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

Johan Fredriksson[1] 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.

*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[1,2]. Recurrent promoter mutations have been identified in many additional genes[3,4] and appear to be particularly common in melanoma[5,6], but convincing functional data such influence on gene expression has been elusive. Here, we show that frequently recurring promoter mutations in melanoma occur almost exclusively at cytosines flanked by an extended non-degenerate 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 mutation signature. Additional analyses solidify this pattern as an extended context-specific mutational signature that mediates an exceptional position-specific elevation in local mutation rate, arguing against positive selection. This finding has implications for the interpretation of somatic mutations in regulatory regions, and underscores 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[7,8]. One important indicator of driver function is recurrence across independent tumors, which can be suggestive of positive selection. However, proper interpretation of recurrent mutations also requires an 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[9-11]. Analyses of mutational signatures have shown the importance of the immediate sequence context for local mutation rates[12]. Additionally, impaired nucleotide excision repair (NER) have been shown to contribute to increased local mutation density in promoter regions and protein binding sites[13,14]. Still, it is not clear to what extent these effects can explain recurrent somatic mutations in promoter regions, which are suggested by previous studies to be particularly frequent in melanoma despite several other cancer types approaching melanoma in terms of total mutation load[5,6].

To characterize somatic promoter mutations in melanoma, we analyzed the sequence context of recurrently mutated individual genomic positions occurring within +/- 500 bp of annotated TSSs, based on 38 melanomas subjected to whole genome sequencing by the Cancer Genome Atlas[6,15]. 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[1,2] (**Table 1**, **Fig. 1b**). The recurrent mutations occurred at cytosines positioned at the 5' end of the motif CTTCCG (**Fig. 1c**) and were normally C>T transitions (**Table 1**). Similar to most mutations in melanoma they were thus C>T substitutions in a dipyrimidine context, compatible with UV-induced damage through cyclobutane pyrimidine dimer (CPD) formation[12,16]. 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 (**Table 1**). Many sites recurrent in 3/38 tumors (8%) also showed the same pattern (**Supplementary Table 1**). We thus find that recurrent promoter mutations are common in melanoma, but also that they consistently adhere

56 to a distinct extended sequence signature, arguing against positive selection as a major
57 causative factor.

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

70 Next, we considered additional properties that could support or argue against a
71 functional role for the recurrent mutations. We first noted a general lack of known cancer-
72 related genes among the affected promoters, with *TERT* as one of few exceptions (**Table 1**
73 and **Supplementary Table 1**, indicated in blue). Secondly, the recurrent promoter mutations
74 were not associated with differential expression of the nearby genes (**Table 1** and
75 **Supplementary Table 1**). This is in agreement with earlier investigations of a few of these
76 mutations, which gave no conclusive evidence regarding influence on gene expression[5,20].
77 Lastly, we found that when comparing different tumors there was a strong positive correlation
78 between the total number of the established hotspot positions that were mutated and the
79 genome-wide mutation load, both in melanoma (**Fig. 2a**) and in cSCC (**Supplementary**
80 **Table 3**). This is again compatible with a passive model involving elevated mutation
81 probability in the affected positions, and contrasted sharply with most of the major driver
82 mutations in melanoma, which were detected also in tumors with lower mutation load (**Fig.**
83 **2b**, **Supplementary Table 3**). These different findings further reinforce the CTTCCG motif
84 as a strong mutational signature in melanoma.

85 We next investigated whether the observed signature would be relevant also outside of
86 promoter regions. As expected, numerous mutations occurred in CTTCCG sequences across
87 the genome, but notably we found that recurrent mutations involving this motif were always
88 located close to actively transcribed TSSs (**Fig. 3abc**). This shows that the signature is

89 relevant only in the context of active promoters, suggesting that a binding partner is required.
90 We further compared the frequencies of mutations occurring at cytosines in the context of the
91 motif to all possible trinucleotide contexts, an established way of describing mutational
92 signatures in cancer[12]. As expected, on a genome-wide scale, the mutation probability for
93 cytosines in CTTCCG-related contexts was only marginally higher compared to
94 corresponding trinucleotide contexts (**Fig. 4a**). However, close to highly transcribed TSSs, the
95 signature conferred a striking elevation in mutation probability compared to related
96 trinucleotides, in particular for cytosines at the 5' end of the motif (**Fig. 4b-d**). Recurrent
97 promoter mutations in melanoma thus conform to a distinct sequence signature manifested
98 only in the context of active promoters, suggesting that a specific binding partner is required
99 for the element to confer elevated mutation probability.

100 The sequence CTTCCG matches the consensus binding motif of transcription factors
101 (TFs) of the ETS family, and similar elements in various individual promoters have
102 previously been shown to be bound by ETS factors including ETS1, GABPA and ELF1[21],
103 ELK4[22], and E4TF1[23]. This suggests that the recurrently mutated CTTCCG elements could be
104 substrates for ETS TFs. As expected, matches to CTTCCG in the JASPAR database of TF
105 binding motifs were mainly ETS-related (**Supplementary Table 6**). Notably, recurrently
106 mutated CTTCCG sites were evolutionarily conserved to a larger degree than non-recurrently
107 mutated but otherwise similar control sites, further supporting that they constitute functional
108 ETS binding sites (**Supplementary Fig. 2**). This was corroborated by analysis of top
109 recurrent CTTCCG sites in relation to ENCODE ChIP-seq data for 161 TFs, which showed
110 that the strongest and most consistent signals were for ETS factors (GABPA and ELF1)
111 (**Supplementary Fig. 3**).

112 The distribution of mutations across tumor genomes is shaped both by mutagenic and
113 DNA repair processes. Binding of TFs to DNA can increase local mutation rates by impairing
114 NER, and strong increases have been observed in predicted sites for several ETS factors[13,14].
115 It is also established that contacts between DNA and proteins modulate DNA damage patterns
116 by altering conditions for UV photoproduct formation[24-27]. In upstream regions of XPC -/-
117 cSCC tumors lacking global NER, we found that the CTTCCG signature still conferred
118 strongly elevated mutation probabilities compared to relevant trinucleotide contexts
119 (**Supplementary Fig. 4**), although to a lesser extent than in melanomas with functional NER
120 (**Fig. 4**). The signal was independent of strand orientation relative to the downstream gene,
121 and is thus unlikely due to transcription coupled NER which is a strand-specific process[16]

122 (**Supplementary Fig. 4**). The signature described here may thus be due to a combination of
123 impaired DNA repair and elevated sensitivity to UV-induced damage at cytosines at the 5'
124 end of ETS-bound CTTCCG elements.

125 In summary, we demonstrate that recurrent promoter mutations are common in
126 melanoma, but also that they adhere to a distinct sequence signature in a strikingly consistent
127 manner, arguing against positive selection as a major driving force. This model is supported
128 by several additional observations, including lack of cancer-relevant genes, lack of obvious
129 effects on gene expression, presence of the signature exclusively in UV-exposed samples of
130 diverse cellular origins, and strong positive correlation between genome-wide mutation load
131 and mutations in the affected positions. Our results will allow better interpretation of somatic
132 mutations in regulatory DNA and point to limitations in conventional genome-wide derived
133 trinucleotide models of mutational signatures.

134 Methods

135 Mapping of somatic mutations

136 Whole-genome sequencing data for 38 metastatic skin cutaneous melanoma tumors (SKCM)
137 was obtained from the Cancer Genome Atlas (TCGA) together with matching RNA-seq and
138 copy-number data. Mutations were called using samtools[28] (command *mpileup* with default
139 settings and additional options *-q1* and *–B*) and VarScan[29] (command *somatic* using the
140 default minimum variant frequency of 0.20, minimum normal coverage of 8 reads, minimum
141 tumor coverage of 6 reads and the additional option *–strand-filter 1*). Mutations where the
142 variant base was detected in the matching normal were not considered for analysis. The
143 resulting set of mutations was further processed by removing mutations overlapping germline
144 variants included in the NCBI dbSNP database, Build 146. The genomic annotation used was
145 GENCODE[30] release 17, mapped to GRCh37. The transcription start site of a gene was
146 defined as the 5'most annotated transcription start. Somatic mutation status for known driver
147 genes was obtained from the cBioPortal[31,32].

148 RNA-seq data processing

149 RNA-seq data was analyzed with resepect to the GENCODE[30] (v17) annotation using HTSeq-
150 count (http://www-huber.embl.de/users/anders/HTSeq) as previously described [33].
151 Differential gene expression between tumors with and without mutations in promoter regions
152 was evaluated using two-sided Wilcoxon rank sum tests.

153 **Analyzed genomic regions**

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

158 **Mutation probability calculation**

159 The February 2009 assembly of the human genome (hg19/GRCh37) was downloaded from
160 the UCSC Genome Bioinformatics site. Sequence motif and trinucleotide frequencies were
161 obtained using the tool *fuzznuc* included in the software suite EMBOSS[34]. The mutation
162 probability was calculated as the total number of observed mutations in a given sequence
163 context across all tumors divided by the number of instances of this sequence multiplied by
164 the number of tumors.

165 **Evolutionary conservation data**

166 The evolutionary conservation of genome regions was evaluated using phastCons scores[35]
167 from multiple alignments of 100 vertebrate species retrieved from the UCSC genome
168 browser. The analyzed regions were 30 bases upstream and downstream of the motif
169 CTTCCG located less than 500 bp from TSS.

170 **ChIP-seq data**

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

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

175 Whole genome sequencing data from sun-exposed skin, eye-lid epidermis, was obtained from
176 Martincorena *et al.*, 2015[19]. Samtools[28] (command *mpileup* with a minimum mapping quality
177 of 60, a minimum base quality of 30 and additional option *–B*) was used to process the data
178 and VarScan[29] (command *mpileup2snp* counting all variants present in at least one read, with
179 minimum coverage of one read and the additional strand filter option disabled) was used for
180 mutation calling.

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

182 Whole genome sequencing data from 8 cSCC tumors and matching peritumoral skin samples
183 was obtained from Durinck *et al.*, 2011[36]. Whole genome sequencing data from cSCC tumors
184 and matching peritumoral skin from 5 patients with germline DNA repair deficiency due to
185 homozygous frameshift mutations (C_{940}del-1) in the *XPC* gene was obtained from Zheng *et*
186 *al.*, 2014[18]. Samtools[28] (command *mpileup* with a minimum mapping quality of 30, a
187 minimum base quality of 30 and additional option *–B*) was used to process the data and
188 VarScan[29] (command *mpileup2snp* counting all variants present in at least one read, with
189 minimum coverage of two reads and the additional strand filter option disabled) was used for
190 mutation calling. For the mutation probability analysis of cSCC tumors with NER deficiency
191 an additional filter was applied to only consider mutations with a total coverage of at least 10
192 reads and a variant frequency of at least 0.2. The functional impact of mutations in driver
193 genes was evaluated using PROVEAN[37] and SIFT[38]. Non-synonymous mutations that were
194 considered deleterious by PROVEAN or damaging by SIFT were counted as driver
195 mutations.

196 ## Acknowledgements

197 The results published here are in whole or part based upon data generated by The Cancer
198 Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and
199 the investigators and institutions who constitute the TCGA research network can be found at
200 "http://cancergenome.nih.gov". We are most grateful to the patients, investigators, clinicians,
201 technical personnel, and funding bodies who contributed to TCGA, thereby making this study
202 possible. This work was supported by grants from the Knut and Alice Wallenberg
203 Foundation, the Swedish Foundation for Strategic Research, the Swedish Medical Research
204 Council, the Swedish Cancer Society, the Åke Wiberg foundation, the Lars Erik Lundberg
205 Foundation for Research and Education. Computations were in part performed on resources
206 provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational
207 Science (UPPMAX) under project b2012108.

208 ## Author contributions

209 J.F and E.L. conceived the study; J.F performed bioinformatics analyses; J.F and E.L. wrote
210 the paper.

211 ## Competing financial interests

212 The authors declare no competing financial interests or other conflict of interest.

213 ## References

214 1. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma.
215 *Science* **339**, 957-9 (2013).
216 2. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science*
217 **339**, 959-61 (2013).
218 3. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis
219 of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-5 (2014).
220 4. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in
221 regulatory regions of human cancer genomes. *Nat Genet* **47**, 710-6 (2015).
222 5. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types
223 highlights a rich landscape of functional molecular alterations. *Nat Genet* **48**, 117-25
224 (2016).
225 6. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of
226 noncoding somatic mutations and gene expression alterations across 14 tumor types.
227 *Nat Genet* **46**, 1258-63 (2014).
228 7. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**,
229 93-108 (2016).
230 8. Poulos, R.C., Sloane, M.A., Hesson, L.B. & Wong, J.W. The search for cis-regulatory
231 driver mutations in cancer genomes. *Oncotarget* **6**, 32509-25 (2015).
232 9. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape
233 of cancer. *Nature* **518**, 360-364 (2015).
234 10. Lawrence, M. *et al.* Mutational heterogeneity in cancer and the search for new cancer-
235 associated genes. *Nature* **499**, 214 - 218 (2013).
236 11. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human
237 cancer genome. *Nature* **463**, 191-196 (2010).
238 12. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
239 415 - 421 (2013).
240 13. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active
241 promoters in cancer genomes. *Nature* **532**, 259-263 (2016).

242 14. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N.
243 Nucleotide excision repair is impaired by binding of transcription factors to DNA.
244 *Nature* **532**, 264-267 (2016).
245 15. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer
246 analysis project. *Nat Genet* **45**, 1113-20 (2013).
247 16. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational
248 signatures in human cancers. *Nat Rev Genet* **15**, 585-98 (2014).
249 17. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations.
250 *Nature* **485**, 502-506 (2012).
251 18. Zheng, Christina L. *et al.* Transcription Restores DNA Repair to Heterochromatin,
252 Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9**, 1228-1234
253 (2014).
254 19. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic
255 mutations in normal human skin. *Science* **348**, 880-886 (2015).
256 20. Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget*
257 (2015).
258 21. Hollenhorst, P.C. *et al.* DNA Specificity Determinants Associate with Distinct
259 Transcription Factor Functions. *PLoS Genet* **5**, e1000778 (2009).
260 22. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions
261 bound by 119 human transcription factors. *Genome Research* **22**, 1798-1812 (2012).
262 23. Tanaka, M. *et al.* Cell-cycle-dependent regulation of human aurora A transcription is
263 mediated by periodic repression of E4TF1. *J Biol Chem* **277**, 10719-26 (2002).
264 24. Gale, J.M., Nissen, K.A. & Smerdon, M.J. UV-induced formation of pyrimidine
265 dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases.
266 *Proc Natl Acad Sci U S A* **84**, 6644-8 (1987).
267 25. Brown, D.W., Libertini, L.J., Suquet, C., Small, E.W. & Smerdon, M.J. Unfolding of
268 nucleosome cores dramatically changes the distribution of ultraviolet photoproducts in
269 DNA. *Biochemistry* **32**, 10527-10531 (1993).
270 26. Pfeifer, G.P., Drouin, R., Riggs, A.D. & Holmquist, G.P. Binding of transcription
271 factors creates hot spots for UV photoproducts in vivo. *Molecular and Cellular*
272 *Biology* **12**, 1798-1804 (1992).
273 27. Tornaletti, S. & Pfeifer, G.P. UV Light as a Footprinting Agent: Modulation of UV-
274 induced DNA Damage by Transcription Factors Bound at the Promoters of Three
275 Human Genes. *Journal of Molecular Biology* **249**, 714-728 (1995).

276 28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
277 2078-2079 (2009).
278 29. Koboldt, D.C. *et al.* VarScan 2: Somatic mutation and copy number alteration
279 discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576 (2012).
280 30. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The
281 ENCODE Project. *Genome Research* **22**, 1760-1774 (2012).
282 31. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles
283 Using the cBioPortal. *Sci. Signal.* **6**, pl1- (2013).
284 32. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring
285 Multidimensional Cancer Genomics Data. *Cancer Discovery* **2**, 401-404 (2012).
286 33. Akrami, R. *et al.* Comprehensive Analysis of Long Non-Coding RNAs in Ovarian
287 Cancer Reveals Global Patterns and Targeted DNA Amplification. *Plos One* **8**(2013).
288 34. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology
289 Open Software Suite. *Trends in Genetics* **16**, 276-277.
290 35. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and
291 yeast genomes. *Genome Research* **15**, 1034-1050 (2005).
292 36. Durinck, S. *et al.* Temporal Dissection of Tumorigenesis in Primary Cancers. *Cancer*
293 *Discovery* **1**, 137-143 (2011).
294 37. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the Functional
295 Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7**, e46688 (2012).
296 38. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous
297 variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**, 1073-1081
298 (2009).
299 39. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in
300 human cancer. *Nucleic Acids Research* **43**, D805-D811 (2015).

301

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 mutations within +/- 500 bp from transcription start sites 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. The mutated base is highlighted in gray. The motif CTTCCG is highlighted in yellow. [f]Distance from mutation to the 5' most transcription start site 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 [39]) are highlighted in blue. [h]Genes were sorted by increasing mean expression and assigned to expression tiers 1 to 3. [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 transcription start site in GENCODE 17. Negative values indicate upstream location of mutation. [k]Second closest gene. [l]Genes were sorted by increasing mean expression and assigned to expression tiers 1 to 3. [m]*P*-values from a two sided Wilcoxon rank sum test of differential expression of the gene between tumors with and without the mutation. [n]Significant differential expression could not be seen when the analysis was repeated in a larger dataset[6].

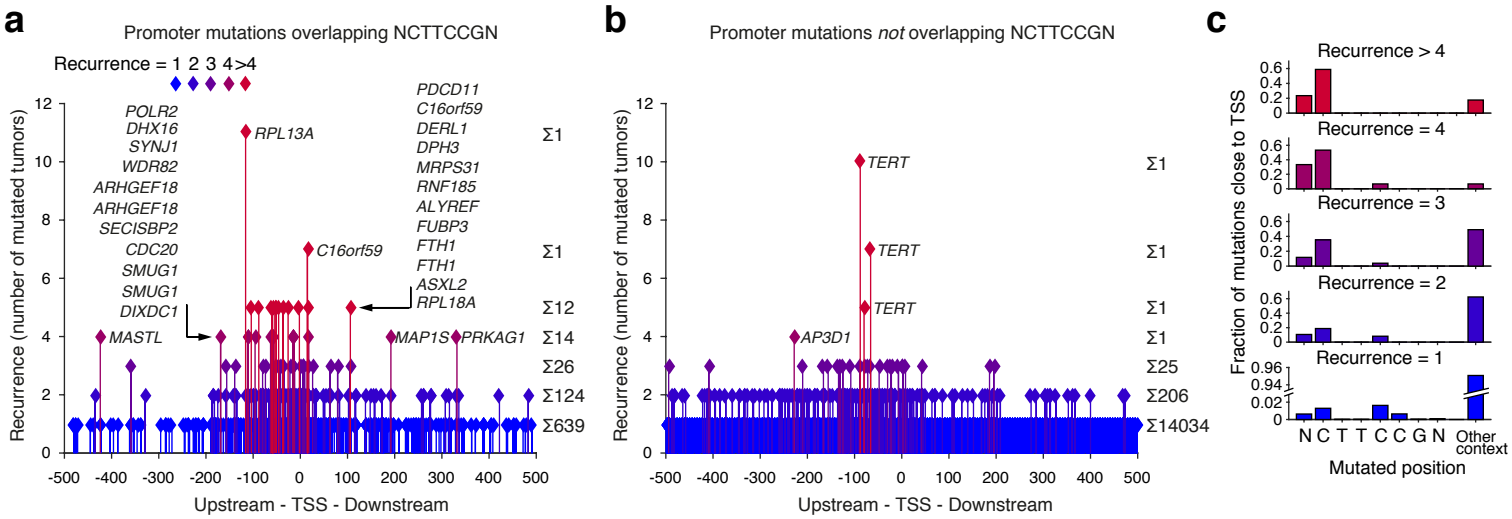


Figure 1 | Recurrent somatic mutations in promoter regions in melanoma are characterized by a distinct sequence signature. Whole genome sequencing data from 38 melanomas were analyzed for individual recurrently mutated bases in promoter regions, and most highly recurrent positions were found to share 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**.

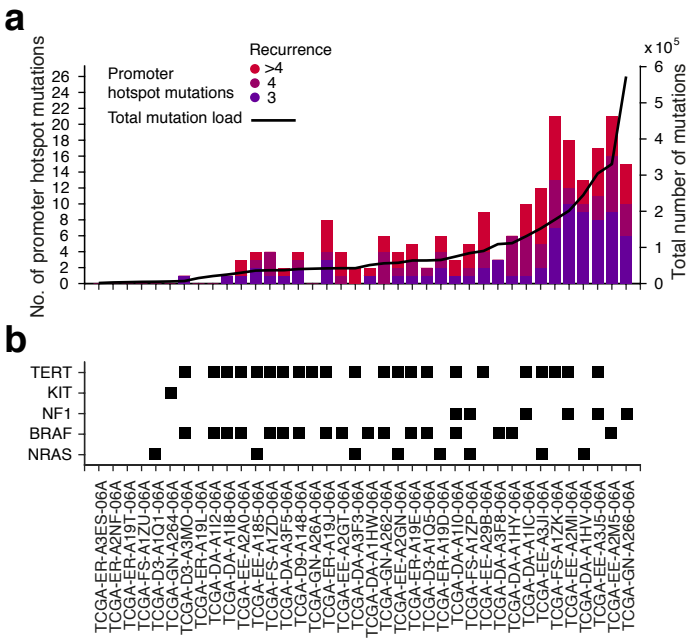


Figure 2 | Positive correlation between promoter hotspot mutations and total mutational load across melanomas. (**a**) Bars, left axis: Number of mutations occurring in the established recurrent CTTCCG-related promoter positions (>= 3 tumors) in each of the 38 samples. Line, right axis: Total mutational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

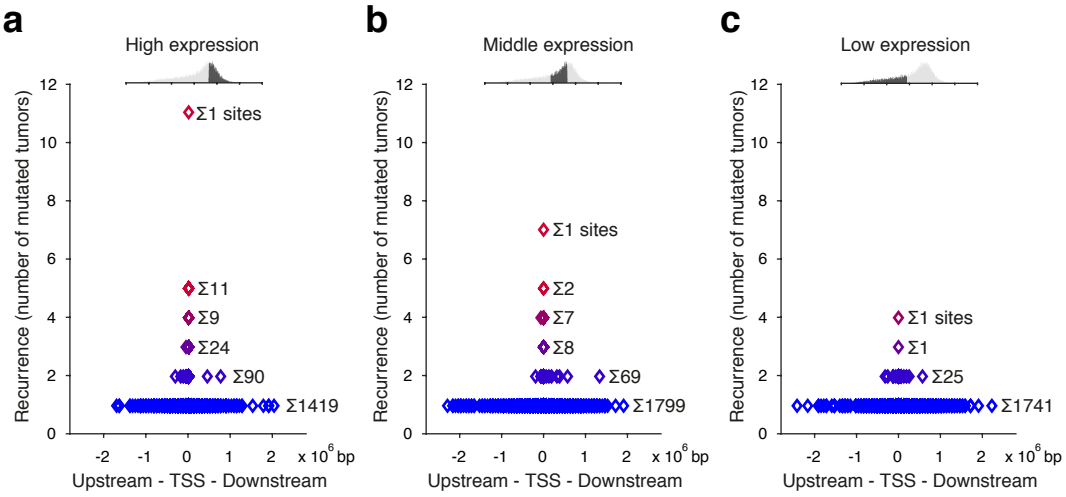


Figure 3 | Recurrent mutations at CTTCCG sites are observed only near active promoters. (**a-c**) Genes were assigned to three expression tiers by increasing mean expression across the 38 melanomas. The graphs show, on the x-axis, the distance to the nearest annotated TSS for all mutations overlapping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

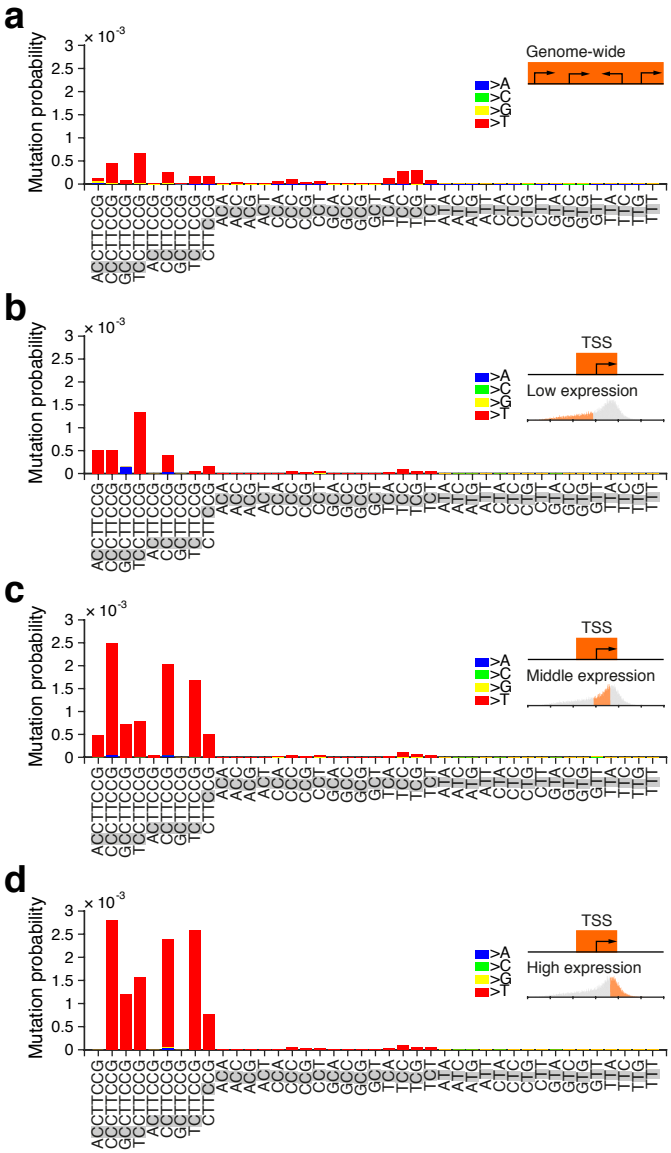
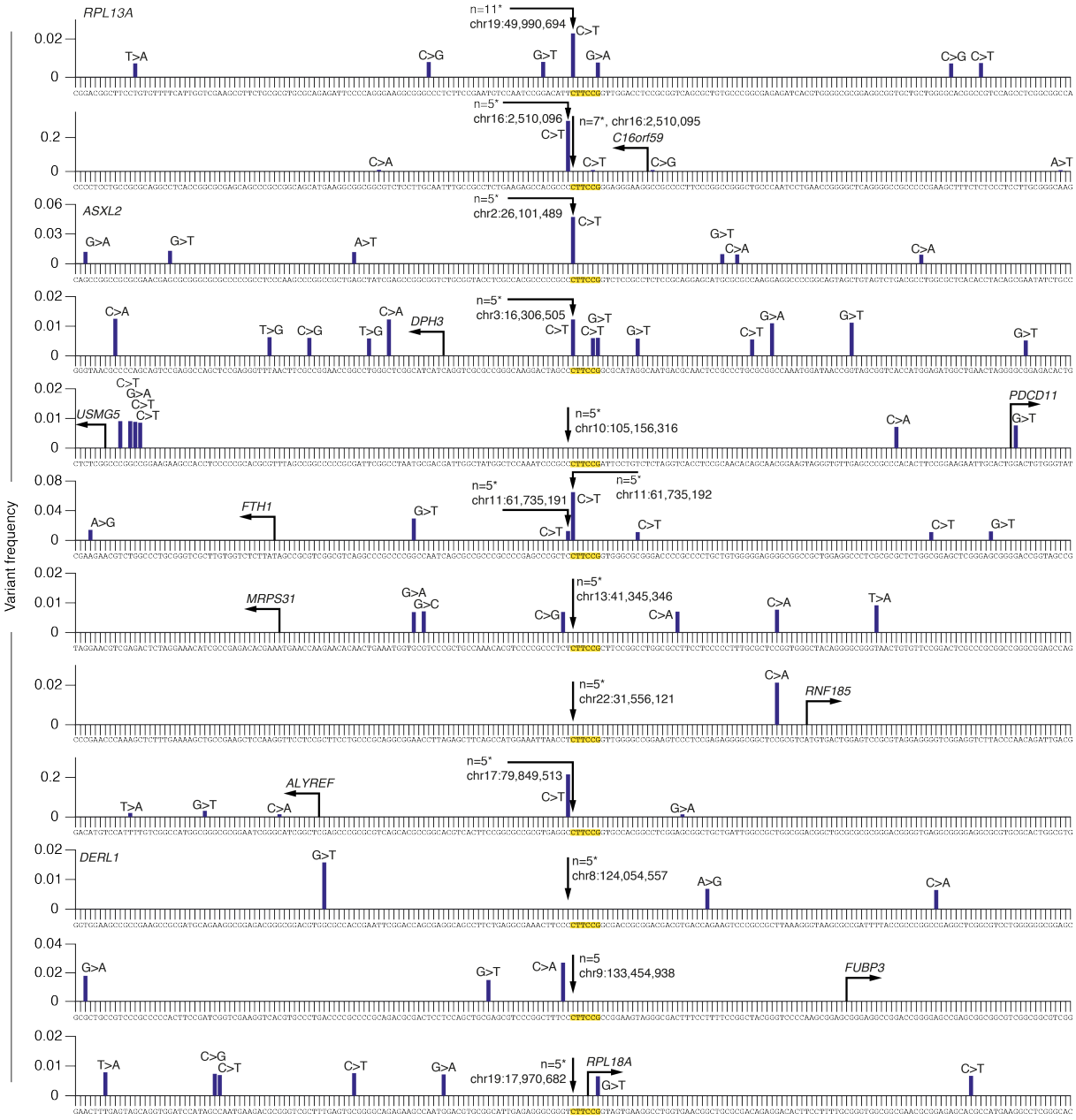
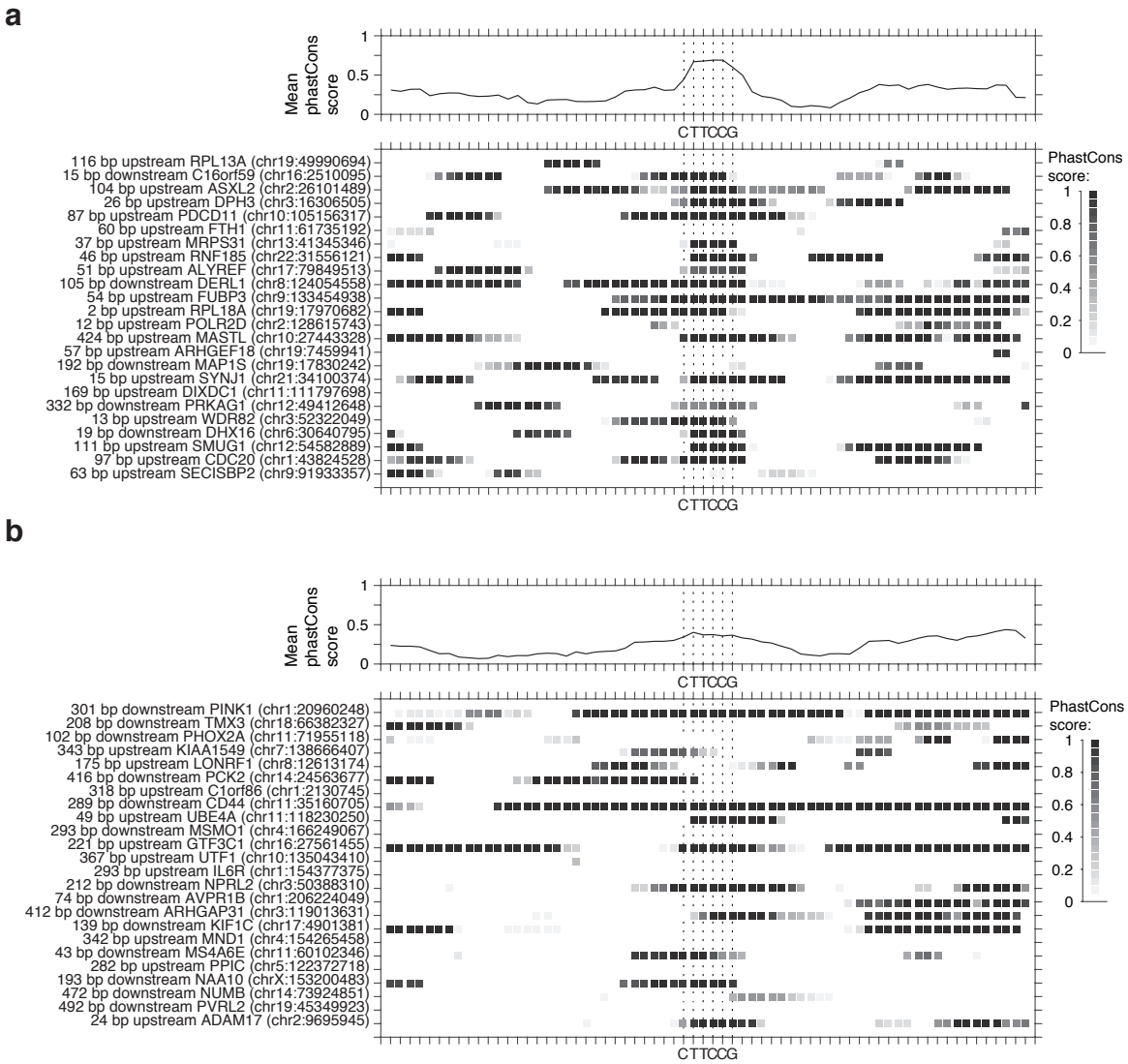


Figure 4 | Mutation probabilities for CTTCCG-related sequence contexts compared to trinucleotides. The mutated position in each sequence context is shaded in gray. Bar colors indicate the substituting bases (mainly C>T). Mutation probabilities were calculated genome-wide (**a**), or only considering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

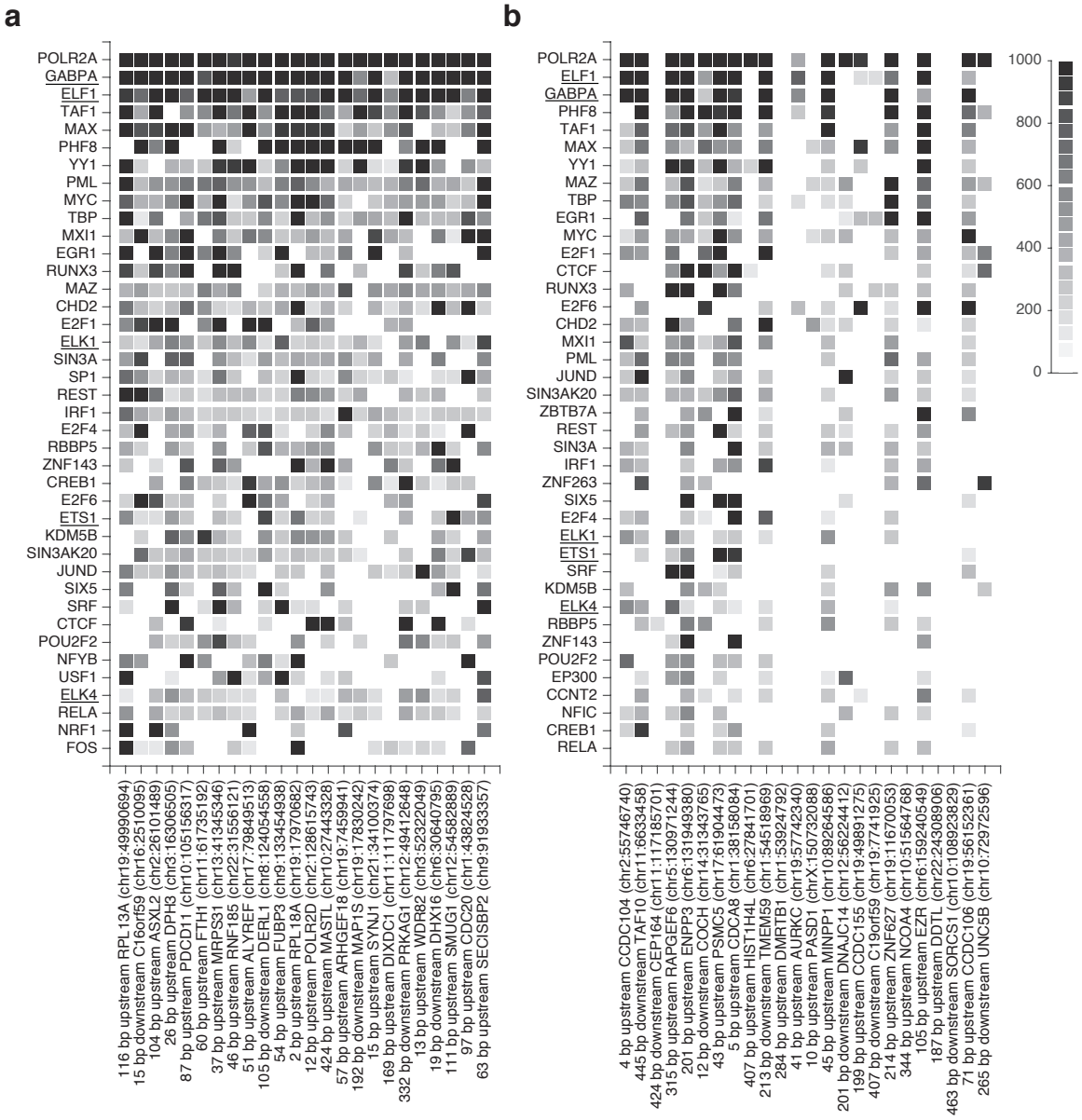
Supplementary figures



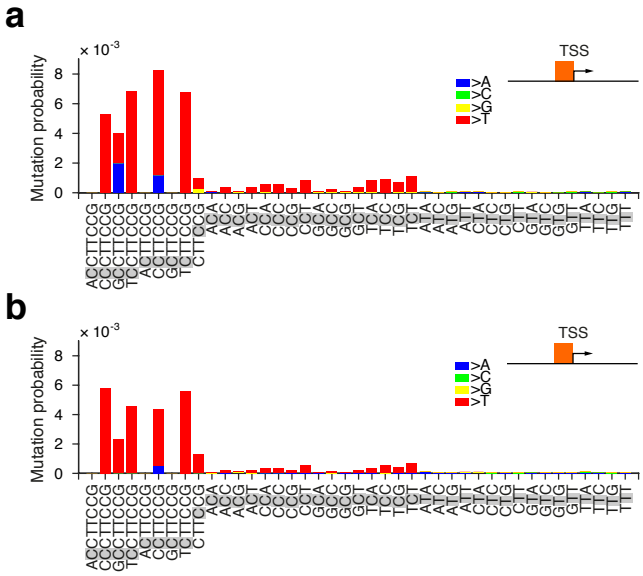
Supplementary Figure 1 | Melanoma promoter hotspot positions are often mutated in sun-exposed skin. Recurrent CTTCCG-related promoter hotspot sites identified in melanoma (mutated in >=5/38 TCGA tumors) were examined for mutations in a sample of sun-exposed normal skin. The graphs show variant allele 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.*[1]. 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[2] 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.* [3]. [c]0.08 was previously obtained using a different calling pipeline applied to the same data[4] 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[5]. 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[5], 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.* [2]. [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[2] 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.* [2]. [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.* [2]. 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[10]. 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[2]. 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.* [2]. [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.

Supplementary references

1. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
2. Zheng, Christina L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9**, 1228-1234 (2014).
3. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506 (2012).
4. 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).
5. Durinck, S. *et al.* Temporal Dissection of Tumorigenesis in Primary Cancers. *Cancer Discovery* **1**, 137-143 (2011).
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. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-339 (2013).