nature genetics

Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations

Carlos L Araya[1,4], Can Cenik[1,4], Jason A Reuter[1], Gert Kiss[2], Vijay S Pande[2], Michael P Snyder[1] & William J Greenleaf[1,3]

Cancer sequencing studies have primarily identified cancer driver genes by the accumulation of protein-altering mutations. An improved method would be annotation independent, sensitive to unknown distributions of functions within proteins and inclusive of noncoding drivers. We employed density-based clustering methods in 21 tumor types to detect variably sized significantly mutated regions (SMRs). SMRs reveal recurrent alterations across a spectrum of coding and noncoding elements, including transcription factor binding sites and untranslated regions mutated in up to ~15% of specific tumor types. SMRs demonstrate spatial clustering of alterations in molecular domains and at interfaces, often with associated changes in signaling. Mutation frequencies in SMRs demonstrate that distinct protein regions are differentially mutated across tumor types, as exemplified by a linker region of PIK3CA in which biophysical simulations suggest that mutations affect regulatory interactions. The functional diversity of SMRs underscores both the varied mechanisms of oncogenic misregulation and the advantage of functionally agnostic driver identification.

In cancer, driver mutations alter functional elements of diverse nature and size. For example, melanoma drivers include hyperactivating mutations mapping to single amino acid residues (for example, BRAF Val600; ref. 1), inactivating mutations along tumor-suppressor exons (for example, in *PTEN*[1]) and regulatory mutations (for example, in the *TERT* promoter[2]). Cancer genomics projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have substantially expanded our understanding of the landscape of somatic alterations by identifying frequently mutated protein-coding genes[3–5]. However, these studies have focused little attention on systematically analyzing the positional distribution of coding mutations or characterizing noncoding alterations[6].

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [2]Department of Chemistry, Stanford University, Stanford, California, USA. [3]Department of Applied Physics, Stanford University, Stanford, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to C.L.A. (claraya@stanford.edu), M.P.S. (mpsnyder@stanford.edu) or W.J.G. (wjg@stanford.edu).

Received 17 June; accepted 20 November; published online 21 December 2015; doi:10.1038/ng.3471

Algorithms to identify cancer driver genes often examine nonsynonymous-to-synonymous mutation rates across the gene body or recurrently mutated amino acids called mutation hotspots[5], as observed in BRAF[7], IDH1 (ref. 8) and DNA polymerase ε (encoded by *POLE*)[9]. Yet, these analyses ignore recurrent alterations in the vast intermediate scale of functional coding elements, such as those affecting protein subunits or interfaces. Moreover, where mutation clustering within genes has been examined[10–12], analyses have employed windows of fixed length or identified clusters of nonsynonymous mutations, assuming that driver mutations exclusively influence protein sequence and ignoring the importance of exon-embedded regulatory elements[13–18].

A significant proportion of regulatory elements in the genome are located proximal to or even in exons[15,19], suggesting that many may be captured by whole-exome sequencing. Efforts to characterize noncoding regulatory variation in cancer genomes have primarily examined either (i) pan-cancer whole-genome sequencing data or (ii) predefined regions—such as ETS-binding sites, splicing signals, promoters and UTRs—or mutation types[20–23]. These approaches either presume the relevant targets of disruption or disregard the established heterogeneity among cancer types at the level of driver genes and pathways[5,24,25] as well as in nucleotide-specific mutation probabilities[3,4]. Yet, systematic analyses of genomic regulatory activity in animals have identified substantial tissue and developmental stage specificity[26–28], suggesting that mutations in cancer type–specific regulatory features may be significant noncoding drivers of cancer.

To address these diverse limitations, we employed density-based clustering techniques using cancer-, mutation type– and gene-specific mutation models to identify regions of recurrent mutation in 21 cancer types. This approach permitted the unbiased identification of variably sized genomic regions recurrently altered by somatic mutation, which we term significantly mutated regions (SMRs). We identified SMRs in numerous well-established cancer drivers as well as in new genes and functional elements. Moreover, SMRs were associated with noncoding elements, protein structures, molecular interfaces, and transcriptional and signaling profiles, thereby providing insight into the molecular consequences of accumulating somatic mutations in these regions. Overall, SMRs identified a rich spectrum of coding and noncoding elements recurrently targeted by somatic alterations that complement gene- and pathway-centric analyses.

RESULTS

Multiscale detection of SMRs

We examined ~3 million previously identified[5] somatic single-nucleotide variants (SNVs) from 4,735 tumors of 21 cancer types, recording[29] their impact on protein-coding sequences, transcripts and adjacent regulatory regions (**Supplementary Fig. 1**). We note that 79.0% ($n = 2{,}431{,}360$) of these somatic mutations do not alter protein-coding sequences or their splicing and thus were not previously considered in the analysis of cancer driver mutations[5] (**Fig. 1a**).

To discover both coding and noncoding cancer drivers, we applied an annotation-independent density-based clustering technique[30] to identify 198,247 variably sized clusters of somatic mutations within exon-proximal domains of the human genome (**Fig. 1b** and Online Methods). We included synonymous mutations because functionally important noncoding features can be embedded within coding regions[13–18].

Mutation density scores within each identified cluster were derived as the Fisher's combined P value of the individual binomial probabilities of observing k or more mutations for each mutation type within the region in each cancer type (Online Methods). We evaluated mutation density for each cluster using gene-specific and genome-wide models of mutation probability (**Supplementary Fig. 2**), which were well correlated (**Supplementary Fig. 3a**), selecting the more conservative estimate for each cluster as the final density score (Online Methods). Gene-specific mutation probability models accounted for sequence composition (GC content) as well as differences in local gene expression and replication timing, which have been shown to correlate with somatic mutation rate[4]. To avoid skewed mutation probability estimates due to selection pressure on exons, we applied a Bayesian framework to derive gene-specific mutation probabilities given intronic mutation probabilities in cancer whole-genome sequencing data[3,20] while controlling for differences in sensitivity in whole-exome and whole-genome sequencing (Online Methods).

Although many known cancer-related genes did not display signals of high mutation density, increasing density scores correlated with stronger enrichments (up to 120×) for somatic SNV–driven cancer genes ($n = 158$), as determined by the Cancer Gene Census (CGC; **Supplementary Fig. 3b**,**c**)[31,32]. Moreover, ~10% of genes associated with SMRs in the quintile with the top density scores were not found previously in a gene-level analysis[5] or in the CGC. Thus, high density scores are enriched for known cancer genes but also nominate potentially new drivers.

We applied Monte Carlo simulations to select density score thresholds controlling the false-discovery rate (FDR) to ≤5% (**Supplementary Fig. 4** and **Supplementary Table 1**). We identified 872 significantly mutated regions (SMRs; **Fig. 1c**) that were altered in ≥2% of patients in 20 cancer types for further characterization (**Fig. 1d**). SMRs spanned 735 genomic regions, which were assigned unique SMR codes (for example, TP53.1). Note that some SMRs ($n = 120$) appeared in more than one cancer type.

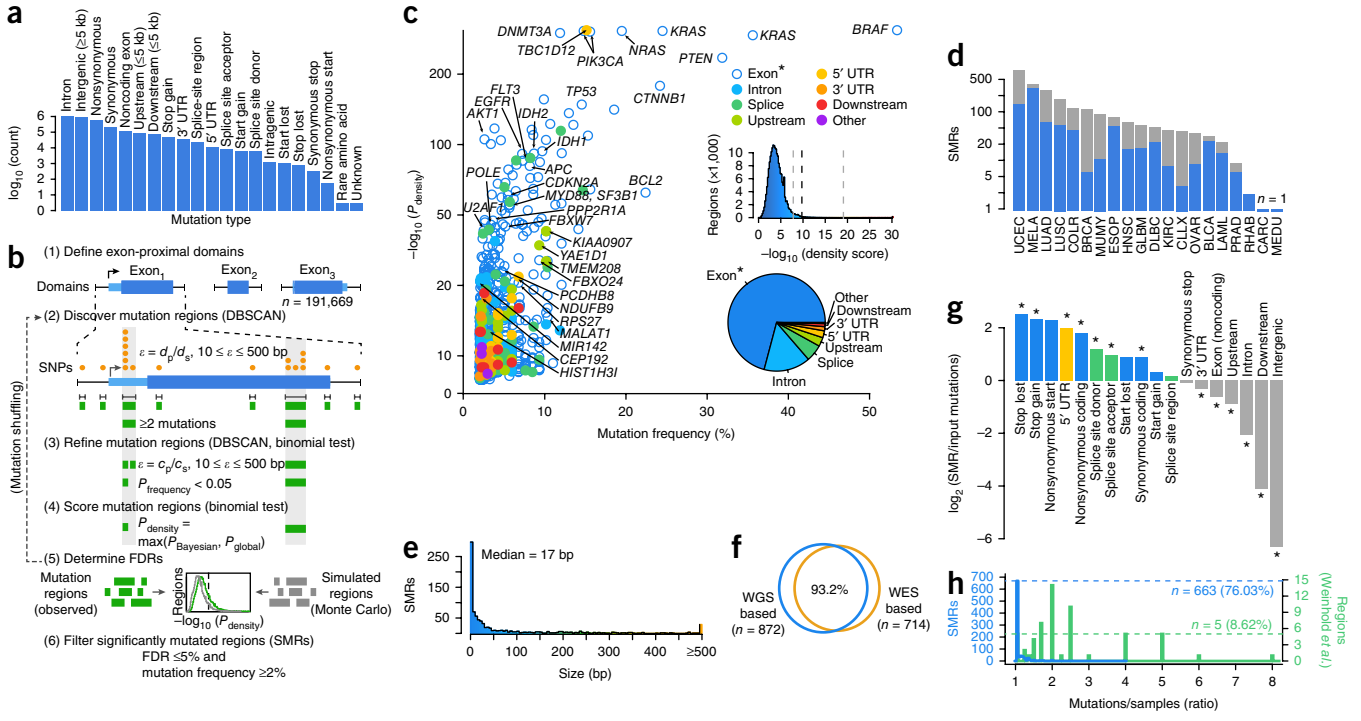

Figure 1 Identification of SMRs in 21 cancer types across a broad spectrum of functional elements. (**a**) Pan-cancer distribution of mutation types for $n = 3{,}078{,}482$ somatic SNV calls. (**b**) Exons and exon-proximal domains (±1,000 bp) were scanned for clusters of somatic mutations (orange; DBSCAN). The distance parameter ε is dynamically defined as the average distance of mutated positions (d_p) in the domain size (d_s). Clusters (green) are divided if subclusters with higher mutation densities ($P < 0.05$, binomial test) were found in a second-pass analysis with ε defined as the average distance of mutated positions (c_p) within the cluster of size c_s (see the Online Methods for details on density scoring and FDR calculation). (**c**) Per-cancer mutation frequency and density scores for the SMRs discovered (color-coded by type and labeled by associated gene). The distribution of density scores in evaluated regions (top) and the distribution of SMR region types (bottom) are shown in insets. Dashed lines indicate the minimum, median and maximum density score FDR (5%) thresholds. The "Exon*" label refers to coding exons and noncoding genes. (**d**) The number of SMRs with FDR ≤5% and mutation frequency ≥2% per cancer type. Gray bars represent SMRs with FDR ≤5% but mutation frequency <2%. (**e**) SMR size distribution. (**f**) Concordance between SMRs discovered by employing background models derived from whole-genome sequencing (WGS based) or whole-exome sequencing (WES based). (**g**) Categories with significant fold change in mutation type representation between SMR-associated and input mutations are denoted (*$P < 0.01$). (**h**) Distribution of the number of mutations per sample in SMRs (blue) and 58 recurrently altered noncoding regions[20] (green). Horizontal lines indicate the number of regions where mutations derive from distinct samples (that is, where mutations/samples equals 1).

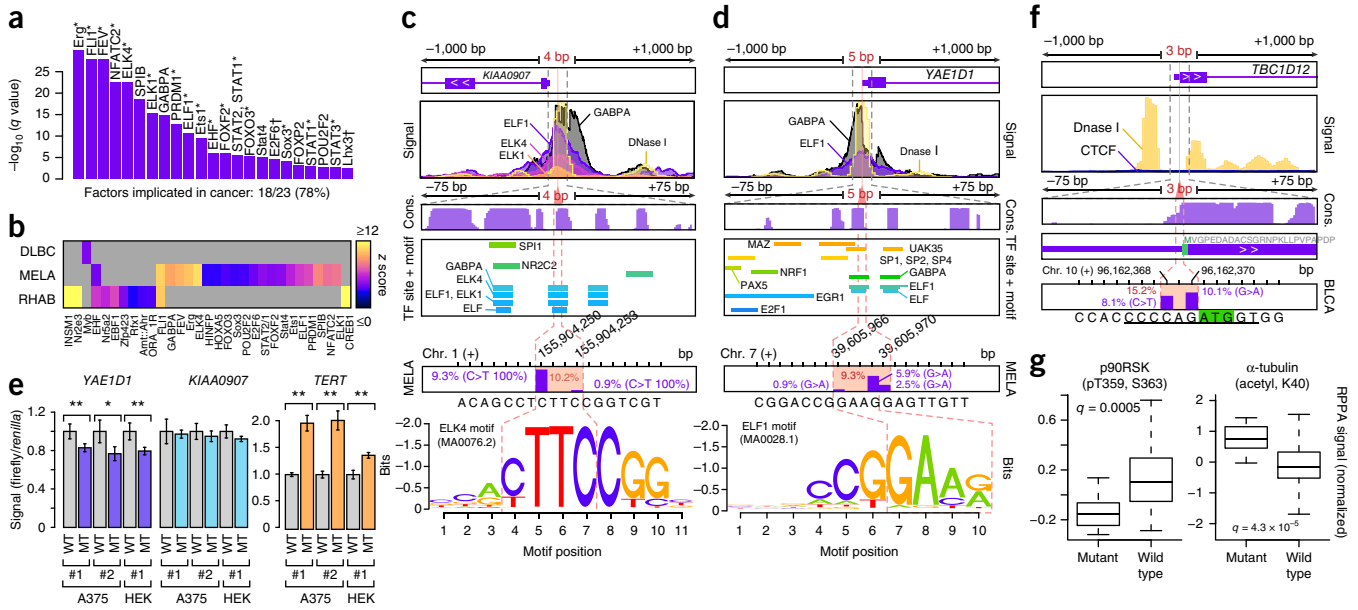

Figure 2 Noncoding SMRs recurrently alter promoters and 5′ UTRs. (**a**) Transcription factors with enriched ($q < 0.01$) motifs in small SMRs (≤25 bp) across all cancer types are shown. Eighteen of the 23 transcription factors are known to be associated with cancer (*) or to be associated with cell cycle control or have developmental roles (†). (**b**) Cancer-specific motif enrichment analysis. (**c**,**d**) Gene structure, ENCODE chromatin immunoprecipitation and sequencing (ChIP-seq) and DNase I signals, vertebrate conservation (Cons.; phastCons 100-way), Factorbook transcription factor (TF) binding sites and motif occurrences, and somatic mutation frequencies at melanoma SMRs in *KIAA0907* (**c**) and *YAE1D1* (**d**) promoter regions are shown on multiple scales (±1,000, ±75 and ±7 bp). Mutation frequencies within each SMR (red) and at each position (purple bars) are shown. Motifs of ETS family binding sites that overlap the SMRs are highlighted. (**e**) Luciferase reporter signal from wild-type (WT) and mutant (MT) promoters in three experiments performed in melanoma (A375) and HEK293T (HEK) cells with independent plasmid DNA preps (1 and 2). For each experiment, three replicates were performed. Firefly/*renilla* luciferase signals are shown and were normalized by the mean signal with wild-type promoter for each experiment. $^{**}P < 0.05$, $^{*}P < 0.1$, two-sided *t* test. Error bars, s.d. (**f**) Gene structure, ENCODE CTCF and DNase I signals, and vertebrate conservation (phastCons 100-way) at the bladder cancer SMR in the 5′ UTR of *TBC1D12* are shown on multiple scales. The position of the start codon is highlighted in green, and the Kozak sequence is underlined. (**g**) Relative protein and post-translational modification signals of wild-type ($n = 78$) and mutant (TBC1D12.1 SMR altered; $n = 14$) bladder tumors. The central band, box boundaries and whiskers correspond to the median, the interquartile range and the highest and lowest points within 1.5 × the interquartile range, respectively. pT359, S363, phosphorylation at Thr359 and Ser363; acetyl K40, acetylation at lysine 40.

We classified SMRs into high-, medium- and low-confidence sets on the basis of their density scores and contribution from mutator samples (Online Methods and **Supplementary Table 2**). We observed correspondingly high (63.3×; $P = 2.5 \times 10^{-46}$), medium (6.2×; $P = 2.6 \times 10^{-10}$) and low (5.0×; $P = 5.0 \times 10^{-4}$) enrichments for somatic SNV–driven cancer genes in these sets. To control for unaccounted processes that could result in clusters of mutations with no selective advantage in cancer, we leveraged single-nucleotide and trinucleotide density scores from intronic mutation clusters under the assumption that these clusters are non-functional (Online Methods). This procedure identified 205 'robust' SMRs that passed a false-discovery threshold (FDR ≤ 5%) in these secondary tests or were found in multiple cancer types. Fully 95.0% of high-confidence SMRs in the cancer types where these tests could be applied satisfied these stringent alternate criteria (**Supplementary Fig. 5**). Over 87% of SMRs were contained within mappable (100-bp) regions of the genome, and an analysis of 6,179 recently published breakpoints[33] yielded a single SMR (in *PTEN*) within 50 bp of a resolved breakpoint, suggesting that the observed mutation density in SMRs is not attributable to mapping artifacts.

SMRs display a wide range of sizes (**Fig. 1e**; median = 17 bp, range = 1–2,041 bp), are robust to distinct mutation background models (**Fig. 1f** and Online Methods), are not driven by unaccounted mutation contexts (**Supplementary Fig. 6**), and are enriched in protein-coding, 5′ UTR and splice-site mutations ($P < 0.01$; **Fig. 1g**). Notably, SMRs are not driven by samples that contribute large numbers of mutations per region (**Fig. 1h**). This is in contrast to recently proposed regions of recurrent alteration[20] where as few as five regions were driven exclusively by distinct samples ($P = 6.0 \times 10^{-45}$, Wilcoxon rank-sum test). Thus, we have identified a diverse set of variably sized SMRs targeted by recurrent somatic alterations, and we sought to characterize their relevance to functional elements and cancer-associated genes.

SMRs enrich for known cancer genes and implicate new ones

SMRs are predicted to have diverse impacts on 610 genes and are 8.35-fold enriched in known genes with somatic cancer-associated alterations (Lawrence *et al.*[5] or CGC, $P = 8.1 \times 10^{-49}$, hypergeometric test), affecting a total of 91 known drivers, including canonical oncogenes (for example, *BRAF*, *KRAS*, *NRAS*, *PIK3CA* and *CTNNB1*) and tumor suppressors (for example, *PTEN*, *TP53* and *APC*). SMR-associated genes also include 17 CGC genes previously undetected in a gene-level analysis[5], such as established oncogenes like *BCL2* and *PIM1* and the cancer-associated noncoding gene *MALAT1*. Most coding SMRs are driven by nonsynonymous mutations (**Supplementary Fig. 7**), demonstrating that SMRs capture positive selection primarily acting on protein alterations. The presence of SMRs implicates 26 known cancer genes in 31 gene × cancer type associations not uncovered by gene-level analysis[5] (**Supplementary Table 3**). We note, however, that most known cancer genes do not harbor regions of dense mutation recurrence within these data (**Supplementary Fig. 8** and **Supplementary Note**), suggesting that SMR identification complements gene-level approaches.

We discovered SMRs in multiple new cancer driver genes, including the breast cancer–associated antigen and putative transcription

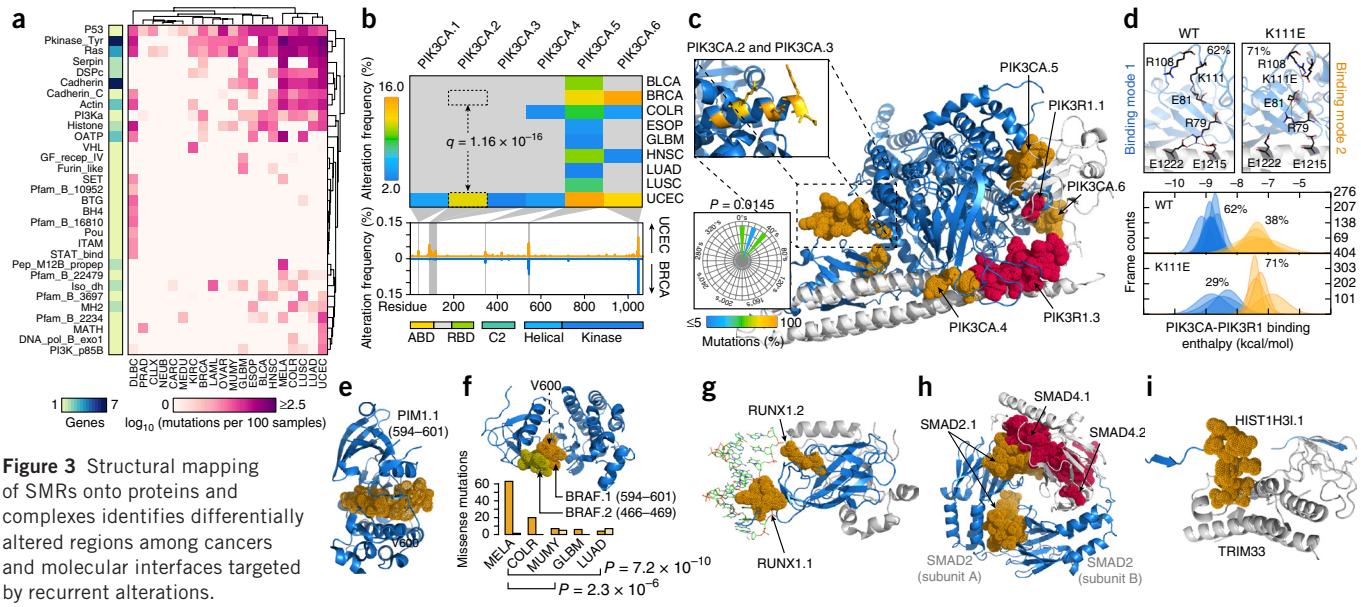

Figure 3 Structural mapping of SMRs onto proteins and complexes identifies differentially altered regions among cancers and molecular interfaces targeted by recurrent alterations. (**a**) Nonsynonymous mutation frequency per PFAM protein domain per cancer, per residue. The number of genes per domain is shown (left). (**b**) Alteration frequency matrix of PIK3CA SMRs across cancer types and comparison of per-residue alteration frequencies of PIK3CA domains[46] in endometrial (UCEC; orange) and breast (BRCA; blue) cancer samples. Gray bars represent SMRs in PIK3CA. (**c**) Co-crystal structure of the PIK3CA (p110α; blue) and PIK3R1 (p85α; gray) interaction (Protein Data Bank (PDB), 2RD0, 2IUG and 3HIZ). Residues within endometrial cancer SMRs in PIK3CA (orange) and PIK3R1 (red) are rendered as solvent-accessible surfaces. Insets display mutated residues in the PIK3CA.2 and PIK3CA.3 SMR in the α helix (yellow; top) and their corresponding side-chain dihedral angles (bottom). (**d**) Mutations within the PIK3CA.2, PIK3CA.3 SMR α-helix interfere with Arg79-binding contacts at the PIK3R1 interface, as shown in the wild-type protein (WT) and Lys111Glu mutant. (**e**–**i**) Molecular structures are shown of spatially clustered alterations (diffuse large B cell lymphoma) (**e**) and SMRs (multiple myeloma) (**f**), a DNA (green) interface SMR (**g**), reciprocal protein interface SMRs (**h**) and a histone H3.1 SMR in the TRIM33 interface (**i**). Structural alignments and molecular visualizations were prepared with PyMOL (Schrödinger). The relative proportions of BRAF.1 and BRAF.2 missense mutations per cancer type are shown in **f**. The PDB codes for **e**–**i** are 3CXW, 1UWH, 1H9D, 1U7V and 3U5N, respectively.

factor *ANKRD30A*[34], in which ~21% of melanomas harbored mutations within one or more of three SMRs. Mutations in these SMRs were validated in whole-genome sequencing data for six of 17 cutaneous melanomas[3,20]. Within the entire gene body, 27 of 118 whole-exome and ten of 17 whole-genome sequencing data sets from patients with melanoma harbored somatic protein-altering mutations in *ANKRD30A*. Overall, of the 185 high-confidence SMRs, 16 were associated with new cancer driver genes (**Supplementary Table 4**). As expected on the basis of methodological differences, these putative new cancer drivers are primarily (~81%) driven by noncoding alterations, as discussed in the next section.

SMRs implicate diverse noncoding regulatory features

A significant proportion (31.2%; $P < 2.2 \times 10^{-16}$, proportions test) of SMRs are not predicted to affect protein sequences, highlighting the potential to discover pathological noncoding variation in whole-exome sequencing data. In total, 130 SMRs lay within open chromatin[28] and were enriched in promoter (4.9×; $q = 4.0 \times 10^{-9}$) and 5′ UTR (6.0×; $q = 4.4 \times 10^{-10}$) features (**Supplementary Table 5**). Three promoter SMRs coincided with regions deemed significantly mutated in a pan-cancer analysis of whole-genome sequencing data[20]. Across all cancer types, small (≤25-bp) noncoding SMRs were enriched in binding sequences for ETS oncogene family (7.4×; $q = 2.6 \times 10^{-6}$) and winged-helix repressor (3.2×; $q = 2.0 \times 10^{-4}$) transcription factors (**Fig. 2a** and **Supplementary Table 6**). We also detected cancer-specific transcription factor motif enrichments within SMRs from diffuse large B cell lymphoma, melanoma and rhabdosarcoma (**Fig. 2b** and **Supplementary Table 7**).

We discovered 4-bp and 5-bp SMRs within open chromatin sites of the *KIAA0907* and *YAE1D1* promoters that were altered in 10.2% and 9.3%, respectively, of melanomas analyzed by whole-exome sequencing (**Fig. 2c,d**). Somatic mutations in these SMRs were confirmed in whole-genome sequencing data for melanomas ($n = 1$ for *KIAA0907* and $n = 2$ for *YAE1D1* of $n = 17$ cases)[3,20]. Yet, these regions did not reach significance in a pan-cancer analysis[20], highlighting cancer specificity in noncoding alterations. In both SMRs, mutations altered core recognition sequences within *in vivo* ETS factor binding sites (Encyclopedia of DNA Elements (ENCODE)), with varying effects on ETS primary sequence preferences. *KIAA0907* encodes a putative RNA-binding protein. However, intronic sequences in this gene harbor *SNORA80E* (also known as *SNORA42*), an H/ACA class small nucleolar RNA (snoRNA) with increased expression in lung and colorectal cancers[35,36], suggesting that promoter SMR alterations may enhance transcription at this locus. However, we observed no detectable changes in reporter gene expression with the mutant *KIAA0907* promoter (**Fig. 2e**). Whereas *YAE1D1* promoter mutations reduced reporter gene expression (**Fig. 2e**), RNA-level overexpression of *YAE1D1* has previously been observed in lower crypt-like colorectal cancer[37] and a small cohort of melanoma samples showed increased YAE1D1 protein levels in comparison to untransformed melanocytes[38].

In addition to SMRs that influence promoter regions, we observed 32 SMRs in 5′ and 3′ UTRs, including putative microRNA (miRNA) target sites[39]. Most strikingly, we discovered a 3-bp SMR in the 5′ UTR of *TBC1D12* that was mutated in ~15% of bladder cancers (**Fig. 2f**). Recurrent mutations were positioned near the start codon (Kozak

Table 1 Recurrently altered protein interfaces uncovered by SMRs

Protein (i)	Partner (j)	PDB	Chain (i)	Chain (j)	Region	Average distance (Å)	Distance ratio	*q* value	Status[a]
VHL	TCEB1	3ZUN	I	H	Chr. 3: 10,191,469–10,191,513	7.259	0.395	7.62×10^{-10}	Known
VHL	TCEB2	1LQB	C	A	Chr. 3: 10,191,469–10,191,513	9.867	0.367	7.62×10^{-10}	Known
SPOP	H2AFY	3HQH	A	M	Chr. 17: 47,696,421–47,696,467	7.962	0.462	3.72×10^{-8}	Known
SMAD2	SMAD4	1U7V	A	C	Chr. 18: 45,374,881–45,374,945	9.231	0.460	5.61×10^{-8}	Known
HIST1H2BK	DNA	2CV5	D	J	Chr. 6: 27,114,446–27,114,519	9.730	0.520	3.27×10^{-7}	New
TP53	TP53BP1	1KZY	B	D	Chr. 17: 7,578,369–7,578,556	13.253	0.556	5.13×10^{-7}	Known
SMAD4	SMAD2	1U7V	B	C	Chr. 18: 48,604,665–48,604,797	11.878	0.694	5.13×10^{-7}	Known
DNMT3A	DNMT3L	2QRV	E	F	Chr. 2: 25,463,271–25,463,308	10.112	0.380	5.13×10^{-7}	Known
SMAD4	SMAD3	1U7F	B	C	Chr. 18: 48,604,665–48,604,797	11.883	0.700	1.94×10^{-6}	Known
PIK3CA	PIK3R1	3HHM	A	B	Chr. 3: 178,936,070–178,936,099	9.028	0.335	2.56×10^{-6}	Known
RUNX1	DNA	1H9D	C	H	Chr. 21: 36,231,782–36,231,792	8.957	0.351	0.001	Known
HIST1H3I	TRIM33	3U5N	D	A	Chr. 6: 27,839,651–27,840,062	11.480	0.610	0.001	New
HIST1H2BK	HIST1H4[b]	2CV5	D	F	Chr. 6: 27,114,446–27,114,519	13.680	0.664	0.002	New
PPP2R1A	PPP2R5C	2NPP	D	E	Chr. 19: 52,716,323–52,716,329	7.313	0.247	0.007	Known
HRAS	RASA1	1WQ1	R	G	Chr. 11: 534,283–534,291	5.302	0.350	0.007	Known
PIK3R1	PIK3CA	3HIZ	B	A	Chr. 5: 67,589,138–67,589,149	6.713	0.567	0.008	Known
NFE2L2	KEAP1	2FLU	P	X	Chr. 2: 178,098,799–178,098,815	6.157	0.566	0.009	Known
EGFR	EGF	3NJP	B	A	Chr. 7: 55,233,035–55,233,043	8.763	0.386	0.019	Known
FGFR2	FGF8	2FDB	R	M	Chr. 10: 123,279,674–123,279,677	10.288	0.413	0.036	Known
FBXW7	SKP1	2OVR	B	C	Chr. 4: 153,249,384–153,249,385	9.352	0.346	0.036	Known
FGFR2	FGF2	1EV2	H	A	Chr. 10: 123,279,674–12,327,9677	11.685	0.406	0.037	Known

[a]Whether the SMR-harboring protein (i) corresponds to a known or new cancer driver gene. [b]Multiple component partner proteins identified.

region positions −1 and −3), suggesting a role in translational control. Mutations in this SMR were validated in the whole-genome sequences of seven cancer types, including two of 20 bladder cancers, two of 40 lung adenomas and three of 172 breast cancers[3,20]. Bladder tumors with mutations in this SMR displayed altered RPS6KA1 (p90RSK) phosphorylation ($P = 0.0005$, *t* test, Benjamini-Hochberg), a signal of increased cell cycle proliferation[40], and altered α-tubulin levels ($P = 4.3 \times 10^{-5}$, *t* test, Benjamini-Hochberg), as determined by reverse-phase protein array (RPPA) assays[41] (**Fig. 2g** and Online Methods). These results establish the usefulness of whole-exome sequencing data for identifying recurrently mutated noncoding regions and our SMR identification method in pinpointing potentially functional noncoding alterations in cancer.

SMRs permit high-resolution analysis of coding alterations

As expected, most exome-derived SMRs lay within protein-coding regions. The identification of SMRs across multiple cancer types permitted a systematic analysis of differential mutation frequencies with subgenic and cancer type resolution. Although many protein domains showed high burdens of somatic alteration in multiple cancers, protein domains can show remarkable cancer type specificity in burdens of alteration, as exemplified by VHL in kidney clear cell carcinoma and SET in diffuse large B cell lymphoma (**Fig. 3a**).

Among genes ($n = 94$) with multiple SMRs, we detected 48 SMRs that were differentially mutated across cancer types (**Supplementary Table 8**). A striking example of this differential targeting occurred within the catalytic subunit of the phosphoinositide 3-kinase subunit PIK3CA (p110α), a key oncoprotein implicated in a range of human cancers[42,43]. We detected six SMRs in *PIK3CA* across eight cancer types (**Fig. 3b**), with multiple cancer types displaying SMRs mapping to the helical (PIK3CA.5) and kinase (PIK3CA.6) domains. In contrast, we observed cancer type–specific SMRs (PIK3CA.2 and PIK3CA.3) affecting an α-helical region between the adaptor-binding domain (ABD) and the linker region between the ABD and Ras-binding domain (RBD) of PIK3CA. Up to 14% of uterine corpus endometrial carcinomas harbored alterations in these intron-separated SMRs, although these regions were not highly recurrently altered in other cancers. For example, we observed significant ($q = 1.2 \times 10^{-16}$, proportions test) differences in PIK3CA.2 alteration frequencies in endometrial and breast cancers (**Fig. 3b**) and further validated these differences ($P = 0.02$, proportions test) in whole-genome sequences[3,20]. These findings indicate that previously described differences[44] in total *PIK3CA* mutation frequency between endometrial and breast cancers could, in part, be localized to this region.

Although the oncogenic effects of recurrent mutations mapping to the ABD (PIK3CA.1), C2 (PIK3CA.4), helical (PIK3CA.5) and kinase (PIK3CA.6) domains of PIK3CA have been previously described, mutations affecting the ABD-RBD linker region are poorly understood[45–48]. Interestingly, missense alterations within this region were directionally orientated to one side of the α helix ($P = 0.0145$, Rayleigh test), suggesting changes to a molecular interface (**Fig. 3c**). Large-scale molecular dynamics simulations of PIK3CA-PIK3R1 binding indicate that PIK3CA.2 (p.Lys111Glu) and PIK3CA.3 (p.Gly118Asp) substitutions can alter intermolecular salt-bridge patterns at Arg79, which may result in a loss of 1.8 kcal/mol in binding interactions in comparison to wild-type PIK3CA (**Fig. 3d**, Online Methods and **Supplementary Fig. 9**). Taken together, these results suggest a previously unrecognized mechanism of oncogenic alteration in PIK3CA.

To systematically characterize the location of alterations with respect to three-dimensional protein structures, we leveraged structural information from 428 SMR-associated and known cancer genes. We detected 46 proteins with three-dimensional clustering of missense alterations (**Supplementary Table 9**), as exemplified by PIM1, an SMR-associated serine/threonine kinase proto-oncoprotein (**Fig. 3e** and Online Methods). This approach also identified three-dimensional clustering between BRAF V600 and BRAF P-loop SMRs (**Fig. 3f**), regions where alterations have been shown to function through distinct mechanisms[49]. Moreover, we found that BRAF V600 alterations were more frequent in melanoma and colorectal cancers, whereas BRAF P-loop alterations were more common in multiple myeloma and lung adenomas ($P < 0.01$, proportions test). In total, seven of 16 proteins with multiple SMRs displayed significant SMR three-dimensional

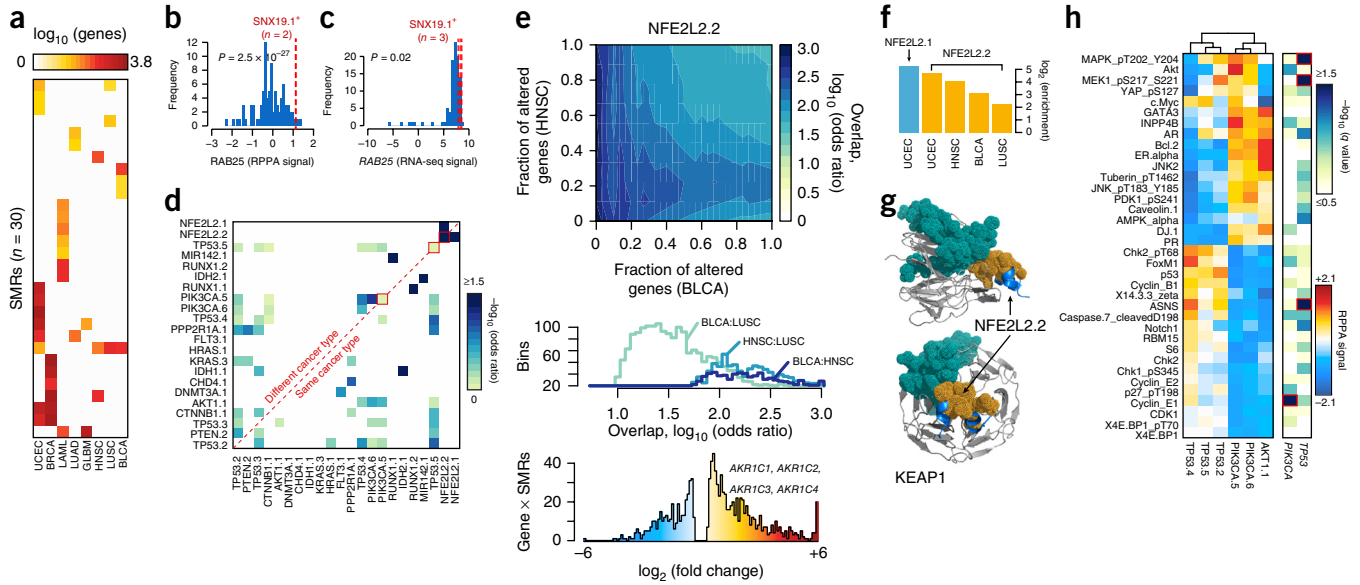

Figure 4 SMRs are associated with distinct molecular signatures. (**a**) Matched RNA-seq data for nine cancers showed that mutations in 30 distinct SMRs associated with ≥10 differentially expressed genes (FDR <5%). (**b**,**c**) Normalized RPPA (**b**) and RNA-seq (**c**) signals for RAB25 are plotted. Red lines represent signals for samples with the mutated *SNX19* SMR. (**d**) Similarity between differentially expressed gene sets associated with mutations in each SMR pair. (**e**) The overlap between differentially expressed genes associated with alteration of the NFE2L2.2 SMR in bladder cancer (BLCA) and head and neck carcinoma (HNSC) is shown (top). Differentially expressed genes are sorted by *P* value, and similarity is quantified by Fisher's exact test odds ratio. The distribution of odds ratios of similarity is summarized for three comparisons (middle). Samples with NFE2L2.2 mutations exhibit highly increased expression of aldo-keto reductase enzymes (bottom). (**f**) Relative enrichment for oxidoreductase activity (GO:0016616) in specific cancer types (**Supplementary Table 13**). (**g**) Structure of the NFE2L2.2 SMR (orange) in the KEAP1-binding domain (PDB, 3WN7). A sector of recurrent alterations in KEAP1 (teal) did not pass our 2% frequency cutoff. (**h**) Patients with breast cancer were grouped by mutations in six SMRs in *PIK3CA*, *AKT1* and *TP53*. Normalized RPPA-based expression data were obtained from the TCPA[41]. The median RPPA signal for 36 markers and the *q* value (Kruskal-Wallis test) of differential expression between SMRs from *TP53* or *PIK3CA* are plotted (red highlights markers with significant intragenic differences, $q < 0.05$).

clustering (**Supplementary Table 10**), which is consistent with frequent spatial coherence for pathogenic alterations.

We next sought to identify SMRs that might affect the molecular interfaces of protein-protein and DNA-protein interactions, a recognized yet understudied mechanism of cancer driver mutation[50–52]. We examined intermolecular distances between SMR residues and interacting proteins or DNA and identified 17 SMRs that likely alter molecular interfaces (**Table 1** and Online Methods). These included 15 molecular interfaces of protein-protein and DNA-protein interactions with established cancer associations, such as the substrate-binding cleft of SPOP[53] and DNA-binding interfaces on RUNX1 (**Fig. 3g**). We detected reciprocal SMRs at all electrostatic interfaces of the SMAD2-SMAD4 heterotrimer in colorectal cancer (**Fig. 3h**), as have been recently described[54], and reciprocal SMRs at the regulatory PIK3CA-PIK3R1 interface in endometrial cancer (**Fig. 3b**). Taken together, these results highlight the robustness of SMRs in detecting validated driver alterations at molecular interfaces (**Supplementary Fig. 10**). In addition, SMRs pinpoint recurrent alterations at the interface between histone H3.1 (**Fig. 3i**) and TRIM33, an E3 ubiquitin ligase, and at the DNA-protein interface of histone H2B (**Supplementary Fig. 11**). These findings underscore and extend recent associations between altered epigenetic regulation and histone alterations in tumorigenesis[55].

Molecular signatures highlight impact of SMR alterations

We sought to determine the potential functional impact of SMR alterations by their association with molecular signatures. We leveraged RNA sequencing (RNA-seq), RPPA and clinical data to ask whether (i) SMR alterations associate with distinct molecular signatures or survival outcomes, (ii) SMR alterations correlate with similar molecular profiles in distinct cancers, and (iii) SMR alterations in the same gene associate with similar or different molecular signatures.

We found that mutations in SMRs were associated with diverse changes in RNA expression, signaling pathways and patient survival (**Fig. 4a**, Online Methods and **Supplementary Tables 11–14**)[56]. These analyses identified previously unappreciated connections between recurrent somatic mutations and molecular signatures, which highlight recurrent GSK3 pathway alterations in endometrial cancer and recurrent mTOR as well as EIF4 and epidermal growth factor (EGF) pathway alterations in glioblastoma (**Supplementary Table 15**). For example, synonymous point mutations in a bladder cancer SMR in *SNX19* (encoding sorting nexin 19) were associated with significant increases in the protein expression levels of RAB25 ($P = 2.5 \times 10^{-27}$, *t* test; **Fig. 4b** and **Supplementary Table 12**), a RAS family GTPase that promotes ovarian and breast cancer progression[57,58]. These increases are consistent with RNA expression differences in *RAB25* ($P = 0.02$, Wilcoxon rank-sum test; **Fig. 4c**). Intriguingly, both SNX19 and RAB25 are implicated in intracellular trafficking, but the mechanism by which synonymous mutations in *SNX19* correlate with RAB25 expression remains to be determined. In both *SNX19* and *NDUFA13*, SMRs with clusters of synonymous mutation overlapped open chromatin sites[28], suggesting potential regulatory effects.

We identified concordant changes in gene expression for SMR pairs, suggesting potential functional relationships, for 23 SMRs from 17 genes (**Fig. 4d**). These included multiple well-established mechanistic relationships, many of which were supported by RPPA measurements[41], such as a relationship between *PIK3CA* and *AKT1*. Furthermore, this analysis indicated that mutations in the same SMR in different cancers can elicit similar molecular profiles in distinct cancers. For instance,

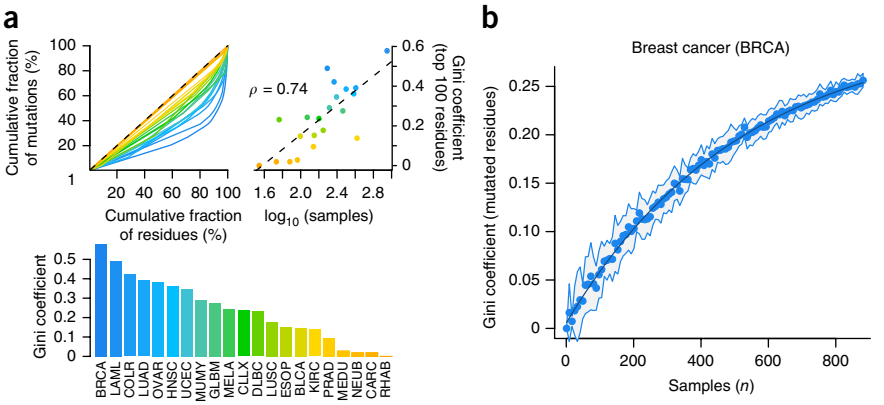

Figure 5 Structure in the distribution of cancer mutations remains largely uncharacterized. Gini coefficients of dispersion were calculated as the fraction of nonsynonymous mutations contained per residue, across ~19,000 proteins. (**a**) Lorenz curves (top left), Gini coefficients (bottom) and the correlation of these coefficients with tumor sample numbers (top right) are shown. (**b**) Gini coefficients of nonsynonymous mutation frequency in breast cancer as a function of (bootstrapped) sample size. The line of exponential fit is shown in dark blue. For comparisons between cancer types in **a**, the Gini coefficients were computed exclusively on the 100 most mutated residues per cancer.

we found that SMR alterations in the oncogenic transcription factor *NFE2L2* (ref. 59) were associated with large, concordant transcriptomic changes in four distinct cancer types (bladder, endometrial, lung squamous cell carcinoma, and head and neck cancers; **Fig. 4e**). The four genes with the highest increases in gene expression among endometrial cancer samples with alterations in *NFE2L2.1* were the aldo-keto reductases *AKR1C1–AKR1C4* (**Fig. 4e**), which contribute to altered androgen metabolism and have been implicated in multiple cancer types[60–62]. Across all four cancer types, genes with expression changes in patients with mutations in *NFE2L2* SMRs were highly enriched for oxidoreductases acting on the CH-OH group of donors with NAD^+ or NADP as acceptors (4.9–39.0×; $P \leq 0.001$, Benjamini-Hochberg; **Fig. 4f**). Mutations in *KEAP1*, encoding an NFE2L2 binding partner, recapitulated the expression changes observed in patients with mutations in *NFE2L2* SMRs ($P < 0.01$, Benjamini-Hochberg; **Fig. 4g** and **Supplementary Fig. 12**).

The identified SMRs also permitted interrogation of mutations in different regions of a given gene with respect to associated molecular signatures. For example, in breast cancer, alterations in distinct SMRs within *TP53* were associated with highly similar changes in protein levels. Yet, we observed SMR-specific differences in ASNS levels and MAPK and MEK1 phosphorylation among samples with altered *TP53* SMRs ($q < 0.01$; **Fig. 4h**). These results establish differences in the molecular signatures associated with alterations of SMRs in the same gene and are consistent with pleiotropy in established oncogenes and tumor suppressors[63,64].

The structure of cancer mutations remains largely unseen

SMR analysis leverages structure in the distribution of somatic driver mutations to identify cancer-associated regions. We sought an alternative metric to assess structure in the distribution of the somatic coding mutations analyzed here by measuring the Gini coefficient of amino acid substitutions per residue in each cancer type (**Fig. 5a**). Gini coefficients of dispersion were well correlated with sample numbers (Spearman's $\rho = 0.74$). Subsampling demonstrated that, even with sample numbers >850, a large proportion of the structure of protein-altering mutations in breast cancer remains unseen (**Fig. 5b**). These findings highlight the value of increasing cancer sample sizes in assessing the landscape of driver mutations.

DISCUSSION

With few exceptions, studies of disease-associated variation have focused on identifying predefined functional units with recurrent alterations. This approach not only assumes accurate annotations but ignores the largely uncharacterized spectrum of functional elements that may be the targets of pathological variants. Our approach avoids these limitations and complements existing gene-level and pathway-based strategies for discovering cancer drivers by identifying variably sized SMRs (**Supplementary Table 16**). SMR-associated genes include known cancer-related genes, such as *PIM1* and *MIR142*, that were missed by gene-level analyses, as well as multiple genes with potentially novel roles in cancer development.

Cancer-associated SMRs target a diverse spectrum of functional elements in the genome, including single amino acids, complete coding exons and protein domains, miRNAs, 5′ UTRs, splice sites and transcription factor binding sites, among others. This functional diversity underscores both the varied mechanisms of oncogenic misregulation and the advantage of functionally agnostic detection approaches. Notably, several of the most frequently altered SMRs lay within noncoding regions. Strikingly, 17 of 39 promoter and 5′ UTR melanoma SMRs overlap the core recognition sequences of *in vivo* ETS family binding sites (odds ratio = 15.2, $P = 1.5 \times 10^{-11}$, Fisher's exact test). In addition, ~15% of patients with bladder cancer harbor 5′ UTR alterations in *TBC1D12*. Together, these results extend the support for noncoding drivers in cancer[20,23,65] and establish the potential for discovering noncoding variation in whole-exome sequencing.

The identification of SMRs provides a subgenic, cancer type–specific analysis of somatic mutations and associated molecular signatures. Differences among cancer types in SMR mutation frequencies within *BRAF*, *EGFR* and a mechanistically uncharacterized α helix in PIK3CA demonstrate substructure in the distribution of somatic mutations across cancers, a property that may arise from pleiotropic functions. The close geometric proximity and directional uniformity of alterations in the PIK3CA helix suggest that mutations in the PIK3CA.2 and PIK3CA.3 SMRs function through similar mechanisms. Moreover, biophysical simulations indicate that mutations in both SMRs result in elevated basal signaling activity of catalytic PIK3CA by way of weakened interactions with the regulatory PIK3R1 protein. These findings are concordant with recent biochemical evidence[48]. Consistent with pleiotropic dependencies, alterations to SMRs within a single gene can be associated with distinct molecular signatures, as exemplified by *TP53* SMRs in breast cancers. Together, these results provide robust support for subgenic functional targeting in distinct cancers and genes, and future efforts to examine SMR mutations in conjunction with clinical data in much larger patient cohorts may permit assessment of the prognostic value of SMRs.

SMR detection would benefit from further improvements of somatic mutation models. Here we have applied cancer type–specific models that take into account variation in somatic mutation rates throughout the genome. We controlled for mutational effects stemming from

differences in replication timing and gene expression[4,66]. In addition, our models capture nucleotide-specific mutation probabilities[3], account for strand specificity[67], leverage whole-genome sequencing mutation frequencies to limit effects from purifying selection on exons and control mutation processes that may result in mutation clustering and trinucleotide mutation biases[3]. However, tumor-specific DNA repair defects[3,66,68,69] and cell type–specific chromatin context[70] also contribute to somatic mutation rates. Mutation models that account for cell type–specific expression and chromatin context at refined scales may require sequencing cohorts of matched normal tissue and increased sample sizes.

Although the sequencing of additional cancer genomes will likely further the identification of new cancer driver genes[5], characterizing the biochemical and cellular consequences of individual mutations is critical. We demonstrate that identifying the spatial distribution of mutation recurrence in the genome, when combined with additional genomic, biophysical, structural or phenotypic information, often enhances mechanistic insights. Applying recently developed high-throughput approaches[71–73] to directly interrogate variation within SMRs may allow further understanding of the molecular mechanisms driving cancer and facilitate diagnostics and therapeutics development.

URLs. Data from Lawrence *et al.*[5] were obtained from TumorPortal through http://www.tumorportal.org/. TCGA Data Portal, https://tcga-data.nci.nih.gov/tcga; UCSC Cancer Browser, http://genome-cancer.ucsc.edu/.

METHODS

Methods and any associated references are available in the online version of the paper.

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

ACKNOWLEDGMENTS

We thank the TCGA, ICGC and TCPA for making these large-scale cancer data sets available to the scientific community. We thank H. Tang for discussions regarding statistical analyses. We thank M.M. Winslow, D.M. Fowler, S. Fields and D.E. Webster for critical reading and suggestions to the manuscript. C.L.A. was supported by US National Institutes of Health (NIH) grants 3U54DK10255602 and 1P50HG00773501. C.C. was supported by the Child Health Research Institute, the Lucile Packard Foundation for Children's Health and US NIH Clinical and Translational Science Award grant UL1TR000093. J.A.R. was supported by the Damon Runyon Cancer Research Foundation and US NIH award 1U01HG007919-01. G.K. acknowledges support from the Lawrence Scholars Program, the US NIH Simbios Program (U54GM072970) and the Center for Molecular Analysis and Design at Stanford University. Biophysical simulations were supported by the Blue Waters project via US National Science Foundation awards OCI-0725070 and ACI-1238993 and the state of Illinois. Further support was provided by the National Center for Multiscale Modeling of Biological Systems (P41GM103712-S1) through Anton-1 resources provided by the Pittsburgh Supercomputing Center under grant PSCA13072P. This work was supported by the Rita Allen Foundation.

AUTHOR CONTRIBUTIONS

C.L.A. and W.J.G. conceived of the project, and all authors designed experiments and methods. C.L.A. and C.C. developed methods for the detection and analysis of SMRs. C.L.A. constructed uniform annotations and non-Bayesian mutation probability models and performed density-based clustering, scoring and empirical false-discovery estimation (simulations), as well as regulatory (noncoding), structural (coding), frequency and whole-genome sequencing recurrence analyses. C.C. constructed Bayesian mutation probability models and performed RNA-seq, RPPA and survival outcome analyses. J.A.R. designed and performed luciferase assays. G.K. carried out biophysical simulations, performed hidden Markov model–based state decompositions and computed binding enthalpies with supervision from V.S.P. C.L.A., C.C., J.A.R., G.K., M.P.S. and W.J.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

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

1. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
2. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
3. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
4. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
5. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
6. Ding, L., Wendl, M.C., McMichael, J.F. & Raphael, B.J. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* **15**, 556–570 (2014).
7. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
8. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
9. Kane, D.P. & Shcherbakova, P.V. A common cancer-associated DNA polymerase ε mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. *Cancer Res.* **74**, 1895–1901 (2014).
10. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
11. 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).
12. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
13. Schnall-Levin, M., Zhao, Y., Perrimon, N. & Berger, B. Conserved microRNA targeting in *Drosophila* is as widespread in coding regions as in 3′ UTRs. *Proc. Natl. Acad. Sci. USA* **107**, 15751–15756 (2010).
14. Cenik, C. *et al.* Genome analysis reveals interplay between 5′ UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS Genet.* **7**, e1001366 (2011).
15. Stergachis, A.B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367–1372 (2013).
16. Wolfe, A.L. *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**, 65–70 (2014).
17. Xiong, H.Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
18. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
21. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
22. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
23. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
24. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
25. Leiserson, M.D.M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
26. Araya, C.L. *et al.* Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature* **512**, 400–405 (2014).
27. Stergachis, A.B. *et al.* Conservation of *trans*-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).
28. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
29. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)* **6**, 80–92 (2012).
30. Martin, E., Kriegel, H.P., Jörg, S. & Xiaowei, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **96**, 226–231 (1996).
31. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
32. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R. & Cooper, C.S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).

33. Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* **23**, 762–776 (2013).
34. Jäger, D. *et al.* Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library. *Cancer Res.* **61**, 2055–2061 (2001).
35. Mei, Y.-P. *et al.* Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* **31**, 2794–2804 (2012).
36. Okugawa, Y. *et al.* Clinical significance of *SNORA42* as an oncogene and a prognostic biomarker in colorectal cancer. *Gut* http://dx.doi.org/10.1136/gutjnl-2015-309359 (15 October 2015).
37. Budinska, E. *et al.* Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **231**, 63–76 (2013).
38. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
39. Vejnar, C.E. & Zdobnov, E.M. MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.* **40**, 11673–11683 (2012).
40. Lara, R., Seckl, M.J. & Pardo, O.E. The p90 RSK family members: common functions and isoform specificity. *Cancer Res.* **73**, 5301–5308 (2013).
41. Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).
42. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
43. Thorpe, L.M., Yuzugullu, H. & Zhao, J.J. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer* **15**, 7–24 (2015).
44. Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
45. Miled, N. *et al.* Mechanism of two classes of cancer mutations in the phosphoinositide 3-kinase catalytic subunit. *Science* **317**, 239–242 (2007).
46. Huang, C.-H. *et al.* The structure of a human p110α/p85α complex elucidates the effects of oncogenic PI3Kα mutations. *Science* **318**, 1744–1748 (2007).
47. Gkeka, P. *et al.* Investigating the structure and dynamics of the *PIK3CA* wild-type and H1047R oncogenic mutant. *PLoS Comput. Biol.* **10**, e1003895 (2014).
48. Burke, J.E., Perisic, O., Masson, G.R., Vadas, O. & Williams, R.L. Oncogenic mutations mimic and enhance dynamic events in the natural activation of phosphoinositide 3-kinase p110α (*PIK3CA*). *Proc. Natl. Acad. Sci. USA* **109**, 15259–15264 (2012).
49. Haling, J.R. *et al.* Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* **26**, 402–413 (2014).
50. Kar, G., Gursoy, A. & Keskin, O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput. Biol.* **5**, e1000601 (2009).
51. Ghersi, D. & Singh, M. Interaction-based discovery of functionally important genes in cancers. *Nucleic Acids Res.* **42**, e18 (2014).
52. Cheng, F. *et al.* Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol. Biol. Evol.* **31**, 2156–2169 (2014).
53. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
54. Fleming, N.I. *et al.* *SMAD2*, *SMAD3* and *SMAD4* mutations in colorectal cancer. *Cancer Res.* **73**, 725–735 (2013).
55. Yuen, B.T.K. & Knoepfler, P.S. Histone H3.3 mutations: a variant path to cancer. *Cancer Cell* **24**, 567–574 (2013).
56. Hornbeck, P.V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–D270 (2012).
57. Cheng, K.W. *et al.* The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. *Nat. Med.* **10**, 1251–1256 (2004).
58. Zhang, J. *et al.* Overexpression of Rab25 contributes to metastasis of bladder cancer through induction of epithelial-mesenchymal transition and activation of Akt/GSK-3β/Snail signaling. *Carcinogenesis* **34**, 2401–2408 (2013).
59. DeNicola, G.M. *et al.* Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* **475**, 106–109 (2011).
60. Ji, Q. *et al.* Selective loss of AKR1C1 and AKR1C2 in breast cancer and their potential effect on progesterone signaling. *Cancer Res.* **64**, 7610–7617 (2004).
61. Stanbrough, M. *et al.* Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer Res.* **66**, 2815–2825 (2006).
62. Rižner, T.L., Šmuc, T., Rupreht, R., Šinkovec, J. & Penning, T.M. AKR1C1 and AKR1C3 may determine progesterone and estrogen ratios in endometrial cancer. *Mol. Cell. Endocrinol.* **248**, 126–135 (2006).
63. Zhao, L. & Vogt, P.K. Helical domain and kinase domain mutations in p110α of phosphatidylinositol 3-kinase induce gain of function by different mechanisms. *Proc. Natl. Acad. Sci. USA* **105**, 2652–2657 (2008).
64. Wu, X. *et al.* Activation of diverse signalling pathways by oncogenic *PIK3CA* mutations. *Nat. Commun.* **5**, 4961 (2014).
65. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
66. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
67. Reijns, M.A.M. *et al.* Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).
68. Lord, C.J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012).
69. Roberts, S.A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
70. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
71. Araya, C.L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA* **109**, 16858–16863 (2012).
72. Buenrostro, J.D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
73. Guenther, U.-P. *et al.* Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature* **502**, 385–388 (2013).

ONLINE METHODS

Scientific computing was performed within Python[74,75] and R environments. Data structure and genomic interval operations were performed with PANDAS[76] and Pybedtools[77], respectively. Statistical computing was performed with SciPy and NumPy[78], and machine learning methods were implemented with SciKit Learn[79]. Structural and sequence alignment analyses were performed with BioPython[80], PyMOL (Schrödinger) modules and custom scripts. RPPA, RNA-seq and survival analyses were performed in R using open source packages (as described below).

Uniform variant annotation. 3,185,590 uniformly processed[5] whole-exome sequencing somatic variant calls from 21 cancer types were downloaded from the indicated URL. We applied snpEff[29] to uniformly annotate n = 3,078,482 (96.6%) SNV calls from 4,735 tumors, recording (GRCh37.66) mutation impact in protein-coding regions, transcribed regions (coding plus noncoding exons, introns, 5′ UTRs and 3′ UTRs) and gene-associated regions (transcribed 5 kb upstream and 5 kb downstream) and standardize gene name assignments. These procedures standardized gene name assignments at multiple scales and removed gene assignments to "?" (n = 64) and "—" (n = 130,728) in the original file. In addition, this procedure reduced variant calls unassigned to any genes ("Unknown", n = 1,239,475) to n = 899,731 intergenic calls (>5 kb from annotated exons). This procedure was also applied to annotate n = 11,461,951 whole-genome sequencing somatic SNV calls from 23 cancer types[3,20].

Mutation probability models. For each tumor type and gene, we calculated multiple distinct mutation probabilities. First, we calculated the frequency of transitions and transversions within the mappable, exonic regions of each gene to derive 'exonic' mutation probabilities for each gene in the hg19 human genome assembly using whole-exome sequencing data. Specifically, these probabilities indicate the fraction of mappable (100-bp), exonic reference bases (for example, adenines) in each gene that were somatically mutated to a specific base (for example, cytosine) per sample, in the cohort of tumor-specific whole-exome sequencing data.

Because expression levels and replication timing have been shown to be major covariates of somatic mutation probability in the genome, we sought to refine our mutation probability models for each gene using this information. For each gene and in each tumor type, we identified the set of genes most similar in expression, replication time and GC content (gene-level features). We used previously compiled[4] expression and replication timing data and derived feature-specific weights defined as the rank correlation between gene features and the observed exonic mutation probabilities in each tumor type. We then converted gene features into their percentile ranks. Genes were sorted sequentially on the basis of the gene feature weights, and the neighborhood of the 500 closest genes was selected for each query gene. We then measured the sum of correlation-weighted, absolute feature distances between gene pairs within the 500-gene rank neighborhood. For each gene, we selected the ≤200 most similar genes with a normalized distance score ≤1. Lastly, we averaged the 'exonic' mutation probability per transition/transversion to derive a set of 'matched' mutation probabilities.

To avoid skewed mutation probabilities due to increased selection pressure on exons, we used pan-cancer whole-genome sequencing[3,20] data in conjunction with cancer-specific whole-exome sequencing data. We employed a Bayesian framework to derive posterior mutation probabilities for each transition and transversion per gene in each of the analyzed cancer types. Specifically, we modeled the likelihood of observing a mutation as a binomial distribution. We placed a prior beta distribution on the mutation probability for each mutation type. The prior distribution was parameterized with parameters $\alpha = \mu \times \nu$ and $\beta = (1 - \mu) \times \nu$, where μ is the per-base mutation probability in the whole-exome sequencing data and ν is the number of exome sequencing samples in each cancer type. This parameterization enables the variance of the prior distribution to scale inversely with sample size. We used the set of genes (≤200) that were matched to the analyzed gene as described above. We used all observed intronic whole-genome sequencing mutations in this cancer-specific matched set to calculate the posterior mutation probability for the analyzed gene. In this framework, the posterior distribution is also another beta distribution. We then assigned the expected value of the posterior probability distribution as the estimate of the mutation probability for each transition/transversion (n = 12). Finally, we calibrated the posterior mutation probabilities by the cancer-specific transition/transversion rates such that the median 'Bayesian' mutation probability was equal to the mean cancer-specific 'exonic' mutation rate.

We computed a 'global' mutation probability per tumor type as the average probability of transitions and transversions across all genes as observed in 'exonic' mutation probabilities in each tumor type. The distributions of whole-exome sequencing–derived ('exonic', 'matched' and 'global') as well as whole-genome sequencing–derived ('Bayesian') mutation probabilities varied strongly between cancer types (**Supplementary Fig. 2a**) and among genes within individual cancer types, highlighting the importance of such cancer- and gene-specific treatment of background mutation probabilities[3,4]. Complementary mutation probabilities were well correlated (**Supplementary Fig. 2b**). The 'Bayesian' and 'matched' mutation probabilities were well correlated among genes (**Supplementary Fig. 2c**), although 'Bayesian' mutation probabilities were better correlated (**Supplementary Fig. 2d**) with the observed whole-genome sequencing intronic mutation densities. These 'Bayesian' (whole-genome sequencing–based) and 'matched' (whole-exome sequencing–based) mutation probabilities were used for the comparison presented in **Figure 1f**.

Lastly, to account for trinucleotide biases[3,4] in diverse mutation processes and cancer types, we computed 'trinucleotide' mutation probability models for each tumor type. Specifically, 'trinucleotide' mutation probabilities were calculated as the fraction of mappable (100-bp), exonic reference bases (for example, adenines) within specific trinucleotide contexts (for example, CAG) that were somatically mutated to a specific base (for example, cytosine, CAG>CCG) per sample, in the cohort of tumor-specific whole-exome sequencing data.

Mutation domain definition. We extended Ensembl (75) exonic regions by 0 bp and 1,000 bp and merged regions to define n = 305,145 'concise' (C) and n = 191,669 'expanded' (E) genomic domains in which mutation clusters were evaluated (see below). We identified the n = 279,979 'concise' and n = 175,228 'expanded' domains in which over ≥90% of positions were fully mappable with single-end 100-bp reads (ENCODE, UCSC Genome Browser). For each set of domains, we computed the number of possible genomic ranges (start, stop), which for the 'expanded' set amounted to 1,005,774,400,023 ranges ($10^{12.0025}$). In addition, we removed 'blacklisted' regions of the human genome previously defined by the ENCODE Project[81].

Mutator sample identification. Samples harboring aberrantly high burdens of mutations in each tumor type were detected using median absolute deviation (MAD) outlier detection on the distribution of mutations (log n) per sample. As a threshold for consistency, mutator (outlier) samples were selected as those exceeding 2 s.d.

Mutation cluster identification. We deployed density-based spatial clustering of applications with noise (DBSCAN) to detect clusters of ≥2 SNVs within exonic domains (above), evaluating density reachability within ε base pairs in each cancer type. The reachability parameter ε was dynamically defined with $\varepsilon = d_p/d_s$ where d_p and d_s refer to the number of mutated positions (base pairs) and the base-pair size of the domain d, thresholded to $10 \leq \varepsilon \leq 500$ bp. In contrast to sliding window approaches or k-means spatial clustering, DBSCAN is not confined to evaluating predefined cluster sizes or numbers and tolerates noise in spatial density, whereby distal mutations are not assigned to clusters. Detected mutation clusters were refined where subclusters of ≥2 SNVs with significantly higher ($P < 0.05$, binomial test) mutation densities (mutated tumor samples per kilobase) existed.

Mutation cluster scoring. The significance of the observed mutation densities in each cluster was determined as Fisher's combined binomial probability of sampling the observed number (k) or more mutations for each mutation type within the region. For each region, we computed the above density scores with the previously described 'exonic', 'matched', 'Bayesian' and 'global' somatic mutation probabilities. As the primary density score (P_{density}), we selected the most conservative of the 'Bayesian' and 'global' density scores, $\max(P_{\text{Bayesian}}, P_{\text{global}})$.

doi:10.1038/ng.3471

Finally, we computed a trinucleotide mutation density score ($P_{trinucleotide}$) for each region using the 'trinucleotide' somatic mutation probabilities.

Mutation cluster thresholding. We applied the procedures above to detect and evaluate mutation clusters in two sets of 'concise' (C) and 'expanded' (E) query domains (described in "Mutation domain definition"). 117,148/198,718 of the mutation clusters identified in E query domains fell within the C query domains, respectively, indicating a 1.7× increase in clusters within the 1,000 bp–expanded domains.

Empirical FDRs were calculated from ten simulations performed by randomizing mutations within C domains in each tumor type, simulating a total of 30,784,820 mutations across cancer types. In each simulation, the positions of the observed mutations in each domain and tumor type were randomized while maintaining reference base identity to retain the observed 'global' mutation probabilities per transition and transversion ($n = 12$). In each iteration, mutation cluster detection, refinement and scoring procedures were repeated as above. For each simulation, we computed the density score ($P_{density}$) threshold that guarantees FDR ≤5%, whereby false and true discoveries are computed as the number of clusters from simulated (randomized) and observed domain mutations, respectively. We excluded clusters with outlier density scores from the false-discovery set if the clusters were associated with CGC genes ($n = 522$)[31,32], as these regions would not represent false discoveries. For each tumor type, the expectation value (average) of FDR ≤5% simulation thresholds was defined as the final tumor-specific FDR threshold. To control FDRs to ≤5% in the E domains, where mutations cannot be randomized owing to the decreased certainty of whole-exome sequencing coverage, we adjusted FDRs from C domains by the 1.7× increase in E/C clusters in each tumor type. E domain 5% FDR thresholds per tumor type are provided in **Supplementary Table 1**.

To assess the robustness of the FDR cutoffs, we expanded the number of simulations to 90× and confirmed a 99.2% overlap (Jaccard index) in the 5% FDR–thresholded clusters (**Supplementary Fig. 4e–g**).

We reiterated mutation cluster FDR estimation and filtering using an alternate, conservative density score, $P_{alternate} = \max(P_{matched}, P_{global})$, resulting in 714 regions. Fully 93.2% of these regions were identified as SMRs on the basis of the primary density scores ($P_{density}$).

Mutation cluster filtering. As a final step in calling SMRs, we selected clusters with density scores ($P_{density}$) at the 5% FDR threshold and that were mutated in ≥2% of samples in each cancer type. Lastly, clusters associated with pseudogenes, olfactory receptors and other repetitive gene classes were removed. This procedure resulted in 872 SMRs, from 735 unique genomic regions, in 20 distinct cancer types.

Mutation cluster annotation. SMRs were annotated on the basis of mutation impact on coding, transcribed and gene-associated regions (see "Uniform variant annotation"). For SMRs associated with multiple genes (overlapping annotations), we preferentially assigned SMRs to (i) previously known cancer driver genes (as defined by Lawrence *et al.* or the CGC) or (ii) the gene affected by the most severe type of mutation. Where mutation impact was insufficient to resolve assignment to multiple genes, we selected the gene affected by the largest number of mutations within the SMR. On this basis, we assigned each SMR to a single gene, recording the types of mutation impacts on the gene and the class of region affected. Region classes included exon (coding region and noncoding gene), intron, splice, upstream, 5′ UTR, 3′ UTR, downstream and other (intergenic). Mutation impacts (from snpEff) included, in order of severity, rare amino acid, splice site acceptor, splice site donor, start lost, stop lost, stop gain, nonsynonymous coding, splice-site branch U12, nonsynonymous start, nonsynonymous stop, splice-site region, splice-site branch, start gain, synonymous coding, synonymous start, synonymous stop, noncoding gene (exon), 3′ UTR, 5′ UTR, miRNA, intron, upstream, downstream and intergenic.

Mutation cluster classification. SMRs were classified into high-, medium- and low-confidence sets as follows. First, SMRs in which alterations fell below the 2% mutation frequency threshold following mutator sample (as defined above) removal were labeled as mutator driven SMRs. Among SMRs robust to mutator removal, those with FDR-corrected density scores significant at adjusted $P < 0.05$ following Bonferroni correction ($P_{density} \leq 5.2 \times 10^{-17}$) were classified as high confidence. Mutator-driven SMRs were classified as low confidence. SMRs that did not meet the high-confidence or low-confidence criteria were deemed medium confidence.

To control for unaccounted mutation processes that could result in clusters of mutations with no selective advantage in cancer, we introduced the assumption that intronic mutations are primarily composed of passenger mutations and treated intronic clusters as false discoveries. For each cancer type, the distribution of density scores from intronic mutation clusters was modeled with Gaussian kernel density estimation (KDE) to derive *P*-value and *q*-value (FDR) estimates that limit the FDR to ≤5%. This approach is limited to the ten cancer types with sufficient intronic mutation clusters to permit KD estimates of their distribution of mutation density scores (**Supplementary Fig. 5**). A threshold of $n \geq 100$ intronic mutation clusters was determined on the basis of the stability of FDR thresholds as determined by subsampling intronic mutation clusters in melanoma (data not shown). We applied this approach to control FDRs on two metrics. First, to account for unaccounted mutation clustering, we applied this approach on our expression-, replication timing– and sequence (GC) composition–controlled single-nucleotide probabilities ($P_{density}$). Second, to account for biases in trinucleotide mutation frequencies in each cancer type, we applied this approach on trinucleotide density scores ($P_{trinucleotide}$). SMRs discovered in multiple cancer types and non-mutator-driven SMRs compliant with intron-based FDR ≤ 5% thresholds (both $P_{density}$ and $P_{trinucleotide}$) were classified as 'robust'.

Mutation cluster labeling. SMRs with higher than expected prevalence for APOBEC mutation signatures[69] were labeled (**Supplementary Fig. 6d**). Finally, we annotated SMRs with respect to their 35-bp uniqueness and alignability with 50-, 75- and 100-bp single-end reads. SMR coordinates and corresponding annotations are provided in **Supplementary Table 2**.

Mutation trinucleotide analysis. We evaluated the frequency of trinucleotide sequence contexts, as a subset of these (TCW) have been previously shown to differ significantly in mutation frequencies from other single-nucleotide contexts owing to APOBEC mutational processes[69]. Although APOBEC mutation signatures are identifiable in the data, our SMRs are depleted for such signatures (**Supplementary Fig. 6a**), suggesting that the background models conservatively control for this mutation signature. Moreover, we extended these analyses to examine two important metrics: (i) unaccounted trinucleotide biases measured as the deviation in the observed trinucleotide mutation frequencies on the basis of single-nucleotide frequencies and (ii) fold change in frequencies of trinucleotide contexts in the SMR mutations as compared to the input mutations.

We observed a low correlation between the unaccounted trinucleotide biases and the fold change in trinucleotide contexts in diverse cancer types (**Supplementary Fig. 6b**), further supporting the conclusion that SMRs are not driven by unaccounted trinucleotide mutation signatures. These analyses were restricted to cancer types ($n = 6$) that had ≥250 SMR mutation sites to prevent noise from cancer types with low numbers of SMR mutations. These cancer types encompassed 79% of SMRs. Across cancer types, unaccounted trinucleotide frequencies made up only ~7.9% of SMR sequences. For completeness, we have calculated within each SMR the fraction of mutations that are consistent with APOBEC signatures (**Supplementary Fig. 6c**). As shown in **Supplementary Figure 6d**, only 4% of SMRs had higher than expected APOBEC mutation signatures following Holmes-Bonferroni correction. Raw (uncorrected) *P* values would indicate that 12% of SMRs have higher than expected APOBEC mutation signatures.

For additional methods describing (i) transcription factor motif enrichments, (ii) protein structure mapping, (iii) mutation spatial clustering, (iv) mutation dihedral angles, (v) molecular dynamics of PIK3CA-PIK3R1 binding, (vi) RNA-seq analysis, (vii) RPPA analysis, (viii) functional enrichment analysis, (ix) survival analysis, (x) miRNA target site analysis and (xi) luciferase assays, please see the **Supplementary Note**.

Code availability. The Python and R scripts to process the data and conduct the analyses described herein are available from the authors by request. Standalone scripts for molecular and transcriptome enrichment overlap visualization are available at http://www.github.com/claraya/SMRx.

74. Oliphant, T.E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).
75. Millman, K.J. & Aivazis, M. Python for scientists and engineers. *Comput. Sci. Eng.* **13**, 9–12 (2011).
76. McKinney, W. in *Proc. 9th Python Sci. Conf.* (eds. van der Walt, S. & Millman, J.) 51–56 (2010). ISBN-13: 978-1-4583-4619-3.
77. Dale, R.K., Pedersen, B.S. & Quinlan, A.R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
78. Van der Walt, S., Colbert, S.C. & Varoquaux, G. The NumPy Array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
79. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
80. Cock, P.J.A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
81. Boyle, A.P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).

doi:10.1038/ng.3471