PNAS
www.pnas.org

Supplementary Information for

Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light

Markus Lindberg, Martin Boström, Kerryn Elliott, Erik Larsson

Erik Larsson
Email: erik.larsson@gu.se

This PDF file includes:

Figures S1 to S6
Tables S1 to S2
SI References

Other supplementary materials for this manuscript include the following:

None

www.pnas.org/cgi/doi/10.1073/pnas.1909021116

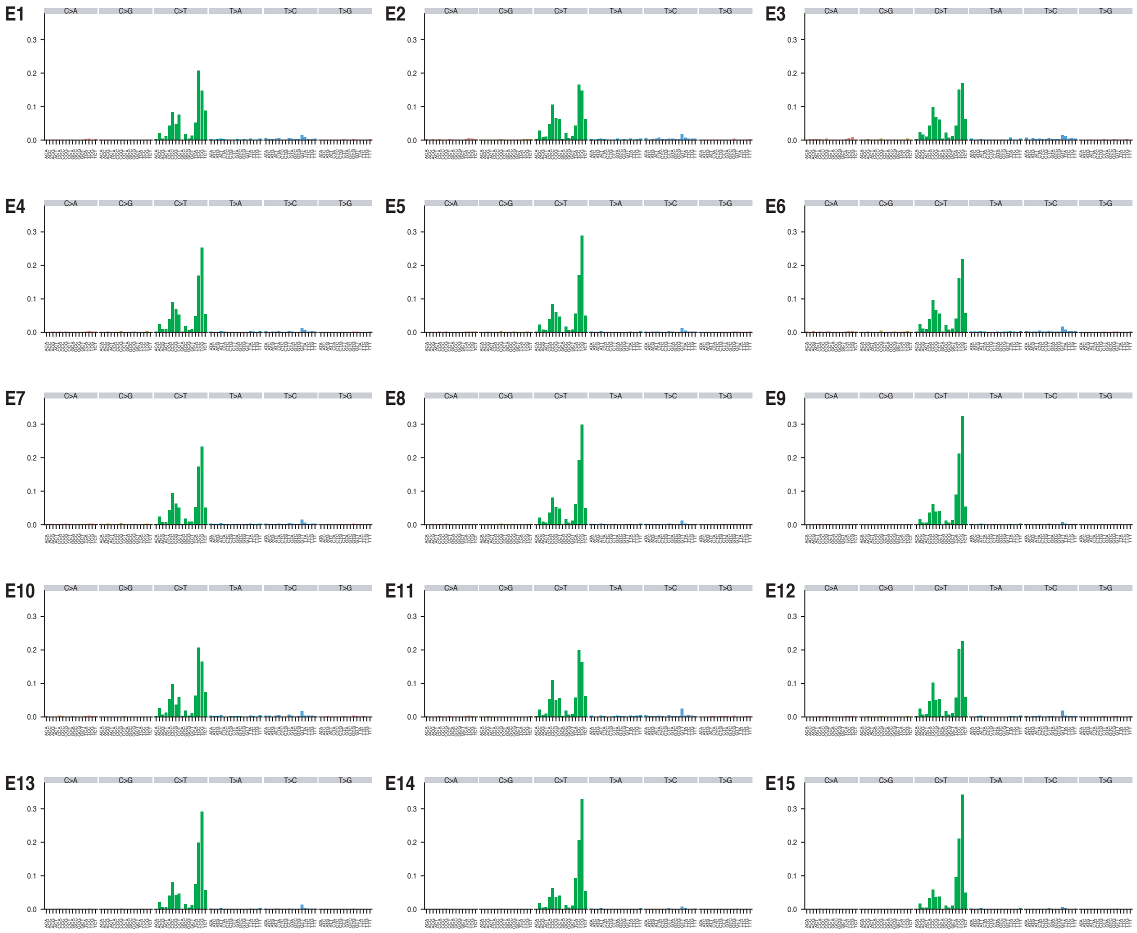

Fig. S1. UV trinucleotide signature in different ChromHMM chromatin states. The signatures were normalized for variable trinucleotide sequence content in the respective regions and further normalized to sum up to one. Each bar thus shows the probability of mutagenesis at a given trinucleotide in relative terms, compared to other trinucleotides.

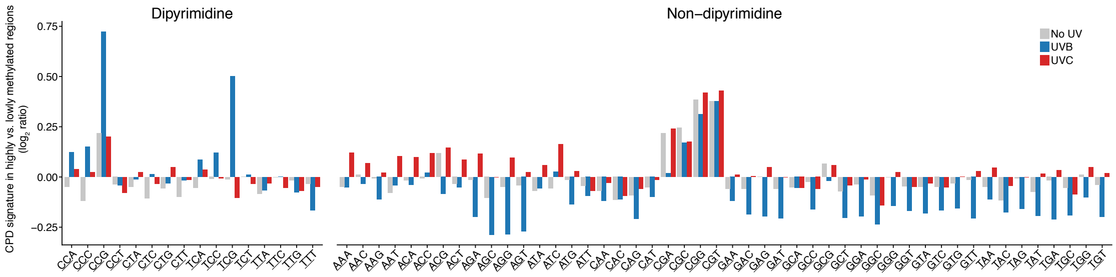

Fig. S2. CPD formation signature in highly vs. lowly methylated genomic regions. The CPD trinucleotide signature (relative formation frequency per genomic site) in highly (>80% CpG methylation) vs. lowly (>20% CpG methylation) methylated regions (1 kb genomic bins) were compared by means of a $\log_2$ ratio, confirming methylation-dependent elevation in CPD formation at CpG-flanking dipyrimidines specifically in response to UVB (blue). Examined patterns include all trinucleotides, where the first two bases (underscored) represent the position at which a CPD was detected. Notably, CPD detections at CG dinucleotides, which cannot form CPDs and thus represent false positives, were found to be methylation-dependent. Although frequencies for these events were low compared to actual CPD-forming dipyrimidines (**Fig. 4**), such detections were observed in all conditions including no-UV-controls, suggesting occasional cleavage by T4 endonuclease V at methylated CpGs independently of CPD formation. Signatures were normalized with respect to genomic sequence content in the respective regions and further normalized to sum to one. Trinucleotides are presented in alphabetical order, separated by CPD-forming and non-CPD-forming patterns. Results for UVB, UVC and no UV controls, all pooled, are shown as separate bars.

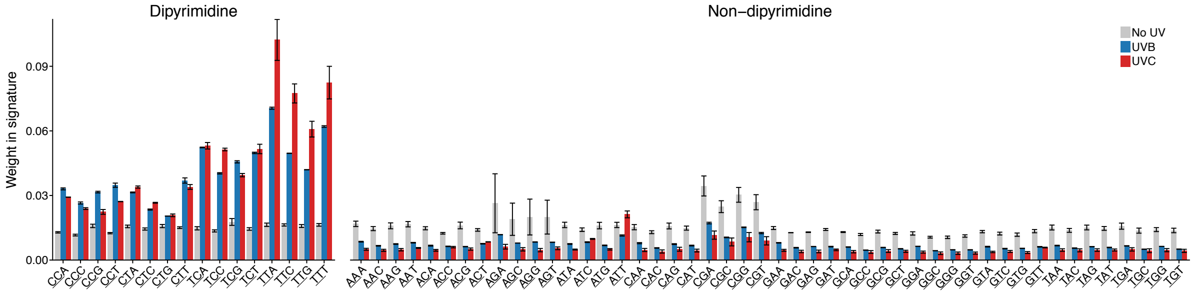

Fig. S3. Genome-wide CPD trinucleotide signature for UVB and UVC. Examined patterns include all trinucleotides, where the first two bases (underscored) represent the position at which a CPD was detected. Signatures were normalized with respect to genomic sequence content in the respective regions and further normalized to sum to one. Trinucleotides are presented in alphabetical order, separated by CPD-forming and non-CPD-forming patterns. Results for UVB, UVC and no UV controls, are shown, and error bars indicate SD (*n* = 2).

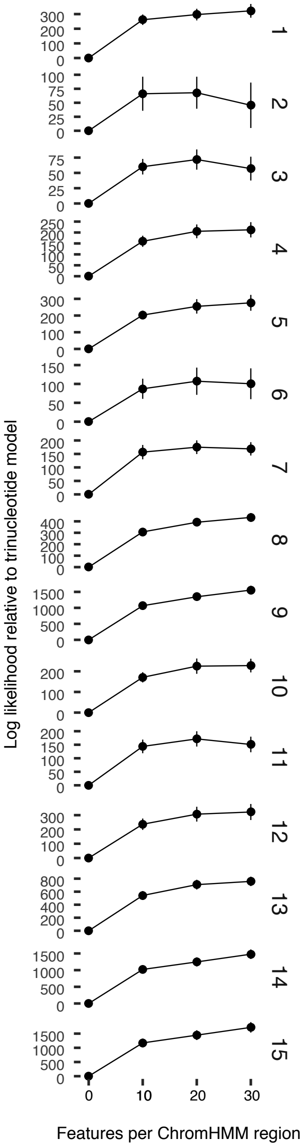

Fig. S4. Likelihood of observed mutation data for the extended signature model as a function of the number of long pentamer features. The x-axis indicates the number of pentamer features contributed from each ChromHMM region during feature selection, which was performed using Fisher's exact test on 500 kb random subsets of cytosine positions in each region. Selected features from each region (top ranking positive or negatively associated motifs) were pooled into one final set used for training models using logistic regression. This was repeated 10 times for each region on separate 500 kb random subsets. For each model, the likelihood was evaluated based on observed mutation data in a separate random 500 kb subset. Error bars indicate standard deviation.

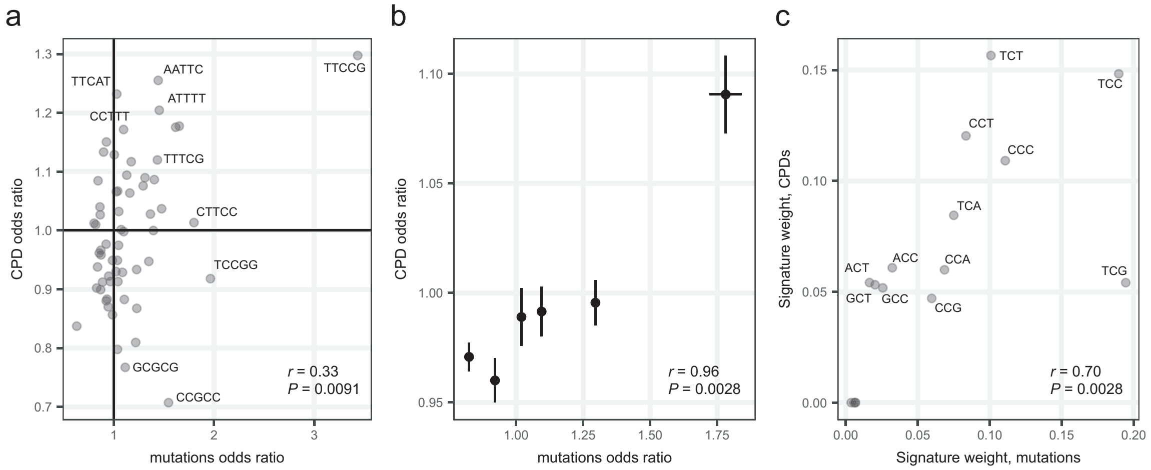

Fig. S5. Evaluation of extended signature patterns in relation to UVB CPD data. The same logistic regression model used for mutations was applied to UVB CPD data, using the same set of features including trinucleotides. This allows for evaluation of CPD formation frequencies at cytosines in relation to the uncovered pentamer signature patterns, while compensating for the effect of the central trinucleotide. For each position in the genome, we first determined whether it overlapped with a detected CPD or not, thus providing a binary variable for regression similar to the mutational analysis. Only cytosines were considered, as the mutational analysis was restricted to C>T transitions. (**a**) Scatter plot comparing regression weights (odds ratios) for mutations (x-axis) and CPDs (y-axis) for pentamer contextual patterns in the E1 region. (**b**) Results from panel a binned along the x-axis (10 patterns in bins 1-5 and 11 patterns in bin 6). (**c**) Same comparison for trinucleotides. The trinucleotide weights are shown as normalized frequencies summing to one for consistency with other figures. Note that, for trinucleotides consisting purely of pyrimidines, CPDs can form in two positions overlapping with the central position, thus elevating the CPD frequencies for such patterns. Pearson’s correlation coefficient is indicated in all panels.

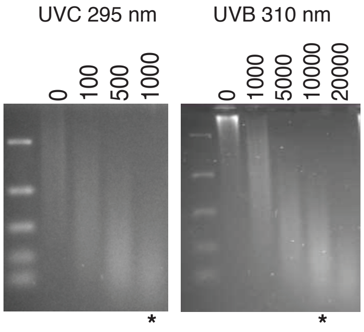

Fig. S6. T4 endonuclease digestion of DNA treated with different doses of UV light. To ensure that CPDs were generated at similar frequencies compared to previously generated UVC data (1), A375 cells were treated with a range of UVC (295nm) and UVB (310 nm) doses. DNA was extracted using the QIAgen Blood Mini kit, and 1 μg of DNA was digested with T4 endonuclease V (NEB). DNA was next purified by phenol/chloroform extraction and ethanol precipitation. DNA was resuspended in alkaline sample buffer and run overnight on a 1% alkaline gel. A UVB dose of 10,000 J/m^2 was judged to be appropriate, as indicated by the asterisks.

Table S1. List of included samples.

ID	SNV burden	C>T burden	DiPy C>T burden
SP124298	375872	328020	326512
SP124307	110715	97493	96979
SP124305	48660	41560	41115
SP124302	51460	44647	44420
SP124316	79306	68376	67951
SP124319	192274	173702	173261
SP124326	60559	52013	51430
SP124329	14895	12258	12049
SP124333	46788	38341	37735
SP124334	66789	56477	56006
SP124362	221175	200047	199236
SP124364	107422	90236	89494
SP124396	48407	40024	39322
SP124409	248586	222953	221924
SP124372	102774	87696	86749
SP124415	565302	500498	499653
SP124386	119838	106509	105606
SP124365	75874	62115	61388
SP124367	85946	74177	73867
SP124369	145545	128208	127448
SP124376	57466	50903	50377
SP124377	52806	43880	43549
SP124380	154507	133439	132809
SP124382	409281	345158	343586
SP124389	72622	65421	65077
SP124394	109490	98231	97791
SP124399	145632	124018	123274
SP124401	439770	378966	377151
SP124406	74238	61167	60644
SP124412	66924	54735	54181
SP124418	104129	86645	86155
SP124420	72313	63554	62991
SP124423	18566	16585	16474
SP124425	42697	35404	35043
SP124428	14706	13361	13308
SP124431	95459	78943	78439
SP124434	87746	76141	75448

SP124439	330466	289757	288958
SP124447	459049	388059	386208
SP124449	142725	125412	124643
SP124452	91282	79265	78647
SP124454	59772	52023	51351
SP124456	46832	39493	38852
SP124458	169152	145097	144143
SP124460	170162	148367	147636
SP124462	118149	101868	101143
SP128037	761170	645515	642748
SP128069	23049	19368	18992
SP128085	16114	14156	14026
SP128089	60373	50257	49865
SP128094	183745	156860	156147
SP128097	51435	41674	41429
SP128114	209183	187732	186779
SP128129	33547	27927	27755
SP128137	73502	63340	62863
SP128146	94098	84680	84088
SP128150	192559	160519	159364
SP128166	43438	35242	34928
SP128172	86790	75994	75539
SP128182	339573	304819	303894
SP128191	29781	25109	24698
SP128316	42559	36197	35871
SP128330	103243	85940	85339
SP128340	32093	28107	27716
SP128347	97035	82379	81724
SP128351	69140	57618	56892
SP128363	95795	83346	82693
SP128372	164529	144904	144170
SP128381	90008	76821	76413
SP128393	92006	76920	76249
SP128404	113589	95515	94972
SP128413	170401	150100	149477
SP128423	68195	59671	59135
SP128429	36128	30071	29625
SP128438	103894	91660	90944
SP128448	56935	49677	49247

SP128469	85080	72552	71989
SP128484	73541	66806	66250
SP128521	236611	202393	201537
SP128529	335628	297244	296405
SP128572	60320	52588	52238
SP128601	28061	22919	22555
SP128608	535940	478552	477506
SP128616	72620	62907	62490
SP128625	71958	61173	60693
SP128639	277301	239370	238258
SP128642	145958	127149	126050
SP128658	383889	350003	349149
SP128672	37585	31499	31131
SP128678	357369	324947	324053
SP128705	105739	94414	93651
SP128718	123161	105508	104859
SP128751	122230	104344	103525
SP128774	684506	589170	587084
SP128799	79382	71070	70478
SP128818	50213	45203	44834
SP128829	65104	54760	54381
SP128843	289461	244352	243101
SP128868	125260	108471	108011
SP128873	33829	29310	28868
SP128900	319351	282071	280887
SP129103	617593	508336	506825
SP129146	134365	114803	113922
SP129177	360487	319829	318595
SP129208	793951	732843	731349
SP129215	322291	278797	277416
SP129230	92693	80774	80266
SP129257	26431	23905	23633
SP129272	25284	20908	20584
SP129287	19258	16290	16136
TCGA-DA-A1HV-06A	269316	238306	237372
TCGA-DA-A1HW-06A	55586	46893	46461
TCGA-DA-A1HY-06A	119232	101543	100976
TCGA-DA-A1I0-06A	93147	81155	80571
TCGA-DA-A1IC-06A	136929	113854	112997

TCGA-DA-A3F3-06A	46865	39305	38744
TCGA-DA-A3F8-06A	174706	159515	158876
TCGA-EE-A29B-06A	94411	79783	79150
TCGA-EE-A2A0-06A	45784	38869	38432
TCGA-EE-A2GN-06A	59230	51518	51028
TCGA-EE-A2GT-06A	57712	47891	47335
TCGA-EE-A2MI-06A	218967	196452	195616
TCGA-EE-A3JI-06A	156660	132750	131884
TCGA-ER-A19D-06A	72730	61646	61040
TCGA-ER-A19J-06A	45522	38639	37964
TCGA-FS-A1ZK-06A	194311	163495	162407
TCGA-FS-A1ZP-06A	91622	82002	81533
TCGA-GN-A262-06A	58752	48985	48501
TCGA-GN-A266-06A	613758	537374	536353
TCGA-GN-A26A-06A	49463	41429	40982

Table S2. Oligonucleotide sequences for CPD-seq. Illumina P5 and *P7* adapters are indicated underlined and italicized respectively, and **indexes** are shown in bold and underline. Oligo 5' modifications are also indicated. All oligos were from Integrated DNA technologies (Coralville, IA).

Primers	Sequence
ARC141/142	5´-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT*T-3´
	5´-/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3AmMO/-3´
ARC143/144	5´-/5Biosg/ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNN/3AmMO/-3´
	5´-/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT/3AmMO/-3´
ARC154	5´-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´
ARC49	5´-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´
ARC78	5´-*CAAGCAGAAGACGGCATACGAGAT***CGTGAT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3´
ARC87	5´-*CAAGCAGAAGACGGCATACGAGAT***CACTGT**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3´
ARC88	5´-*CAAGCAGAAGACGGCATACGAGAT***ATTGGC**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3´

References

1. K. Elliott *et al.*, Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet* **14**, e1007849 (2018).