Genome-wide analysis of noncoding regulatory mutations in cancer

Nils Weinhold1,4, Anders Jacobsen1,2,4, Nikolaus Schultz1, Chris Sander1 & William Lee1,3

Cancer primarily develops because of somatic alterations in the genome. Advances in sequencing have enabled large-scale sequencing studies across many tumor types, emphasizing the discovery of alterations in protein-coding genes. However, the protein-coding exome comprises less than 2% of the human genome. Here we analyze the complete genome sequences of 863 human tumors from The Cancer Genome Atlas and other sources to systematically identify noncoding regions that are recurrently mutated in cancer. We use new frequency- and sequence-based approaches to comprehensively scan the genome for noncoding mutations with potential regulatory impact. These methods identify recurrent mutations in regulatory elements upstream of PLEKHS1, WDR74 and SDHD, as well as previously identified mutations in the TERT promoter. SDHD promoter mutations are frequent in melanoma and are associated with reduced gene expression and poor prognosis. The non-protein-coding cancer genome remains widely unexplored, and our findings represent a step toward targeting the entire genome for clinical purposes.

Large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA)1 and the International Cancer Genome Consortium (ICGC)2 have expended substantial effort characterizing the cancer genome. So far, these projects have put their focus on genomic variation in the coding sequences of tumor genomes and have identified a number of new alterations, such as recurrent mutations affecting the exonuclease domain of DNA polymerase ε (refs. 3,4). As most studies rely heavily on targeted exome sequencing, the understanding of somatic variation in coding regions has improved substantially. However, the protein-coding component of the genome accounts for less than 2% of the total sequence, and there is very little information on how noncoding variation affects cancer development. Furthermore, even well-studied cancer types such as non-small-cell lung cancer still have major subpopulations with no observable ‘driver’ mutation5,6.

The Encyclopedia of DNA Elements (ENCODE) Project estimates that roughly 80% of the human genome has some sort of biochemical functionality7. It is well known that somatic mutations in noncoding regions are frequent8, but their effect is poorly understood. Recent efforts to understand noncoding variation in the human population have shown that disease-associated genomic variation is commonly located in regulatory elements9,10. Taking this evidence together, it is reasonable to expect that a substantial portion of the recurrent noncoding somatic mutations observed in cancer could have a regulatory effect. Thus far, the most notable example of this type of variation comes from the recent discovery of mutations in the promoter of the TERT gene11,12. Other computational approaches to systematically characterize noncoding variation have primarily focused on nucleotide conservation9,13. Further progress in this area has been hampered by the relatively high cost of whole-genome sequencing for large numbers of tumor samples, which are necessary to screen various regulatory regions for significant events. Indeed, previous studies on noncoding mutations in cancer have been limited by sample size12,14. The maturation of sequencing technologies now allows the systematic sequencing of whole genomes, and it is thus only now that researchers can begin to appreciate the role that noncoding mutations might have in the formation and development of cancer.

We performed a comprehensive analysis of somatic mutations in the whole-genome sequences of 863 individuals with cancer collected from TCGA and other public sources15 (Fig. 1a and Supplementary Tables 1 and 2). Our approach targets genomic variation in the noncoding part of the genome, which is poorly characterized and rarely implicated in cancer. We called somatic mutations in tumor-normal pairs across the whole genome and annotated the mutations, focusing on those most likely to affect regulatory elements. We then used multiple independent approaches to identify functional noncoding alterations.

RESULTS
Assessing the genomic landscape of noncoding mutations
The genome-wide mutation burden varied among different cancer types (Supplementary Fig. 1), and the trend was generally consistent with previous observations in exome sequencing studies16. In the tumors from TCGA, most genomes had between 1,000 and 50,000 total somatic mutations. Mutations in transcribed regions, including coding sequences, introns, and 3′ and 5′ UTRs, were observed at similar frequencies (Fig. 1b). This observation is consistent with previous studies, suggesting a role for transcription-coupled repair in the acquisition of mutations8,17. Interestingly, promoter and enhancer

1Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. 2Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark. 3Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, New York, USA. 4These authors contributed equally to this work. Correspondence should be addressed to W.L. (leew1@mskcc.org).

Received 11 June; accepted 3 September; published online 28 September 2014; doi:10.1038/ng.3101
regions were mutated at a rate similar to the transcribed genic regions. In contrast, intergenic regions, which by and large should be less implicated in gene regulation and possibly under weaker selective constraint, carried the highest mutational burden across all regions investigated here (Mann-Whitney P < 2.2 × 10^{-16}). Taken together, these observations suggest a functional role for a subset of mutations in annotated regulatory regions.

We used three distinct approaches to identify noncoding mutations that might have a role in tumor development and progression (Fig. 1c). First, a ‘hotspot’ analysis identified small regions with frequent mutation by detecting clusters of mutations within 50 bp of each other (see the Online Methods for details). This approach returns very focal regions that are significantly recurrently mutated in comparison to a random distribution of mutations across the genome. Next, we targeted annotated regulatory regions that were mutated more frequently than expected by chance using a regional recurrence approach. This method takes into account the length and replication timing of different regulatory regions. It computes two measures of significance by comparing the observed frequency of mutations within a region to the expected frequency under a random distribution of mutations across the genome. Next, we targeted annotated regulatory regions that were mutated more frequently than expected by chance using a regional recurrence approach. This method takes into account the length and replication timing of different regulatory regions. It computes two measures of significance by comparing the observed frequency of mutations within a region to the expected frequency under a random distribution of mutations across the genome.

Figure 1 Summary of data and methods. (a) Tumor samples by disease type. Tumor types from TCGA are labeled in boldface; other published samples15 are shown in regular type. ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia. (b) Mean mutation frequency and 95% confidence interval across samples (n = 858) by type of genomic region. CDS, coding sequence. (c) Workflow for the identification of recurrent noncoding mutations in regulatory regions of interest. Our approach integrates mutation calls from 863 tumor-normal pairs and regulatory regions of interest, which are tested for noncoding mutations using 3 distinct analyses. Hotspot analysis detects recurrent mutations that are often very focal. Regional recurrence analysis identifies annotated regions of interest that are enriched for mutation throughout the entire region. Transcription factor analysis searches for regions that contain recurrent mutations within transcription factor binding sites.

genes using Fisher’s exact one-tailed test), suggesting a possible functional role for these mutations in cancer.

Mutation hotspot in the PLEKHS1 promoter

TERT promoter mutations constituted the most significant hotspot in our data (P = 1.1 × 10^{-27}; Fig. 2a). The hotspot analysis identified a focal region in the promoter of the TERT gene, encoding the catalytic subunit of telomerase, which was mutated in 56 samples. This hotspot contained two highly recurrent mutations at chr. 5: 1,295,228 and chr. 5: 1,295,250, which were found in 38 and 15 samples, respectively. Both sites had C>T substitutions, a finding consistent with previous reports of TERT promoter mutation11,12. Here we found mutations at these two recurrent sites in 7 cancer types (glioblastoma (16 samples), melanoma (10 samples), bladder (10 samples), low-grade glioma (9 samples), liver (5 samples), medulloblastoma (2 samples) and lung (1 sample)), at frequencies similar to those described in other recent reports20,21. We also observed a corresponding increase in TERT gene expression in mutated samples (Supplementary Fig. 2). Our results suggest that these mutations might be among the most prevalent functional noncoding mutations across all cancers.

In addition to the hotspot in the TERT promoter, hotspot analysis identified a number of other recurrently mutated hotspots (Fig. 2a).
The next most significant was a small hotspot in the promoter of PLEKHS1 (P = 4.6 × 10^{-88}). This hotspot contained 23 mutations distributed over 20 samples, with 2 mutated sites at the far end of the promoter (~50 bp into the first intron). These 2 sites were mutated in 11 (chr. 10: 115,511,590) and 12 (chr. 10: 115,511,593) samples, and both were predominantly C>T transitions (Fig. 2b). Interestingly, these two mutations are flanked by stretches of 10 bp on both sides that are palindromic to each other (Supplementary Fig. 3). This hotspot was found in five cancer types, and mutated samples appeared to have lower expression of PLEKHS1 (Supplementary Fig. 4). In bladder cancer, 40% of samples were affected by a mutation in this hotspot (8 of 20 samples). PLEKHS1 is a largely uncharacterized gene that has not previously been linked to tumorigenesis. The gene contains sequence encodings a pleckstrin homology domain, which suggests a role for the protein in intracellular signaling22.

This analysis identified several other significant hotspots linked to STAG3, BCL2, TCL1A, AGAP5, TRMT10C, TNK2 and WDR74 (P < 1 × 10^{-10}; Supplementary Tables 5–8). Interestingly, many of these genes have been associated with cancer previously23–26. In contrast to the hotspots for PLEKHS1 and TERT, which mostly had one hotspot mutation per sample, hotspots in the promoter and 5′ UTR of BCL2 were significant but seemed to occur as clusters of several mutations within the same sample (average of 2.2 mutations per mutated sample). Closer examination showed that these mutations were all in B cell lymphoma samples, and they are likely a result of targeted somatic hypermutation at hypervariable regions27.

The WDR74 promoter is frequently mutated

The protein-coding regions of many tumor suppressor genes display frequent inactivating somatic mutations, not at specific sites but instead distributed across the entire ORF. To identify genes with frequent mutations across an entire regulatory region, we developed a statistical framework that evaluates the mutation rates of annotated regulatory regions in both a local and global genomic context. Briefly, the local approach compares regional mutation rates to the overall mutation frequency in the immediate genomic neighborhood, whereas the global approach compares mutation rates for regions in the same category (for example, promoter or 3′ UTR) and with similar DNA replication timing (Online Methods).

This approach identified larger, more frequently mutated genomic regions, thus complementing the hotspot analysis, which focused on much smaller regions. Apart from frequent promoter mutations in TERT (P < 1.3 × 10^{-17}), we observed a number of regulatory regions that were significantly enriched for noncoding mutations (Fig. 3a). In particular, the 5′ UTR (P < 5.1 × 10^{-8}) and promoter (P < 3.6 × 10^{-9}) of WDR74 were highly enriched for mutations. In contrast to the hotspot mutations in PLEKHS1, the mutations in WDR74 were broadly distributed across numerous positions (Fig. 3b), and WDR74 transcript levels were not significantly different in mutated samples (Supplementary Fig. 5). Although the coding sequence of WDR74 did not contain any mutations, our analysis identified 36 noncoding mutations in a ~1-kb window near the 5′ end of the WDR74 gene, most of which clustered at the start of the 5′ UTR.

WDR74 contains sequence encoding a WD40 repeat, which has enzymatic activity and has been shown to be involved in a variety of biological processes, including cell cycle control and apoptosis28. The promoter region of WDR74 was previously found to be under purifying selection and to likely be sensitive to mutation9. Khurana et al. also reported WDR74 promoter mutations in 2 of the 20 prostate cancer genomes analyzed. Here we demonstrate that mutations in this region are more common than previously known. We identified a total of 40 mutations in the promoter region of WDR74 (Fig. 3b), including 4 distinct single
nucleotides with recurrent mutations in up to 4 samples. Overall, 35 of 858 samples (4%) harbored at least one mutation in the region.

Other frequently mutated regions were found in noncoding regions of genes such as SGK1, DHX16 and SDHD (Supplementary Tables 9–12). Interestingly, the 5' end of the SDHD gene contained multiple mutations in putative ETS (E26 transformation specific) family transcription factor binding sites. We next used transcription factor analysis to specifically assess the significance of mutations in ETS response elements on a genome-wide scale.

Promoter mutations in ETS binding sites alter regulation of SDHD

As mentioned above, hotspot analysis identified two known, highly recurrent sites in the promoter of TERT\(^{11,12}\). Both hotspot mutations create novel binding sites for ETS transcription factors by substituting a cytosine nucleotide with a thymine nucleotide (C>T), thereby generating the TTCC response element, which is highly conserved for ETS transcription factors. Both elements are located on the minus strand, which is the coding strand for TERT (TTCC>TTTC at chr. 5: 1,295,228 and TTCC>TTC at chr. 5: 1,295,250, with the mutated nucleotide underlined). Here we systematically screened regulatory regions of interest for mutations that either created novel ETS binding sites or disrupted existing ones (see the Online Methods for details). Apart from TERT promoter mutations, promoter mutations in ANKRD53 were the most significant mutations that created novel ETS binding sites (\(P < 0.0049\)). Several regulatory regions contained a significant (\(P < 0.05\)) number of mutations that disrupted ETS binding sites, including MEF2C (promoter), KRT4 (promoter), ERLIN2 (5' UTR), TAF11 (5' UTR) and SDHD (5' UTR), among others (Supplementary Tables 13–16). SDHD, which encodes subunit D of the succinate dehydrogenase complex, was also observed in the regional recurrence analysis above. SDHD promoter mutations (C>T) occurred exclusively in melanoma samples and potentially disrupted two separate putative ETS binding sites in a small genomic region upstream of the coding sequence (Fig. 4a). The recurrent mutations were located at chr. 11: 111,957,523 (TTCC>TTTC) and chr. 11: 111,957,541 (TTCC>TTTC), close enough to the start codon to allow further examination using TCGA melanoma whole-exome data, which exist for a larger number of samples. The exome data showed a third putative ETS binding site in the SDHD promoter, located at chr. 11: 111,957,544, which had a mutation just outside of the core response element, converting CTTC to TTTC. The mutated base is not conserved in all ETS family transcription factors but is highly conserved in ETS1 (Supplementary Fig. 6),

![Image](259x399 to 540x502)

Figure 3 Regional recurrence analysis. (a) Significance of recurrent mutations in regulatory regions of interest. Regulatory regions for individual genes are shown according to local (y axis) and global (x axis) measures of statistical significance (FDR-adjusted \(P\) value). Colors indicate the type of regulatory region. (b) Strong enrichment of mutations in the promoter region of WDR74 in contrast to the remainder of the gene sequence. The bar chart summarizes the frequency of the hotspot mutation in individual cancer types (colors correspond to those in Fig. 1a).

...the only ETS transcription factor whose gene expression correlated with SDHD gene expression and that also binds the SDHD promoter, according to ENCODE data\(^7,29\). Out of the 128 samples with a read depth of 15 or greater, 13 had a mutation in the promoter region (10%), 10 of which had recurrent mutations in ETS binding sites. In contrast to recurrent mutations in the TERT promoter, which create a novel ETS binding site, mutations in the SDHD promoter damaged existing ETS binding sites. Because TERT promoter mutations led to increased expression of the TERT gene, we expected expression of SDHD to be lower in comparison to a group of ‘wild-type’ melanoma samples without SDHD promoter mutation. Using whole-exome sequencing data and gene expression data from TCGA, we compiled a set of 42 samples that did not have promoter mutations in SDHD (Online Methods). Analysis of expression data showed that tumors with SDHD promoter mutations indeed had significantly reduced expression of the SDHD gene (\(P = 0.0044; \text{Fig. 4b}\)). On the basis of chromatin immunoprecipitation and sequencing (CHIP-seq) data from the ENCODE Project, we were able to identify three ETS family transcription factors with binding activity in the SDHD promoter (EHF, ELF1 and ETS1). Among these three transcription factors, only ELF1 expression exhibited significant positive correlation with the SDHD expression data in the subset of 42 SDHD-proficient samples without promoter mutation (\(P < 0.0035; \text{Fig. 4c and Supplementary Fig. 7}\)), indicating that SDHD could be under the control of the ELF1 transcription factor under normal circumstances. Interestingly, tumor samples with SDHD promoter mutation did not exhibit a correlation between SDHD and ELF1 mRNA levels (\(P = 0.35\), suggesting a possible adverse effect for SDHD promoter mutation on transcriptional regulation by ELF1 (Fig. 4c). In addition to the apparent changes in gene expression, we observed that samples with...
Survival is significantly lower for samples with mutations in ETS transcription factor binding sites that were highly conserved and highlighted in blue. These three methods used clearly delineated regions with ETS transcription factor binding sites that were mutated numerous mutations; and transcription factor analysis nominates regions with ETS transcription factor binding sites that were disrupted or created by mutation. These three methods used clearly distinct approaches, and as a result in general found different regions of interest. However, the most significant findings, which are highlighted in this report, were identified by multiple methods. Promoter mutations in the TERT gene were found by all three methods. Hotspot analysis identified highly recurrent mutations in PLEKHS1, which contains a sequence encoding a pleckstrin homology domain. The mutations occur at the center of a perfectly palindromic sequence. This observation is striking, even though it is not known if this particular palindrome is functional. However, it is known that transcription factor binding sites can be palindromic. The finding of SDHD promoter mutation was moderately significant in regional recurrence analysis but was subsequently substantiated by transcription factor binding site analysis. Recurrent mutations in three distinct ETS response elements were associated with loss of correlation with ETS transcription factor (ELF1) expression at the mRNA level and with shorter survival times for the affected individuals. Although our study focused exclusively on ETS transcription factor binding sites, we believe that such an approach will be valuable when carefully applied to all known conserved binding sites.

SDHD mutation had significantly shorter overall patient survival times in comparison to a reference group of 88 melanoma samples \( (P = 0.005; \text{Fig } 4d) \). SDHD, which encodes subunit D of the succinate dehydrogenase tetramer, is of particular interest because succinate dehydrogenase is the only protein that participates in the citric acid cycle as well as the electron transport chain. It has been shown that SDHD mutations can cause paraganglioma, a benign tumor of the head and neck. Previous studies suggested that SDHD acts like a tumor suppressor, a hypothesis consistent with our observation of reduced mRNA expression in tumor samples with SDHD promoter mutation.

**DISCUSSION**

Here we present a comprehensive analysis of whole-genome sequencing data from 863 individuals with cancer to characterize the landscape of noncoding mutations in cancer. We show that intergenic regions are more often affected by mutation than transcribed regions in close proximity to the coding sequence, such as introns, promoters, enhancers and UTRs. In addition, our data suggest that regulatory regions at the 5′ end of genes, such as promoters and 5′ UTRs, are recurrently mutated more often than 3′ UTRs or distal enhancers (Fig. 3a).

We used three complementary types of analysis to identify regions of interest that are significantly affected by mutation: hotspot analysis focused on small regions that frequently contained mutations; regional recurrence analysis identified annotated regions that contained numerous mutations; and transcription factor analysis nominated regions with ETS transcription factor binding sites that were disrupted or created by mutation. These three methods used clearly distinct approaches, and as a result in general found different regions of interest.
important tumorigenic mutations occur in noncoding regions, even though large numbers of passenger mutations exist in these regions as well, and the interpretation of such mutations remains a challenge. However, interrogation and interpretation of noncoding mutation will become more accurate and more important as the availability of whole-genome sequencing data increases.

URLs. CGHub, https://cghub.ucsc.edu/; Broad Genome Data Analysis Center (GDAC) Firehose, http://gdac.broadinstitute.org/; data from Alexandrov et al.15, ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAL.

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

The authors thank TCGA as well as numerous other groups who have made their data available for public analysis. The authors additionally thank the CGHub team, members of the Memorial Sloan Kettering Cancer Center Bioinformatics Core, G. Rätsch and J. Chodera for the setup and administration of high-performance computing resources. This work was supported in part by a TCGA GDAC grant from the US National Institutes of Health for N.W., N.S., C.S. for the Bioinformatics Core, G. Rätsch and J. Chodera for the setup and administration of my multiple myeloma. Nature 471, 467–472 (2011).

Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).

Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).

Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196 (2010).

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).

Futreal, P.A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).

Vinagre, J. et al. Frequency of TERT promoter mutations in human cancers. Nat. Commun. 4, 2185 (2013).

Killela, P.J. et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. USA 110, 6021–6026 (2013).

Mayer, B.J., Ren, R., Clark, K.L. & Baltimore, D. A putative modular domain present in diverse signaling proteins. Cell 73, 629–630 (1993).

Tsujimoto, Y., Finger, L.R., Yunis, J., Nowell, P.C. & Croce, C.M. Cloning of the chromosome breakpoint of neoplastic B cells with the X(14;18) chromosome translocation. Science 226, 1097–1099 (1984).

Virgilio, L. et al. Identification of the TCE1 gene involved in T-cell malignancies. Proc. Natl. Acad. Sci. USA 91, 12530–12534 (1994).

Mahajan, N.P. et al. Activated Cdc42-associated kinase Ack1 promotes prostate cancer progression via androgen receptor tyrosine phosphorylation. Proc. Natl. Acad. Sci. USA 104, 8438–8443 (2007).

Krol, M. et al. Transcriptomic signature of cell lines isolated from canine mammary adenocarcinoma metastases to lungs. J. Appl. Genet. 51, 37–50 (2010).

Schneider, C., Pasqualucci, L. & Dalla-Favera, R. Molecular pathogenesis of diffuse large B-cell lymphoma. Semin. Diagn. Pathol. 28, 167–177 (2011).

Strimmer, C.U., Petsalaki, E., Russell, R.B. & Muller, C.W. WD40 proteins propel cellular networks. Trends Biochem. Sci. 35, 565–574 (2010).

Thurman, R.E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

Schneider, C., Pasqualucci, L. & Dalla-Favera, R. Molecular pathogenesis of diffuse large B-cell lymphoma. Semin. Diagn. Pathol. 28, 167–177 (2011).

Strimmer, C.U., Petsalaki, E., Russell, R.B. & Muller, C.W. WD40 proteins propel cellular networks. Trends Biochem. Sci. 35, 565–574 (2010).

Thurman, R.E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

Schneider, C., Pasqualucci, L. & Dalla-Favera, R. Molecular pathogenesis of diffuse large B-cell lymphoma. Semin. Diagn. Pathol. 28, 167–177 (2011).
ONLINE METHODS

Calling mutations. Whole-genome sequencing data were downloaded from CGHub in the form of tumor and matched normal BAM files. Mutations were called across the whole genome using MuTect\textsuperscript{35} and Strekla\textsuperscript{36} with default parameters. The intersection of the somatic mutation calls made by both programs was used as the mutation list for each sample.

To investigate promoter mutations in SDHD, whole-exome sequencing data for skin cutaneous melanoma (SKCM) were downloaded from CGHub and allele counts were generated for SDHD 5′ coordinates chr. 11: 111,957,523 and chr. 11: 111,957,541. Samples without whole-genome sequences but whose whole-exome sequence data exhibited two or more mutant alleles (T instead of the reference C) at these positions were considered to be mutated.

We excluded 5 samples with more than 500,000 mutations each, limiting the data set to 858 samples. All analyses focused on single-nucleotide substitutions and did not consider insertions, deletions or other structural variants.

Defining noncoding regions of interest. We used gene annotation from Ensembl\textsuperscript{37} (v70) for the transcripts of all protein-coding genes (having at least one annotated ORF). 5′ UTRs and 3′ UTRs were used as defined by Ensembl. Promoter regions were defined as the genomic intervals ranging from 2,000 bp upstream to 200 bp downstream of all transcription start sites. We used 66,944 enhancer regions to gene associations (27,493 unique regions) from a previous comprehensive study\textsuperscript{38} in which the inferred middle positions of the enhancer regions were extended by 200 bp up- and downstream. To avoid mutation bias from protein-coding regions, we removed any regions overlapping ORFs (extended by 5 bp also account for splice sites) from the collection of regions of interest. Furthermore, to avoid bias from immune system–coupled somatic hypermutation, we also removed the regions corresponding to 429 annotated immunoglobulin loci (each region was extended by 50 kb; Ensembl v73).

Identification of hotspot mutations. All mutations within 50 bp of each other were merged using BEDTools into hotspot clusters until no cluster was within 50 bp of another cluster\textsuperscript{39}. Clusters with only one or two mutations were removed from further consideration. A P value was calculated for each cluster using the negative binomial distribution, taking into account the length of the candidate hotspot, the number of mutations in the cluster and a background mutation rate for the cluster. The cluster background mutation rate was calculated as the mean of the background mutation probability for each sample that had a mutation represented in the cluster. The background mutation probability of each sample was calculated as the total number of mutations divided by the genome size. P values were adjusted for multiple testing with the multtest R package\textsuperscript{40} using the Benjamini-Hochberg method, and hotspot clusters were ranked accordingly.

Testing regions of interest for mutation recurrence. All regions of interest longer than 50 bp were tested for the recurrence of mutations using both a global and local statistical approach. Both approaches assumed that the observed number of mutated samples, k, for a given region followed a binomial distribution, binomial (n, p\textsubscript{i}), where n was the total number of samples with mutation data and p\textsubscript{i} was the estimated sample mutation rate for region of interest i under the null hypothesis that the region was not recurrently mutated. We could therefore compute the following P value:

$$P(X \geq k) = 1 - P(X < k) = 1 - \sum_{j=0}^{k-1} \binom{n}{j} p_j^j (1 - p_i)^{n-j}$$

Here we assumed that p\textsubscript{j} depended on the effective length L\textsubscript{j} of the region (with any ORF overlap subtracted; see above) and the estimated nucleotide mutation rate q\textsubscript{i} for the region under the null hypothesis as follows:

$$p_j = 1 - (1 - q_i)^{L_j}$$

The background mutation frequency q\textsubscript{i} was not readily available and needed to be estimated before we could compute a P value using the above equation. We estimated both a local and a global background mutation frequency, which formed the basis for the local and global tests, respectively. For the local approach, we extracted 10-kb flanking regions upstream and downstream of the region of interest, excluding ORFs to reduce mutation bias from nearby protein-coding regions. The local background nucleotide mutation frequency was then estimated by dividing the total number of observed mutations by the effective length of the flanking region. In the global approach, we estimated nucleotide mutation frequencies from other regions of the same category of regions (for example, promoter, 3′ UTR, etc.). Because DNA replication timing has previously been shown to affect somatic mutation rates in tumors\textsuperscript{41,42}, we further stratified regions of interest from the same category by their replication timing in five cancer cell lines (HeLa, K562, HEK293, MCF7 and SKNHS; data from the University of Washington ENCODE group\textsuperscript{43}). We first computed average replication time values in 100-kb bin sizes for each cell line, and for each region of interest we computed a single replication time value (average if spanning more than one bin) for each cell line. For a given region in category C, we identified the top 5% of regions in C with the most similar replication timing profiles (Euclidian distance between vectors of replication time values across the five cell lines). The global background nucleotide mutation frequency was then estimated by dividing the total number of observed mutations in the top 5% of the regions of interest by the effective length of these regions. P values were computed using the equation above and adjusted for multiple testing with the multtest R package using the Benjamini-Hochberg method.

For each region or gene, we selected the maximum FDR for the individual data set to 858 samples. All analyses focused on single-nucleotide substitutions and did not consider insertions, deletions or other structural variants.

Transcription factor analysis. All mutations were annotated if they affected ETS transcription factor binding sites. Mutations were considered to create ETS transcription factor binding sites if the nucleotide substitution created a novel ETS transcription factor core response element on either strand (for example, TGCC>TGCC). Mutations were considered to disrupt ETS transcription factor binding sites if they altered an existing ETS core response element (for example, TTCC>TGCC). Using the regions of interest defined above, we then calculated a count statistic for each region of interest by summing the number of mutations that created or disrupted ETS transcription factor binding sites within each region. For each region of interest that contained more than one mutation in an ETS binding site, an empirical P value was computed by comparing the observed count statistic (number of mutations creating or disrupting ETS binding sites within the region of interest) to a reference distribution of count statistics. Reference distributions were generated for each region of interest by iteratively calculating the above count statistic on the same data and randomizing (binary annotations of whether or not mutations created novel binding sites were randomized) during each iteration. A P value was derived by comparing the observed count statistic of a given region of interest to the distribution of count statistics for its corresponding reference distribution (based on 10,000 iterations) and was defined as the fraction of the count statistics in the reference distribution greater or equal to the observed count statistic. P values were adjusted for multiple testing using the Benjamini-Hochberg method.

Expression analysis. Expression analysis was performed using RNA sequencing raw counts from TCGA. P values are reported using a negative binomial test from the edgeR package\textsuperscript{43}, which is available through Bioconductor\textsuperscript{44}. In-depth analyses of SDHD promoter mutations (Fig. 4b,c) were performed on a set of melanoma samples from TCGA (v20130923, level 3) for which exome sequencing data were available. The set of reference (wild-type) samples consisted of melanoma samples that had a read depth of 15 or more in the SDHD promoter region. Wild-type samples with putative copy number alterations at the SDHD locus were excluded. Survival analysis (Fig. 4d) was performed on a set of 88 samples with read depth of 15 or more in the SDHD promoter region using the clinical data file for melanoma from TCGA.

35. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
36. Saunders, C.T. et al. Strekla: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
37. Hubbard, T. et al. The Ensembl genome database project. Nucleic Acids Res. 30, 38–41 (2002).
38. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
39. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
40. Pollard, K.S., Dudoit, S. & Laan, M.J.v.d. in Bioinformatics Computational Biology Solutions Using R and Bioconductor (Springer, New York, 2005).
41. Chen, C.L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20, 447–457 (2010).
42. Hansen, R.S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. USA 107, 139–144 (2010).
43. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).
44. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004).