Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature

Nils Johan Fredriksson[1], Kerryn Elliott[1], Stefan Filges[2], Jimmy Van den Eynden[1], Anders Ståhlberg[2], Erik Larsson[1]*

1 Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, 2 Sahlgrenska Cancer Center, Department of Pathology and Genetics, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

* erik.larsson@gu.se

Abstract

Sequencing of whole tumor genomes holds the promise of revealing functional somatic regulatory mutations, such as those described in the *TERT* promoter. Recurrent promoter mutations have been identified in many additional genes and appear to be particularly common in melanoma, but convincing functional data such as influence on gene expression has been more elusive. Here, we show that frequently recurring promoter mutations in melanoma occur almost exclusively at cytosines flanked by a distinct sequence signature, TTCCG, with *TERT* as a notable exception. In active, but not inactive, promoters, mutation frequencies for cytosines at the 5' end of this ETS-like motif were considerably higher than expected based on a UV trinucleotide mutational signature. Additional analyses solidify this pattern as an extended context-specific mutational signature that mediates an exceptional position-specific vulnerability to UV mutagenesis, arguing against positive selection. We further use ultra-sensitive amplicon sequencing to demonstrate that cell cultures exposed to UV light quickly develop subclonal mutations specifically in affected positions. Our findings have implications for the interpretation of somatic mutations in regulatory regions, and underscore the importance of genomic context and extended sequence patterns to accurately describe mutational signatures in cancer.

Author summary

Cancer is caused by somatic mutations that alter cell behavior. While such mutations typically occur in protein-coding genes, recent studies describe individual positions in gene regulatory regions (promoters) that are recurrently mutated in many independent tumors. This suggests that positive selection could be acting on these non-coding mutations, and that they may contribute to carcinogenesis. However, proper interpretation of recurrent mutations requires a detailed understanding of how such mutations arise in the absence of selection pressures, referred to as mutational heterogeneity. In this paper, we

describe a distinct sequence signature that characterizes nearly all highly recurrent promoter mutations in melanoma. Additional analyses support that this sequence mediates an exceptional local vulnerability to UV-induced mutagenesis, explaining why mutations are frequently observed in these positions. Importantly, cultured cells exposed to UV light quickly developed mutations specifically in the expected sites. Our results have important implications for the interpretation of recurrent somatic mutation patterns in non-coding DNA.

Introduction

A major challenge in cancer genomics is the separation of functional somatic driver mutations from non-functional passengers. This problem is relevant not only in coding regions, but also in the context of non-coding regulatory regions such as promoters, where putative driver mutations are now mappable with relative ease using whole genome sequencing[1,2]. One important indicator of driver function is recurrence across independent tumors, which can be suggestive of positive selection. However, proper interpretation of recurrent mutations requires a detailed understanding of how somatic mutations occur in the absence of selection pressures. Somatic mutations are not uniformly distributed across tumor genomes, and regional variations in mutation rates have been associated with differences in transcriptional activity, replication timing as well as chromatin accessibility and modification[3–5]. Impaired nucleotide excision repair (NER) has been shown to contribute to increased local mutation density in promoter regions and protein binding sites[6,7]. Additionally, analyses of mutational processes and their sequence signatures have shown the importance of the immediate sequence context for local mutation rates[8]. Still, our understanding of mutational heterogeneity is incomplete, and it is not clear to what extent such effects can explain recurrent somatic mutations in promoter regions, which are suggested by some studies to be particularly frequent in melanoma despite several other cancer types approaching melanoma in terms of total mutation load[9,10].

Results/Discussion

To characterize somatic promoter mutations in melanoma, we analyzed the sequence context of recurrently mutated individual genomic positions occurring within +/- 500 bp of annotated transcription start sites (TSSs), based on 38 melanomas subjected to whole genome sequencing by the Cancer Genome Atlas[10,11]. Strikingly, of 17 highly recurrent promoter mutations (recurring in at least 5/38 of tumors, 13%), 14 conformed to an identical 6 bp sequence signature (Fig 1a and 1b). Importantly, the only exceptions were the previously described *TERT* promoter mutations at chr5:1,295,228, 1,295,242 and 1,295,250[12,13] (Fig 1c). The recurrent mutations occurred at cytosines positioned at the 5' end or one base upstream of the motif CTTCCG (Fig 1d), and were normally C>T or CC>TT transitions (Fig 1a). Similar to most mutations in melanoma they thus occurred in a dipyrimidine context and were compatible with UV-induced damage through cyclobutane pyrimidine dimer (CPD) or 6–4 photoproduct formation[8,14]. Out of 15 additional positions recurrently mutated in 4/38 tumors (11%), 13 conformed to the same pattern, while the remaining two showed related sequence contexts (Fig 1a). Many less recurrent sites also showed the same pattern (S1 Table). The signature described here matches the consensus binding sequence of ETS family transcription factors (TFs)[15], and the results are consistent with recent reports showing that ETS promoter sites are often recurrently mutated in melanoma[9] and that such mutations preferably occur at
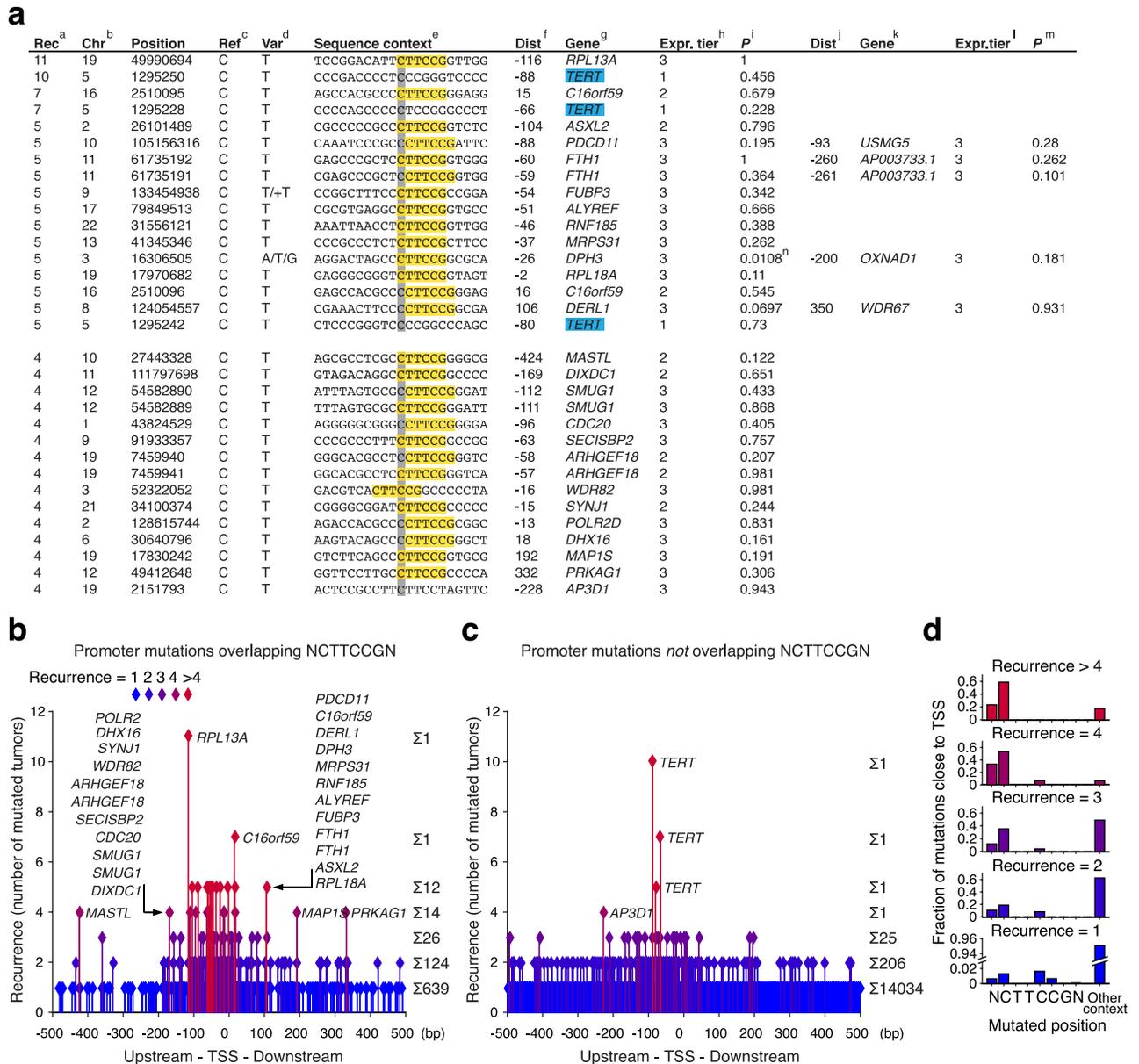
Fig 1. A distinct sequence signature characterizes nearly all highly recurrent somatic promoter mutations in melanoma. Whole genome sequencing data from 38 melanomas were analyzed for individual recurrently mutated bases in promoter regions. (**a**) All highly recurrent mutations within +/- 500 bp from TSSs ordered by recurrence (number of mutated tumors). [a]Recurrence of each mutation. [b]Chromosome. [c]Reference base. [d]Variant base. [e]Sequence context, showing pyrimidine-containing strand with respect to the central mutated base (gray). The motif CTTCCG is highlighted in yellow. [f]Distance to the nearest TSS in GENCODE 17. [g]Closest gene. Cancer gene census genes [42] in blue. [h]Genes were sorted by mean expression (all samples) and assigned to tiers 1 to 3 with 3 being the highest. [i]P-value for differential expression of the gene comparing tumors with and without the mutation (two-sided Wilcoxon test). [j, k, l, m]Second closest TSS, if within 500 bp. [n]A previous analysis in a larger cohort failed to show significant differential expression [10]. (**b**) All mutations occurring within +/- 500 bp of a TSS while overlapping with the motif NCTTCCGN. The distance to the nearest TSS and the degree of recurrence (number of mutated tumors) is indicated. (**c**) Similar to panel **b**, but instead showing mutations *not* overlapping NCTTCCGN. (**d**) Positional distribution across the sequence NCTTCCGN for mutations listed in panel **a**.

cytosines upstream of the core TTCC sequence[16]. Thus, while recurrent promoter mutations are common in melanoma, they consistently adhere to a distinct sequence signature, which may argue against positive selection as a major causative factor.

The recurrently mutated positions were next investigated in additional cancer cohorts, first by confirming them in an independent melanoma dataset[17] (S2 Table). We found that the identified hotspot positions were often mutated also in cutaneous squamous cell carcinoma (cSCC)[18] (S3 Table) as well as in sun-exposed skin[18,19], albeit at lower variant frequencies (S1 Fig, S4 Table). Additionally, one of the mutations, upstream of *DPH3*, was recently described as highly recurrent in basal cell skin carcinoma[20]. However, we did not detect mutations in these positions in 13 non-UV-exposed cancer types (S5 Table). The hotspots are thus present in UV-exposed samples of diverse cellular origins, but in contrast to the *TERT* promoter mutations they are completely absent in non-UV-exposed cancers. This further supports that recurrent mutations at the 5' end of CTTCCG elements are due to elevated susceptibility to UV-induced mutagenesis in these positions.

Next, we considered additional properties that could support or argue against a functional role for the recurrent mutations. We first noted a general lack of known cancer-related genes among the affected promoters, with *TERT* as one of few exceptions (Fig 1a and S1 Table, indicated in blue). Secondly, the recurrent promoter mutations were not associated with differential expression of the nearby genes (Fig 1a and S1 Table). This is in agreement with earlier investigations of some of these mutations, which gave no conclusive evidence regarding influence on gene expression[9,16,20], although it should be noted that significant association was lacking also for *TERT* in this relatively small cohort. Lastly, we found that when comparing different tumors there was a strong positive correlation between the total number of the established hotspot positions that were mutated and the genome-wide mutation load, both in melanoma (Fig 2a; Spearman's $r = 0.88$, $P = 2.8e-13$) and in cSCC (S3 Table; $r = 0.78$, $P = 0.026$). This is again compatible with a passive model involving elevated mutation probability in the affected positions. Importantly, this contrasted sharply with most of the major driver mutations in melanoma, which were detected also in tumors with lower mutation load (Fig 2b, S3 Table). These different findings further reinforce the CTTCCG motif as a strong mutational signature in melanoma.

We next investigated whether the observed signature would be relevant also outside of promoter regions. As expected, numerous mutations occurred in CTTCCG sequences across the genome, but notably we found that recurrent mutations involving this motif were always located close to actively transcribed TSSs (Fig 3a, 3b and 3c). We further compared the frequencies of mutations occurring at cytosines in the context of the motif to all possible trinucleotide contexts, an established way of describing mutational signatures in cancer[8]. As expected, on a genome-wide scale, the mutation probability for cytosines in CTTCCG-related contexts was only marginally higher compared to corresponding trinucleotide contexts (Fig 4a). However, close to TSSs, the signature conferred a striking elevation in mutation probability compared to related trinucleotides, in particular for cytosines at the 5' end of the motif and most notably near highly expressed genes (Fig 4b–4d). Recurrent promoter mutations in melanoma thus conform to a distinct sequence signature manifested only in the context of active promoters, suggesting that a specific binding partner is required for the element to confer elevated mutation probability.

CTTCCG elements have in various individual promoters been shown to be bound by ETS factors such as ETS1, GABPA and ELF1[21], ELK4[22], and E4TF1[23]. This suggests that the recurrently mutated CTTCCG elements could be substrates for ETS TFs. As expected, matches to CTTCCG in the JASPAR database of TF binding motifs were mainly ETS-related (S6 Table). Notably, recurrently mutated CTTCCG sites were evolutionarily conserved to a larger degree
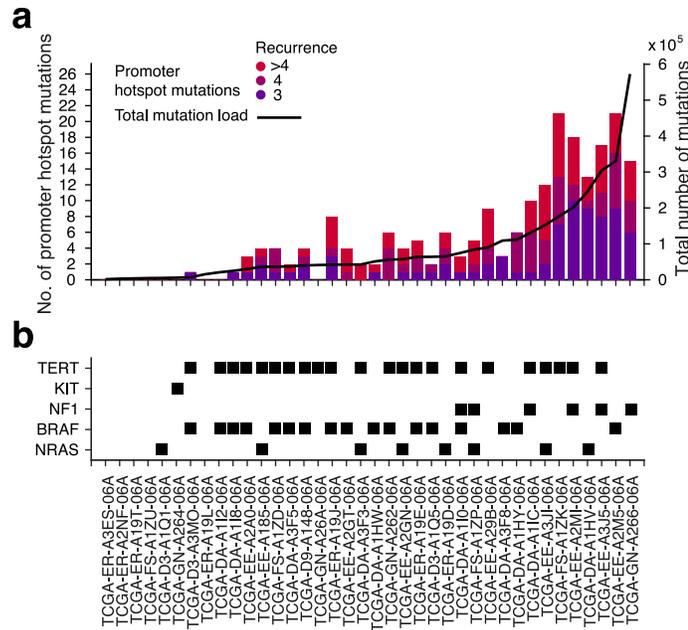
Fig 2. Positive correlation between promoter hotspot mutations and total mutational load across melanomas. (**a**) Bars, left axis: Number of mutations occurring in the established recurrent CTTCCG-related promoter positions (> = 3 tumors) in each of the 38 samples. Line, right axis: Total mutational load per tumor (number of mutations across the whole genome). (**b**) Presence of *TERT* promoter mutations and mutations in known driver genes are indicated for all samples.

than non-recurrently mutated but otherwise similar control sites, further supporting that they constitute functional ETS binding sites (S2 Fig). This was corroborated by analysis of top recurrent CTTCCG sites in relation to ENCODE ChIP-seq data for 161 TFs, which showed that the strongest and most consistent signals were for ETS factors (GABPA and ELF1) (S3 Fig).
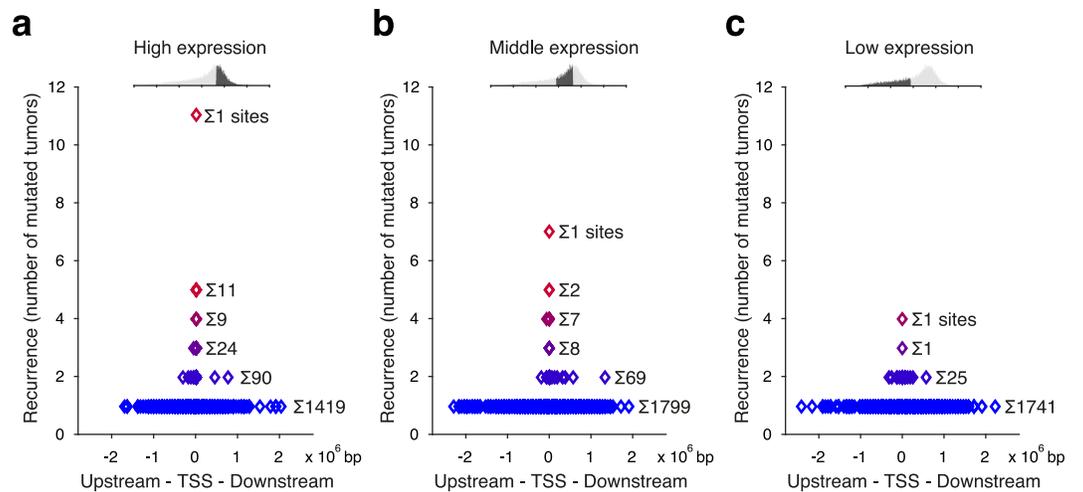
Fig 3. Recurrent mutations at CTTCCG sites are observed only near active promoters. (**a-c**) Genes were assigned to three expression tiers by increasing mean expression across the 38 melanomas. The graphs show, on the x-axis, the distance to the nearest annotated TSS for all mutations overlapping with or being adjacent to the motif CTTCCG across the whole genome, separately for each expression tier. The level of recurrence is indicated on the y-axis.
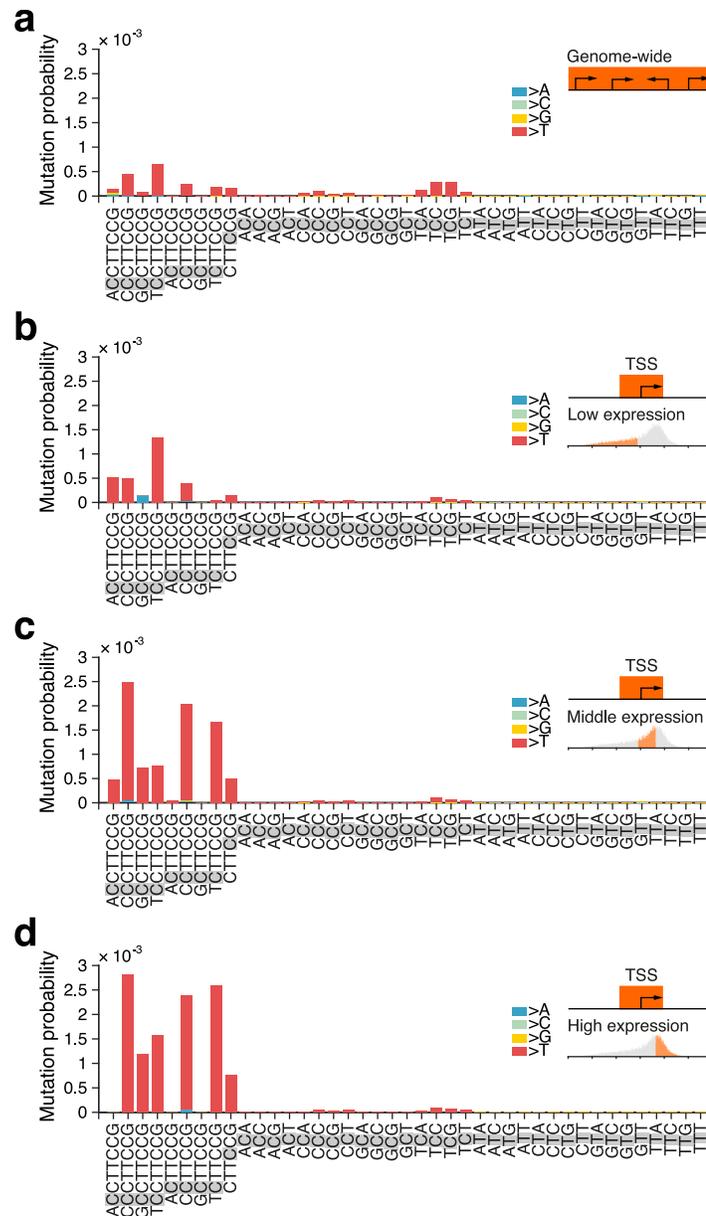
Fig 4. Mutation probabilities for CTTCCG-related sequence contexts compared to trinucleotides in melanoma. The mutated position in each sequence context is shaded in gray. Bar colors indicate the substituting bases (mainly C>T). Mutation probabilities were calculated genome-wide (**a**), or only considering mutations less than 500 bases from TSS of genes with a low (**b**), middle (**c**) or high (**d**) mean expression level.

https://doi.org/10.1371/journal.pgen.1006773.g004

The distribution of mutations across tumor genomes is shaped both by mutagenic and DNA repair processes. Binding of TFs to DNA can increase local mutation rates by impairing NER, and strong increases have been observed in predicted binding sites for several ETS factors[6,7]. It is also established that contacts between DNA and proteins can modulate DNA damage patterns by altering conditions for UV photoproduct formation[24–27]. In upstream regions of XPC -/- cSCC tumors lacking global NER, we found that several of the established hotspot sites were mutated (S7 Table) and that the CTTCCG signature still conferred elevated mutation probabilities compared to relevant trinucleotide contexts (Fig 5), although to a lesser
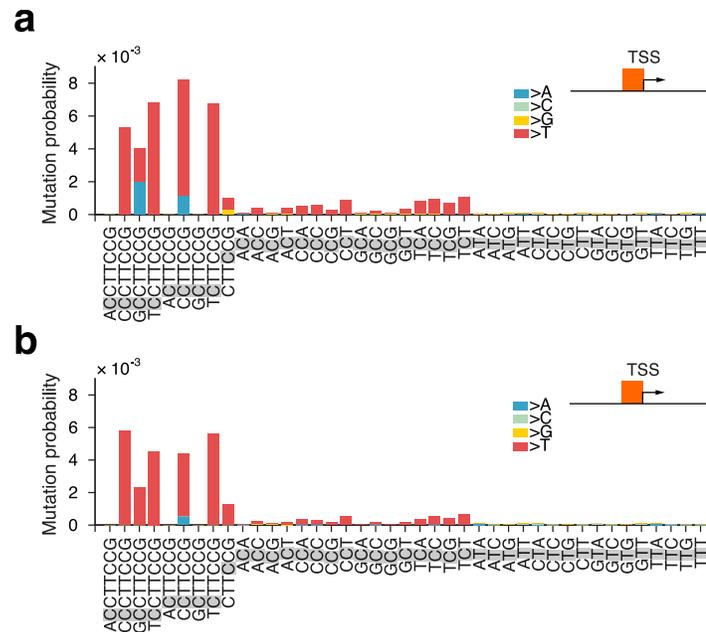
Fig 5. Mutation probabilities for CTTCCG-related sequence contexts compared to trinucleotides in cSCC tumors with NER deficiency. 5 cSCC tumors with defective global NER[18] were screened for mutations within 500 bp upstream of the TSSs, considering only genes in the upper expression tier as defined earlier based on TCGA data. Template (**a**) and non-template strands (**b**), with respect to the transcription direction of the downstream gene, were considered separately. The mutated position in each sequence context is shaded in gray. Bar colors indicate the substituting bases.

https://doi.org/10.1371/journal.pgen.1006773.g005

extent than in melanomas with functional NER (Fig 4). Transcription-coupled NER (TC-NER) may still be active in XPC -/- tumors, and the signature could thus theoretically arise due to blocking of TC-NER at CTTCCG elements. However, only upstream regions, which should not be subjected to this process, were considered in this analysis. Additionally, TC-NER is strand-specific[14], but the signature was present independently of strand orientation relative to the downstream gene in XPC -/- tumors (Fig 5a and 5b). The signature described here is thus unlikely explained by impaired NER alone, and other mechanisms, such as inhibition of other repair-related processes or favorable conditions for UV lesion formation at the 5' end of ETS-bound CTTCCG elements, may contribute.

Finally, we sought to experimentally test our proposed model that the observed promoter hotspots are due to localized vulnerability to mutagenesis by UV light. We subjected human melanoma cells and keratinocytes to daily UV doses for a period of 5 or 10 weeks and used an ultrasensitive error-correcting amplicon sequencing protocol, SiMSen-Seq[28], to assay two of the observed promoter hotpots for mutations: *RPL13A*, the most frequently mutated site in the tumor data, and *DPH3*[10,20] (Fig 6a). Between 36k and 82k error-corrected reads (>20x oversampling) were obtained for each of 16 different conditions (Fig 6b and 6c). Strikingly, subclonal mutations appeared specifically in expected positions at both time points and in both cell lines at a frequency reaching up to 2.9% of fragments (*RPL13A*, 10 weeks of exposure), while being absent in non-exposed control cells (Fig 6d and 6e). As predicted by the tumor data, mutations occurred primarily at cytosines upstream of the TTCCG motif, with lower-frequency mutations occurring also in the central cytosines. Few mutations were observed outside of the TTCCG context despite presence of many cytosines in theoretically vulnerable configurations in the two amplicons (Fig 6d and 6e, underscored). Interestingly, an
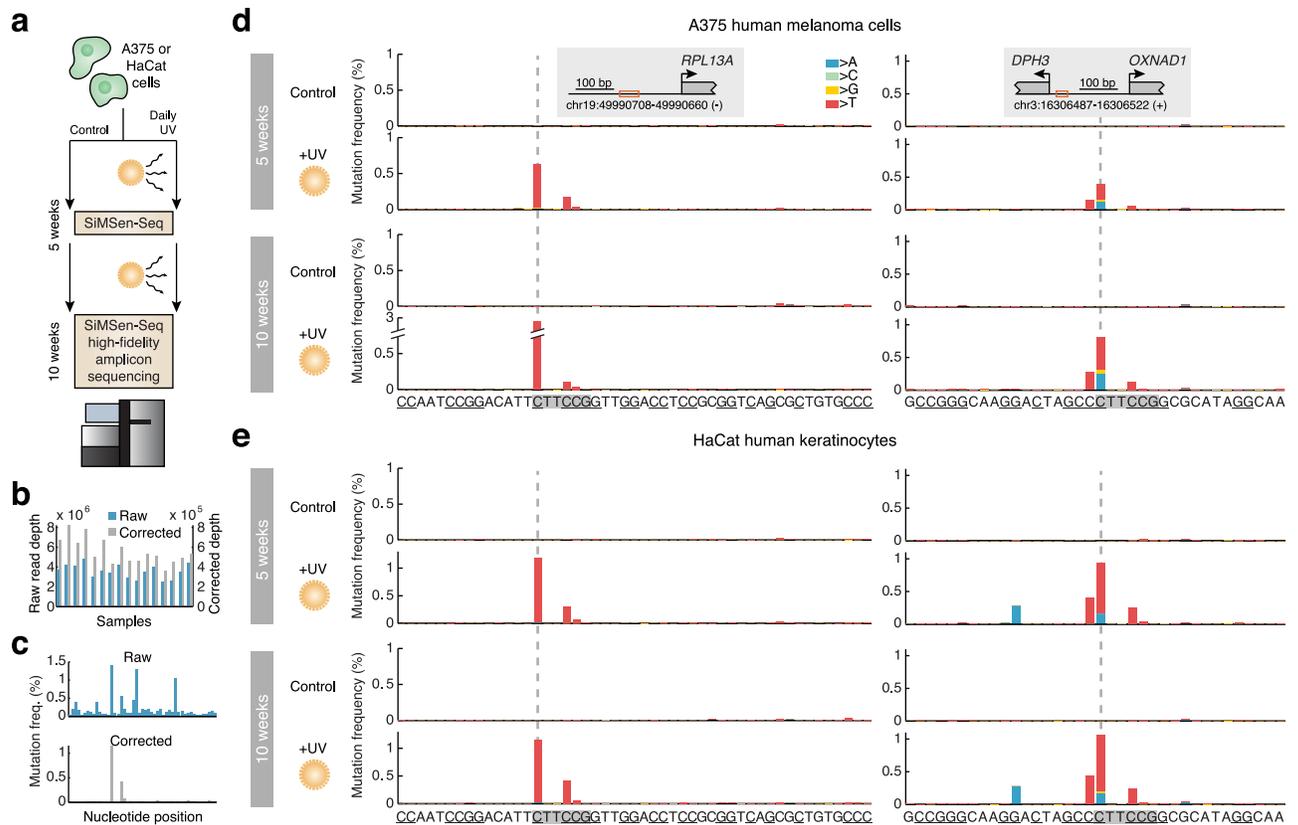
Fig 6. UV exposure of cultured cells induces mutations specifically at CTTCCG-related promoter hotspot sites. (**a**) Human cells (A375 melanoma cells or HaCat keratinocytes) were subjected to daily UV doses (254 nm, 36 J/m^2 once a day, 5 days a week). An ultrasensitive amplicon sequencing protocol, SiMSen-Seq[28], was used to assay for subclonal mutations in two of the established promoter hotspot sites after 5 or 10 weeks. (**b**) 16 different conditions (+/- UV, two regions, two time points, and two cell lines) were sequenced at 2.5M to 4.8M reads per library. Minimum 20 times oversampling was required, resulting in 36k-82k error-corrected reads per library. (**c**) Example of raw and corrected mutation frequencies upstream of *RPL13A* (HaCat cells, 10 weeks UV exposure). (**d-e**) Subclonal mutations at or near CTTCCG hotspots upstream of *RPL13A* or *DPH3*, after 5 or 10 weeks of UV exposure. The CTTCCG elements are indicated, and other possible UV-susceptible sites (cytosines flanking pyrimidines) are underscored. The amplicon sizes were 49 and 36 bp, respectively.

atypical substitution pattern displayed by the *DPH3* hotspot in the tumors, involving C>A and C>G in addition to the expected C>T transitions (Fig 1a), was mirrored also in the UV exposure data (Fig 6d). Our results from UV exposure of cultured cells further reinforce that recurrent mutation hotspots in promoters in melanoma arise due to an exceptional vulnerability to UV mutagenesis in these positions.

In summary, we demonstrate that recurrent promoter mutations are common in melanoma, but also that they adhere to a distinct sequence signature in a strikingly consistent manner, arguing against positive selection as a major driving force. This model is supported by several additional observations, including lack of cancer-relevant genes, lack of obvious effects on gene expression, presence of the signature exclusively in UV-exposed samples of diverse cellular origins, and strong positive correlation between genome-wide mutation load and mutations in the affected positions. Crucially, exposing cells to UV light under controlled conditions efficiently induces mutations specifically in affected sites. These results point to limitations in conventional genome-wide derived trinucleotide models of mutational signatures, and imply that extended sequence patterns as well as genomic context should be taken into account to improve interpretation of somatic mutations in regulatory DNA.

Materials & methods

Mapping of somatic mutations

Whole-genome sequencing data for 38 skin cutaneous melanoma (SKCM) metastases were obtained from the Cancer Genome Atlas (TCGA) together with matching RNA-seq data (dbGap accession phs000178.v9.p8). Mutations were called using SAMtools[29] (command *mpileup* with default settings and additional options *-q1* and*–B*) and VarScan[30] (command *somatic* using the default minimum variant frequency of 0.20, minimum normal coverage of 8 reads, minimum tumor coverage of 6 reads and the additional option *–strand-filter 1*). Mutations where the variant base was detected in the matching normal were not considered for analysis. Mutations overlapping germline variants included in the NCBI dbSNP database, Build 146, were removed. The genomic annotation used was GENCODE[31] release 17, mapped to GRCh37. The TSS of a gene was defined as the 5'most annotated transcription start. Somatic mutation status for known driver genes was obtained from the cBioPortal [32,33].

RNA-seq data processing

RNA-seq data was analyzed with respect to the GENCODE[31] (v17) annotation using HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq) as previously described[34]. Differential gene expression between tumors with and without mutations in promoter regions was evaluated using the two-sided Wilcoxon rank sum test.

Analyzed genomic regions

The SKCM tumors were analyzed across the whole genome or in regions close to TSS, in which case only mutations less than 500 bp upstream or downstream of TSS were included. For the analysis of regions close to TSS the genes were divided in three tiers of equal size based on the mean gene expression level across the 38 SKCM tumors.

Mutation probability calculation

The February 2009 assembly of the human genome (hg19/GRCh37) was downloaded from the UCSC Genome Bioinformatics site. Sequence motif and trinucleotide frequencies were obtained using the tool *fuzznuc* included in the software suite EMBOSS[35]. The mutation probability was calculated as the total number of observed mutations in a given sequence context across all tumors divided by the number of instances of this sequence and by the number of tumors.

Evolutionary conservation data

The evolutionary conservation of genome regions was evaluated using phastCons scores[36] from multiple alignments of 100 vertebrate species retrieved from the UCSC genome browser. The analyzed regions were 30 bases upstream and downstream of the motif CTTCCG located less than 500 bp from TSS.

ChIP-seq data

Binding of transcription factors at NCTTCCGN sites was evaluated using normalized scores for ChIP-seq peaks from 161 transcription factors in 91 cell types (ENCODE track wgEncodeRegTfbsClusteredV3) obtained from the UCSC genome browser.

Analysis of whole genome sequencing data from UV-exposed skin

Whole genome sequencing data from sun-exposed skin, eye-lid epidermis, was obtained from Martincorena *et al.*, 2015[19]. SAMtools[29] (command *mpileup* with a minimum mapping quality of 60, a minimum base quality of 30 and additional option *–B*) was used to process the data and VarScan[30] (command *mpileup2snp* counting all variants present in at least one read, with minimum coverage of one read and the additional strand filter option disabled) was used for mutation calling.

Analysis of whole genome sequencing data from cSCC tumors

Whole genome sequencing data from 8 cSCC tumors and matching peritumoral skin samples was obtained from Durinck *et al.*, 2011[37]. Whole genome sequencing data from cSCC tumors and matching peritumoral skin from 5 patients with germline DNA repair deficiency due to homozygous frameshift mutations (C_{940}del-1) in the *XPC* gene was obtained from Zheng *et al.*, 2014[18]. SAMtools[29] (command *mpileup* with a minimum mapping quality of 30, a minimum base quality of 30 and additional option *–B*) was used to process the data and VarScan[30] (command *mpileup2snp* counting all variants present in at least one read, with minimum coverage of two reads and the additional strand filter option disabled) was used for mutation calling. For the mutation probability analysis of cSCC tumors with NER deficiency, an additional filter was applied to only consider mutations with a total coverage of at least 10 reads and a variant frequency of at least 0.2. The functional impact of mutations in driver genes was evaluated using PROVEAN[38] and SIFT[39]. Non-synonymous mutations that were considered deleterious by PROVEAN or damaging by SIFT were counted as driver mutations.

Cell lines and UV treatments

A375 melanoma cells were a gift from Joydeep Bradbury and HaCaT keratinocyte cells were a gift from Marica Ericson. Cells were grown in DMEM + 10% FCS + gentamycin (A375) or pen/strep (HaCaT) (Thermo Scientific). Cells were treated in DMEM in 10 cm plates without lids with 36 J/m^2 UVC 254 nm (equivalent to 6 hour daily dose at 0.1J/m^2/min[40], CL-1000 UV crosslinker, UVP), 5 days a week for 10 weeks. Cells were split when confluent and reseeded at 1:5. Cells were frozen at -20˚C.

DNA purification

DNA was extracted based on Tornaletti and Pfeifer [41]. Briefly, cell pellets were lysed in 0.5 ml of 20 mM Tris-HCl (pH 8.0), 20 mM NaCl, 20mM EDTA, 1% (w/v) sodium dodecyl sulfate, 600 mg/ml of proteinase K, and 0.5 ml of 150 mM NaCl, 10 mM EDTA. The solution was incubated for two hours at 37˚C. DNA was extracted twice with phenol-chloroform and once with chloroform and precipitated by adding 0.1 vol. 3 M sodium acetate (pH 5.2), and 2.5 volumes of ethanol. The pellets were washed with 75% ethanol and briefly air-dried. DNA was dissolved in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA (TE buffer) (all from Sigma Aldrich). DNA was treated with RNAse for 1 hr at 37˚C and phenol-chloroform extracted and ethanol precipitated before dissolving in TE buffer.

Ultrasensitive mutation analysis

To detect and quantify mutations we applied SiMSen-Seq (Simple, Multiplexed, PCR-based barcoding of DNA for Sensitive mutation detection using Sequencing) as described[28]. Briefly, barcoding of 150 ng DNA was performed in 10 μL using 1x Phusion HF Buffer, 0.1U

Phusion II High-Fidelity polymerase, 200 μM dNTPs (all Thermo Fisher Scientific), 40 nM of each primer (PAGE-purified, Integrated DNA Technologies) and 0.5M L-Carnitine inner salt (Sigma Aldrich). Barcode primer sequences are shown in S8 Table. The temperature profile was 98˚C for 3 min followed by three cycles of amplification (98˚C for 10 sec, 62˚C for 6 min and 72˚C for 30 sec), 65˚C for 15 min and 95˚C for 15 min. The reaction was terminated by adding 20 μL TE buffer, pH 8.0 (Invitrogen, Thermo Fisher Scientific) containing 30 ng/μL protease from *Streptomyces griseus* (Sigma Aldrich) at the beginning of the 65˚C incubation step. Next, 10 μL of the diluted barcoded PCR products were amplified in a 40 μL using 1x Q5 Hot Start High-Fidelity Master Mix (New England BioLabs) and 400 nM of each sequencing adapter primer. Adapter primers are shown in S8 Table. The temperature profile was 95˚C for 3 min followed by 40 cycles of amplification (98˚C for 10 sec, 80˚C for 1 sec, 72˚C for 30 sec and 76˚C for 30 sec, with a ramp rate of 0.2˚C/sec). The 40 μL PCR products were then purified using Agencourt AMPure XP beads (Beckman-Coulter) according to the manufacturers' instructions using a bead to sample ratio of 1. The purified product was eluted in 20 μL TE buffer, pH 8.0. Library concentration and quality was assessed using a Fragment Analyzer (Advanced Analytical). Final libraries were pooled to equal molarity in Buffer EB (10 mM Tris-HCl, pH 8.5, Qiagen) containing 0.1% TWEEN 20 (Sigma Aldrich).

Sequencing was performed on an Illumina NextSeq 500 instrument at TATAA Biocenter (Gothenburg, Sweden) using 150 bp single-end reads. Raw FastQ files were subsequently processed as described[28] using Debarcer Version 0.3.0 (https://github.com/oicr-gsi/debarcer). Sequence reads with the same barcode were grouped into families for each amplicon. Barcode families with at least 20 reads, where ≥ 90% of the reads were identical, were required to compute consensus reads. FastQ files were deposited in the Sequence Read Archive under BioProject ID PRJNA375726.

Supporting information

S1 Fig. Melanoma promoter hotspot positions are often mutated in sun-exposed skin. Recurrent CTTCCG-related promoter hotspot sites identified in melanoma (mutated in > = 5/38 TCGA tumors) were examined for mutations in a sample of sun-exposed normal skin. The graphs show variant allele frequencies for mutations in genomic regions centered on these sites, based on whole genome sequencing data from sun-exposed normal eyelid skin obtained from Martincorena *et al.*[19]. Known population variants were excluded, but all other deviations from the reference sequence are shown regardless of allele frequency. (PDF)

S2 Fig. Conservation in melanoma promoter hotspot sites. PhastCons conservation scores at CTTCCG sites in melanoma promoter hotspot sites **(a)** and in 24 randomly chosen CTTCCG sites less than 500 bp from TSS of highly expressed genes, that were not mutated in any tumor **(b)**. PhastCons conservation scores were derived from multiple alignments of 100 vertebrate species and downloaded from the UCSC genome browser. (PDF)

S3 Fig. Transcription factor binding in melanoma promoter hotspot sites. Normalized scores for ChIP-seq peaks from 161 transcription factors in 91 cell types at NCTTCCGN sites (ENCODE track wgEncodeRegTfbsClusteredV3 obtained from the UCSC genome browser). **(a)** Promoter mutation hotspot sites. **(b)** 24 randomly chosen NCTTCCGN sites less than 500 bp from TSS of highly expressed genes that were not mutated in any tumor. In both panels, factors are ranked by mean signal across the 24 sites, with the 40 top factors being shown. Transcription factors from the ETS transcription factor family are underlined. The given

genomic position for each site, indicated in the x-axis labels, is the location of the motif CTTCCG.
(PDF)

S1 Table. Genomic positions close to transcription start sites recurrently mutated in 3/38 melanomas. The table complements Fig 1a and shows sites with a lower degree of mutation recurrence (3/38 melanomas, 8%), but is otherwise identical to Fig 1a. Approximately 50% of sites at this level of recurrence conform to the CTTCCG pattern.
(PDF)

S2 Table. The identified promoter hotspot positions are frequently mutated also in an independent set of melanomas. [a]Mutation frequency (fraction of tumors having a mutation) in the original analysis based on 38 TCGA tumors, as shown also in Fig 1a. [b]Mutation frequencies for these sites across 25 melanoma tumors as reported by Berger *et al.* [17]. [c]0.08 was previously obtained using a different mutation calling pipeline applied to the same data[10] while 0.04 refers to the calls provided by Berger *et al.* See main Fig 1a for an explanation of remaining columns.
(PDF)

S3 Table. Mutations in promoter hotspots in cSCC tumors. Melanoma hotspot positions were investigated in 8 cSCC tumors[37]. In cases where mutations are present, the variant allele frequency is shown for each individual sample (columns) and site (rows), with variant frequencies below 0.2 given within parentheses. [a]Mutation frequency across the 8 cSCC tumors[37], only considering mutations with a variant frequency of at least 0.2. [b]Mutation frequency across the 38 TCGA melanoma tumors. [c]Total number of called mutations as reported by Zheng *et al.* [18]. [d]Number of promoter hotspot mutations with variant frequency of at least 0.2. [e]Number of deleterious mutations in SCC driver genes with a variant frequency of at least 0.2. Non-synonymous mutations that were considered deleterious by PROVEAN[38] or damaging by SIFT[39] were counted as driver mutations.
(PDF)

S4 Table. Mutations in promoter hotspots in skin samples. Mutations in promoter hotspots were found at low variant frequencies in 8 peritumoral skin samples[18] that were available as matching normals for the cSCC tumors analyzed in S3 Table. In cases where mutations are present, the variant allele frequency is shown for each individual sample (columns) and site (rows), with variant frequencies below 0.2 given within parentheses. [a]Mutation frequency across the 8 samples, only considering mutations with a variant frequency of at least 0.2. [b]Mutation frequency across the 38 TCGA melanoma tumors; [c]Total number of called mutations as reported by Zheng *et al.* [18]. [d]Number of promoter hotspot mutations with variant frequency of at least 0.2.
(PDF)

S5 Table. Mutational characteristics and promoter hotspot mutations in different cancer types. [a]Median number of somatic mutations per tumor derived from whole-genome sequencing data. cSCC counts from Zheng *et al.* [18]. All other counts from Fredriksson *et al.* [10]. [b]UV-radiation as the mutational process driving tumor development. [c]Presence of mutational signatures 2, 7, 11 or 13 [8], all of which have elevated ratios of C to T mutations in CCT or TCT contexts, which allow for mutations of melanoma promoter hotspot sites. [d]Presence of TERT promoter mutations[10]. [e]Presence of melanoma promoter hotspot mutations. [f]Data not available.
(PDF)

S6 Table. Transcription factor motifs matching CTTCCG. Motif search in the JASPAR database using the tool TOMTOM[43]. The motif CTTCCG was compared with motifs in the databases for human transcription factors (HOCOMOCOv10).
(PDF)

S7 Table. Mutations in promoter hotspots and driver genes in cSCC tumors with NER deficiency. Melanoma promoter hotspot positions were investigated in whole genome sequencing data from cSCC tumors from 5 patients with germline NER DNA repair deficiency due to germline homozygous frameshift mutations (C_{940}del-1) in the *XPC* gene[18]. In cases where mutations are present, the variant allele frequency is shown for each individual sample (columns) and site (rows), with variant frequencies below 0.2 given within parentheses. [a]Mutation frequency across the 8 tumors, only considering mutations with a variant frequency of at least 0.2. [b]Mutation frequency across the 38 TCGA melanoma tumors. [c]Total number of called mutations as reported by Zheng *et al.* [18]. [d]Number of promoter hotspot mutations with variant frequency of at least 0.2. [e]Number of non-synonymous mutations in SCC driver genes with a variant frequency of at least 0.2. Non-synonymous mutations that were considered deleterious by PROVEAN[38] or damaging by SIFT[39] were counted as driver mutations.
(PDF)

S8 Table. PCR and adapter primer sequences used for SiMSen ultrasensitive amplicon sequencing.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: NJF EL.

Data curation: NJF JVdE.

Formal analysis: NJF SF JVdE.

Funding acquisition: EL AS.

Investigation: NJF KE AS SF.

Methodology: SF AS.

Resources: KE SF.

Software: NJF.

Supervision: EL AS.

Validation: JVdE.

Visualization: NJF EL.

Writing – original draft: NJF EL.

Writing – review & editing: NJF KE SF AS EL.

References

1. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, et al. (2016) Role of non-coding sequence variants in cancer. Nat Rev Genet 17: 93–108. https://doi.org/10.1038/nrg.2015.17 PMID: 26781813

2. Poulos RC, Sloane MA, Hesson LB, Wong JW (2015) The search for cis-regulatory driver mutations in cancer genomes. Oncotarget 6: 32509–32525. https://doi.org/10.18632/oncotarget.5085 PMID: 26356674

3. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, et al. (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518: 360–364. https://doi.org/10.1038/nature14221 PMID: 25693567

4. Lawrence M, Stojanov P, Polak P, Kryukov G, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218. https://doi.org/10.1038/nature12213 PMID: 23770567

5. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191–196. https://doi.org/10.1038/nature08658 PMID: 20016485

6. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, et al. (2016) Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature 532: 259–263. https://doi.org/10.1038/nature17437 PMID: 27075100

7. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N (2016) Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature 532: 264–267. https://doi.org/10.1038/nature17661 PMID: 27075101

8. Alexandrov L, Nik-Zainal S, Wedge D, Aparicio S, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500: 415–421. https://doi.org/10.1038/nature12477 PMID: 23945592

9. Araya CL, Cenik C, Reuter JA, Kiss G, Pande VS, et al. (2016) Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat Genet 48: 117–125. https://doi.org/10.1038/ng.3471 PMID: 26691984

10. Fredriksson NJ, Ny L, Nilsson JA, Larsson E (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet 46: 1258–1263. https://doi.org/10.1038/ng.3141 PMID: 25383969

11. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45: 1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849

12. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, et al. (2013) Highly recurrent TERT promoter mutations in human melanoma. Science 339: 957–959. https://doi.org/10.1126/science.1229259 PMID: 23348506

13. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, et al. (2013) TERT promoter mutations in familial and sporadic melanoma. Science 339: 959–961. https://doi.org/10.1126/science.1230062 PMID: 23348503

14. Helleday T, Eshtad S, Nik-Zainal S (2014) Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 15: 585–598. https://doi.org/10.1038/nrg3729 PMID: 24981601

15. Wei GH, Badis G, Berger MF, Kivioja T, Palin K, et al. (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J 29: 2147–2160. https://doi.org/10.1038/emboj.2010.106 PMID: 20517297

16. Colebatch AJ, Di Stefano L, Wong SQ, Hannan RD, Waring PM, et al. (2016) Clustered somatic mutations are frequent in transcription factor binding motifs within proximal promoter regions in melanoma and other cutaneous malignancies. Oncotarget 7: 66569–66585. https://doi.org/10.18632/oncotarget.11892 PMID: 27611953

17. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, et al. (2012) Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 485: 502–506. https://doi.org/10.1038/nature11071 PMID: 22622578

18. Zheng Christina L, Wang Nicholas J, Chung J, Moslehi H, Sanborn JZ, et al. (2014) Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. Cell Reports 9: 1228–1234. https://doi.org/10.1016/j.celrep.2014.10.031 PMID: 25456125

19. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, et al. (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348: 880–886. https://doi.org/10.1126/science.aaa6806 PMID: 25999502

20. Denisova E, Heidenreich B, Nagore E, Rachakonda PS, Hosen I, et al. (2015) Frequent DPH3 promoter mutations in skin cancers. Oncotarget.

21. Hollenhorst PC, Chandler KJ, Poulsen RL, Johnson WE, Speck NA, et al. (2009) DNA Specificity Determinants Associate with Distinct Transcription Factor Functions. PLoS Genet 5: e1000778. https://doi.org/10.1371/journal.pgen.1000778 PMID: 20019798

22. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research 22: 1798–1812. https://doi.org/10.1101/gr.139105.112 PMID: 22955990

23. Tanaka M, Ueda A, Kanamori H, Ideguchi H, Yang J, et al. (2002) Cell-cycle-dependent regulation of human aurora A transcription is mediated by periodic repression of E4TF1. J Biol Chem 277: 10719–10726. https://doi.org/10.1074/jbc.M108252200 PMID: 11790771

24. Gale JM, Nissen KA, Smerdon MJ (1987) UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. Proc Natl Acad Sci U S A 84: 6644–6648. PMID: 3477794

25. Brown DW, Libertini LJ, Suquet C, Small EW, Smerdon MJ (1993) Unfolding of nucleosome cores dramatically changes the distribution of ultraviolet photoproducts in DNA. Biochemistry 32: 10527–10531. PMID: 8399198

26. Pfeifer GP, Drouin R, Riggs AD, Holmquist GP (1992) Binding of transcription factors creates hot spots for UV photoproducts in vivo. Molecular and Cellular Biology 12: 1798–1804. PMID: 1549126

27. Tornaletti S, Pfeifer GP (1995) UV Light as a Footprinting Agent: Modulation of UV-induced DNA Damage by Transcription Factors Bound at the Promoters of Three Human Genes. Journal of Molecular Biology 249: 714–728. https://doi.org/10.1006/jmbi.1995.0331 PMID: 7602584

28. Stahlberg A, Krzyzanowski PM, Jackson JB, Egyud M, Stein L, et al. (2016) Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res 44: e105. https://doi.org/10.1093/nar/gkw224 PMID: 27060140

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

30. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22: 568–576. https://doi.org/10.1101/gr.129684.111 PMID: 22300766

31. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research 22: 1760–1774. https://doi.org/10.1101/gr.135350.111 PMID: 22955987

32. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, et al. (2013) Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal 6: pl1-. https://doi.org/10.1126/scisignal.2004088 PMID: 23550210

33. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, et al. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery 2: 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877

34. Akrami R, Jacobsen A, Hoell J, Schultz N, Sander C, et al. (2013) Comprehensive Analysis of Long Non-Coding RNAs in Ovarian Cancer Reveals Global Patterns and Targeted DNA Amplification. Plos One 8.

35. Rice P, Longden I, Bleasby A EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277. PMID: 10827456

36. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050. https://doi.org/10.1101/gr.3715005 PMID: 16024819

37. Durinck S, Ho C, Wang NJ, Liao W, Jakkula LR, et al. (2011) Temporal Dissection of Tumorigenesis in Primary Cancers. Cancer Discovery 1: 137–143. https://doi.org/10.1158/2159-8290.CD-11-0028 PMID: 21984974

38. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE 7: e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405

39. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protocols 4: 1073–1081. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590

40. Harm W (1969) Biological determination of the germicidal activity of sunlight. Radiat Res 40: 63–69. PMID: 4898082

41. Tornaletti S, Pfeifer GP (1995) UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. J Mol Biol 249: 714–728. https://doi.org/10.1006/jmbi.1995.0331 PMID: 7602584

42. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research 43: D805–D811. https://doi.org/10.1093/nar/gku1075 PMID: 25355519

43. Jolma A, Yan J, Whitington T, Toivonen J, Nitta Kazuhiro R, et al. (2013) DNA-Binding Specificities of Human Transcription Factors. Cell 152: 327–339. https://doi.org/10.1016/j.cell.2012.12.009 PMID: 23332764