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

Nils Johan Fredriksson[1], Kerryn Elliot[1], Stefan Filges[2], Anders Ståhlberg[2], and Erik Larsson[1*]

[1]Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, SE-405 30 Gothenburg, Sweden.

[2]Sahlgrenska Cancer Center, Department of Pathology and Genetics, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, SE-405 30 Gothenburg, Sweden.

*Correspondence to erik.larsson@gu.se

Summary

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.

Main text

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].

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 (**Table 1**, **Fig. 1a**). 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] (**Table 1**, **Fig. 1b**). The recurrent mutations occurred at cytosines positioned at the 5' end or one base upstream of the motif CTTCCG (**Fig. 1c**), and were normally C>T or CC>TT transitions (**Table 1**). Similar to most mutations in melanoma they were thus C>T changes in a dipyrimidine context, 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, 13 conformed to the same pattern, while the remaining two showed related sequence contexts (**Table 1**). Many less recurrent sites also showed the same pattern

62 (**Supplementary Table 1**). The signature described here matches the consensus binding
63 sequence of ETS family transcription factors (TFs)[15], and the results are consistent with recent
64 reports showing that ETS promoter sites are often recurrently mutated in melanoma[9] and that
65 such mutations preferably occur at cytosines upstream of the core TTCC sequence[16]. Thus,
66 while recurrent promoter mutations are common in melanoma, they consistently adhere to a
67 distinct sequence signature, which may argue against positive selection as a major causative
68 factor.

69 The recurrently mutated positions were next investigated in additional cancer cohorts,
70 first by confirming them in an independent melanoma dataset[17] (**Supplementary Table 2)**.
71 We found that the identified hotspot positions were often mutated also in cutaneous squamous
72 cell carcinoma (cSCC)[18] (**Supplementary Table 3**) as well as in sun-exposed skin[18,19], albeit
73 at lower variant frequencies (**Supplementary Fig. 1, Supplementary Table 4**). Additionally,
74 one of the mutations, upstream of *DPH3*, was recently described as highly recurrent in basal
75 cell skin carcinoma[20]. However, we did not detect mutations in these positions in 13 non-UV-
76 exposed cancer types (**Supplementary Table 5**). The hotspots are thus present in UV-
77 exposed samples of diverse cellular origins, but in contrast to the *TERT* promoter mutations
78 they are completely absent in non-UV-exposed cancers. This further supports that recurrent
79 mutations at the 5' end of CTTCCG elements are due to elevated susceptibility to UV-
80 induced mutagenesis in these positions.

81 Next, we considered additional properties that could support or argue against a
82 functional role for the recurrent mutations. We first noted a general lack of known cancer-
83 related genes among the affected promoters, with *TERT* as one of few exceptions (**Table 1**
84 and **Supplementary Table 1**, indicated in blue). Secondly, the recurrent promoter mutations
85 were not associated with differential expression of the nearby genes (**Table 1** and
86 **Supplementary Table 1**). This is in agreement with earlier investigations of some of these
87 mutations, which gave no conclusive evidence regarding influence on gene expression[9,16,20],
88 although it should be noted that significant association is lacking also for *TERT* in this
89 relatively small cohort[10]. Lastly, we found that when comparing different tumors there was a
90 strong positive correlation between the total number of the established hotspot positions that
91 were mutated and the genome-wide mutation load, both in melanoma (**Fig. 2a**; Spearman's r
92 = 0.88, P = 2.8e-13) and in cSCC (**Supplementary Table 3**; r = 0.78, P = 0.026). This is
93 again compatible with a passive model involving elevated mutation probability in the affected
94 positions. Importantly, this contrasted sharply with most of the major driver mutations in

95 melanoma, which were detected also in tumors with lower mutation load (**Fig. 2b**,
96 **Supplementary Table 3**). These different findings further reinforce the CTTCCG motif as a
97 strong mutational signature in melanoma.

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

112 CTTCCG elements have in various individual promoters been shown to be bound by
113 ETS factors such as ETS1, GABPA and ELF1[21], ELK4[22], and E4TF1[23]. This suggests that the
114 recurrently mutated CTTCCG elements could be substrates for ETS TFs. As expected,
115 matches to CTTCCG in the JASPAR database of TF binding motifs were mainly ETS-related
116 (**Supplementary Table 6**). Notably, recurrently mutated CTTCCG sites were evolutionarily
117 conserved to a larger degree than non-recurrently mutated but otherwise similar control sites,
118 further supporting that they constitute functional ETS binding sites (**Supplementary Fig. 2**).
119 This was corroborated by analysis of top recurrent CTTCCG sites in relation to ENCODE
120 ChIP-seq data for 161 TFs, which showed that the strongest and most consistent signals were
121 for ETS factors (GABPA and ELF1) (**Supplementary Fig. 3**).

122 The distribution of mutations across tumor genomes is shaped both by mutagenic and
123 DNA repair processes. Binding of TFs to DNA can increase local mutation rates by impairing
124 NER, and strong increases have been observed in predicted sites for several ETS factors[6,7]. It
125 is also established that contacts between DNA and proteins can modulate DNA damage
126 patterns by altering conditions for UV photoproduct formation[24-27]. In upstream regions of
127 XPC -/- cSCC tumors lacking global NER, we found that the CTTCCG signature still

128 conferred strongly elevated mutation probabilities compared to relevant trinucleotide contexts
129 (**Supplementary Fig. 4**), although to a lesser extent than in melanomas with functional NER
130 (**Fig. 4**). Transcription-coupled NER (TC-NER) may still be active in XPC -/- tumors, and the
131 signature could thus theoretically arise due to inhibition of TC-NER at CTTCCG elements.
132 However, the recurrently mutated positions were typically positioned upstream of the TSSs
133 (**Fig. 1**) and should not be subjected to this process. Additionally, TC-NER is strand-specific[14]
134 while the mutations occurred independently of strand orientation relative to the downstream
135 gene (**Supplementary Fig. 4**). The signature described here is thus unlikely explained by
136 impaired NER alone and may instead arise due to inhibition of other repair processes or due
137 to favorable conditions for UV lesion formation at the 5' end of ETS-bound CTTCCG
138 elements.

139 Finally, we sought to experimentally test our proposed model that the observed
140 promoter hotspots are due to localized vulnerability to mutagenesis by UV light. We
141 subjected human melanoma cells and keratinocytes to daily UV doses for a period of 5 or 10
142 weeks **(Fig. 5a)** and used an ultrasensititive error-correcting amplicon sequencing protocol,
143 SiMSen-Seq[28], to assay two of the observed promoter hotpots for mutations: *RPL13A*, the
144 most frequently mutated site in the tumor data (**Table** 1), and *DPH3*[10,20]. Between 36k and
145 82k error-corrected reads (>20x oversampling) were obtained for each of 16 different
146 conditions (**Fig. 5b-c**). Strikingly, subclonal mutations appeared specifically in expected
147 positions at both time points and in both cell lines at a frequency reaching up to 2.9% of
148 fragments (*RPL13A*, 10 weeks of exposure), while being absent in non-exposed control cells
149 (**Fig. 5d-e**). As predicted by the tumor data, mutations occurred primarily at cytosines
150 upstream of the TTCCG motif, with lower-frequency mutations occurring also in the central
151 cytosines. Few mutations were observed outside of the TTCCG context despite presence of
152 many cytosines in theoretically vulnerable configurations in the two amplicons (**Fig. 5d-e**,
153 underscored). These results further reinforce that recurrent mutation hotspots in promoters in
154 melanoma arise due to an exceptional vulnerability to UV mutagenesis in these positions.

155 In summary, we demonstrate that recurrent promoter mutations are common in
156 melanoma, but also that they adhere to a distinct sequence signature in a strikingly consistent
157 manner, arguing against positive selection as a major driving force. This model is supported
158 by several additional observations, including lack of cancer-relevant genes, lack of obvious
159 effects on gene expression, presence of the signature exclusively in UV-exposed samples of
160 diverse cellular origins, and strong positive correlation between genome-wide mutation load

161 and mutations in the affected positions. Crucially, exposing cells to UV light under controlled
162 conditions efficiently induces mutations specifically in affected sites. These results point to
163 limitations in conventional genome-wide derived trinucleotide models of mutational
164 signatures, and imply that extended sequence patterns as well as genomic context should be
165 taken into account to improve interpretation of somatic mutations in regulatory DNA.

166 **Methods**

167 **Mapping of somatic mutations**

168 Whole-genome sequencing data for 38 skin cutaneous melanoma (SKCM) metastases was
169 obtained from the Cancer Genome Atlas (TCGA) together with matching RNA-seq data.
170 Mutations were called using samtools[29] (command *mpileup* with default settings and
171 additional options *-q1* and *–B*) and VarScan[30] (command *somatic* using the default minimum
172 variant frequency of 0.20, minimum normal coverage of 8 reads, minimum tumor coverage of
173 6 reads and the additional option *–strand-filter 1*). Mutations where the variant base was
174 detected in the matching normal were not considered for analysis. The resulting set of
175 mutations was further processed by removing mutations overlapping germline variants
176 included in the NCBI dbSNP database, Build 146. The genomic annotation used was
177 GENCODE[31] release 17, mapped to GRCh37. The TSS of a gene was defined as the 5'most
178 annotated transcription start. Somatic mutation status for known driver genes was obtained
179 from the cBioPortal[32,33].

180 **RNA-seq data processing**

181 RNA-seq data was analyzed with resepect to the GENCODE[31] (v17) annotation using HTSeq-
182 count (http://www-huber.embl.de/users/anders/HTSeq) as previously described[34]. Differential
183 gene expression between tumors with and without mutations in promoter regions was
184 evaluated using the two-sided Wilcoxon rank sum test, avoiding assumptions about
185 distribution or directionality.

186 **Analyzed genomic regions**

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

191 **Mutation probability calculation**

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

198 **Evolutionary conservation data**

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

203 **ChIP-seq data**

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

207 **Analysis of whole genome sequencing data from UV-exposed skin**

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

214 **Analysis of whole genome sequencing data from cSCC tumors**

215 Whole genome sequencing data from 8 cSCC tumors and matching peritumoral skin samples
216 was obtained from Durinck *et al.*, 2011[37]. Whole genome sequencing data from cSCC tumors
217 and matching peritumoral skin from 5 patients with germline DNA repair deficiency due to
218 homozygous frameshift mutations (C_{940}del-1) in the *XPC* gene was obtained from Zheng *et*
219 *al.*, 2014[18]. Samtools[29] (command *mpileup* with a minimum mapping quality of 30, a
220 minimum base quality of 30 and additional option *–B*) was used to process the data and
221 VarScan[30] (command *mpileup2snp* counting all variants present in at least one read, with

222 minimum coverage of two reads and the additional strand filter option disabled) was used for
223 mutation calling. For the mutation probability analysis of cSCC tumors with NER deficiency
224 an additional filter was applied to only consider mutations with a total coverage of at least 10
225 reads and a variant frequency of at least 0.2. The functional impact of mutations in driver
226 genes was evaluated using PROVEAN[38] and SIFT[39]. Non-synonymous mutations that were
227 considered deleterious by PROVEAN or damaging by SIFT were counted as driver
228 mutations.

229 **Cell lines and UV treatments**

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

236 **DNA purification**

237 DNA was extracted based on Tornaletti and Pfeifer [41]. Briefly, cell pellets were lysed in 0.5
238 ml of 20 mM Tris-HCl (pH 8.0), 20 mM NaCl, 20mM EDTA, 1% (w/v) sodium dodecyl
239 sulfate, 600 mg/ml of proteinase K, and 0.5 ml of 150 mM NaCl, 10 mM EDTA. The solution
240 was incubated for two hours at 37°C. DNA was extracted twice with phenol-chloroform and
241 once with chloroform and precipitated by adding 0.1 vol. 3 M sodium acetate (pH 5.2), and
242 2.5 volumes of ethanol. The pellets were washed with 75% ethanol and briefly air-dried.
243 DNA was dissolved in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA (TE buffer) (all from Sigma
244 Aldrich). DNA was treated with RNAse for 1 hr at 37°C and phenol-chloroform extracted and
245 ethanol precipitated before dissolving in TE buffer.

246 **Ultrasensitive mutation analysis**

247 To detect and quantify mutations we applied SiMSen-Seq (Simple, Multiplexed, PCR-based
248 barcoding of DNA for Sensitive mutation detection using Sequencing) as described[28]. Briefly,
249 barcoding of 150 ng DNA was performed in 10 µL using 1x Phusion HF Buffer, 0.1U
250 Phusion II High-Fidelity polymerase, 200 µM dNTPs (all Thermo Fisher Scientific), 40 nM
251 of each primer (PAGE-purified, Integrated DNA Technologies) and 0.5M L-Carnitine inner
252 salt (Sigma Aldrich). Barcode primer sequences are shown in **Supplementary Table 8**. The
253 temperature profile was 98 °C for 3 min followed by three cycles of amplification (98 °C for

254 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
255 reaction was terminated by adding 20 µL TE buffer, pH 8.0 (Invitrogen, Thermo Fisher
256 Scientific) containing 30 ng/µL protease from *Streptomyces griseus* (Sigma Aldrich) at the
257 beginning of the 65 °C incubation step. Next, 10 µL of the diluted barcoded PCR products
258 were amplified in a 40 µL using 1x Q5 Hot Start High-Fidelity Master Mix (New England
259 BioLabs) and 400 nM of each sequencing adapter primer. Adapter primers are shown in
260 **Supplementary Table 8**. The temperature profile was 95 °C for 3 min followed by 40 cycles
261 of amplification (98 °C for 10 sec, 80 °C for 1 sec, 72 °C for 30 sec and 76 °C for 30 sec, with
262 a ramp rate of 0.2 °C/sec). The 40 µL PCR products were then purified using Agencourt
263 AMPure XP beads (Beckman-Coulter) according to the manufacturers' instructions using a
264 bead to sample ratio of 1. The purified product was eluted in 20 µL TE buffer, pH 8.0. Library
265 concentration and quality was assessed using a Fragment Analyzer (Advanced Analytical).
266 Final libraries were pooled to equal molarity in Buffer EB (10 mM Tris-HCl, pH 8.5, Qiagen)
267 containing 0.1% TWEEN 20 (Sigma Aldrich).

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

274 Acknowledgements

275 The results published here are in whole or part based upon data generated by The Cancer
276 Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and
277 the investigators and institutions who constitute the TCGA research network can be found at
278 http://cancergenome.nih.gov. We are most grateful to the patients, investigators, clinicians,
279 technical personnel, and funding bodies who contributed to TCGA, thereby making this study
280 possible. E.L. was supported by the Knut and Alice Wallenberg Foundation, the Swedish
281 Foundation for Strategic Research, the Swedish Medical Research Council, the Swedish
282 Cancer Society, the Åke Wiberg foundation, and the Lars Erik Lundberg Foundation for
283 Research and Education. A.S. was supported by Sahlgrenska Academy-ALF, the Swedish
284 Childhood Cancer Foundation, the Swedish Cancer Society, and the Wallenberg Centre for
285 Molecular and Translational Medicine. Computations were in part performed on resources

286 provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational
287 Science (UPPMAX) under project b2012108.

288 Author contributions

289 J.F and E.L. conceived the study; J.F performed bioinformatics analyses; J.F and E.L. wrote
290 the paper; K.E. performed cell culture and UV irradiation experiments; S.F. and A.S
291 performed ultrasensitive amplicon sequencing.

292 Competing financial interests

293 The applied SiMSen-Seq approach is patent pending (A.S.). The other authors declare no
294 competing financial interests or other conflict of interest.

295 References

296 1. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**,
297 93-108 (2016).
298 2. Poulos, R.C., Sloane, M.A., Hesson, L.B. & Wong, J.W. The search for cis-regulatory
299 driver mutations in cancer genomes. *Oncotarget* **6**, 32509-25 (2015).
300 3. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape
301 of cancer. *Nature* **518**, 360-364 (2015).
302 4. Lawrence, M. *et al.* Mutational heterogeneity in cancer and the search for new cancer-
303 associated genes. *Nature* **499**, 214 - 218 (2013).
304 5. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human
305 cancer genome. *Nature* **463**, 191-196 (2010).
306 6. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active
307 promoters in cancer genomes. *Nature* **532**, 259-263 (2016).
308 7. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N.
309 Nucleotide excision repair is impaired by binding of transcription factors to DNA.
310 *Nature* **532**, 264-267 (2016).
311 8. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
312 415 - 421 (2013).
313 9. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types
314 highlights a rich landscape of functional molecular alterations. *Nat Genet* **48**, 117-25
315 (2016).

316 10. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of
317 noncoding somatic mutations and gene expression alterations across 14 tumor types.
318 *Nat Genet* **46**, 1258-63 (2014).
319 11. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer
320 analysis project. *Nat Genet* **45**, 1113-20 (2013).
321 12. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma.
322 *Science* **339**, 957-9 (2013).
323 13. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science*
324 **339**, 959-61 (2013).
325 14. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational
326 signatures in human cancers. *Nat Rev Genet* **15**, 585-98 (2014).
327 15. Wei, G.H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in
328 vivo. *EMBO J* **29**, 2147-60 (2010).
329 16. Colebatch, A.J. *et al.* Clustered somatic mutations are frequent in transcription factor
330 binding motifs within proximal promoter regions in melanoma and other cutaneous
331 malignancies. *Oncotarget* **7**, 66569-66585 (2016).
332 17. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations.
333 *Nature* **485**, 502-506 (2012).
334 18. Zheng, Christina L. *et al.* Transcription Restores DNA Repair to Heterochromatin,
335 Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9**, 1228-1234
336 (2014).
337 19. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic
338 mutations in normal human skin. *Science* **348**, 880-886 (2015).
339 20. Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget*
340 (2015).
341 21. Hollenhorst, P.C. *et al.* DNA Specificity Determinants Associate with Distinct
342 Transcription Factor Functions. *PLoS Genet* **5**, e1000778 (2009).
343 22. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions
344 bound by 119 human transcription factors. *Genome Research* **22**, 1798-1812 (2012).
345 23. Tanaka, M. *et al.* Cell-cycle-dependent regulation of human aurora A transcription is
346 mediated by periodic repression of E4TF1. *J Biol Chem* **277**, 10719-26 (2002).
347 24. Gale, J.M., Nissen, K.A. & Smerdon, M.J. UV-induced formation of pyrimidine
348 dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases.
349 *Proc Natl Acad Sci U S A* **84**, 6644-8 (1987).

350 25. Brown, D.W., Libertini, L.J., Suquet, C., Small, E.W. & Smerdon, M.J. Unfolding of
351 nucleosome cores dramatically changes the distribution of ultraviolet photoproducts in
352 DNA. *Biochemistry* **32**, 10527-10531 (1993).
353 26. Pfeifer, G.P., Drouin, R., Riggs, A.D. & Holmquist, G.P. Binding of transcription
354 factors creates hot spots for UV photoproducts in vivo. *Molecular and Cellular*
355 *Biology* **12**, 1798-1804 (1992).
356 27. Tornaletti, S. & Pfeifer, G.P. UV Light as a Footprinting Agent: Modulation of UV-
357 induced DNA Damage by Transcription Factors Bound at the Promoters of Three
358 Human Genes. *Journal of Molecular Biology* **249**, 714-728 (1995).
359 28. Stahlberg, A. *et al.* Simple, multiplexed, PCR-based barcoding of DNA enables
360 sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Res* **44**,
361 e105 (2016).
362 29. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
363 2078-2079 (2009).
364 30. Koboldt, D.C. *et al.* VarScan 2: Somatic mutation and copy number alteration
365 discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576 (2012).
366 31. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The
367 ENCODE Project. *Genome Research* **22**, 1760-1774 (2012).
368 32. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles
369 Using the cBioPortal. *Sci. Signal.* **6**, pl1- (2013).
370 33. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring
371 Multidimensional Cancer Genomics Data. *Cancer Discovery* **2**, 401-404 (2012).
372 34. Akrami, R. *et al.* Comprehensive Analysis of Long Non-Coding RNAs in Ovarian
373 Cancer Reveals Global Patterns and Targeted DNA Amplification. *Plos One* **8**(2013).
374 35. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology
375 Open Software Suite. *Trends in Genetics* **16**, 276-277.
376 36. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and
377 yeast genomes. *Genome Research* **15**, 1034-1050 (2005).
378 37. Durinck, S. *et al.* Temporal Dissection of Tumorigenesis in Primary Cancers. *Cancer*
379 *Discovery* **1**, 137-143 (2011).
380 38. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the Functional
381 Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7**, e46688 (2012).

382 39. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous
383 variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**, 1073-1081
384 (2009).
385 40. Harm, W. Biological determination of the germicidal activity of sunlight. *Radiat Res*
386 **40**, 63-9 (1969).
387 41. Tornaletti, S. & Pfeifer, G.P. UV light as a footprinting agent: modulation of UV-
388 induced DNA damage by transcription factors bound at the promoters of three human
389 genes. *J Mol Biol* **249**, 714-28 (1995).
390 42. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in
391 human cancer. *Nucleic Acids Research* **43**, D805-D811 (2015).

392

393

Rec[a]	Chr[b]	Position	Ref[c]	Var[d]	Sequence context[e]	Dist[f]	Gene[g]	Expr. tier[h]	P[i]	Dist[j]	Gene[k]	Expr.tier[l]	P[m]
11	19	49990694	C	T	TCCGGACATTCTTCCGGTTGG	-116	*RPL13A*	3	1				
10	5	1295250	C	T	CCCGACCCCTCCCGGGTCCCC	-88	*TERT*	1	0.456				
7	16	2510095	C	T	AGCCACGCCCCTTCCGGGAGG	15	*C16orf59*	2	0.679				
7	5	1295228	C	T	GCCCAGCCCCCTCCGGGCCCT	-66	*TERT*	1	0.228				
5	2	26101489	C	T	CGCCCCCGCCCTTCCGGTCTC	-104	*ASXL2*	2	0.796				
5	10	105156316	C	T	CAAATCCCGCCCTTCCGATTC	-88	*PDCD11*	3	0.195	-93	*USMG5*	3	0.28
5	11	61735192	C	T	GAGCCCGCTCCTTCCGGTGGG	-60	*FTH1*	3	1	-260	*AP003733.1*	3	0.262
5	11	61735191	C	T	CGAGCCCGCTCCTTCCGGTGG	-59	*FTH1*	3	0.364	-261	*AP003733.1*	3	0.101
5	9	133454938	C	T/+T	CCGGCTTTCCCTTCCGCCGGA	-54	*FUBP3*	3	0.342				
5	17	79849513	C	T	CGCGTGAGGCCTTCCGGTGCC	-51	*ALYREF*	3	0.666				
5	22	31556121	C	T	AAATTAACCTCTTCCGGTTGG	-46	*RNF185*	3	0.388				
5	13	41345346	C	T	CCCGCCCTCTCTTCCGCTTCC	-37	*MRPS31*	3	0.262				
5	3	16306505	C	A/T/G	AGGACTAGCCCTTCCGGCGCA	-26	*DPH3*	3	0.0108[n]	-200	*OXNAD1*	3	0.181
5	19	17970682	C	T	GAGGGCGGGTCTTCCGGTAGT	-2	*RPL18A*	3	0.11				
5	16	2510096	C	T	GAGCCACGCCCCTTCCGGGAG	16	*C16orf59*	2	0.545				
5	8	124054557	C	T	CGAAACTTCCCCTTCCGGCGA	106	*DERL1*	3	0.0697	350	*WDR67*	3	0.931
5	5	1295242	C	T	CTCCCGGGTCCCCGGCCCAGC	-80	*TERT*	1	0.73				
4	10	27443328	C	T	AGCGCCTCGCCTTCCGGGGCG	-424	*MASTL*	2	0.122				
4	11	111797698	C	T	GTAGACAGGCCTTCCGGCCCC	-169	*DIXDC1*	2	0.651				
4	12	54582890	C	T	ATTTAGTGCGCCTTCCGGGAT	-112	*SMUG1*	3	0.433				
4	12	54582889	C	T	TTTAGTGCGCCTTCCGGGATT	-111	*SMUG1*	3	0.868				
4	1	43824529	C	T	AGGGGGCGGGCCTTCCGGGGA	-96	*CDC20*	3	0.405				
4	9	91933357	C	T	CCCGCCCTTTCTTCCGGCCGG	-63	*SECISBP2*	3	0.757				
4	19	7459940	C	T	GGGCACGCCTCCTTCCGGGTC	-58	*ARHGEF18*	2	0.207				
4	19	7459941	C	T	GGCACGCCTCCTTCCGGGTCA	-57	*ARHGEF18*	2	0.981				
4	3	52322052	C	T	GACGTCACTTCCGGCCCCCTA	-16	*WDR82*	3	0.981				
4	21	34100374	C	T	CGGGGCGGATCTTCCGCCCCC	-15	*SYNJ1*	2	0.244				
4	2	128615744	C	T	AGACCACGCCCCTTCCGCGGC	-13	*POLR2D*	3	0.831				
4	6	30640796	C	T	AAGTACAGCCCCTTCCGGGCT	18	*DHX16*	3	0.161				
4	19	17830242	C	T	GTCTTCAGCCCTTCCGGTGCG	192	*MAP1S*	3	0.191				
4	12	49412648	C	T	GGTTCCTTGCCTTCCGCCCCA	332	*PRKAG1*	3	0.306				
4	19	2151793	C	T	ACTCCGCCTTCTTCCTAGTTC	-228	*AP3D1*	3	0.943				

Table 1 | Recurrent somatic mutations in promoter regions in melanoma are characterized by a distinct sequence signature. 38 melanomas were analyzed for individual recurrently mutated bases in promoter regions. The table shows all 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 10 bases upstream and downstream of the mutation. Sequences were reverse complemented as required to always show the pyrimidine-containing strand with respect to the central mutated base (highlighted in gray). The motif CTTCCG is highlighted in yellow. [f]Distance from mutation to the 5' most TSS in GENCODE 17. Negative values indicate upstream location of mutation. [g]Closest gene. Genes included in the Cancer gene census (http://cancer.sanger.ac.uk [42]) are highlighted in blue. [h]Genes were sorted by increasing mean expression (all samples) and assigned to expression tiers 1 to 3 with 3 being the highest. [i]*P*-values from a two-sided Wilcoxon rank sum test of differential expression of the gene between tumors with and without the mutation. [j]Distance from mutation to the second closest 5'-most TSS in GENCODE 17, when present within 500 bp. Negative values indicate an upstream location. [k]Second closest gene. [l]Same as column h for the second gene, when applicable. [m]*P*-values from a two-sided Wilcoxon rank sum test of differential expression of the gene comparing tumors with and without the mutation. [n]Significant differential expression could not be seen when the analysis was repeated in a larger dataset[10].

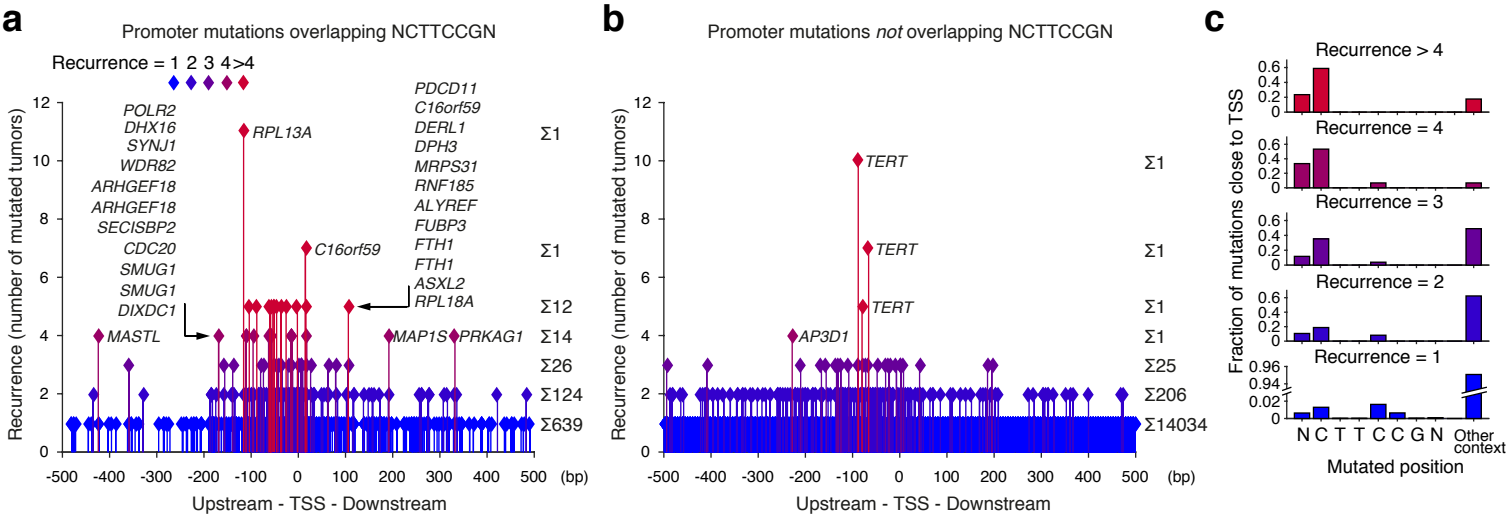


Figure 1 | Recurrent somatic mutations in promoter regions in melanoma are characterized by a distinct sequence signature. Whole genome sequencing data from 38 melanomas were analyzed for individual recurrently mutated bases in promoter regions, and most highly recurrent positions were found to share a distinct sequence context, CTTCCG (see **Table 1**). (**a**) All mutations occurring within +/- 500 bp of a TSS while overlapping with or being adjacent to the motif CTTCCG. The distance to the nearest TSS and the degree of recurrence (number of mutated tumors) is indicated. (**b**) Similar to panel **a**, but instead showing mutations *not* overlapping or adjacent to CTTCCG. (**c**) Positional distribution across the sequence NCTTCCGN for mutations indicated in panel **a**.

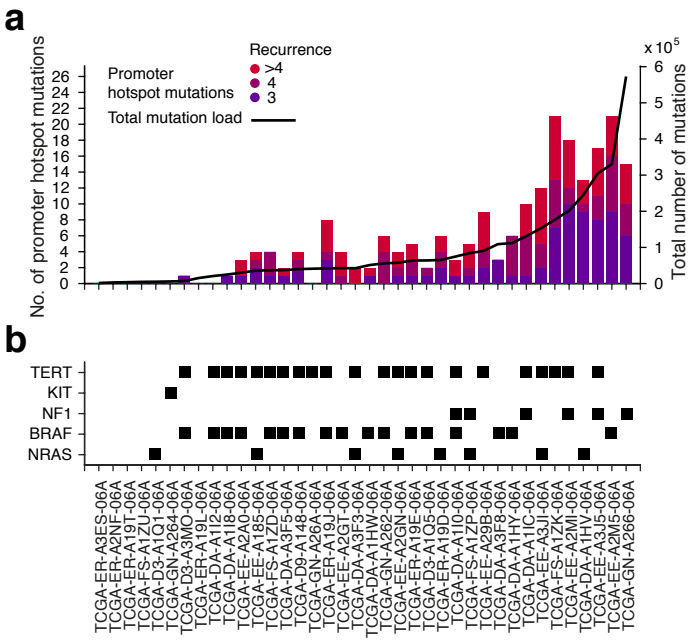


Figure 2 | Positive correlation between promoter hotspot mutations and total mutational load across melanomas. (**a**) Bars, left axis: Number of mutations occurring in the established recurrent CTTCCG-related promoter positions (>= 3 tumors) in each of the 38 samples. Line, right axis: Total mutational 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.

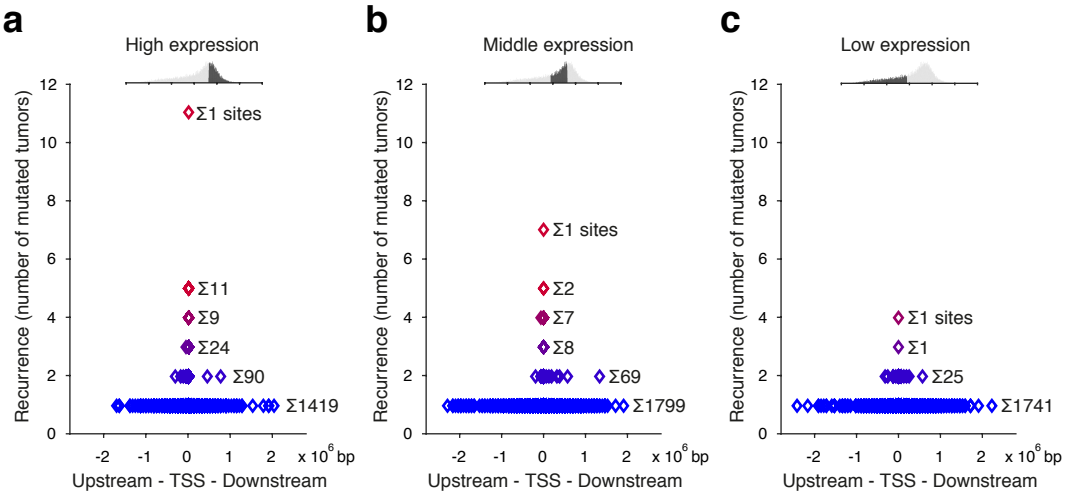


Figure 3 | Recurrent mutations at CTTCCG sites are observed only near active promoters. (**a-c**) Genes were assigned to three expression tiers by increasing mean expression across the 38 melanomas. The graphs show, on the x-axis, the distance to the nearest annotated TSS for all mutations overlapping 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.

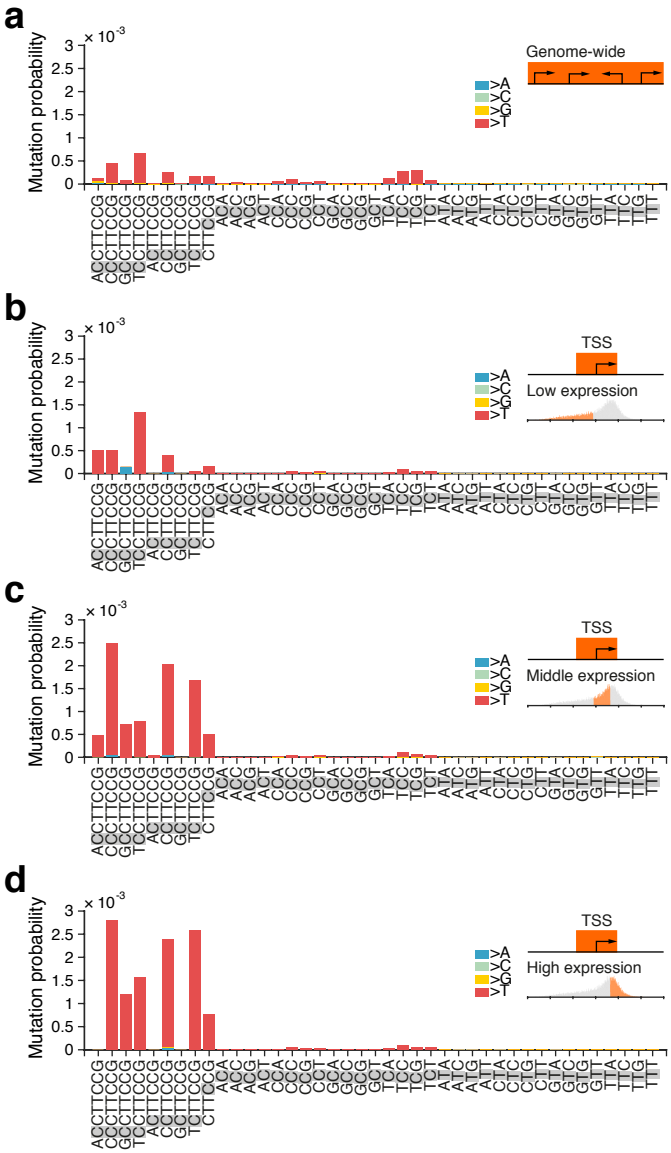


Figure 4 | Mutation probabilities for CTTCCG-related sequence contexts compared to trinucleotides. The mutated position in each sequence context is shaded in gray. Bar colors indicate the substituting bases (mainly C>T). Mutation probabilities were calculated genome-wide (**a**), or only considering mutations less than 500 bases from TSS of genes with a low (**b**), middle (**c**) or high (**d**) mean expression level.

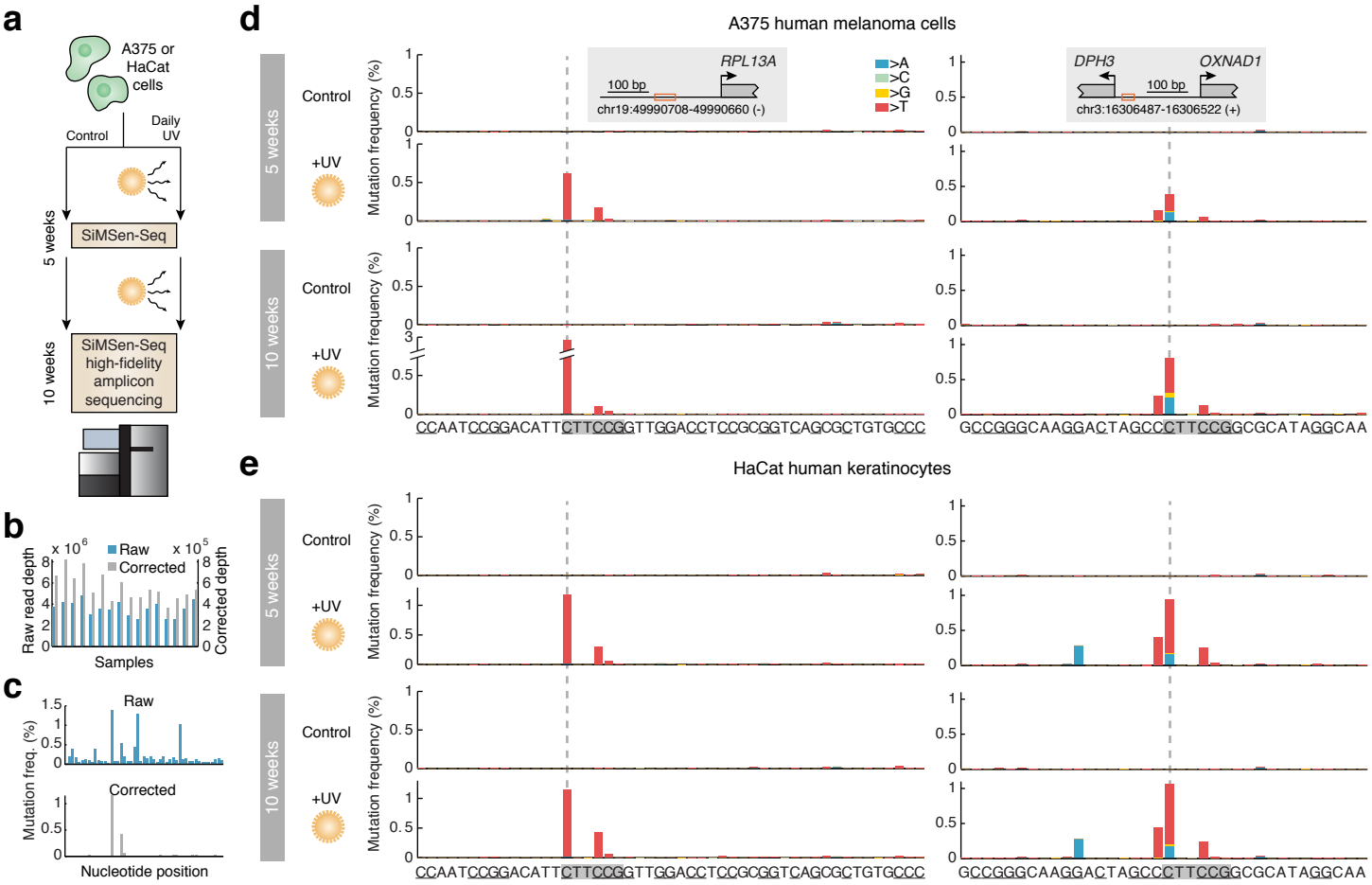
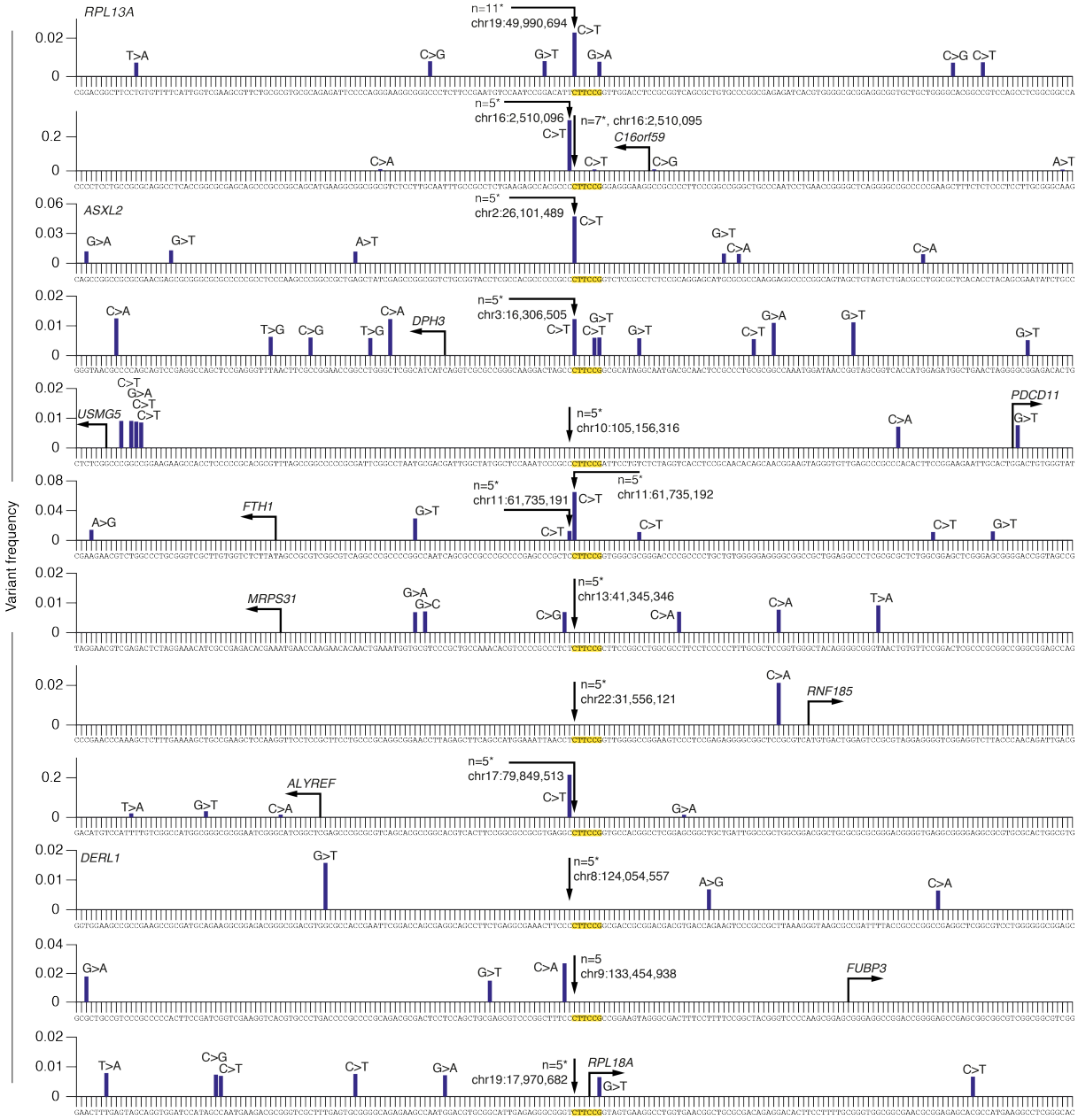
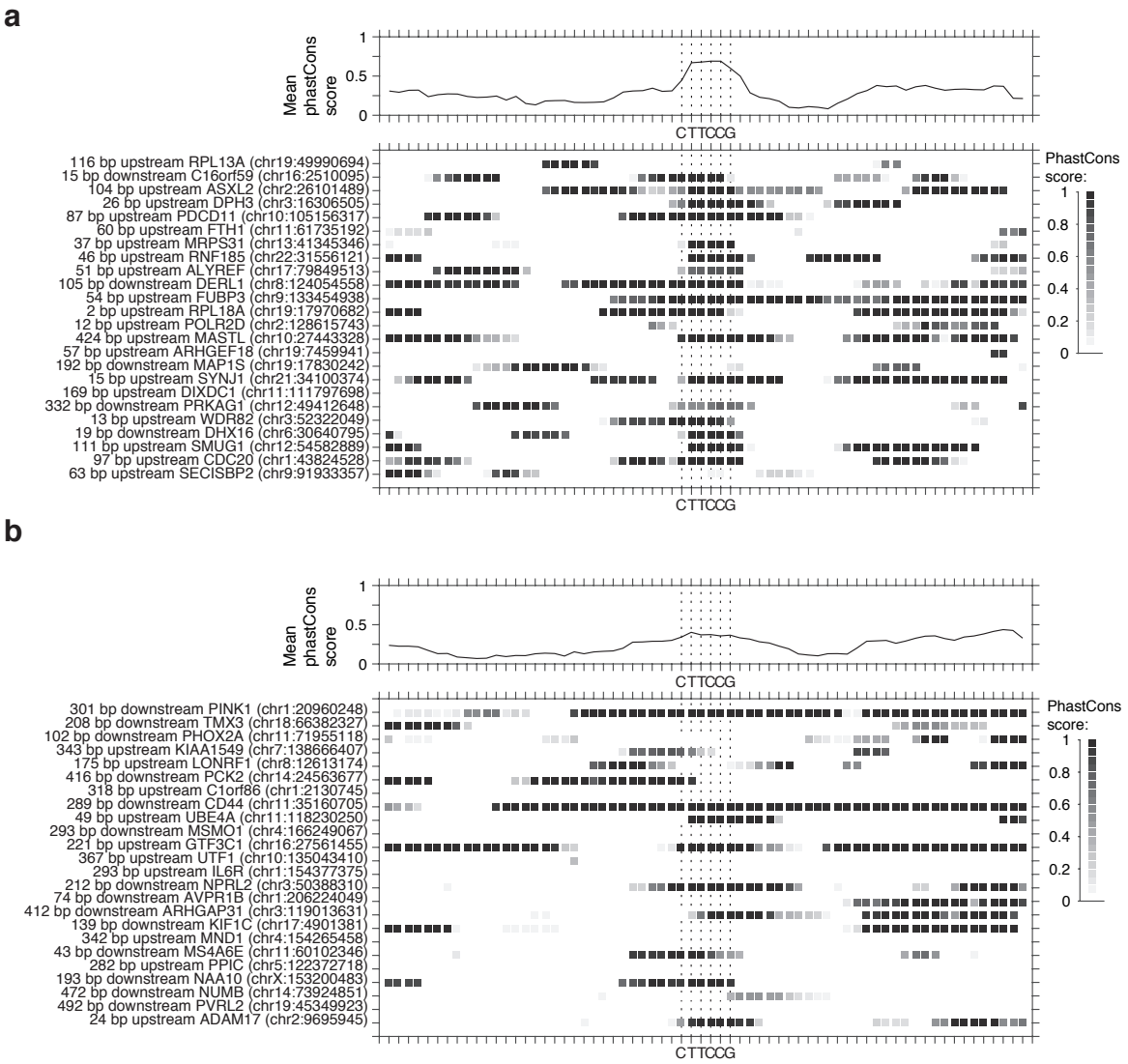


Figure 5 | UV exposure of cultured cells induces mutations specifically at CTTCCG-related promoter hotspot sites. (**a**) Human cells (A375 melanoma cells or HaCat human keratinocytes) were subjected daily UV doses (254 nm, 36 J/m^2 once a day, 5 days a week). An ultrasensitive amplicon sequencing procotol, 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 between 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 hotspot sites are indicated, and other possible UV-susceptible sites (cytosines flanking pyrimidines) are underscored. The amplicon sizes were 49 and 36 bp, respectively.

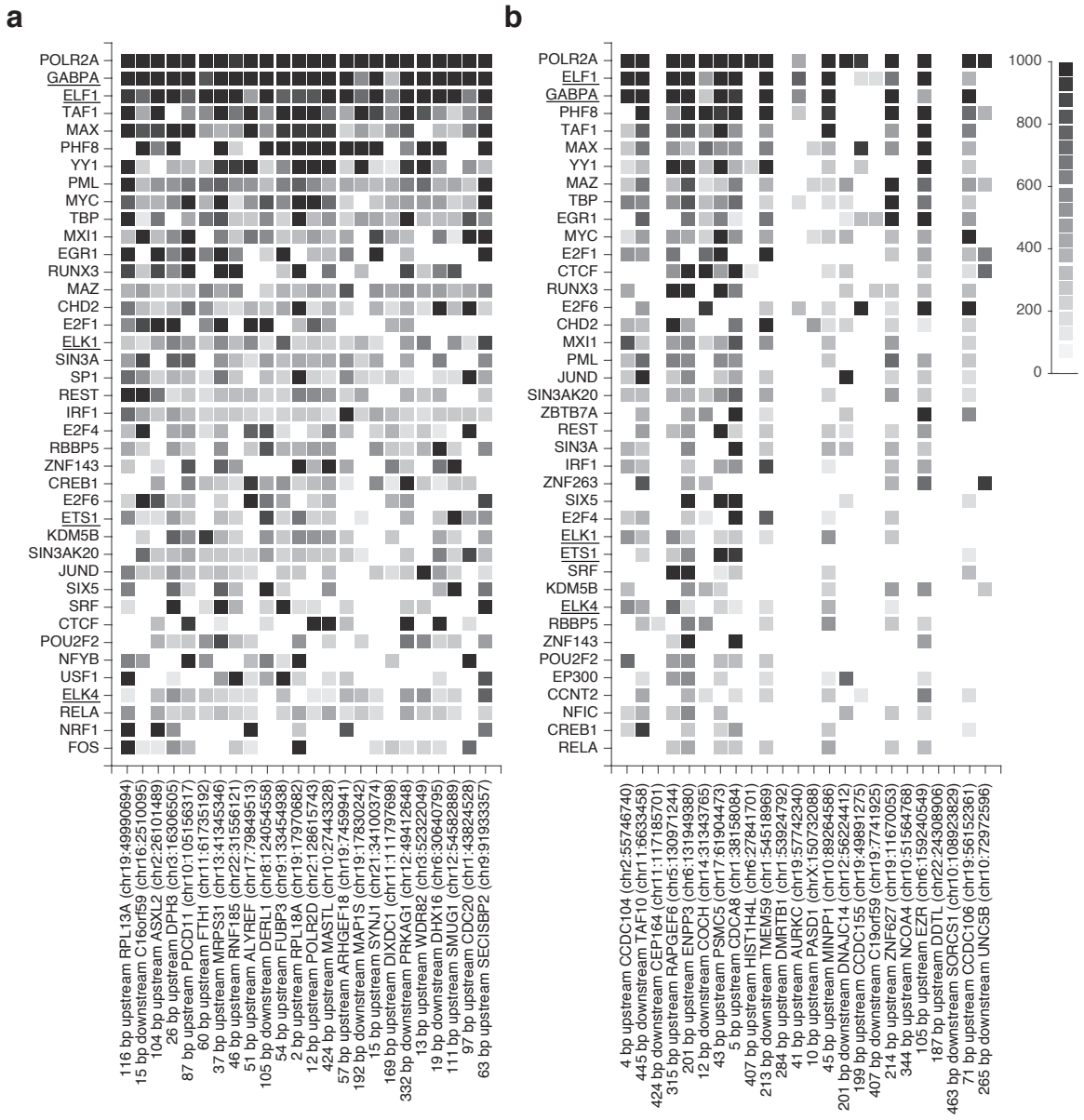
Supplementary figures



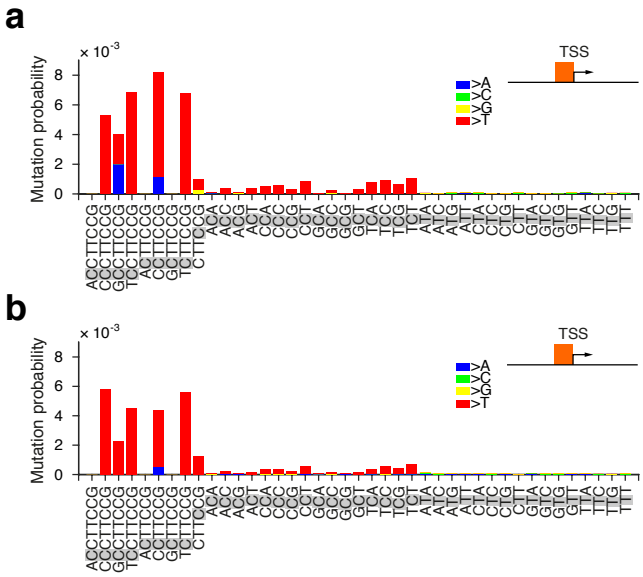
Supplementary Figure 1 | Melanoma promoter hotspot positions are often mutated in sun-exposed skin. Recurrent CTTCCG-related promoter hotspot sites identified in melanoma (mutated in >=5/38 TCGA tumors) were examined for mutations in a sample of sun-exposed normal skin. The graphs show variant allele 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.*[2]. Known population variants were excluded, but all other deviations from the reference sequence are shown regardless of allele frequency.



Supplementary Figure 2 | 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.



Supplementary Figure 3 | 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.



Supplementary Figure 4 | Mutation probabilities for CTTCCG-related sequence contexts compared to trinucleotides in SCC tumors with NER deficiency. 5 SCC tumors from patients with defective global genome NER[3] were screened for mutations within 500 bp upstream of TSSs, considering only genes in the upper mean expression tier level as defined earlier based on TCGA data. Mutation probabilities for different sequence contexts (trinucleotides and CTTCCG-related) were calculated in these regions, considering the template strand **(a)** and non-template strand **(b)** separately. The mutated position in each sequence context is shaded in gray. Bar colors indicate the substituting bases (mainly C>T). Only upstream regions were considered to avoid influence from transcription-coupled repair. The assignment to template and non-template strands was determined by the transcription direction of the downstream gene. Notably, transcription coupled repair is a strand-specific process, but elevated probabilities for CTTCCG-related context compared to trinucleotides were observed regardless of the strand orientation.

Supplementary tables

Rec[a]	Chr[b]	Position	Ref[c]	Var[d]	Sequence context[e]	Dist[f]	Gene[g]	Expr. tier[h]	*P*[i]	Dist[j]	Gene[k]	Expr.tier[l]	*P*[m]
3	6	24721423	C	T	CCCGCCACTCCTTCCGCCCCC	-359	*C6orf62*	3	0.13				
3	12	498776	C	T	GTGACGCTTTCTTCCGGCGCG	-156	*KDM5A*	3	0.787	263	*CCDC77*	3	0.0513
3	9	131038413	C	T	GCCACGCCCCCTTCCGCTTCA	-139	*GOLGA2*	3	0.871				
3	1	25559064	C	T	AGCCCCGCCCCTTCCGGGAGG	-80	*SYF2*	3	0.552				
3	2	73964607	C	T	CCCGCCCATTCTTCCGCCTCC	-80	*TPRKB*	3	1				
3	22	35795975	C	T	ACCTCGCCTCCTTCCGGGCTC	-80	*MCM5*	3	0.234				
3	19	1812349	C	T	CCCCCGGCCCCCTTCCGGGGT	-74	*ATP8B3*	2	0.516				
3	6	31940123	C	T	AAAATAGGGTCTTCCGGCGCA	-54	*DOM3Z*	3	0.588				
3	14	50779320	C	T	CGGCTTCTTTCTTCCGGCTCG	-54	*L2HGDH*	2	0.588	274	*ATP5S*	2	0.482
3	4	2936631	C	T	ACGTTCTCTTCCGGCCGGAGT	-45	*MFSD10*	3	0.0832				
3	1	100598553	C	T	CCATCGGATTCTTCCGGTTCT	-42	*SASS6*	2	0.588	-152	*TRMT13*	2	0.0513
3	3	101280671	C	T	CCGCCCTCTCCCTTCCGGCGC	-34	*TRMT10C*	3	0.482				
3	11	1330918	C	T	GGCACCGCCCCTTCCGGCTCT	-34	*TOLLIP*	3	0.279				
3	22	43011002	C	T	CGTCCCCGCCCCTTCCGGTTC	-34	*POLDIP3*	3	0.516				
3	2	70056751	C	T	TTGCCCCGCCCCTTCCGGAGG	-22	*GMCL1*	2	0.957				
3	13	29233226	C	T	GGACGCACTTCCGGCGGATGT	-14	*POMP*	3	0.745				
3	10	7830002	C	T	GCCCCACCTCCTTCCGCCTCT	-12	*KIN*	2	0.871	-89	*ATP5C1*	2	0.588
3	2	198318145	C	T	CCCCTTCTTCCCTTCCGGGTT	-1	*COQ10B*	3	0.417				
3	7	23338823	C	T	CCAAGTAGCTCTTCCGGGTCA	5	*MALSU1*	2	0.213				
3	6	170893742	C	A/T	CGCCTCTTGCCTTCCGGCCCG	6	*PDCD2*	3	0.871				
3	13	41837733	C	T	TGGTTCACTTCTTCCGGGTTA	9	*MTRF1*	2	0.626				
3	11	46958262	C	T	TCCCGTCCCCCTTCCGCCGCG	15	*C11orf49*	3	1				
3	20	57607411	C	T	CGCCCCGCCTCTTCCGCTTCT	26	*ATP5E*	3	0.588				
3	X	153059915	C	T	ATTCACGCCCCCTTCCGGCGC	63	*IDH3G*	3	0.256				
3	16	83841526	C	T	CCTCGAGGCCCTTCCGGTGCG	79	*HSBP1*	3	0.665				
3	8	124054558	C	T	GAAACTTCCCCTTCCGGCGAC	105	*DERL1*	3	0.914	351	*WDR67*	3	1
3	14	20585071	C	T	CTGTGTTTTTCCTCCTTATCT	-494	*OR4K17*	1	-[o]				
3	5	137800769	C	T	GGCGGGGGATCCTTCCTGCTC	-409	*EGR1*	3	0.705				
3	3	12883297	C	T	GAAGTAAATTCCTCCCTCCAC	-210	*RPL32*	3	0.914				
3	3	122135048	C	T	ATGTTCATTTCCGGTCTTTTT	-166	*WDR5B*	2	0.787				
3	11	47448149	C	T	TTCTCCCTCCCTTTCCGGTTT	-156	*PSMC3*	3	0.957				
3	2	65283371	C	T	CCCCCTACTTCGTCCTGGCCT	-134	*CEP68*	2	0.588				
3	8	67579583	C	T	ACTTGTAGTTCCTTCTGGACT	-131	*VCPIP1*	2	0.482	-267	*SGKL*	2	0.871
3	19	54641318	C	T	TGCCCCCTTTCGCGGATTGGG	-125	*CNOT3*	3	0.705				
3	16	68119138	C	T	CACGTGACTTCCTTCCTTCTC	-108	*NFATC3*	2	0.279				
3	1	38478324	C	T	AGCCGGCTTTCAGGAAGTACG	-89	*UTP11L*	3	0.279				
3	8	57124164	C	T	GCCCACTCCTCCCGCCTCGGC	-80	*CHCHD7*	3	0.745	-305	*PLAG1*	3	0.304
3	8	56987141	C	T	CAGGAAATATCCCGGGCCCTA	-72	*RPS20*	3	0.213				
3	15	90931383	C	T	GGCCCCGCCTCTTTCCGGCCG	-66	*IQGAP1*	3	0.159				
3	19	48867542	C	T	CCCCCTCCCTCTTTCCGCCTA	-48	*TMEM143*	2	0.665	-109	*SYNGR4*	2	0.664
3	19	48248748	C	T	GCGCAGATTTCCCACCCTCTT	-30	*GLTSCR2*	3	0.116				
3	2	234763235	C	T	TCCCTGCCTTCCTCTCGGTTT	-23	*HJURP*	2	0.957				
3	1	234509179	C	T	CTTCCTGTTTCGCTTTTATCT	-22	*COA6*	3	0.516				
3	16	2510073	C	T	AAGGCCGCCCCTTCCCGGCCG	-7	*C16orf59*	2	0.705				
3	14	55658403	C	T	ATTCAAATATCGCACGGAGCA	-7	*DLGAP5*	2	0.829				
3	19	36870099	C	T	GCTCGCAGTTCTTTCCGGCTT	2	*ZFP14*	2	0.256				
3	12	100660920	C	T	GACGTCACTTCCTGCCGGTTC	3	*SCYL2*	3	0.417	-63	*DEPDC4*	3	0.705
3	14	31028336	C	T	GCCGACCGCTCTTTCCGGGTT	8	*G2E3*	2	0.552				
3	6	34855824	C	T	GATCTTACTTCCTGTCGTCGC	42	*TAF11*	3	0.957				
3	15	40675107	C	T	GAGTGCGATTCCCACCCACTT	186	*KNSTRN*	3	0.829				
3	1	242011463	C	T	GACGTCACATCCTCTGGGCGG	195	*EXO1*	2	0.914				

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

Rec	Chr	Pos	Ref	Var	Context	Dist	Gene	Freq[a]	Berger freq.[b]
11	19	49990694	C	T	TCCGGACATTCTTCCGGTTGG	-116	*RPL13A*	0,29	0,12
10	5	1295250	C	T	CCCGACCCCTCCCGGGTCCCC	-88	*TERT*	0,26	0,48
7	16	2510095	C	T	AGCCACGCCCCTTCCGGGAGG	15	*C16orf59*	0,18	0,12
7	5	1295228	C	T	GCCCAGCCCCCTCCGGGCCCT	-66	*TERT*	0,18	0,2
5	2	26101489	C	T	CGCCCCCGCCCTTCCGGTCTC	-104	*ASXL2*	0,13	0,04
5	10	105156316	C	T	CAAATCCCGCCCTTCCGATTC	-88	*PDCD11*	0,13	0,08
5	11	61735192	C	T	GAGCCCGCTCCTTCCGGTGGG	-60	*FTH1*	0,13	0,08
5	11	61735191	C	T	CGAGCCCGCTCCTTCCGGTGG	-59	*FTH1*	0,13	0,04
5	9	133454938	C	T/+T	CCGGCTTTCCCTTCCGCCGGA	-54	*FUBP3*	0,13	0
5	17	79849513	C	T	CGCGTGAGGCCTTCCGGTGCC	-51	*ALYREF*	0,13	0,04
5	22	31556121	C	T	AAATTAACCTCTTCCGGTTGG	-46	*RNF185*	0,13	0,08
5	13	41345346	C	T	CCCGCCCTCTCTTCCGCTTCC	-37	*MRPS31*	0,13	0
5	3	16306505	C	A/T/G	AGGACTAGCCCTTCCGGCGCA	-26	*DPH3*	0,13	0,04[c]
5	19	17970682	C	T	GAGGGCGGGTCTTCCGGTAGT	-2	*RPL18A*	0,13	0,12
5	16	2510096	C	T	GAGCCACGCCCCTTCCGGGAG	16	*C16orf59*	0,13	0,08
5	8	124054557	C	T	CGAAACTTCCCCTTCCGGCGA	106	*DERL1*	0,13	0
5	5	1295242	C	T	CTCCCGGGTCCCCGGCCCAGC	-80	*TERT*	0,13	0
4	10	27443328	C	T	AGCGCCTCGCCTTCCGGGGCG	-424	*MASTL*	0,11	0,04
4	11	111797698	C	T	GTAGACAGGCCTTCCGGCCCC	-169	*DIXDC1*	0,11	0
4	12	54582890	C	T	ATTTAGTGCGCCTTCCGGGAT	-112	*SMUG1*	0,11	0
4	12	54582889	C	T	TTTAGTGCGCCTTCCGGGATT	-111	*SMUG1*	0,11	0,08
4	1	43824529	C	T	AGGGGGCGGGCCTTCCGGGGA	-96	*CDC20*	0,11	0,08
4	9	91933357	C	T	CCCGCCCTTTCTTCCGGCCGG	-63	*SECISBP2*	0,11	0
4	19	7459940	C	T	GGGCACGCCTCCTTCCGGGTC	-58	*ARHGEF18*	0,11	0,08
4	19	7459941	C	T	GGCACGCCTCCTTCCGGGTCA	-57	*ARHGEF18*	0,11	0,08
4	3	52322052	C	T	GACGTCACTTCCGGCCCCCTA	-16	*WDR82*	0,11	0
4	21	34100374	C	T	CGGGGCGGATCTTCCGCCCCC	-15	*SYNJ1*	0,11	0,04
4	2	128615744	C	T	AGACCACGCCCCTTCCGCGGC	-13	*POLR2D*	0,11	0,04
4	6	30640796	C	T	AAGTACAGCCCCTTCCGGGCT	18	*DHX16*	0,11	0
4	19	17830242	C	T	GTCTTCAGCCCTTCCGGTGCG	192	*MAP1S*	0,11	0
4	12	49412648	C	T	GGTTCCTTGCCTTCCGCCCCA	332	*PRKAG1*	0,11	0
4	19	2151793	C	T	ACTCCGCCTTCTTCCTAGTTC	-228	*AP3D1*	0,11	0

Supplementary Table 2 | 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 main **Table 1**. [b]Mutation frequencies for these sites across 25 melanoma tumors as reported by Berger *et al.* [4]. [c]0.08 was previously obtained using a different calling pipeline applied to the same data[5] while 0.04 refers to the calls provided by Berger *et al*. See main **Table 1** for an explanation of remaining columns.

Sample	WT9	WT11	WT12	WT10	WT13	WT8	WT6	WT7	Total mut. freq. [a]	TCGA SKCM mut. freq. [b]
RPL13A chr19:49990694	(0.19)	(0.083)	-	-	0.33	0.54	0.47	(0.051)	0.38	0.29
C16orf59 chr16:2510095	(0.08)	-	-	-	-	-	-	-	0	0.18
ASXL2 chr2:26101489	-	-	-	0.62	-	-	0.32	(0.16)	0.25	0.13
PDCD11 chr10:105156316	0.36	-	-	-	-	-	-	0.46	0.25	0.13
FTH1 chr11:61735192	-	-	1	-	0.43	-	-	0.41	0.38	0.13
FTH1 chr11:61735191	(0.059)	0.75	0.67	-	-	0.7	(0.12)	0.33	0.5	0.13
FUBP3 chr9:133454938	-	-	-	-	-	-	-	-	0	0.13
ALYREF chr17:79849513	-	0.21	-	0.41	-	-	-	0.28	0.38	0.13
RNF185 chr22:31556121	-	-	-	-	-	-	0.39	-	0.12	0.13
MRPS31 chr13:41345346	-	-	-	-	-	-	-	-	0	0.13
DPH3 chr3:16306505	-	-	-	0.25	-	(0.16)	0.57	-	0.25	0.13
RPL18A chr19:17970682	-	(0.14)	-	-	-	-	-	-	0	0.13
C16orf59 chr16:2510096	-	-	-	(0.025)	0.45	-	-	(0.054)	0.12	0.13
DERL1 chr8:124054557	-	-	-	0.23	-	-	-	-	0.12	0.13
MASTL chr10:27443328	-	-	-	-	-	-	-	-	0	0.11
DIXDC1 chr11:111797698	-	-	-	-	-	-	0.56	-	0.12	0.11
SMUG1 chr12:54582890	-	-	-	-	-	-	0.41	(0.16)	0.12	0.11
SMUG1 chr12:54582889	-	(0.17)	-	-	-	-	0.42	-	0.12	0.11
CDC20 chr1:43824529	-	-	-	0.24	-	(0.026)	0.78	0.2	0.25	0.11
SECISBP2 chr9:91933357	-	-	-	-	0.8	-	-	-	0.12	0.11
ARHGEF18 chr19:7459940	0.21	-	0.88	-	0.21	0.23	0.47	0.63	0.75	0.11
ARHGEF18 chr19:7459941	-	-	0.83	-	-	-	-	0.3	0.25	0.11
WDR82 chr3:52322052	-	-	-	-	-	-	-	-	0	0.11
SYNJ1 chr21:34100374	-	-	-	-	-	-	-	0.52	0.12	0.11
POLR2D chr2:128615744	(0.033)	-	-	0.55	-	-	-	-	0.12	0.11
DHX16 chr6:30640796	-	-	-	-	-	-	-	-	0	0.11
MAP1S chr19:17830242	-	-	0.67	(0.029)	-	-	-	(0.023)	0.12	0.11
PRKAG1 chr12:49412648	-	-	-	-	-	(0.069)	-	(0.04)	0	0.11
Total no. of mutations [c]	24961	64326	85537	88427	116673	119549	224931	267306		
Total no. of promoter hotspot mutations [d]	2	2	5	6	5	3	9	7		
NOTCH1 [e]	1	1	3	1	0	1	3	1		
NOTCH2	0	2	2	1	2	1	4	2		
CDKN2A	0	1	0	0	1	0	1	1		
TP53	0	0	1	1	1	0	1	2		
Total no. of driver mutations	1	4	6	3	4	2	9	6		

Supplementary Table 3 | Mutations in promoter hotspots in cSCC tumors. Melanoma hotspot positions were investigated in 8 cSCC tumors[1]. 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[1], 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.* [3]. [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[6] or damaging by SIFT[7] were counted as driver mutations.

Sample	WT9	WT11	WT12	WT10	WT13	WT8	WT6	WT7	Total mut. freq. [a]	TCGA SKCM mut. freq. [b]
RPL13A chr19:49990694	-	-	-	(0.1)	-	-	-	-	0	0.29
C16orf59 chr16:2510095	-	-	-	-	-	-	-	-	0	0.18
ASXL2 chr2:26101489	-	-	-	-	-	(0.05)	-	(0.038)	0	0.13
PDCD11 chr10:105156316	-	-	-	-	-	-	-	-	0	0.13
FTH1 chr11:61735192	-	-	-	-	-	-	-	-	0	0.13
FTH1 chr11:61735191	-	-	-	-	-	-	-	-	0	0.13
FUBP3 chr9:133454938	-	-	-	-	-	-	-	-	0	0.13
ALYREF chr17:79849513	-	-	-	-	-	-	-	-	0	0.13
RNF185 chr22:31556121	-	-	-	-	-	-	(0.053)	-	0	0.13
MRPS31 chr13:41345346	-	-	-	-	-	(0.033)	-	(0.028)	0	0.13
DPH3 chr3:16306505	-	-	-	-	-	-	-	(0.03)	0	0.13
RPL18A chr19:17970682	-	-	-	-	(0.12)	-	(0.12)	-	0	0.13
C16orf59 chr16:2510096	-	-	-	-	-	-	(0.071)	-	0	0.13
DERL1 chr8:124054557	-	-	-	-	-	-	-	-	0	0.13
MASTL chr10:27443328	-	-	-	-	-	-	-	-	0	0.11
DIXDC1 chr11:111797698	-	-	-	-	-	-	-	-	0	0.11
SMUG1 chr12:54582890	-	-	-	-	-	-	-	0.23	0.12	0.11
SMUG1 chr12:54582889	-	-	-	-	-	-	-	-	0	0.11
CDC20 chr1:43824529	-	-	-	-	-	-	-	(0.034)	0	0.11
SECISBP2 chr9:91933357	-	-	-	-	(0.18)	(0.034)	-	-	0	0.11
ARHGEF18 chr19:7459940	-	-	-	-	-	-	-	(0.036)	0	0.11
ARHGEF18 chr19:7459941	-	-	-	-	-	-	-	(0.036)	0	0.11
WDR82 chr3:52322052	-	-	-	-	-	-	-	-	0	0.11
SYNJ1 chr21:34100374	-	-	-	-	-	-	-	-	0	0.11
POLR2D chr2:128615744	-	-	-	-	-	-	-	-	0	0.11
DHX16 chr6:30640796	-	-	-	-	-	-	-	-	0	0.11
MAP1S chr19:17830242	-	-	-	-	-	-	-	-	0	0.11
PRKAG1 chr12:49412648	-	-	-	-	-	-	-	-	0	0.11
Total no. of mutations [c]	24961	64326	85537	88427	116673	119549	224931	267306		
Total no. of promoter hotspot mutations [d]	0	0	0	0	0	0	0	1		

Supplementary Table 4 | Mutations in promoter hotspots in skin samples. Mutations in promoter hotspots were found at low variant frequencies in 8 peritumoral skin samples[3] that were available as matching normals for the cSCC tumors analyzed in **Supplementary Table 3**. 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.* [3]. [d]Number of promoter hotspot mutations with variant frequency of at least 0.2.

Cancer	Mutation load[a]	UV radiation[b]	Mutational signatures[c]	TERT promoter mutations[d]	Melanoma promoter hotspots[e]
Prostate, PRAD	1361				
Thyroid, THCA	2055		2	+	
Low-grade glioma, LGG	2873			+	
Kidney (chrom.), KICH	5147				
Breast, BRCA	6194		2, 13		
Kidney (clear), KIRC	7234				
Head & neck, HNSC	7324		2, 7		
Uterus, UCEC	8352		2		
Glioblastoma, GBM	9240		11	+	
Bladder, BLCA	16011		2, 13	+	
Lung (adeno), LUAD	18942		2	+	
Colorectal, CRC	21994				
Lung (squamous), LUSC	37741		2		
Melanoma, SKCM	52663	+	7, 11	+	+
Skin, cSCC	102550	+	-[f]	-[f]	+

Supplementary Table 5 | 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.* [3]. All other counts from Fredriksson *et al.* [8]. [b]UV-radiation as the mutational process driving tumor development. [c]Presence of mutational signatures 2, 7, 11 or 13 [9], 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[8]. [e]Presence of melanoma promoter hotspot mutations. [f]Data not available.

Rank	Name	p-value	E-value	q-value	Overlap	Offset	Orientation
1	ETV6	5.06e-05	3.25e-02	4.07e-02	6	2	Reverse Complement
2	GABPA	6.75e-05	4.33e-02	4.07e-02	6	3	
3	ELK1	1.14e-04	7.33e-02	4.07e-02	6	1	Reverse Complement
4	ELK4	1.28e-04	8.22e-02	4.07e-02	6	3	
5	GABP1	1.80e-04	1.16e-01	4.58e-02	6	3	Reverse Complement
6	ELF2	2.56e-04	1.64e-01	5.43e-02	6	6	Reverse Complement
7	ELF1	3.96e-04	2.54e-01	7.18e-02	6	3	Reverse Complement
8	ERG	5.63e-04	3.61e-01	8.93e-02	6	3	Reverse Complement
9	EHF	1.15e-03	7.35e-01	1.62e-01	6	2	Reverse Complement
10	ETV1	1.52e-03	9.75e-01	1.81e-01	6	10	Reverse Complement
11	ETS1	1.57e-03	1.01e+00	1.81e-01	6	1	Reverse Complement
12	FLI1	1.88e-03	1.21e+00	1.99e-01	6	5	Reverse Complement
13	ETS2	2.23e-03	1.43e+00	2.03e-01	6	3	Reverse Complement
14	STAT3	2.23e-03	1.43e+00	2.03e-01	6	0	Reverse Complement
15	ETV4	2.42e-03	1.55e+00	2.05e-01	6	1	Reverse Complement
16	ELK3	3.70e-03	2.37e+00	2.94e-01	6	3	Reverse Complement
17	SPIB	4.26e-03	2.73e+00	3.04e-01	6	0	Reverse Complement
18	SPDEF	4.32e-03	2.77e+00	3.04e-01	6	4	Reverse Complement
19	ETV5	4.87e-03	3.12e+00	3.22e-01	6	4	Reverse Complement
20	STAT4	5.08e-03	3.26e+00	3.22e-01	6	0	Reverse Complement
21	ELF5	9.79e-03	6.28e+00	5.92e-01	6	2	Reverse Complement
22	ETV7	1.17e-02	7.52e+00	6.77e-01	6	6	Reverse Complement

Supplementary Table 6 | Transcription factor motifs matching CTTCCG. Motif search in the JASPAR database using the tool TOMTOM[11]. The motif CTTCCG was compared with motifs in the databases for human transcription factors (HOCOMOCOv10).

Sample	XPC1	XPC2	XPC3	XPC4	XPC5	Total mut. freq. [a]	TCGA SKCM mut. freq. [b]
RPL13A chr19:49990694 [c]	-	-	-	-	-	0	0.29
C16orf59 chr16:2510095	0.57	-	0.62	-	-	0.4	0.18
ASXL2 chr2:26101489	-	-	-	-	0.6	0.2	0.13
PDCD11 chr10:105156316	-	(0.14)	-	-	-	0	0.13
FTH1 chr11:61735192	-	-	-	-	0.75	0.2	0.13
FTH1 chr11:61735191	-	-	-	-	-	0	0.13
FUBP3 chr9:133454938	-	-	-	-	-	0	0.13
ALYREF chr17:79849513	-	-	-	-	-	0	0.13
RNF185 chr22:31556121	-	-	-	-	-	0	0.13
MRPS31 chr13:41345346	-	-	(0.19)	-	-	0	0.13
DPH3 chr3:16306505	-	0.64	-	-	-	0.2	0.13
RPL18A chr19:17970682	-	-	-	-	-	0	0.13
C16orf59 chr16:2510096	0.69	-	0.57	-	-	0.4	0.13
DERL1 chr8:124054557	-	-	-	-	-	0	0.13
MASTL chr10:27443328	-	0.45	-	-	-	0.2	0.11
DIXDC1 chr11:111797698	-	-	-	-	-	0	0.11
SMUG1 chr12:54582890	-	-	-	-	-	0	0.11
SMUG1 chr12:54582889	-	-	-	-	-	0	0.11
CDC20 chr1:43824529	-	-	-	-	-	0	0.11
SECISBP2 chr9:91933357	-	-	-	-	-	0	0.11
ARHGEF18 chr19:7459940	-	-	-	0.8	-	0.2	0.11
ARHGEF18 chr19:7459941	-	-	-	-	-	0	0.11
WDR82 chr3:52322052	(0.024)	-	-	-	-	0	0.11
SYNJ1 chr21:34100374	-	-	-	-	-	0	0.11
POLR2D chr2:128615744	-	-	-	-	-	0	0.11
DHX16 chr6:30640796	-	-	-	-	-	0	0.11
MAP1S chr19:17830242	-	-	-	-	-	0	0.11
PRKAG1 chr12:49412648	0.6	-	-	-	-	0.2	0.11
Total no. of mutations [c]	260487	300932	407399	708800	757189		
Total no. of promoter hotspot mutations [d]	3	2	2	1	2		
NOTCH1 [e]	3	6	1	1	1		
NOTCH2	2	5	1	2	3		
CDKN2A	3	1	0	0	2		
TP53	6	6	3	2	0		
Total no. of driver mutations	14	18	5	5	6		

Supplementary Table 7 | 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[3]. 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.* [3]. [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[6] or damaging by SIFT[7] were counted as driver mutations.

Gene	Position	Forward primer 5' - 3'	Reverse primer 5' - 3'	Amplicon length
RPL13A	chr19:49990642-49990705	CCCACGTGATCTCTCGCC	GCCCTCTTCCGAATGTCCAA	83
DPH3	chr3:16306469-16306523	TCGGCATCATCAGGTCGC	GGGCGGAGTTGCGTCA	70

Universal fwd primer	GGACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNATGGGAAAGAGTGTCC-fwd target primer
Universal Rev primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-rev target primer
Illumina fwd primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Illumina rev primer with index	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Index primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Sequencing primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Supplementary Table 8 | PCR and adapter primer sequences used for SiMSen ultrasensitive amplicon sequencing.

Supplementary references

1. Durinck, S. *et al.* Temporal Dissection of Tumorigenesis in Primary Cancers. *Cancer Discovery* **1**, 137-143 (2011).
2. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
3. Zheng, Christina L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9**, 1228-1234 (2014).
4. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506 (2012).
5. 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**, 1258-63 (2014).
6. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7**, e46688 (2012).
7. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**, 1073-1081 (2009).
8. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics* **46**, 1258-63 (2014).
9. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 - 421 (2013).
10. Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget* (2015).
11. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-339 (2013).