). We further sorted publications with the same relevance score by the number of citations, which we retrieved through the Elink tool (NCBI E-utilities; link name: 'pubmed_ pubmed_citedin'). We used the publication date as a third criterion.

Steps 4a/b. We manually reviewed the abstracts to examine whether the publication reported experimental data for the gene–tumor pair. In particular, we excluded publications that only co-mentioned the tumor type and the gene name in the abstract or reported the presence of a somatic mutation without any functional validation. In addition, we excluded all publications that reported germline mutations—for example, associated the gene with increased cancer risk or heritability. As a negative control, we ran the entire literature search pipeline for 2,500 randomly chosen gene–tumor pairs—that is, randomly chosen combinations of arbitrary genes in the RefSeq database and an arbitrary cancer type examined in this study.

More details on these steps as well as a visualization of our search strategy can be found in the Supplementary Note.

Analysis of mutations in unusual nucleotide contexts. In Fig. 2 we visualized the unusualness of nucleotide contexts for mutations in ten known melanoma genes and five noncancer genes. To quantify whether a position contained more mutations than expected on the basis of its surrounding nucleotide context, we counted the number v_i of mutations in each position i and compared these counts with the mutational likelihood λ_i in position i. For each position with v_i mutations, we determined the probability of observing v_i or more mutations in position i by chance according to a binomial distribution

$$p_i := \sum_{v_i \le k \le v} \binom{v}{k} \cdot \left(\frac{\lambda_i}{\lambda}\right)^k \cdot \left(1 - \frac{\lambda_i}{\lambda}\right)^{v-k}$$

where $v = \sum v_i$ and $\lambda = \sum \lambda_i$ denote the sum of counts and mutational likelihoods, respectively, across all positions in the gene.

We then adjusted these probabilities p_i for multiple testing. We randomly distributed v mutations across the gene by using a multinomial distribution with probabilities λ_i/λ. For each position, we determined a P value with the same equation as above. We repeated this procedure 100 times to generate a cumulative distribution function of the expected distribution of P values. For each observed P value p_i, we determined the expected P value $\tilde{p}_i$ at the same rank based on the distribution of simulated P values. We then determined the fraction f_i of simulated P values that were smaller than p_i. Similarly, we computed the fraction $\tilde{f}_i$ of simulated P values that were smaller than $\tilde{p}_i$. We then derived the ratio $f_i/\tilde{f}_i$ and defined the q value of p_i as the minimum of that ratio and all following q values. For each nonsynonymous mutation, we then plotted the q value of its position against its genomic coordinate in the gene, where we used an FDR cutoff of 0.1 to classify a mutation as usual ($q \ge 0.1$) versus unusual ($q < 0.1$).

Analysis of physical interactions between driver genes. For Fig. 6, we used experimental data from the STRING database[49] to study physical interactions between driver genes in our catalog. The STRING database collects experimental interaction data from the BIND[85], DIP[86], GRID[87], HPRD[88], IntAct[89], MINT[90] and PID[91] datasets, and assigns a unified score between 0 (no interaction) and 1 (strong interaction) to each interaction. We examined whether physical interactions with established driver genes might inform the characterization of less well-established driver genes.

We visualized physical interactions between driver genes in our catalog as a minimum spanning tree based on Kruskal's algorithm. In brief, Kruskal's algorithm starts with a separate unconnected component for each gene. The algorithm then goes through all physical interactions in descending order. If a physical interaction connects two unconnected components, it is added as an edge to the graph, and otherwise it is ignored. This procedure is analogous to hierarchical clustering with single linkage. We then used force-directed graph drawing (Fruchterman–Reingold algorithm) to align nodes and physical interactions between them.

Visualization of mutations using protein crystal structures. Protein structures were visualized using the PyMOL Molecular Graphics System, v.2.0 Schrödinger, LLC and the respective publicly available coordinate files derived from the Protein Data Bank (PDB). The X-ray diffraction crystal structures of CEBPA (PDB 1NWQ)[92], GATA3 (PDB 4HCA)[93], RUNX1 (PDB 1H9D)[94] and SOX17 (PDB 4A3N superposed with 3F27)[95,96], as well as electron microscopy structures of ANAPC1 (PDB 5G05)[97] and POLR2A (PDB 5IYB)[98] were utilized. All protein sequences were *Homo sapiens* except for CEBPA, which was *Rattus norvegicus* (sequence homology of 93.1% to *H. sapiens* CEBPA). The crystalized human sequence of SOX17 was superimposed with the crystal structure of *Mus musculus* SOX17 in complex with DNA. No structural differences between human (no DNA) and mouse SOX17 (plus DNA) were observed. HDAC4 is a co-crystalized structure with a selective class IIa HDAC inhibitor (not shown) occupying the active site of the deacetylase.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A complete MAF of the sequencing data used in this study is available on www.cancer-genes.org and in the Supplementary Information.

Code availability

MutPanning can be downloaded as an interactive software package from www.cancer-genes.org and from the Supplementary Information (including Supplementary Data 1–4). MutPanning can be run on a local computer with at least one CPU, 8 GB memory and 2.5 GB hard drive. In addition, an online version of MutPanning is available through the GenePattern platform (http://www.genepattern.org/modules/docs/MutPanning and http://bit.ly/mutpanning-gp). The MutPanning source code is available on GitHub (https://github.com/vanallenlab/MutPanningV2). MutPannig is distributed under the BSD-3-Clause open source license.

References

81. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
82. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
83. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
84. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
85. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–53 (2016).
86. Xenarios, I. et al. DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).

87. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–9 (2006).
88. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
89. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–5 (2004).
90. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–61 (2012).
91. Schaefer, C. F. et al. PID: the pathway interaction database. *Nucleic Acids Res.* **37**, D674–9 (2009).
92. Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z. & Johnson, P. F. Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein α. *J. Biol. Chem.* **278**, 15178–15184 (2003).
93. Chen, Y. et al. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep.* **2**, 1197–1206 (2012).
94. Bravo, J., Li, Z., Speck, N. A. & Warren, A. J. The leukemia-associated AML1 (Runx1)–CBFβ complex functions as a DNA-induced molecular clamp. *Nat. Struct. Biol.* **8**, 371–378 (2001).
95. Gao, N. et al. Structural basis of human transcription factor Sry-related box 17 binding to DNA. *Protein Pept. Lett.* **20**, 481–488 (2013).
96. Palasingam, P., Jauch, R., Ng, C. K. & Kolatkar, P. R. The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *J. Mol. Biol.* **388**, 619–630 (2009).
97. Zhang, S. et al. Molecular mechanism of APC/C activation by mitotic phosphorylation. *Nature* **533**, 260–264 (2016).
98. He, Y. et al. Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359–365 (2016).

Acknowledgements

We thank G. Getz and C. Cotsapas for their valuable comments and suggestions. We thank M. Reich and T. Liefeld for adding MutPanning as a module to the GenePattern platform. The results presented in this study are in part based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga. F.D. was supported by the EMBO Long-Term Fellowship Program (grant no. ALTF 502-2016), the Claudia Adams Barr Program for Innovative Cancer Research and the AWS Cloud Credits for Research Program. E.M.V.A. and S.R.S received funding from the National Institutes of Health (grants nos K08 CA188615, R01 CA227388 and R21 CA242861 to E.M.V.A. and grants nos R01 MH101244, R35 GM127131 and U01 HG009088 to S.R.S.). E.M.V.A acknowledges support through the Phillip A. Sharp Innovation in Collaboration Award. F.D. and E.M.V.A. were further supported through the ASPIRE Award of The Mark Foundation for Cancer Research.

Author contributions

F.D., D.W., A.R., E.S.L., E.M.V.A. and S.R.S. wrote the manuscript and prepared the figures, which all authors reviewed. F.D., D.W., B.R., D.L., E.M.V.A. and S.R.S. designed and performed the bioinformatics analyses for driver-gene identification, and designed and performed the bioinformatics analyses for method comparison and stratification of the driver-gene catalog. F.D., D.W., A.T.-W., A.R., B.R., D.L., E.S.L., E.M.V.A. and S.R.S. performed a review of the findings and biological follow-up analyses. F.D., D.W., A.T.-W., B.R., D.L., E.S.L., E.M.V.A. and S.R.S. contributed to the development of the method and its implementation.

Competing interests

E.M.V.A. is a consultant for Tango Therapeutics, Genome Medical, Invitae, Foresite Capital, Dynamo and Illumina. E.M.V.A. received research support from Novartis and BMS as well as travel support from Roche and Genentech. E.M.V.A. is an equity holder of Syapse, Tango Therapeutics and Genome Medical.

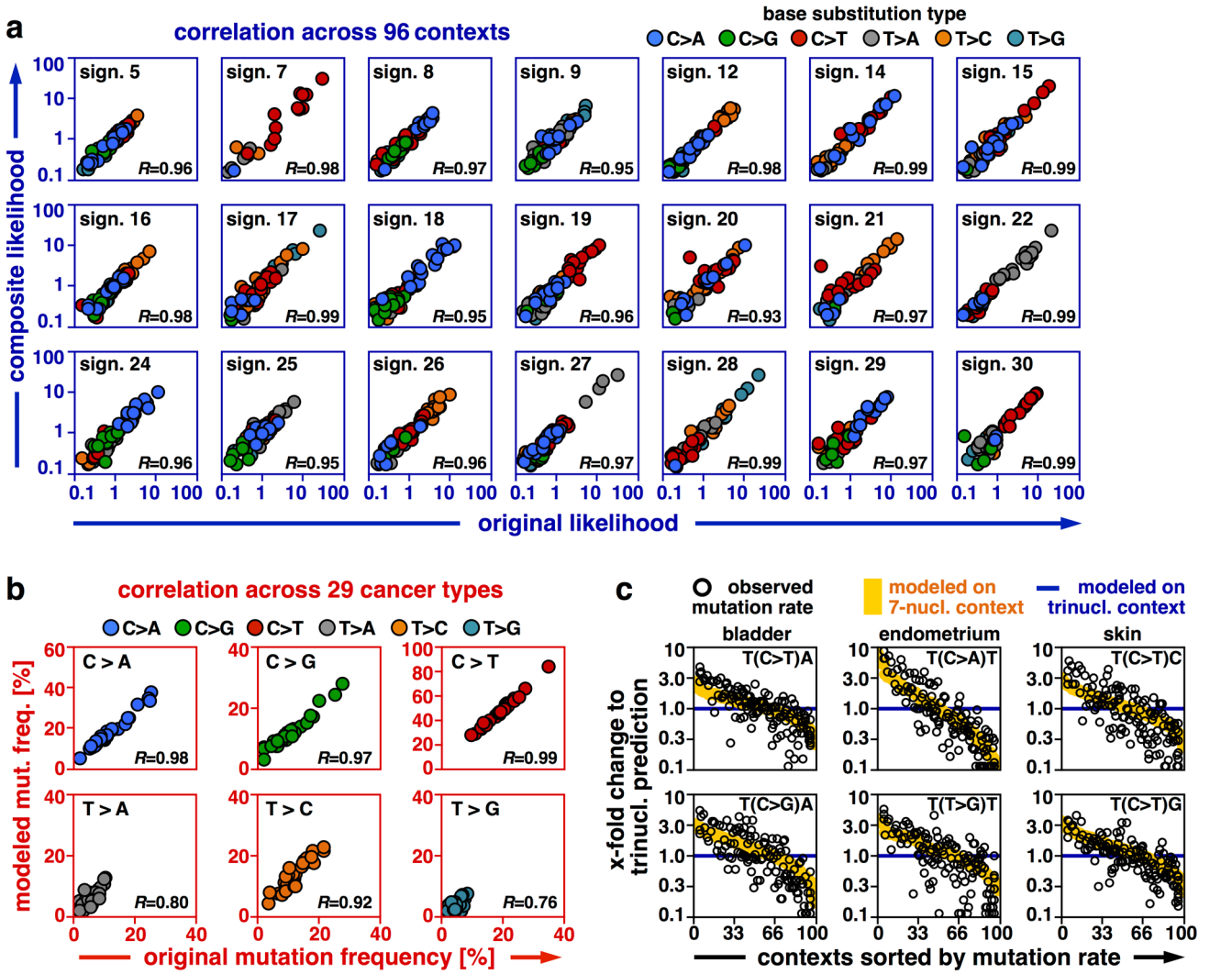
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-019-0572-y.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0572-y.

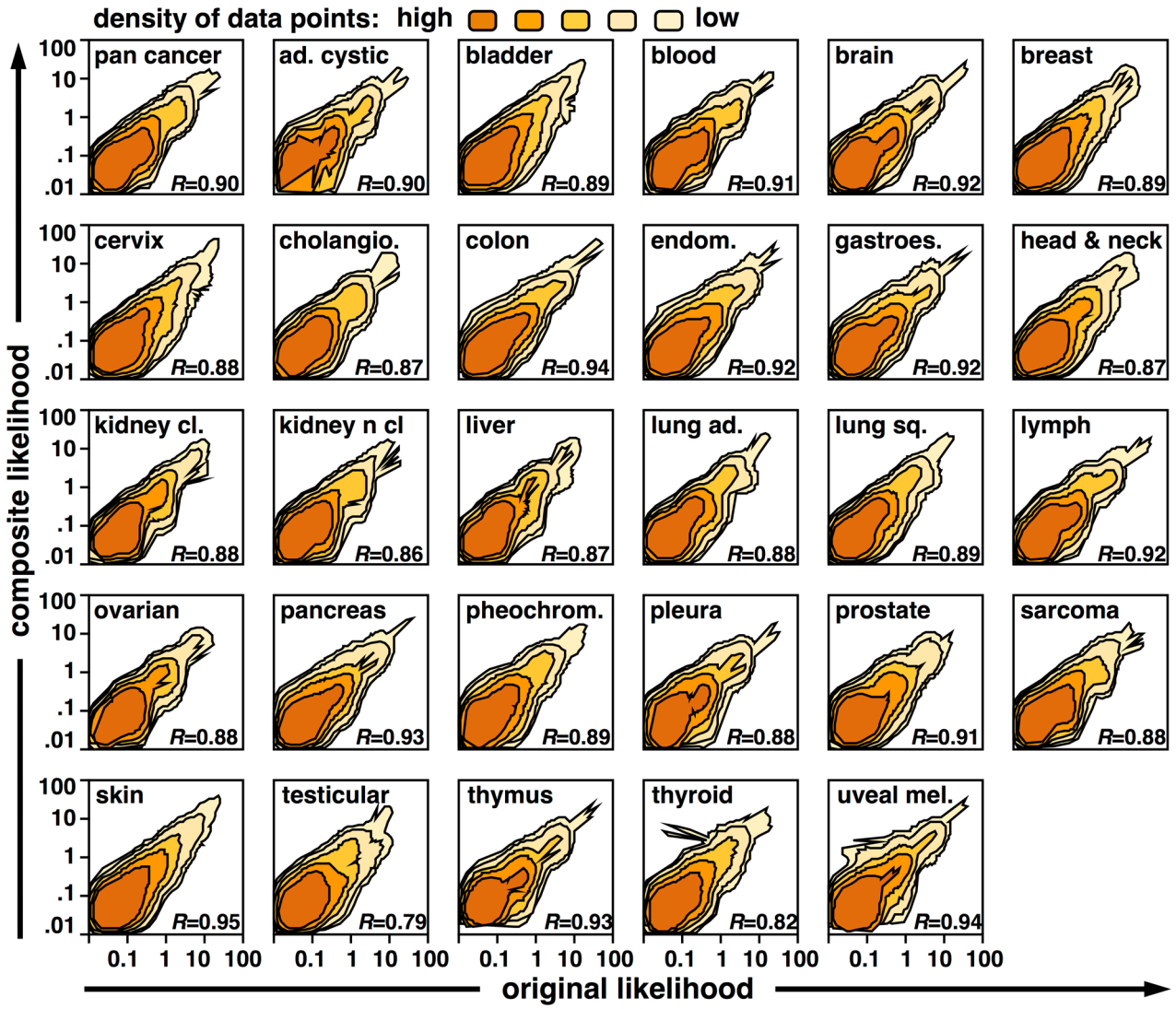
Correspondence and requests for materials should be addressed to F.D., E.M.V.A. or S.R.S.

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

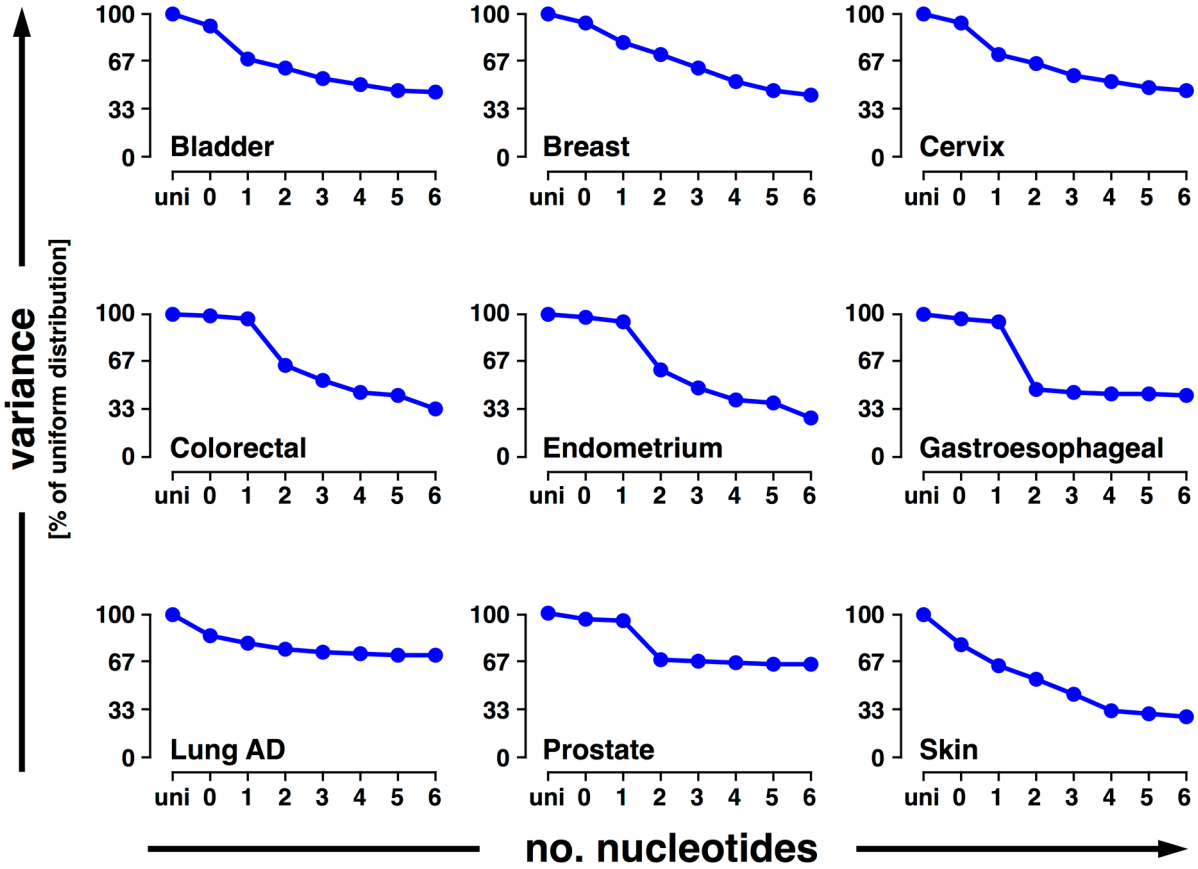


Extended Data Fig. 1 | Modeling of mutation probabilities based on extended nucleotide contexts. a, We applied the composite likelihood model to COSMIC mutation signatures. For each trinucleotide context, we compared the original mutation frequency against the mutation frequency returned by the composite likelihood model based on Pearson correlation. Dot colors reflect base substitution types. **b**, For six base substitution types, we plotted the original mutation probability (based on 11873 samples) against the prediction of the composite likelihood model, which we derived as the product of the mutational likelihood of its reference nucleotide and its substitution type. Each dot represents a cancer type. Pearson correlations are annotated at the bottom right. The number of samples per cancer type can be found in Extended Data Fig. 5. **c**, For three cancer types (bladder, $n = 317$ samples; endometrium, $n = 327$; skin, $n = 582$) we examined whether nucleotides outside the trinucleotide context affected mutation probabilities. For this purpose, we compared mutation probabilities, modeled based on tri- (blue) and 7-nucleotide contexts (yellow), with original mutation probabilities based on context-specific mutation counts. Data points are sorted according to the modeled mutation rates, derived from the 7-nucleotide context (x-axis). Black circles indicate ratios between the observed probabilities and the corresponding trinucleotide-specific likelihoods (y-axis). Similarly, the orange line displays the ratio between the likelihoods, derived from the 7-nucleotide and trinucleotide contexts, respectively (y-axis). Local mutation probabilities vary across positions surrounded the same trinucleotide context. Accounting for extended nucleotide contexts reduces this heterogeneity.

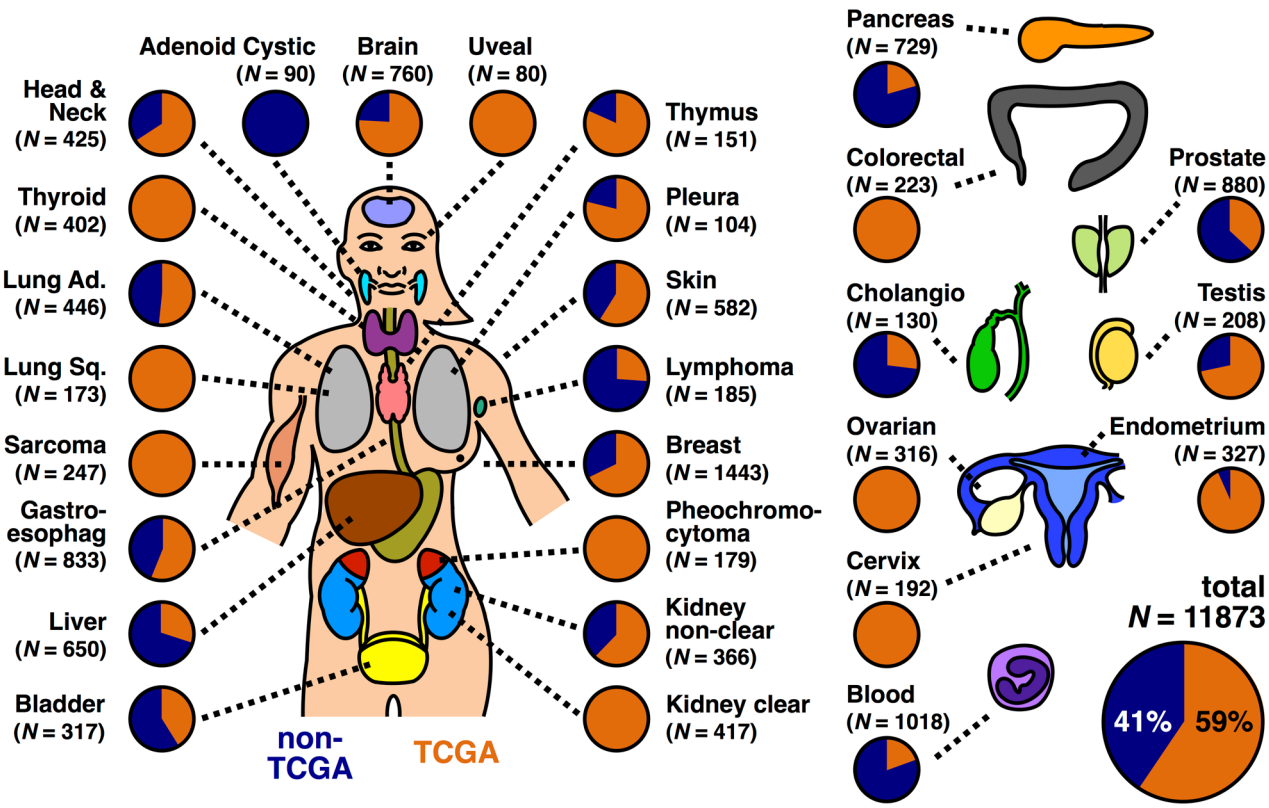
Extended Data Fig. 2 | Evaluation of the composite likelihood model applied to extended nucleotide contexts. To test the independence assumption of the composite likelihood model, we examined the interaction between any two positions (25 possible combinations) in the 11-nucleotide context around mutations of eight cancer types (bladder, $n = 317$ samples; breast, $n = 1443$; colorectal, $n = 223$; endometrium, $n = 327$; gastroesophageal, $n = 833$; head and neck, $n = 425$; lung adeno, $n = 446$; skin, $n = 582$). For any two positions, there are 96 possible nucleotide contexts and we plotted the observed mutation count of each nucleotide context (x-axis) against the predictions of the composite likelihood model (y-axis). Pearson correlation coefficients between observed and predicted data served as a measure of interaction. Each position pair is visualized in a separate correlation plot, and positions are annotated at the bottom right of the plot. For instance, pair (-1,1) refers to the trinucleotide context. Dot colors indicate the base substitution types.



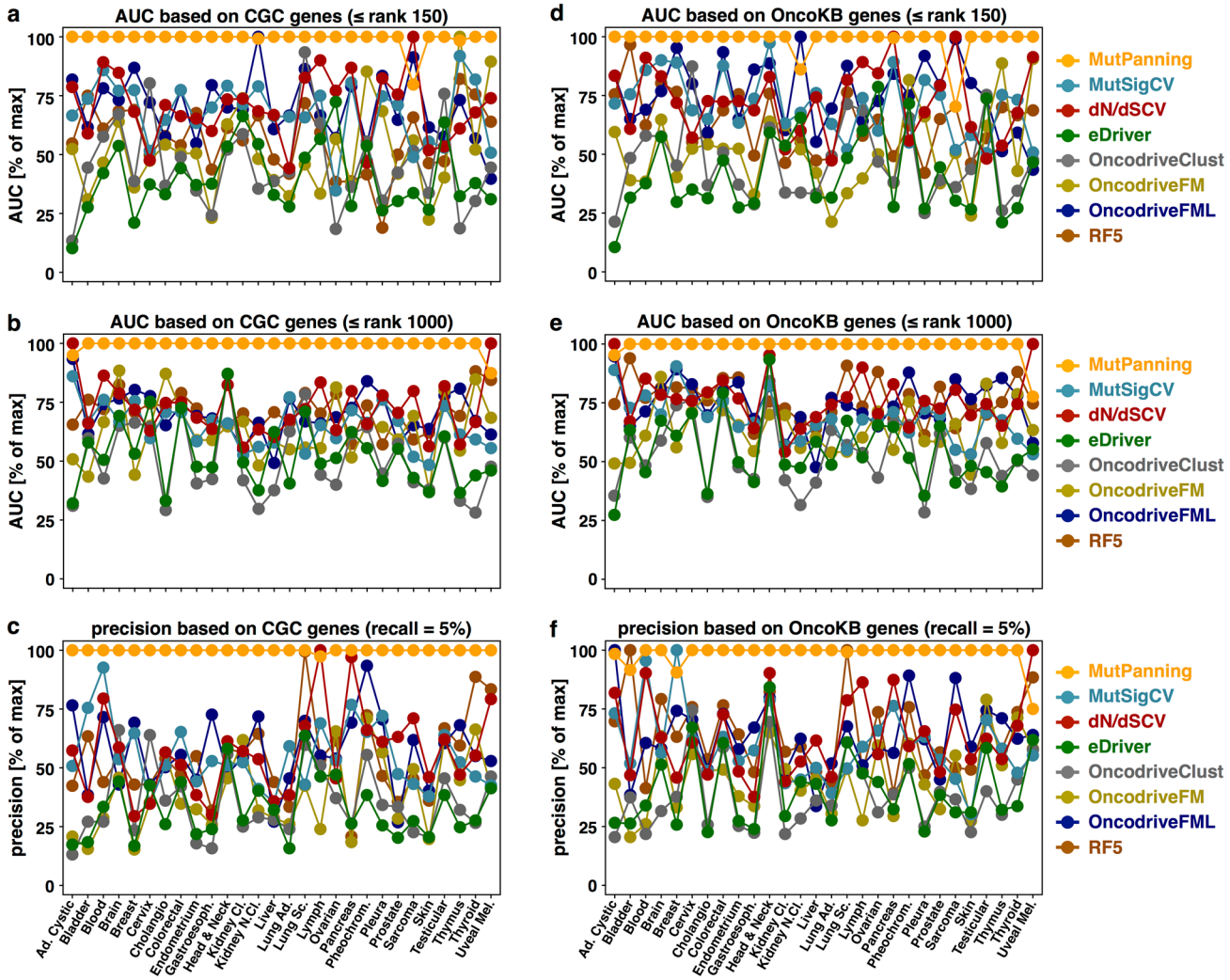
Extended Data Fig. 3 | Generalization of the composite likelihood model to extended nucleotide contexts. We counted the number of mutations in each possible nucleotide context of length ≤7 based on the sequencing data of 11,873 samples. The exact number of samples per cancer type included in this analysis is shown in Extended Data Fig. 5. We compared these counts with the mutability scores returned by the composite likelihood model (218,448 different nucleotide contexts). Since the number of possible nucleotide contexts was too large to be visualized directly, we plotted the data point density. The Pearson correlation coefficient (*R*) of each plot is annotated at the bottom right.



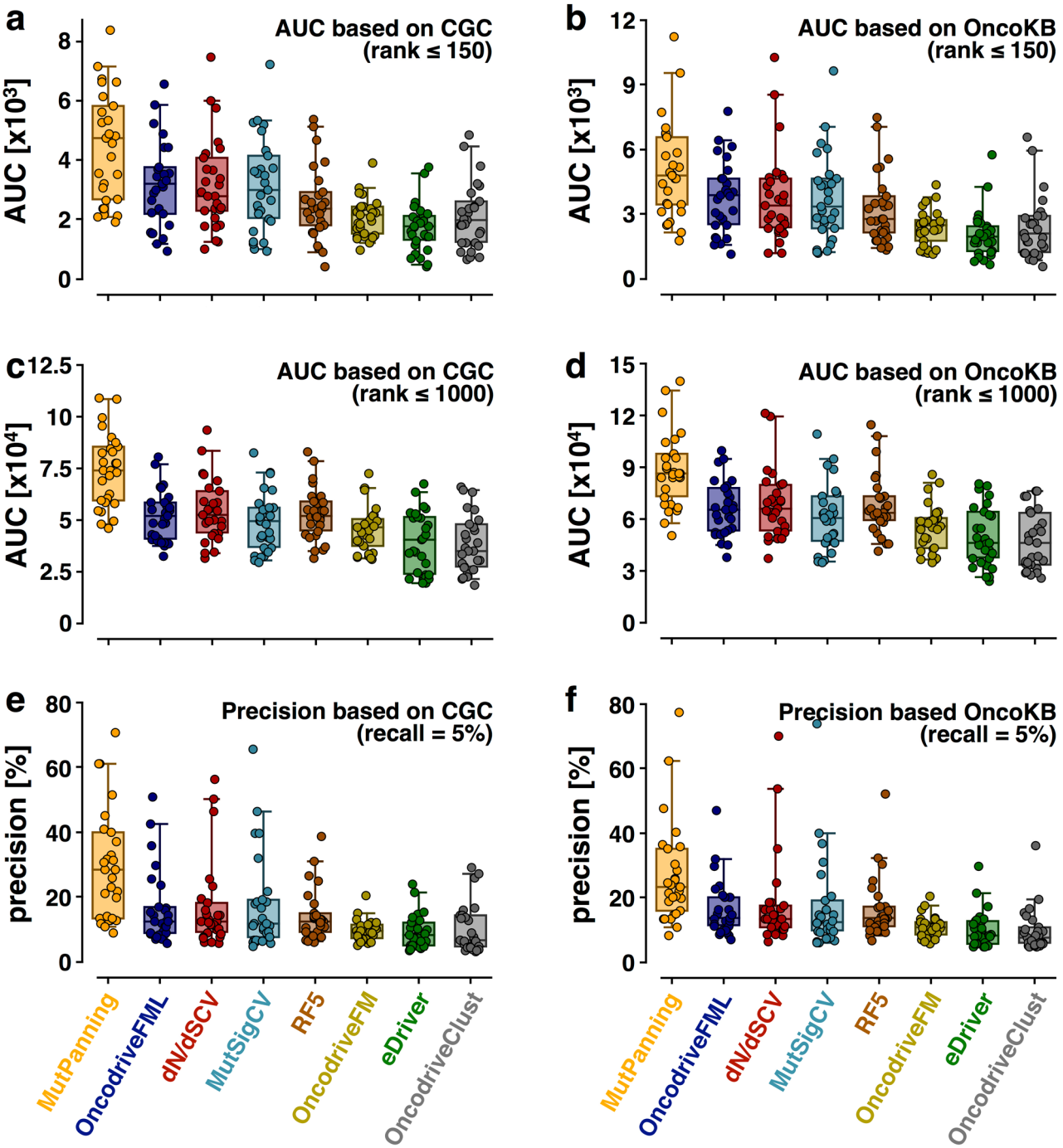
Extended Data Fig. 4 | Extended nucleotide contexts contribute to the performance of the composite likelihood model. We examined whether accounting for extended contexts beyond trinucleotide contexts improved the fit of the composite likelihood model. To this end, we varied the number of nucleotides in the composite likelihood model between 0 (i.e. only substitution types) and 6 (i.e. 7-nucleotide contexts). We computed the residual sum of squared differences between observed mutation counts and the predictions of the composite likelihood model. As a negative control, we determined the residual sum of squares for a uniform distribution. This baseline was used to normalize the residual sum of squares for each cancer type. For some cancer types with 'flat' mutation signatures, nucleotide contexts only had minor impact on the fit of the model, but did not decrease the performance of the model (for example, lung adeno., $n = 446$ samples). For other cancer types, the fit of the model largely depended on the trinucleotide context, but not on the extended nucleotide context (e.g., prostate cancer, $n = 880$). For most cancer types with high background mutation rates, the fit of the composite likelihood model strongly depended on the extended nucleotide context (e.g., bladder, $n = 317$; breast, $n = 1443$; cervical, $n = 192$; colorectal, $n = 223$; endometrial cancer, $n = 327$; melanoma, $n = 582$).



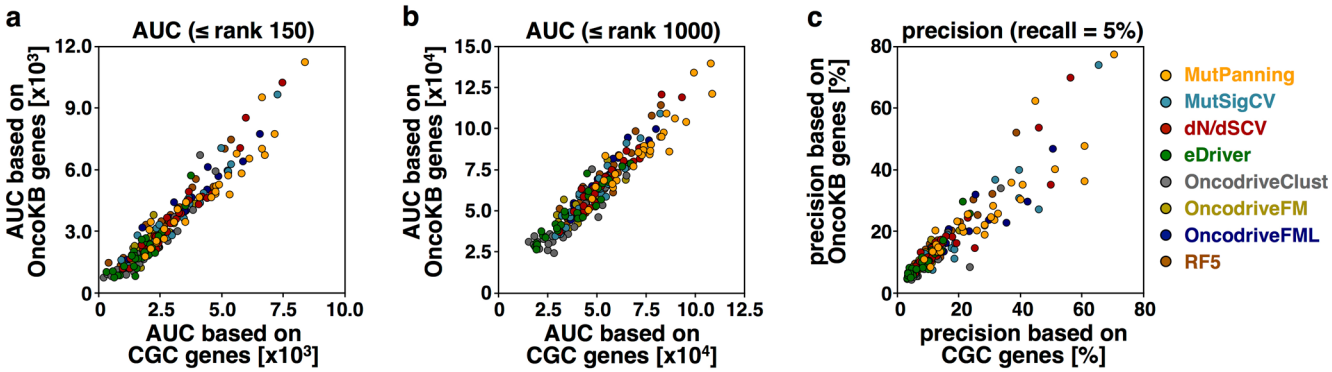
Extended Data Fig. 5 | A large-scale cohort of whole-exome sequencing data to identify rare cancer genes. To systematically identify candidate cancer genes, we analyzed sequencing data from 11,873 individual tumor samples using the statistical framework that we had developed in this study. Our study cohort contained whole-exome sequencing data from 32 TCGA-related (orange) and 55 TCGA-independent (blue) projects.



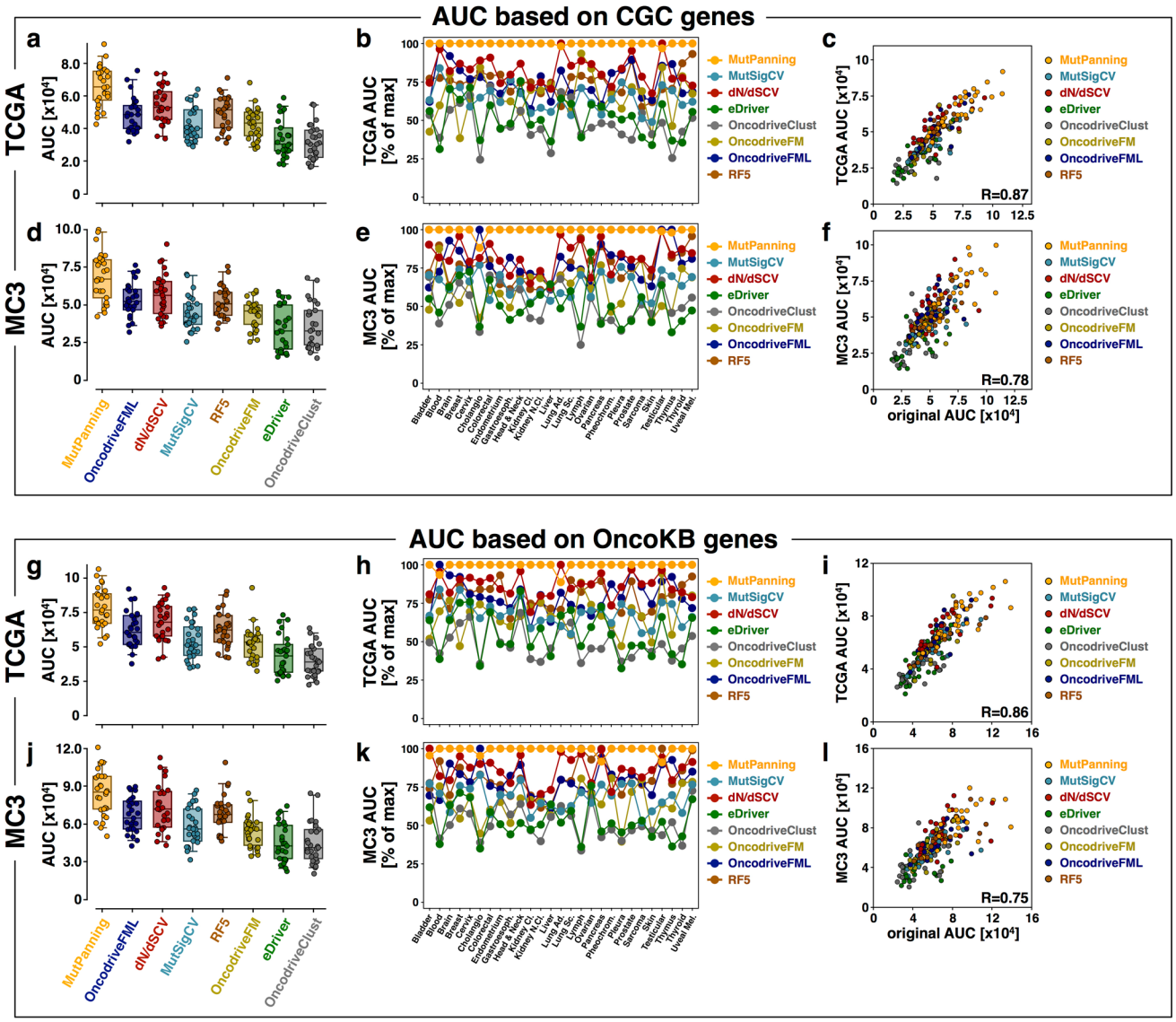
Extended Data Fig. 6 | Benchmarking of the performance of MutPanning for cancer gene identification. We benchmarked the performance of our method against 7 previously published methods for cancer gene identification based on the sequencing data of 11,873 samples spanning 28 different cancer types. The exact number of samples per cancer type can be found in Extended Data Fig. 5. To benchmark the performance of a method, we sorted genes according to the significance values (adjusted for multiple testing) returned by the method. As a conservative approximation of the true-positive rate we used Cancer Gene Census (CGC) genes (**a, b, c**) and OncoKB genes (**d, e, f**) to derive ROC and precision-recall curves. We quantified the performance of each method as the area under the ROC curve (AUC) for the top 150 (**a, d**) or 1000 (**b, e**) non-CGC/OncoKB genes, respectively. Further, we determined the precision at 5% recall for each method (**c, f**). We normalized these measures to the maximum within each cancer type.



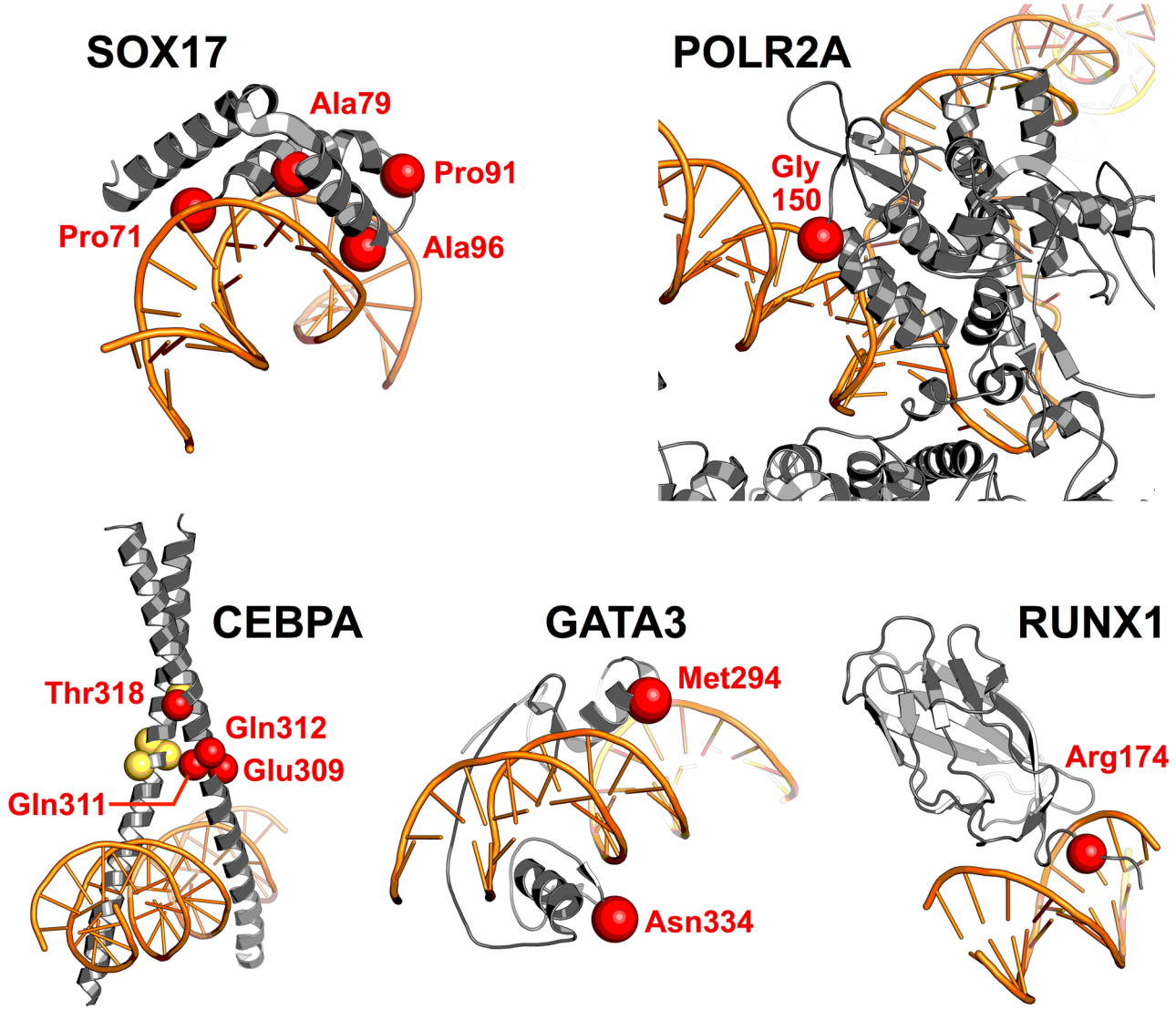
Extended Data Fig. 7 | Comparison of different methods for cancer-gene identification. We benchmarked the performance of our method against 7 previously published methods for cancer gene identification based on the sequencing data of 11,873 samples spanning 28 different cancer types. To benchmark the performance of a method, we sorted genes according to the significance values (adjusted for multiple testing) returned by the method. As a conservative approximation of the true-positive rate we used Cancer Gene Census (CGC) genes (**a, c, e**) and OncoKB genes (**b, d, f**) to derive ROC and precision-recall curves. We quantified the performance of each method as the area under the ROC curve (AUC) for the top 150 (**a, b**) or 1000 (**c, d**) non-CGC/OncoKB genes, respectively. Further, we determined the precision at 5% recall for each method (**e, f**). Box plots indicate the distribution of these performance measures for each method across cancer types. Each cancer type is represented by a dot. Boxes indicate the 25%/75% interquartile range, whiskers extend to the 5%/95%-quantile range. The median of each distribution is indicated as a vertical line.



Extended Data Fig. 8 | Comparison of performance measures derived from CGC versus OncoKB. We benchmarked the performance of our method against 7 previously published methods for cancer gene identification based on the sequencing data of 11,873 samples spanning 28 different cancer types. To benchmark the performance of a method, we sorted genes according to the significance values (adjusted for multiple testing) returned by the method. As a conservative approximation of the true-positive rate we used Cancer Gene Census (CGC) genes and OncoKB genes to derive ROC and precision-recall curves. We quantified the performance of each method as the area under the ROC curve (AUC) for the top 150 (**a**) or 1000 (**b**) non-CGC/OncoKB genes, respectively. Further, we determined the precision at 5% recall for each method (**c**). This figure compares the performance measures derived from the CGC (x-axis) and OncoKB (y-axis) databases. Each dot represents the AUC/precision of a different method (dot color) for an individual cancer type. The concordance between CGC and OncoKB measures suggests that our measure of performance does not entirely depend on the dataset used to approximate the true-positive rate.



Extended Data Fig. 9 | Comparison of methods in two homogeneously processed datasets. We compared the performance of MutPanning with 7 other methods on two independently processed datasets (TCGA subcohort (**a-c**, **g-i**), $n = 7060$ samples; MC3 dataset (**d-f**, **j-l**), $n = 9079$). We used the Cancer Gene Census (CGC) (**a-f**) and OncoKB (**g-l**) for benchmarking. We quantified the performance by the AUC of the ROC curve of the top 1,000 non-CGC/OncoKB genes returned by each method. **a**, **d**, **g**, **j**, Box plots indicate the distribution of performance measures for each method. Boxes indicate the 25%/75% interquartile range, whiskers extend to the 5%/95%-quantile range. Distribution medians are indicated as vertical lines. Each dot represents an AUC for one of the 27 cancer types in the TCGA and MC3 datasets. **b**, **e**, **h**, **k**, We normalized AUCs by the maximum AUC within each tumor type. We then compared these normalized AUCs between methods across cancer types. **c**, **f**, **i**, **l**, We compared the AUCs obtained from our original study cohort with the AUCs from TCGA and MC3 based on Pearson correlation. Each dot reflects a cancer type/method. Cohort sizes for TCGA/MC3 datasets: bladder: 130/386; blood: 197/139; brain: 576/821; breast: 975/779; cervix: 192/274; cholangio: 35/34; colorectal: 223/316; endometrium: 305/451; gastroesophageal: 467/529; head&neck: 279/502; kidney clear: 417/368; kidney non-clear: 227/340; liver: 194/354; lung adenocarcinoma: 230/431; lung squamous: 173/464; lymph: 48/37; ovarian: 316/408; pancreas: 149/155; pheochromocytoma: 179/179; pleura: 82/81; prostate: 323/477; sarcoma: 247/204; skin: 342/422; testicular: 149/145; thymus: 123/121; thyroid: 402/492; uveal melanoma: 80/80.



Extended Data Fig. 10 | Recurrent mutations in domains of protein–DNA interaction. Significance values in this figure legend were computed using MutPanning and adjusted for multiple testing (false discovery rate, FDR). Recurrent SOX17 mutations in endometrial cancer ($n=327$ samples, FDR $=8.77\times10^{-3}$) are located in the high-mobility-group box domain at the SOX17–DNA interface (PDB: 4A3N superposed with 3F27). POLR2A harbors recurrent mutations in lung adenocarcinoma ($n=446$, FDR $=9.28\times10^{-6}$) at the end of an alpha helical segment that is directly pointed at the major groove of the double stranded DNA (PDB: 5IYB). The open complex of a cryo-EM multicomponent structure where the melted single-stranded template DNA is inserted into the active site and RNA polymerase II locates the transcription start site is visualized. CEBPA harbors recurrent mutations in hematological malignancies ($n=1{,}018$, FDR $=1.16\times10^{-7}$) at the cross-over interface of the two CEBPA homodimers (PDB: 1NWQ). GATA3 (PDB: 4HCA) harbors recurrent mutations in breast cancer ($n=1{,}443$, FDR $<10^{-20}$) at Asn334, which is located in the GATA-type 2 zinc finger (res317–res341), as well as the residue Met294, which is located peripheral to the GATA-type 1 zinc finger domain (res263–res287). RUNX1 harbors recurrent mutations in breast cancer ($n=1{,}443$, FDR $=2.22\times10^{-4}$) and hematological malignancies ($n=1018$, FDR $=1.94\times10^{-5}$). Arg174 plays an important role for DNA recognition and facilitates the formation of hydrogen bond interactions to a guanosine base from the consensus DNA binding sequence of RUNX1 (PDB: 1H9D).

nature research

Corresponding author(s): Shamil R. Sunyaev, Eliezer M. Van Allen, Felix Dietlein

Last updated by author(s): Nov 7, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

from the Online Methods (sections "Sequencing data curation and variant filtering" and "Visualization of mutations using protein crystal structures") and the Supplementary Note (sections "Selection of sequencing studies" and "Additional data used in this study"):
"[...] we compiled the whole-exome sequencing data from 32 TCGA-related projects (7,091 samples), as well as from 55 TCGA-independent publications (4,856 samples). Data were manually curated to fulfill the following criteria:
• whole-exome sequencing data only, in particular no whole-genome sequencing data, no targeted sequencing data
• patient samples only, in particular no cell lines, mouse models or patient-derived xenograft models
• sequencing data had been aligned against the Hg19 human reference genome
• all tumor samples had been sequenced against a matched normal, and studies had filtered out germline variants from the matched normal, as well as common germline variants
• sequencing results were available as a standard mutation annotation file (MAF) or as a comparable format
• studies had applied filters for common sequencing artifacts, including artifacts introduced by DNA oxidation during sequencing, low-confidence mutations with strand bias, and low quality variant calls
For studies where only a subset of samples satisfied all these criteria, we manually selected those samples for inclusion in this study. Further, we discarded samples that had been flagged for low quality in either of these studies. Mutation annotation files (MAF) for TCGA-related projects were directly obtained from the TCGA Gene Data Analysis Center (GDAC) data portal hosted by the Broad Institute (gdac.broadinstitute.org, latest data version from 01/28/2016, doi:10.7908/C11G0KM9). MAF files for TCGA-independent studies were either downloaded from the cBioPortal platform (cbioportal.org) or - if not available there - directly form the supplement of the publications. [...]
Finally, mutations from this combined MAF file were processed through a homogeneous filtering step, in order to minimize sequencing artifacts, mutation calling errors, and germline variants that might have slipped through the variant filters applied in each study. Our variant filtering pipeline included the following filters:
• Filtering of common germline variants: Each mutation was compared against the Exome Aggregation Consotrium (ExAC) database, which reports germline variants of 60,706 individuals. As similarly described previously, we removed all variants from the MAF file that occurred more than 10 times in any of the 7 ExAC subpopulations.
• Removal of OxoG and strand bias sequencing artifacts: The 8-oxoguanine (OxoG) artifact results from excessive oxidation during sequence library preparation, whereas the strand bias artifact produces disparities between G>T and C>A mutation counts at low variant

allele frequencies. We used the MC3 dataset in order to eliminate OxoG and strand bias artifacts from our MAF file, which were identified by the DetOxoG tool.
• Removal of low quality samples: Samples for which >10% of the somatic mutations were flagged as artifacts or germline variants were entirely removed from the study. In total, this resulted in the removal of 0.62% (N=74) of all samples.
In this way, we arrived at a study cohort of 11,873 tumor samples, spanning 28 different cancer types. The final MAF file, which we used for all subsequent analyses in this study, is available online (www.cancer-genes.org). [...]
The Hg19 human reference exome sequence and the Blat alignment tool were downloaded from the UCSC genome browser (https://genome.ucsc.edu). Genomic coordinates of exon/intron boundaries for each gene were annotated using the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). The coverage files of all TCGA tumor samples were obtained in a wig file format (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/). Sequencing data for the TCGA validation was part of our original study cohort and obtained from gdac.broadinstitute.org (cf. above). The data for the MC3 dataset were obtained from Ellrot et al. via https://gdc.cancer.gov/about-data/publications/mc3-2017 (publicly available Maf file). We then excluded samples that were hypermutated and that were flagged based on pathology exactly as described in Bailey et al. ("Data preparation" section of their paper). That way, we arrived at the same Maf file of 9,079 samples as in Bailey et al.10 Details on the underlying variant calling and filtering pipeline can be found in Ellrot et al., 2018. Details on the underlying variant calling and filtering pipeline of the TCGA dataset can be found on http://gdac.broadinstitute.org/ [...] Protein structures were visualized using [...] publicly available coordinate files derived from The Protein Data Bank (PDB)."

Data analysis

The Online Methods (section "Statistical analyses to identify driver genes") and the Supplementary Note (sections "A composite likelihood model to quantify the mutability of genomic positions based on nucleotide contexts" and "A statistical framework for the identification of cancer genes") provide a detailed description of the statistical framework that we developed and used to identify significant gene-tumor pairs. Furthermore, a brief overview of the statistical framework is provided in the results of the main text ("A framework for identifying driver genes in cancer"). In brief, our statistical model consists of the following six major steps:
1.) Individual tumor samples were clustered according to their context-dependent distribution of somatic passenger mutations (Bayesian hierarchical clustering to guarantee compatibility of the clusters with the statistics used in the subsequent steps).
2.) For each cluster, the broad nucleotide context around its somatic mutations was characterized (composite likelihood model).
3.) For each position in the human exome, its local mutation probability was determined based on its surrounding nucleotide context (composite likelihood model) as well as the regional background mutation rate (Bayesian model calibrated with the help of synonymous mutations).
4.) We compared the nucleotide context around mutations with the characteristic nucleotide context around passenger mutations (Multinomial distribution), thereby determining whether the mutation occured in a "usual" nucleotide context (high chance of being a passenger mutation) or "unusual" nucleotide context (lower chance of being a passenger mutation).
5.) Based on this comparison, we identified genes harboring a significant excess of mutations in unusual nucleotide contexts that differed from the characteristic context around passenger mutations (Monte Carlo simulation). The rationale behind this test was that a shift from usual to unusual nucleotide contexts reflects the shift of driver mutations from functionally neutral towards functionally important positions without prior knowledge of the location of functional positions (cf. introduction in the main text for a more detailed explanation of this rationale).
6.) In addition to unusual nucleotide contexts, our method further exploits signals used by previous methods to identify driver genes. For instance, we searched for genes with an increased number of nonsynonymous mutations compared with the local background mutation rate (mutational recurrence). For this purpose, we followed a Bayesian strategy and modeled the fluctuation of the background mutation rate across the exome based on the number of synonymous mutations in each gene (prior distribution). Based on the number of synonymous mutations in a gene, we then estimated the local background mutation rate (posterior distribution). Lastly, we tested whether the observed number of nonsynonymous mutations in the same gene exceeded the expectation of the local background mutation rate significantly.
We incorporated our complete statistical framework in a computational tool called MutPanning. The source code of MutPanning was written in Java and is publicly available on the GitHub repository (https://github.com/vanallenlab/MutPanningV2).
Protein structures in were visualized using The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC and the respective publicly available coordinate files derived from The Protein Data Bank (PDB). In detail, X-ray diffraction crystal structures of CEBPA (PDB: 1NWQ), GATA3 (PDB: 4HCA), HDAC4 (PDB: 4CBY), RUNX1 (PDB: 1H9D), and SOX17 (PDB: 4A3N superposed with 3F27), as well as electron microscopy structures of ANAPC1 (PDB: 5G05), and POLR2A (5IYB) were utilized (cf. section 4.1 for more details).
Publications supporting the significant gene-tumor pairs reported in this study were identified through the NCBI Entrez Programming Utilities (Esearch, Efetch, and Elink tools) (Online Methods, section "Stratification of driver genes based on literature support", and Supplementary Note, section "Stratification of driver genes based on literature support"). We compared the performance of our approach with six current methods, which are widely used to identify driver genes and cover a wide range of different signals used for driver gene detection (MutSigCV, dNdScv, OncodriveCLUST, OncodriveFM, OncodriveFML, e-Driver) (cf. Supplementary Note, section "Method Comparison"). Further, we used experimental data from the STRING database to study physical interactions between driver genes in our catalog, and to cluster driver genes into signaling pathways (Online Methods, "Analysis of physical interactions between driver genes").

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The manuscript describes the availability of the sequencing data of the full study cohort in the Data availability section:
"Data availability
A complete mutation annotation file of the sequencing data used in this study is available on www.cancer-genes.org and in the Supplementary Information."

Furthemore, the availability of the MutPanning software and the source code is described in the Code availability section:
"Code availability

MutPanning can be downloaded as an interactive software package from www.cancer-genes.org and from the Supplementary Information. MutPanning can be run on a local computer with at least 1 CPU, 8 GB memory, and 2.5 GB hard drive. In addition, an online version of MutPanning is available through the GenePattern platform (http://www.genepattern.org/modules/docs/MutPanning and http://bit.ly/mutpanning-gp). The MutPanning source code is available on GitHub (https://github.com/vanallenlab/MutPanningV2). MutPannig is distributed under the BSD-3-Clause open source license."

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

In total, we examined sequencing data from 11,873 individual tumor samples from 32 TCGA-related projects (7,091 samples), as well as from 55 TCGA-independent publications (4,856 samples). As part of our filtering pipeline, we removed of low quality samples, i.e. samples for which >10% of the somatic mutations were flagged as artifacts or germline variants. In total, this resulted in the removal of 74 samples, thereby arriving at a cohort size of 11,873 tumor samples. In addition, we restricted ourselves to whole-exome sequencing data, as we did not want technical differences (e.g. coverage fluctuation in whole-genome sequencing data) to confound our analysis (cf. Supplementary Note, section "Selection of sequencing studies", for more details).
No statistical methods were used to predetermine the sample size of our cohort. Previous studies have pointed out that the statistical power to identify driver genes increases with the cohort size (Lawrence et al. 2014). A power analysis in our study (Supplementary Figure 24) suggests that the statistical power also depends on a myriad of other factors (e.g. selection frequency, background mutation rate, epigenomic covariants, deviation of driver mutations from passenger mutation nucleotide contexts). Since these factors are tumor type-intrinsic, the rationale behind our study design was to aggregate as many whole-exome sequencing data as possible based on the sequencing data currently available in the literature. Compared with other pan-cancer studies for driver gene identification, our cohort size was larger (n=11,873 samples vs. Bailey et al. 2018, n=9,423, Martincorena et al. 2017, n=7,664, Lawrence et al. 2014, n=4,742). Furthermore, we used two homogeneously processed study cohorts for validation purposes (TCGA subcohort, n=7,060, MC3 dataset, n=9,079). Our study identified driver genes at mutation frequencies as low as ~1%. However, we cannot assume that the size of our study cohort was sufficient to comprehensively identify all driver genes that exist, especially in tumor types with high background mutation rates and a low context dependency (e.g. lung cancer). It is very likely that additional driver genes exist with mutation frequencies <1% that were missed by our study based on the sequencing data available in the literature.

Data exclusions

We included data from whole-exome sequencing studies into our study cohort that fulfilled the following criteria (cf. Online Methods, section "Sequencing data curation and variant filtering", and Supplementary Note, section "Selection of sequencing studies", for more details):
• whole-exome sequencing data only, in particular no whole-genome sequencing data, no targeted sequencing data
• patient samples only, in particular no cell lines, mouse models or patient-derived xenograft models
• sequencing data had been aligned against the Hg19 human reference genome
• all tumor samples had a matched normal
• sequencing results were available as a standard mutation annotation file (MAF) or as a comparable format
• samples which had been flagged for bad quality in the study were discarded
For studies where only a subset of samples satisfied all these criteria, we manually selected those samples for inclusion in this study. Finally, mutations from this combined MAF file were processed through a homogeneous filtering step in order to minimize sequencing artifacts, mutation calling errors, and germline variants that might have slipped through the variant filters applied in each study. Our variant filtering step included the following filters:
• Filtering of common germline variants: Each mutation was compared against the Exome Aggregation Consotrium (ExAC) database, which reports germline variants of 60,706 individuals. As similarly described previously, we removed all variants from the MAF file that occurred more than 10 times in any of the 7 ExAC subpopulations.
• Removal of OxoG and strand bias sequencing artifacts: The 8-oxoguanine (OxoG) artifact results from excessive oxidation during sequence library preparation, whereas the strand bias artifact produces disparities between G>T and C>A mutation counts at low variant allele frequencies. We used the MC3 dataset in order to eliminate OxoG and strand bias artifacts from our MAF file, which were identified by the DetOxoG tool.
• Removal of low quality samples: Samples for which >10% of the somatic mutations were flagged as artifacts or germline variants were entirely removed from the study. In total, this resulted in the removal of 0.62% (N=74) of all samples.
Mutation annotation files (MAF) for TCGA-related projects were directly obtained from the TCGA Gene Data Analysis Center (GDAC) data portal hosted by the Broad Institute (gdac.broadinstitute.org, latest data version from 01/28/2016, doi:10.7908/C11G0KM9). MAF files for TCGA-independent studies were either downloaded from the cBioPortal platform (cbioportal.org) or - if not available there - directly form the supplement of the publications. We integrated all MAF files into a combined MAF file and removed duplicate patients from the combined MAF file. Apart from these criteria, no samples were excluded. Similar criteria had been used in other large-scale sequencing studies (Lawrence et al. 2014, Bailey et al. 2018, Ellrott et al. 2018, The AACR Project GENIE Consortium 2017), but these exclusion criteria had not been pre-established prior to this study.

Replication

To allow independent scientists to reproduce our results and apply our statistical model to their own data, we provide source code and compiled versions of our statistical tool on www.cancer-genes.org, genepattern.org, the GitHub repository (https://github.com/vanallenlab/MutPanningV2). Further, MutPanning can be downloaded as an interactive software package from www.cancer-genes.org and from the supplement. This version can be run on a local computer with at least 1 CPU, 8 GB memory, and 2.5 GB hard drive, preferably with Windows or MacOS. In addition, an online version of MutPanning is available through the GenePattern platform (http://www.genepattern.org/modules/docs/MutPanning for the documentation, http://bit.ly/mutpanning-gp to directly get to the MutPanning module one GenePattern, sign in required). Executing MutPanning through the interactive desktop version locally or through the GenePattern platform online does not

require any computational expertise,.
To guarantee that our instructions all the steps required for successful reproduction of our results, we asked two graduate students in our lab to reproduce our results. Both students were able to reproduce the significantly mutated melanoma genes reported in this study. All attempts at replication were successful.

Randomization

Our study included data from 32 TCGA-related projects (7,091 samples), as well as from 55 TCGA-independent publications (4,856 samples). We grouped tumor samples according to their cancer types, based on the cancer types reported in the original sequencing studies. Most of these tumor types were defined as in the TCGA marker papers (27/28 tumor types). In this way, we arrived at a study cohort of 11,873 tumor samples, spanning 28 different cancer types and including the sequencing data from 87 sequencing projects (cf. Online Methods, section "Sequencing data curation and variant filtering", and Supplementary Note, section "Selection of sequencing studies", for more details).
A more detailed description of the cancer types can be found in the Supplementary Note. Further, Supplementary Table 1 provides literature references to the original sequencing studies. As we analyzed samples within the same cancer types as those noted in the publications from which the sequencing data was obtained, randomization of samples was not applicable to our study design. Hence, no randomization step was used to allocate tumor samples to their experimental groups (i.e., their cancer types).

Blinding

Investigators were not blinded to group allocation (i.e., the cancer type of each sample) during data collection and data analysis, as blinding was not relevant for the design of this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging