Ultra-deep sequencing differentiates patterns of skin clonogenic mutations associated with sun-exposure status

Classification: Biological Sciences, Genetics; Cancer risk; Early detection;

Lei Wei[1], Megan Fitzgerald[2], James Graham[1], Nicholas Hutson[1], Chi Zhang[3], Ziyun Huang[4], Qiang Hu[1], Fenglin Zhan[1,5], Jun Xie[6], Jianmin Zhang[7], Song Liu[1], Eva Remenyik[8], Emese Gellen[8], Oscar R. Colegio[9,10], Sean Christensen[11], Haifan Lin[12], Michael Bax[9], Jinhui Xu[13], Wendy J. Huss[14*], Barbara A. Foster[14*], Gyorgy Paragh[2,9*]

[*] These authors contributed equally

Author affiliations:

[1]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY
[2]Department of Cell Stress Biology, Roswell Park Comprehensive Cancer Center, Buffalo, NY
[3]School of Biological Sciences Center for Plant Science and Innovation, University of Nebraska, Lincoln, NE
[4]Department of Computer Science and Software Engineering, Penn State Erie, The Behrend College
[5]PET/CT center, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230001, P.R. China
[6]Department of Statistics, Purdue University, West Lafayette, IN
[7]Department of Cancer Genetics and Genomics, Roswell Park Comprehensive Cancer Center, Buffalo, NY
[8]Department of Dermatology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary
[9]Department of Dermatology, Roswell Park Comprehensive Cancer Center, Buffalo, NY
[10]Department of Immunology, Roswell Park Comprehensive Cancer Center, Buffalo, NY
[11]Department of Dermatology, Yale University School of Medicine, New Haven, CT
[12]Yale Stem Cell Center, Yale University School of Medicine, New Haven, CT
[13]Department of Computer Science and Engineering, State University of New York at Buffalo
[14]Department of Pharmacology and Therapeutics, Roswell Park Comprehensive Cancer Center, Buffalo, NY

Corresponding authors: Lei.Wei@RoswellPark.org and Gyorgy.Paragh@RoswellPark.org

Email addresses:
LW: Lei.Wei@RoswellPark.org
MF: Megan.Fitzgerald@RoswellPark.org
JG: james.graham@stonybrookmedicine.edu
NH: ndhutso@gmail.com
CZ: czhang5@unl.edu
ZH: zxh201@psu.edu
QH: qiang.hu@roswellpark.org
FZ: zhan209@hotmail.com
JX: junxie@purdue.edu
JZ: jianmin.zhang@roswellpark.org
SL: song.liu@roswellpark.org
ER: remenyik@med.unideb.hu

EG: emesegellen@med.unideb.hu
OC: Oscar.Colegio@RoswellPark.org
SC: Sean.Christensen@Yale.edu
HL: haifan.lin@yale.edu
MB: Michael.Bax@RoswellPark.org
JX: jinhui@buffalo.edu
WJH: Wendy.Huss@RoswellPark.org
BAF: Barbara.Foster@RoswellPark.org
GP: Gyorgy.Paragh@RoswellPark.org

Keywords: ultraviolet light, clonogenic mutation, photocarcinogenesis, sun exposure, ultra-deep sequencing, skin cancer risk

Abstract

Exposure to ultraviolet (UV) radiation results in growth of clonal somatic mutation (CM) harboring cell groups in clinically normal appearing skin and can cause non-melanoma skin cancer (NMSC), the most common cancers in humans. CMs can be considered the earliest detectable manifestations of photocarcinogenesis. Systematic evaluation of CMs is crucial to understand early photo-carcinogenesis. Differences between CMs in sun-exposed (SE) and non-SE (NE) areas are unknown. To address these problems, we developed a robust method to quantitatively measure CMs in skin using targeted (5,500bp), ultra-deep (64,730X median coverage), high-throughput sequencing in 450 matched SE and NE epidermal punch biopsies obtained from clinically normal skin (n=13 individuals). In total we identified 638 mutations including 298 UV-signature mutations (USMs). SE samples showed 3-fold higher rate of USM and higher USM variant allele frequencies than NE. We identified areas in *TP53*, *NOTCH1* and *GRM3* where mutation burden was significantly associated with UV-exposure. Six mutations were almost exclusively present in SE epidermis and accounted for 42% of the overall difference between SE and NE mutation burden. Cumulative Relative Clonal Area (CRCA) was defined as the overall relative percentage of the sampled skin area affected by CM. UV-induced DNA damage is a clinically relevant criterion and CRCA is a single metric of UV-damage. CRCA was dramatically elevated (mean = 11.2 fold) in SE samples compared to NE. This study represents the first analysis of clonogenic mutations in SE and NE human skin. The findings unveiled previously unknown patterns of early mutations introduced by UV exposure.

Significance statement:

In UV radiation exposed skin, mutations fuel clonal cell growth. We established a sequencing-based method to objectively assess the mutational differences between sun-exposed and non-sun-exposed areas of normal human skin. Striking differences, in both the numbers of mutations

and variant allele frequencies, were found between sun-exposed and -protected areas. Furthermore, we identified specific genomic regions where mutation burden is significantly associated with UV-exposure status. These findings revealed previously unknown mutational patterns associated with UV-exposure, providing important insights into UV radiation's early carcinogenic effects. The CRCA can be developed for quantitatively measuring subclinical skin UV damage and field carcinogenesis.

Background

Ultraviolet (UV) light is responsible for over 5 million cases of skin cancer annually in the US, which is more human malignancies than all other environmental carcinogens combined[1,2]. The preferential location of UVB induced DNA lesions results in a specific pattern of so-called UV signature mutations at dipyridine sites (C>T, CC>TT)[3]. In mammals, nucleotide excision repair eliminates DNA lesions, but this mechanism of repair is error prone resulting in frequent mutations. In most skin cancers the burden of UV signature driver mutations is high[3,4]. In melanocytes, hypersensitive genomic regions related to UV exposure have been recently identified[5]. While some cutaneous squamous cell carcinoma (cSCC) arise from visible precancerous lesions, actinic keratoses (AKs), many cSCC arise in apparently "normal" skin areas from precursors that are clinically invisible[6]. Therefore, visible precancers are an ominous sign but not a good early measure of photocarcinogenesis.

TP53 mutations were detected by immunohistochemistry as p53 variants are dysfunctional and accumulate due to impaired degradation[7,8]. These UV-induced *TP53* mutations facilitate clonal expansion of cells harboring them and therefore behave as early clonogenic mutations (CMs)[9]. For two decades *TP53* mutant keratinocyte cell clones were considered the earliest manifestations of skin carcinogenesis[7,8,10]. Because p53 clonal immunopositivity could not be efficiently quantified in human skin, detection of mutant *TP53* for assessment of photocarcinogenesis in clinical dermatology practice has been unattainable. The low abundance

of clonal DNA vs background DNA sequences previously limited efficient detection of early mutated cell groups. However, with improved high throughput sequencing technology we have finally reached the lower end of this threshold and efficient detection of CMs have become possible. Using high throughput ultra-deep sequencing, CMs were found to be abundant in clinically normal skin from sun-exposed sites in *NOTCH1, NOTCH2, FAT1* and several other genes besides *TP53*[11]. Prior attempts to establish a main-stream quantitative method for assessing photodamage had limited success[12,13]. A method that enables quantitative evaluation of early photodamage is expected to help optimize personalized sun-protective measures and may also serve as a tool for assessing the need and efficacy of early preventative treatment interventions.

In the current work we present a bioinformatics analysis method for evaluating focused ultra-deep sequencing to quantify low variant allele frequency CMs in the epidermis and show differences in CMs between sun-exposed and non-sun-exposed skin areas. We also present evidence that limited genomic areas harbor most of the mutation burden difference between sun-exposed and protected sites.

Methods

Samples:

A total of 464 normal human skin samples were collected from 13 Caucasian post-mortem donors over the age of 55 years using Roswell Park's Rapid Tissue Acquisition Program under a Roswell Park approved IRB protocol within 24 hours of death from frequently sun-exposed (SE) sites (left dorsal forearm) and non-sun-exposed (NE) sites (left medial buttock). Exclusion criteria included any visible skin abnormalities in the tissue areas. Eligible donors were identified and clinically normal appearing skin was harvested. Skin samples were kept in tissue

preservation medium, Belzer UW cold storage solution (Bridge to Life, USA) at 4°C until processed. All samples that could be processed within 36 hours or less after death were included in the study. The mean age of the donors was 72.3 years (SD: ±8.2 years; range 60-80 years). The male to female gender ratio was 7:6, and 12/13 donors had no history of skin cancer.

The adipose tissue was removed from each human skin sample using sterile scissors. The samples were cut into strips wide enough to harvest 6 mm punches. The epidermis was separated from the dermis by placing the strips in tubes containing 10 ml of 5U/ml Dispase II (Stem Cell Technologies, USA) and incubated at 4°C overnight and at 37°C for 2-3 hours. After Dispase digestion the specimens were placed in a petri dish containing a small amount of 1x DPBS (Corning, USA) and using sterile tweezers, the epidermis was carefully removed from the dermis. Using disposable biopsy punches, 1, 2, 3, 4 and 6 mm diameter epidermal pieces were taken from the epidermal sheets and punched epidermal pieces were placed into a sterile 1.5 mL vials. In addition to the epidermal punches, large bulk pieces of dermis were also removed from the skin samples using a disposable #15 blade and placed into a sterile 1.5 mL vial for use as a germline control.

DNA isolation:

DNA samples were extracted using Purelink™ Genomic DNA mini kit (Invitrogen, USA). Epidermal samples were digested using Proteinase K at 55°C heating block overnight following the manufacturers recommendations. DNA was eluted with 28 µL of Molecular Biology Grade Water (Corning, USA) for 1 and 2 mm punches or 36 µL of Molecular Biology Grade Water for 3, 4, and 6 mm punches. The isolated genomic DNA was stored at -20°C and the DNA concentration of each extraction was measured using a Qubit fluorometer (Invitrogen, USA).

Ultra-deep Targeted Sequencing:

The sequencing libraries were generated using the TruSeq Custom Amplicon kit (Illumina, USA) using 50 ng of gDNA. Amplicons of ~150bp in length were designed using Illumina Design Studio Software. Custom oligo capture probes that flank the regions of interest were hybridized to the gDNA. A combined extension/ligation reaction completed the region of interest between these flanking custom oligo probes. PCR was then performed to add indices and sequencing adapters. The amplified final libraries were cleaned up using AmpureXP beads (Beckman Coulter). Purified libraries were run on a Tapestation DNA1000 screentape chip to verify desired size distribution, quantified by KAPA qPCR (KAPA Biosystems) and pooled equal molar in a final concentration of 2 nM. Pooled libraries were loaded on an Illumina HiSeq Rapid Mode V2 flow cell following standard protocols for 2x100 cycle sequencing.

Bioinformatics analysis:

The bioinformatics analysis began by utilizing high quality paired-end reads passing Illumina RTA filter aligned to the NCBI human reference genome (GRCh37) using BWA[14]. PCR duplicated reads were marked and removed by using Picard (http://picard.sourceforge.net/). The coverage quality control requires at least 80% of the targeted region covered by a minimum of 1,000X coverage. Putative mutations, including single nucleotide variants (SNVs) and small insertions/deletions (Indels), were initially identified by running variation detection module of Strelka[15] on each SE or NE epidermis sample paired with the matched dermal sample. From the detected SNVs, dinucleotide variants (DNV) or cluster of single nucleotide variants (CSNV) were recognized by running Multi-Nucleotide Variant Annotation Corrector (MAC) [16] on the original sequences. The putative mutations detected from all samples were consolidated into a list of unique mutations. Every unique mutation was re-visited in all samples to calculate the numbers of mutant/wildtype reads, as well as variant allele frequency (VAF) in each sample as previously described [17].

To distinguish mutations from background errors, we modelled each mutation's background error rate distributions using VAFs from all 14 control (dermal) samples. For each mutation, we started by fitting a *Weibull* distribution to VAFs from all control samples following a previously published method[18], then every SE or NE epidermal sample's VAF was compared to the fitted distribution. A positive sample was defined as the sample's VAF of a mutation was significantly above background ($p < 0.05$, after Bonferroni correction). All mutations including SNVs, DNVs, CSNVs and Indels were annotated using a customized program with NCBI RefSeq database.

Cumulative Relative Clonal Area (CRCA), defined as the overall percentage of biopsied skin area covered by UV-signature mutations (USMs) in a patient skin punch, was calculated as following:

$$CRCA = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (\pi r_i^2 * 2VAF_j)}{\sum_{i=1}^{n} \pi r_i^2}$$

with n = the total number of punches collected in the patient; r_i = the size (radius) of each punch; m_i = the number of mutations in punch i; VAF_j = the variant allele fraction of a specific mutation j. Here, the calculation of CRCA was based on the assumption that all mutations occur in one chromosome of regular diploid genomic regions. Additionally, although we did not consider the situation when multiple mutations occur in the same cell, we did identify mutations that occur on the same reads and combined them into one mutation using MAC [16].

Statistics:

The overall mutation numbers and VAFs between SE and NE groups were compared using the Wilcoxon test. Group-specific markers, including mutations, genes, regions and signatures were identified using the Fisher's exact test where the two variables in the contingency table were the samples' sun-exposure status and mutational status.

Results

Ultra-deep sequencing of epidermal DNA samples using a customized focused panel

To generate a focused sequencing panel, targeting the most commonly mutated sequences in normal human skin, we selected an area of focus based on a previous dataset[11]. All previous mutations were assigned to 100 bp genomic segments, after sorting the segments by number of mutations, we designed a panel to capture the top 55 segments from 12 genes (5.5 kb in total, **Table S1**). The majority (65%) of the targeted segments came from the following 3 genes: *NOTCH1*, *NOTCH2*, and *TP53*. Additionally, 79% of the targeted segments lie in protein coding regions, and the remaining segments were mostly in introns. Furthermore, in the previous dataset 87% of the samples harbored at least one mutation covered by this panel.

We used the focused panel to sequence a discovery cohort of 374 human skin samples from 13 post-mortem donors: 360 epidermal samples, equally acquired from both sun-exposed (SE) and non-sun-exposed regions (NE) using 1 mm, 2 mm, 3 mm, 4 mm or 6 mm punch sizes. From the same 13 donors, DNA from bulk NE dermis (n=14, 1 donor contributed 2 samples) was isolated for germline controls. After initial analysis to determine the optimal punch size, we tested a separate validation cohort of 90 epidermal samples from 9 of the 13 donors using the most effective punch size (2 mm, as detailed in results “Optimization of punch size for USM detection”). In total, the dataset contains 464 samples: 225 SE, 225 NE, and 14 dermal samples as controls **(Table 1)** from 13 individuals. After sequencing, 85% of samples reached a minimum of 10,000X coverage in at least 80% of the targeted region. The median of average coverage across all samples was 64,730 **(Table S2)**, with only one sample exclusion (NE sample) due to sequencing failure.

Differentiating somatic mutations by SE and NE skin site type

To identify somatic mutations in both SE and NE specimens, we compared each epidermal sample with its patient-matched dermal sample followed by an in silico error suppression to remove germline polymorphisms and low-frequency technical artifacts. Moreover, we identified dinucleotide and other complex mutation by re-analyzing the raw reads. Through these processes a total of 638 mutations were identified, predominantly single nucleotide variants (SNVs, n = 614 or 96.2%) or dinucleotide variants (DNVs, n = 20 or 3.1%). The median variant allele frequency (VAF) of all mutations was 2.1% (range 0.1% - 36.6%), and only 3% mutations reached a VAF greater than 10%, a level that can generally be detected by conventional sequencing **(Table S3)**.

Among the 55 genomic segments, mutations were detected in 50 segments with an average of 7.1 and 4.7 mutations per segment in SE and NE samples, respectively **(Figure 1)**. Two segments were significantly ($p<0.001$) associated with UV-exposure status, approximately corresponding to *TP53* amino acids 227-261 ("*TP53-3*", SE/NE = 38/0) and *NOTCH1* p.449-481 ("*NOTCH1-9*", SE/NE = 30/4). Interestingly, an adjacent region in *NOTCH1* p.419-449 ("*NOTCH1-10*"), appeared to be non-differentiating between SE (n=48) and NE (n=40), even though it harbored the highest number of mutations among all segments. An additional three segments were also significantly enriched in SE samples ($p<0.05$), two in *NOTCH1* ("*NOTCH1-14*" and "*NOTCH1-19*") and one in *GRM3* ("*GRM3-2*"). On the gene level, mutations in SE samples were significantly enriched in two genes ($p<0.05$): *TP53* and *GRM3*, with SE/NE mutation ratios of 6.3 and 4.3, respectively. Mutations identified in nine other genes did not exhibit significant association with sun-exposure status either on the gene or segment level: *NOTCH2, ARID1A, SALL1, SCN1A, ERBB4, FAT4, FGFR3, ADGRB3 and PPP1R3A*.

In order to identify potential hotspot mutations and regions associated with UV-exposure, all mutations were ordered by their genomic locations. Upon ordering, one specific region in NE samples, *TP53* (p.217-280), appeared to be "mutation exempt" in comparison to surrounding

regions. In contrast however, this region became highly mutated in SE samples **(Figure 2a)**. In another recent study including RNA sequences of both SE and NE normal skin samples, four mutations were found in this region, all from SE samples (Table SX)[19]. On the mutation level, six mutations were significantly enriched in SE samples: *TP53* R248W, *NOTCH1* P460L, *NOTCH1* S385F, *NOTCH1* E424K, *TP53* G245D and *NOTCH1* P460S, and nearly all of them were exclusively found in SE samples ($p<0.05$, **Figure 2b**). Five of the six enriched mutations were found in both discovery and validation cohorts, suggesting they were unlikely to be caused by batch-effect. Unexpectedly, one specific mutation (*NOTCH1* E424K) was associated with significantly elevated VAFs (median = 10%, $p<0.001$, Wilcoxon test), about five-fold higher than other mutations (median = 2.1%). Of the 19 mutations with VAFs above 0.1, five were *NOTCH1 E424K.* Through protein structure modelling **(Figure 2c)**, we found that the *NOTCH1* E424K mutation is predicted to disrupt the binding of *NOTCH1* to delta-like canonical ligand 4 (*DLL4*), a negative regulator of the Notch signaling pathway[19]. By prohibiting formation of a salt bridge between *NOTCH1* E424 and *DLL4* K189/R191, the mutation E424K creates a repulsive force that inhibits *DLL4* binding [20]. Based on the biological role of *DLL4* and *NOTCH1*, we expect the mutation to promote epithelial proliferation [21,22]. The overall prevalence of the *NOTCH1* E424K mutation in our dataset is 2.7%. For comparison, in GINIE cBioPortal[23], *NOTCH1* E424K is mutated in 1.3% of cutaneous SCCs, 0.04% in melanomas, and is rarer in other cutaneous or non-cutaneous malignancies **(Table S5)**.

UV-signature mutations are exclusively enriched in SE samples

To determine whether the current mutations can be further classified based on known patterns of mutations caused by UV, we compared the identified mutations with previously known UV-signature mutations (USMs), i.e., C>T transition at dypyrimidines [4]. Among all 638 mutations in SE and NE samples, 298 matched to the USMs. Of these 298 USMs, 76% were present in SE

samples. Conversely, non-UV-signature mutations (NUSMs) appeared equally in SE (47%) and NE (53%) groups **(Figure 3a)**, confirming these mutations were not associated with UV-exposure. Together, USMs were significantly enriched in SE samples ($p<0.01$, Fisher's exact test). Especially for high-VAF mutations, 18 of 19 mutations with VAFs above 0.1 were from SE samples, and 13 of them were USMs. To explore a potential association between UV-signature and functional effects, we classified all 638 mutations into four effect-groups: nonsense, missense, silent and noncoding. The classification further split each effect-group by USMs and NUSMs **(Figure 3b)**. In two of four effect-groups, nonsense and missense, the SE mutations were significantly enriched (<0.05, Fisher's exact test) in USMs compared to NUSMs. Moreover, the ratios of SE/NE in USMs were 9 and 4.2, for nonsense and missense, respectively. Since NUSMs were not correlated with UV exposure, further development of a sequencing-based measurement of UV-induced DNA damage included only USMs.

UV-induced DNA damage level quantification by UV-signature mutations

Next the feasibility of using USMs to quantify UV-induced DNA damage was explored based on the hypothesis that SE samples harbor more USMs and are associated with higher VAFs compared to NE samples. To avoid the potential bias introduced by different punch sizes, only the most abundant 2 mm sample size ($n = 90$ and 89, SE and NE, respectively) **(Figure 4a)** was analyzed. The average USM rate per sample had a three-fold difference ($p <0.001$, Wilcoxon test) of 1.2 in SE and 0.4 in NE. Multiple USMs were found in 33% of SE samples but only 9% of NE samples. Additionally, the identified USMs had significantly higher VAFs in SE (mean = 3.7%) than NE (mean = 2.1%) samples, ($p < 0.01$, Wilcoxon test), suggesting the presence of larger clones in SE samples. In an expanded comparison that included all punch sizes, the pattern was consistent with 34% of SE and only 6% of NE samples having multiple USMs

(Table S4) and three-fold higher average USMs per sample in SE (1.0) than NE (0.3) samples ($p < 0.001$, Wilcoxon test).

To help capture the mutational burden of a patient in a single metric, an overall measurement of UV-induced DNA damage was developed by distilling the VAFs and number of USMs: Cumulative Relative Clonal Area (CRCA), defined as the overall percentage of biopsied skin area covered by USMs in a patient skin punch **(Figure 4b)**. Here, we disregard whether mutations occurred independently or were present in the same clone. Hence, CRCA does not provide an exact measure of the mutated cell population, but rather serves as a relative measure of the mutation burden in the sampled area. To minimize the potential for repeat counting of co-occurring mutations in the same cells, co-occurring mutations were identified, primarily dinucleotide CC>TT mutations, and counted as one mutation. When counted separately by sun-exposure status, the median CRCA, i.e. the relative mutational load of biopsied skin areas, across the 13 patients was 6.1% (range 1.4-14.2%) in SE and 1.4% (range 0.1-4.0%) in NE samples. In matched samples from the same individuals comparing SE versus NE skin, the average CRCA was 11.2-fold higher (range = 1.4 - 55.0-fold). The CRCAs were calculated using only USMs. If all CMs were included, the all-CM-CRCA would be only 2.2-fold higher (range = 0.8 - 5.6-fold) in SE skin (data not shown).

The effect of punch size on USM detection

In the discovery cohort, we sought to compare different punch sizes to determine the most efficient one for detecting USMs. Theoretically, although larger punches likely contain more clones, they tend to become less effective for detecting smaller clones due to a dilutional effect by other clones harboring no mutations or different mutations **(Figure 5a)**. Overall, USMs were detected in 54% of the SE, which was significantly greater than the detection rate of 21% of the

NE ($p < 0.001$, Fisher's exact test). Among all 5 punch sizes, 2 mm punches were most likely to capture CMs, positive rate of 64%**,** and had the most significant difference between SE and NE samples ($p < 0.0001$, **Figure 5b**). Thus, only 2mm punches were collected in the 90-sample validation cohort. In the validation cohort, similarly, a higher number of USMs were identified in the SE skin and the positive rate of USMs (69%) was consistent with the discovery cohort.

When combining the discovery and validation cohorts, the SE samples had the highest positive rate of 67% for USMs in 2 mm samples and were significantly higher than NE samples ($p < 0.001$), followed by 60% in 4 mm ($p < 0.05$), and 54% in 3 mm ($p < 0.05$). Interestingly, the largest punch size of 6 mm and the smallest, 1 mm, showed lower positive USM rates in SE samples of 53% (6mm) and 36% (1mm), respectively. The 2 mm punches were selected for analysis because the detection rate of SE skin USMs in 2 mm punches was significantly higher than 1 mm ($p <0.01$) and results in less scaring than larger punches. In all NE samples, positive USM rates ranged from 17-30% (**Figure 5c**). Moreover, the punch size also affected the detected VAFs of the mutations. Specifically, in SE samples, larger punches were associated with smaller VAFs. The VAFs' standard deviation was the highest in 1 mm punches (8.9%) and decreased with punch size: 2 mm (4.3%), 3 mm (2.8%), 4 mm (2.6%) and 6 mm (1.7%). This trend, between VAF range and punch size, did not occur in NE samples **(Figure 5d)**. Considering the high positive USM rate and VAF, the 2 mm punch size was the most effective in detecting USMs.

Mutation nucleotide contexts associated with UV-exposure

The nucleotide context of all mutations was analyzed to determine their relative contribution to USMs. Based on each mutation's trinucleotide context (for DNVs only dinucleotide context was included), all mutations were separated into 83 contexts of which 13 contexts matched

previously described USMs [4]. None of the 70 non-USM contexts were differentially detected in SE versus NE samples. The 13 previously identified USM contexts did not equally differentiate SE and NE samples. Only 6 of the 13 contexts were significantly enriched in SE samples ($p<0.05$); including the dinucleotide CC>TT context, which was exclusively found in SE samples **(Figure 6a)**. The remaining seven UV-signature contexts were not significantly enriched in SE samples. Of particular note G[C>T]C, which was the most frequent context by total number of mutations within SE and NE samples, was not associated with sun-exposure status **(Figure 6b)**. These findings in early CMs of normal tissues are overall consistent with the previous "Mutational Signature #7" reported in skin cancers [24]. In both cases, T[C>T]C was the highest ranked context associated with UV exposure.

Discussion

Most cancers are initiated by accumulation of somatic mutations[25,26]. However, early mutations in normal tissues are difficult to detect due to lack of extensive clonal expansion. Several recent studies demonstrated the feasibility of detecting clonal mutations (CMs) using high-throughput sequencing in various tissue types[11,19,27]. However, the contribution of these CMs to cancer remains poorly understood in several important areas: what types of mutations are generated by each specific carcinogen, how the CMs are selected by cellular microenvironment, and which mutations directly contribute/lead to the development of cancer. Indeed, all tissue types are under the influence of multiple intrinsic and extrinsic factors that vary greatly by individual's lifestyle and environment. Therefore, studying the CMs generated by one specific carcinogen requires comparative studies of matched sample types.

This study represents a comprehensive analysis of individual-matched human skin samples to detect UV radiation's mutational effects. We optimized our approach for detecting UV-induced

CMs by: 1) acquiring matched SE/NE skin samples of each individual to control for aging and other environmental factors; 2) separating epidermal from dermal layers decreases non-epidermal background DNA quantity; 3) ultra-deep DNA sequencing for maximized sensitivity, followed by error-suppression using matched dermal samples to exclude sequencing and alignment errors. Consistent with previous studies [11,19], CMs were widespread in epidermal samples. As expected mutation burden and VAFs were significantly elevated in SE samples. The mutational signatures of the current CMs are consistent with those previously found in skin cancers, confirming the validity of the detected CMs. Markedly, our unique approach allowed us to identify several important new insights about epidermal CMs: 1) the existence of "mutation-exempt" regions in human genomes. Although mutations frequently occur across most of the sequenced regions in NE skin, presumably due to metabolism and aging related intrinsic or non-UV extrinsic factors, the mutation-exempt regions harbor no mutations prior to UV exposure. Interestingly, the mutation-exempt property of these regions is not maintained upon exposure to UV radiation. Further studies are warranted to determine whether the mutation-exempt regions are caused by active protection or by a passive selective advantage, and to explore how this mechanism is bypassed after exposure to UV radiation. 2) in SE skin, USMs were found to be significantly enriched in Glutamate Metabotropic Receptor 3 (*GRM3*), which was previously identified as a potential therapeutic target in melanoma [28], but not reported as a cancer driver in cutaneous SCC; 3) we identified six mutations that were almost exclusively mutated in SE skin. All six mutations had been previously reported in human cancers. Three of these mutations in *TP53* R248W and G245D were highly recurrent with hundreds of occurrences reported in *the Catalogue Of Somatic Mutations In Cancer* [29] and all six of these mutations are amongst the most common mutations found in human cutaneous squamous cell carcinomas in the cBioPortal suggesting that the presence of these mutations may be representative of an early phase of carcinogenesis.

Future clinical utility directed our approach, and these results demonstrate the feasibility of using a small panel of genomic regions (5.5 kb) to quantitatively measure UV-induced CMs. We established Relative Cumulative Clonal Area (CRCA) as a combined measure of mutation burden and clone sizes, that showed an overall 11.2-fold difference between patient-matched SE/NE samples. The most effective punch size in capturing CMs in the current study was 2 mm and is clinically favorable as it leaves relatively smaller scars due to the small diameter punch. However, a non-invasive skin sampling method may provide even wider accessibility to epidermal sampling. In addition, the efficiency of this panel is related to the performance of sequencing method and mutation calling algorithm, which will likely be improved with adoption of more sensitive future methods focusing on the genomic hotspots that are most sensitive to UV exposure.

The current study focused on the most frequently mutated regions in sun-exposed skin samples defined by the mutations in a previous study [11]. However, we note that many of these regions are mutated in both sun-exposed and non-sun-exposed skin samples, suggesting many mutations in these regions were unrelated to UV exposure. In fact, only 6 of 55 original regions were found to harbor significantly enriched mutations in SE samples. Future studies, including much larger targeted regions, are needed to systematically identify UV-sensitive genomic regions. The skin samples were all from clinically normal skin; therefore, they do not provide longitudinal information about clone initiation and progression. Future studies including biopsies of both SCC and adjacent normal skin acquired at multiple time points are warranted to unveil the complete role of these CMs in cancer.

Conclusions

In summary, this study provides individual-matched ultra-deep sequencing based genomic analysis of UV's mutational effects in human skin. Our results revealed previously unknown mutational patterns associated with UV-exposure status, providing important insights into UV radiation's early carcinogenic effects. The quantification of CMs has the potential to become the cornerstones for future development of quantitative measures of UV-induced DNA damage, as measured by CRCA, in the clinical setting to monitor early carcinogenesis and highlight the importance of sun protection.

List of abbreviations

Ethics Statements

All specimens were collected from post-mortem donors collected in collaboration with Buffalo's local organ procurement organization (Unyts) the Roswell Park's Rapid Tissue Acquisition Program under a Roswell Park approved IRB protocol.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding authors upon request.

Competing interests

None.

Funding

This work was supported by National Cancer Institute (NCI) grant P30CA016056 to Roswell Park Comprehensive Cancer Center, the Roswell Park Alliance Foundation, NIH U24 CA232979 and 5U24ES026465.

Acknowledgments

The authors thank the excellent technical help provided by Paula Pera, MS.

Common Abbreviations:

UV – Ultraviolet

CM – Clonogenic mutation

NMSC – Nonmelanoma skin cancer

SE – Sun-exposed

NE – Non-sun-exposed

USM – UV-signature mutation

NUSM – Non-UV-signature mutation

CRCA – Cumulative Relative Clonal Area

cSCC – Cutaneous squamous cell carcinoma

AK – Actinic keratosis

SNV – Single nucleotide variant

Indels – Insertions/deletions

DNV – Dinucleotide variant

CSNV – Cluster of single nucleotide variant

MAC – Multi-Nucleotide Variant Annotation Corrector

VAF – Variant allele frequency

Figure Legends

Figure 1. Graph shows the number of mutations identified within 100-bp genomic target areas in SE and NE skin samples. Top: the overall gene-level percentage of mutations from SE and NE samples in six genes containing more than one 100-bp genomic target area (*NOTCH2, TP53, FAT4, FGFR3, GRM3, NOTCH1*), with the left and right boundaries aligned to the x-axis regions of the gene. Red stars indicate the segments or genes where mutations are significantly enriched in the SE samples. Note, *NOTCH1* had three genomic regions with significant difference between SE and NE mutational burden, but overall gene level mutational burden was not significantly different.

Figure 2. Mutational patterns associated with UV-exposure status. a. All mutations are ordered by their genomic locations. X-axis: the order of the mutation's genomic location. Y-axis: variant allele fraction (VAF). Colors demonstrate genes harboring the mutations. The three genes with significant difference between SE and NE, either on the gene level or segment level, were labeled on top (*TP53, GRM3, NOTCH1*). One specific mutation with elevated VAFs (*NOTCH1* E424K) is indicated with a red arrow. b. The VAF of the six individual mutations that are significantly enriched in SE vs NE epidermis in the discovery (green) and validation (red) data sets. The dotted red line represents median VAF of all mutations and black lines indicate the median of each group. c. The predicted protein complex structure of NOTCH1 and DLL4 to show the position of the mutant E424K and the interacting partners, DLL4 K189/R191, in wild type.

Figure 3. Summary of distribution of mutations in SE and NE samples. Only UV-signature mutations are associated with sun-exposure status. a. Left: higher numbers of USMs are found in SE than NE skin. Right: NUSMs are almost equally distributed in SE and NE samples. Red

dotted line indicates high-VAF (>0.1). b. The numbers of mutations by each amino-acid-change type found in SE and NE skin, grouped by USMs and NUSMs.

Figure 4. Quantitative features of USMs. a. Overall distribution of the numbers of USMs per sample (top) and the VAFs of the mutations (bottom) using the 2 mm punch size. Inside the violin plots: black dots - original data points from individual samples; yellow dot with bar - averaged value with standard deviation. SE samples are associated with higher numbers of USMs, as well as higher VAFs indicating potential larger clones. b. To provide an overall measurement score for UV-induced skin damage, a Cumulative Relative Clonal Area (CRCA) was developed, to represent the overall percentage of the biopsied skin areas that are covered by clonal mutations, assuming all mutations occurred independently. In all 13 donors tested, CRCAs were higher in SE than in the matched NE group, with the ratios of SE/NE ranged from 1.4 to 55.0 (mean = 11.2).

Figure 5. Optimization of punch size for detecting USMs. a. A representative figure showing one punch of each collection size with the most mutations. Each mutation is plotted as a dot with its size calculated to match the clonal area harboring the mutation. 3 mm was not shown as it was obtained by cutting a 6 mm punch into quarters. b. In the discovery cohort, 2 mm was found to be the most efficient size in differentiating CRCA from SE and NE skin samples by p value. c. Distribution of numbers of USMs per sample at each punch size, combining both the discovery and validation cohorts. d. VAF of USM detected in different size punches. The size of the dot indicates the approximate relative area of cells containing the mutation. VAFs of USMs detected from larger punches are associated with smaller variations, which was only observed in SE samples.

Figure 6. Mutational contexts associated with UV-exposure. a. Each dot represents a specific mutation context of SNVs (with the 5' and 3' bases) and dinucleotide mutations. X-axis: the total numbers of mutations of each context; y-axis: p value of the context for differentiating SE and

NE skin, shown as -log(p). b). For all USM contexts, the numbers of mutations in SE and NE skin. Mutation contexts are ordered by the p value of SE vs NE in an increasing order from left to right.

Tables

Table 1. Patient and sample cohort

Patient	Control (dermis)	Epidermis SE/NE pairs						
		1 mm	2 mm	3 mm*	4 mm	6 mm	2 mm#	Total SE/NE pairs
Pt1	1		5	4	1	1		11
Pt2	1	3	5	4	1	1		14
Pt3	1	3	5	4	1	1		14
Pt4	1	3	3					6
Pt5	1	3	3	3	3	3	5	20
Pt6	1	3	3	3	3	3	5	20
Pt7	1	3	3	3	3	3	5	20
Pt8	1	3	3	3	3	3	5	20
Pt9	1	3	3	3	3	3	5	20
Pt10	1	3	3	3	3	3	5	20
Pt11	1	3	3	3	3	3	5	20
Pt12	1	3	3	3	3	3	5	20
Pt13	2	3	3	3	3	3	5	20
Total	14	36	45	39	30	30	45	225

* 3 mm punches were obtained by cutting 6 mm punches into quarters

Validation cohort containing only 2 mm punches

References

1 Rogers, H. W., Weinstock, M. A., Feldman, S. R. & Coldiron, B. M. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. *JAMA Dermatol* **151**, 1081-1086, doi:10.1001/jamadermatol.2015.1187 (2015).

2 Koh, H. K., Geller, A. C., Miller, D. R., Grossbart, T. A. & Lew, R. A. Prevention and early detection strategies for melanoma and skin cancer. Current status. *Arch Dermatol* **132**, 436-443 (1996).

3 Brash, D. E. UV signature mutations. *Photochem Photobiol* **91**, 15-26, doi:10.1111/php.12377 (2015).

4 Wikonkal, N. M. & Brash, D. E. Ultraviolet radiation induced signature mutations in photocarcinogenesis. *J Investig Dermatol Symp Proc* **4**, 6-10 (1999).

5 Premi, S. *et al.* Genomic sites hypersensitive to ultraviolet radiation. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.1907860116 (2019).

6 Marks, R., Rennie, G. & Selwood, T. S. Malignant transformation of solar keratoses to squamous cell carcinoma. *Lancet* **1**, 795-797, doi:10.1016/s0140-6736(88)91658-3 (1988).

7 Ling, G. *et al.* Persistent p53 mutations in single cells from normal human skin. *Am J Pathol* **159**, 1247-1253, doi:10.1016/S0002-9440(10)62511-4 (2001).

8 Brash, D. E. Cancer. Preprocancer. *Science* **348**, 867-868, doi:10.1126/science.aac4435 (2015).

9 Urano, Y. *et al.* Frequent p53 accumulation in the chronically sun-exposed epidermis and clonal expansion of p53 mutant cells in the epidermis adjacent to basal cell carcinoma. *J Invest Dermatol* **104**, 928-932 (1995).

10 Williams, C. *et al.* Clones of normal keratinocytes and a variety of simultaneously present epidermal neoplastic lesions contain a multitude of p53 gene mutations in a xeroderma pigmentosum patient. *Cancer Res* **58**, 2449-2455 (1998).

11 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).

12 Gamble, R. G. *et al.* Sun damage in ultraviolet photographs correlates with phenotypic melanoma risk factors in 12-year-old children. *J Am Acad Dermatol* **67**, 587-597, doi:10.1016/j.jaad.2011.11.922 (2012).

13 Creidi, P. *et al.* Profilometric evaluation of photodamage after topical retinaldehyde and retinoic acid treatment. *J Am Acad Dermatol* **39**, 960-965 (1998).

14 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

15 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817, doi:10.1093/bioinformatics/bts271 (2012).

16 Wei, L. *et al.* MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC genomics* **16**, 569, doi:10.1186/s12864-015-1779-7 (2015).

17 Wei, L. *et al.* Accurate Quantification of Residual Cancer Cells in Pelvic Washing Reveals Association with Cancer Recurrence Following Robot-Assisted Radical Cystectomy. *J Urol* **201**, 1105-1114, doi:10.1097/JU.0000000000000142 (2019).

18 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547-555, doi:10.1038/nbt.3520 (2016).

19 Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, doi:10.1126/science.aaw0726 (2019).

20 Luca, V. C. *et al.* Structural biology. Structural basis for Notch1 engagement of Delta-like 4. *Science* **347**, 847-853, doi:10.1126/science.1261093 (2015).

21 Blanpain, C., Lowry, W. E., Pasolli, H. A. & Fuchs, E. Canonical notch signaling functions as a commitment switch in the epidermal lineage. *Genes Dev* **20**, 3022-3035, doi:10.1101/gad.1477606 (2006).

22 Lefort, K. & Dotto, G. P. Notch signaling in the integrated control of keratinocyte growth/differentiation and tumor suppression. *Semin Cancer Biol* **14**, 374-386, doi:10.1016/j.semcancer.2004.04.017 (2004).

23 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1, doi:10.1126/scisignal.2004088 (2013).

24 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

25 Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43-47, doi:10.1038/nature16166 (2016).

26 Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81, doi:10.1126/science.1260825 (2015).

27 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science*, doi:10.1126/science.aau3879 (2018).

28 Kunz, M. The genetic basis of new treatment modalities in melanoma. *Curr Drug Targets* **16**, 233-248, doi:10.2174/1389450116666150204112138 (2015).

29 Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008).

Figure 1. Distribution of somatic mutations by genomic regions

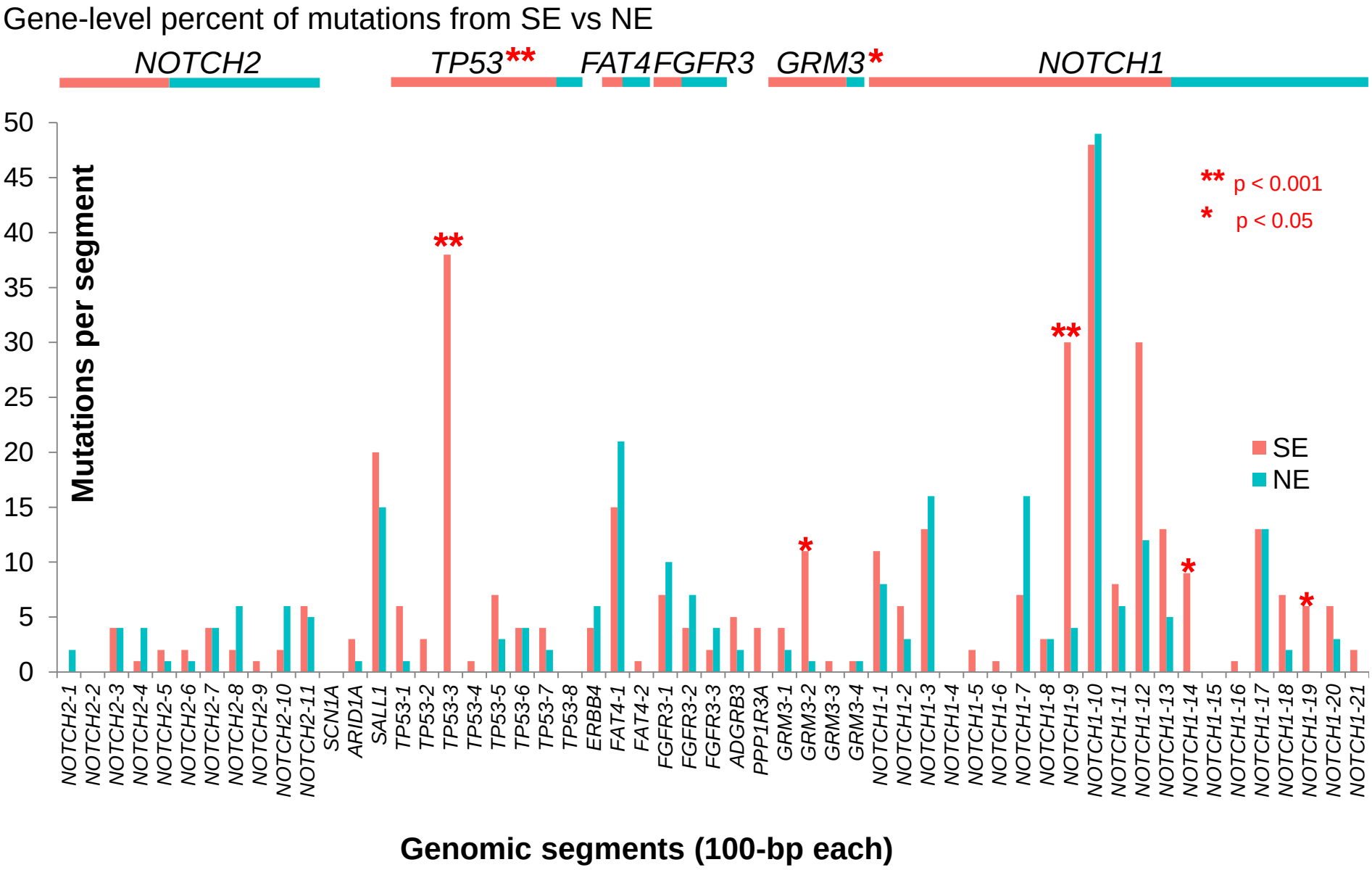


Figure 2. Mutational patterns associated with UV-exposure status

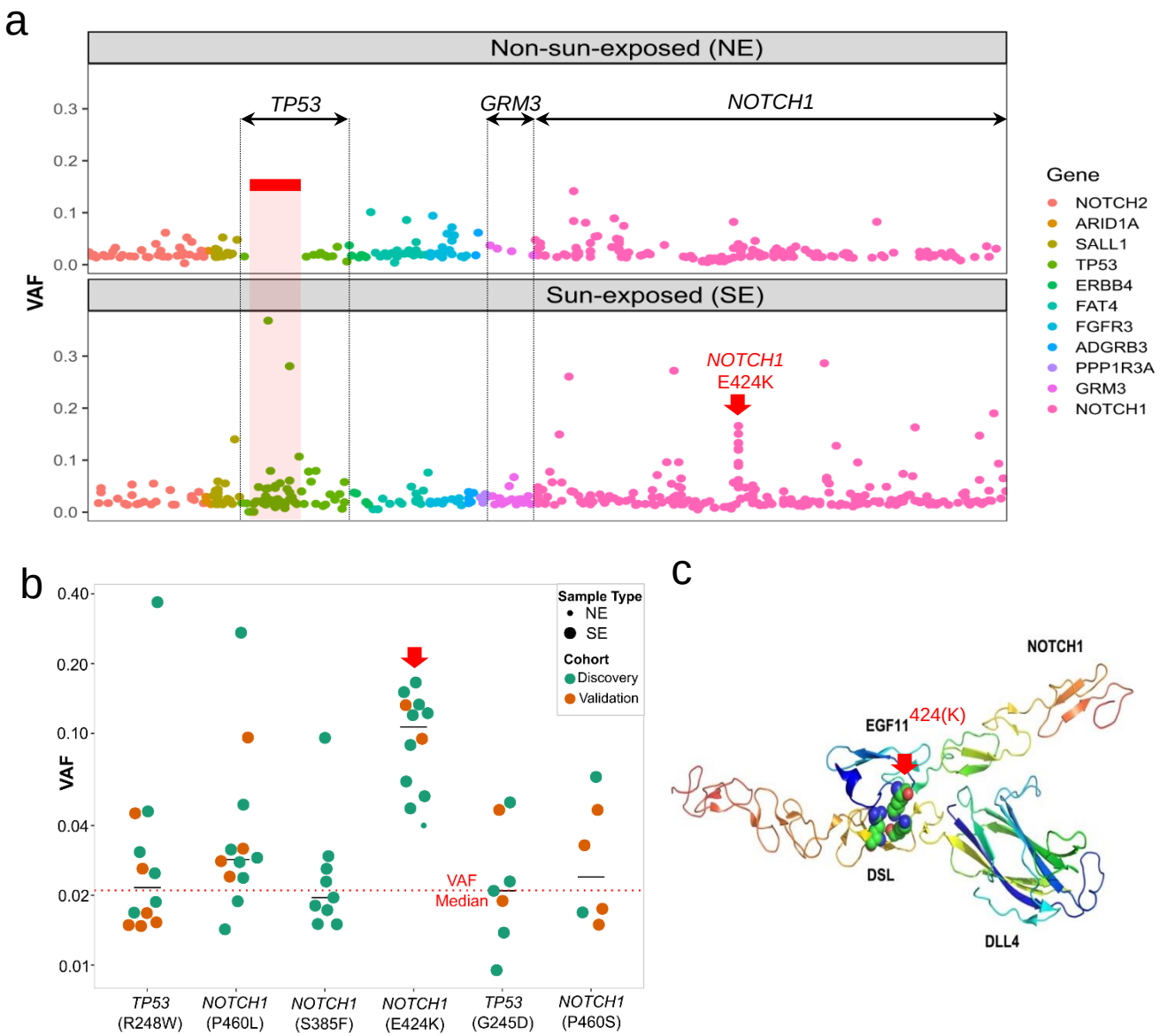


Figure 3. Only UV-signature mutations are associated with sun-exposure

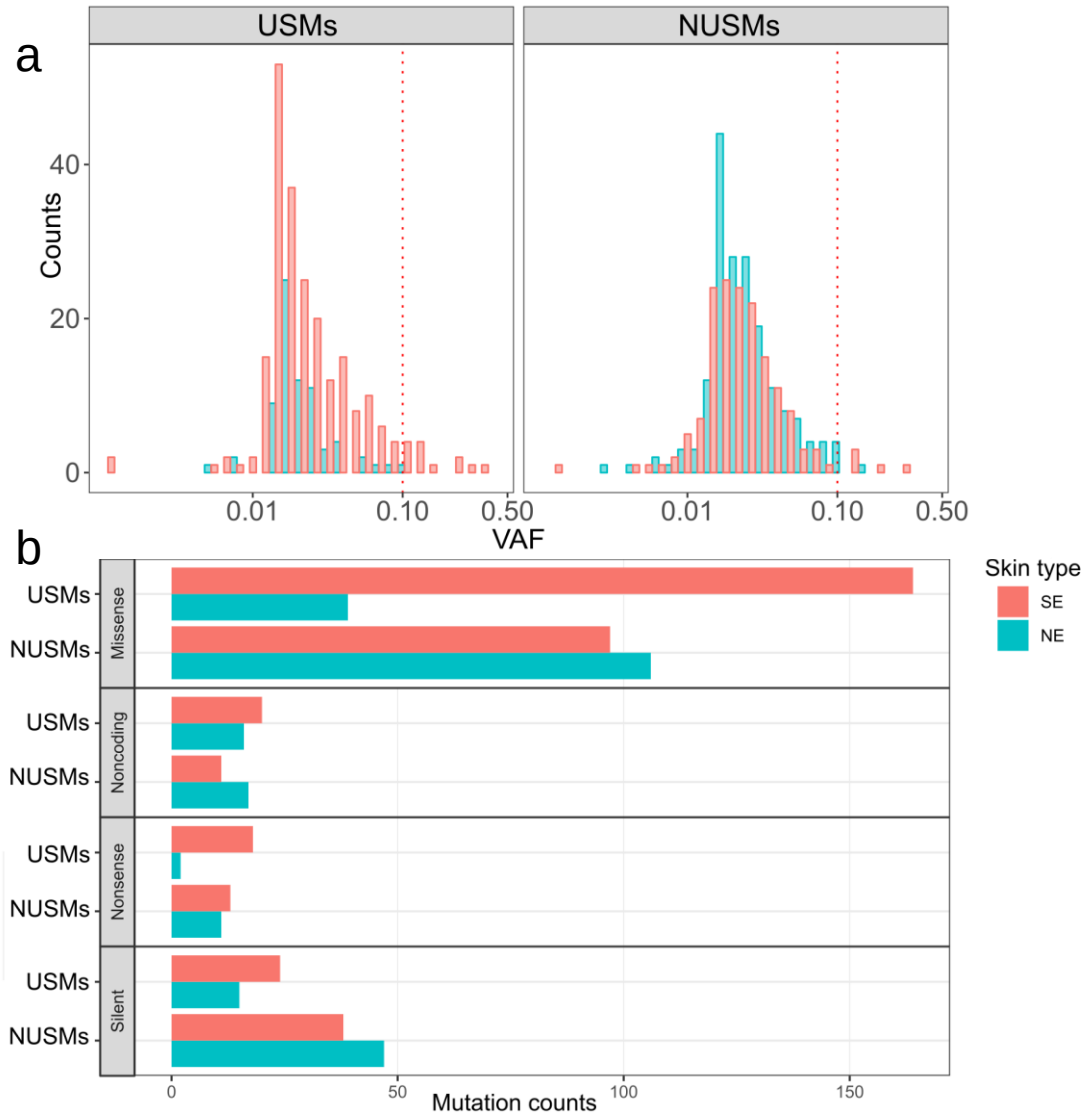
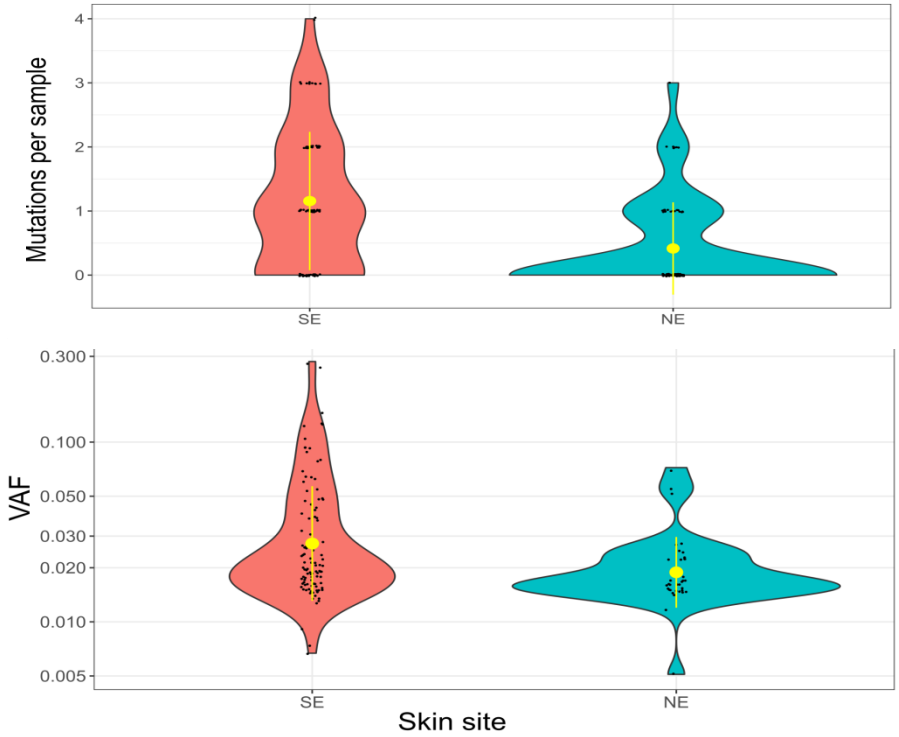


Figure 4. Quantitatively measure DNA damage using USMs

a



b

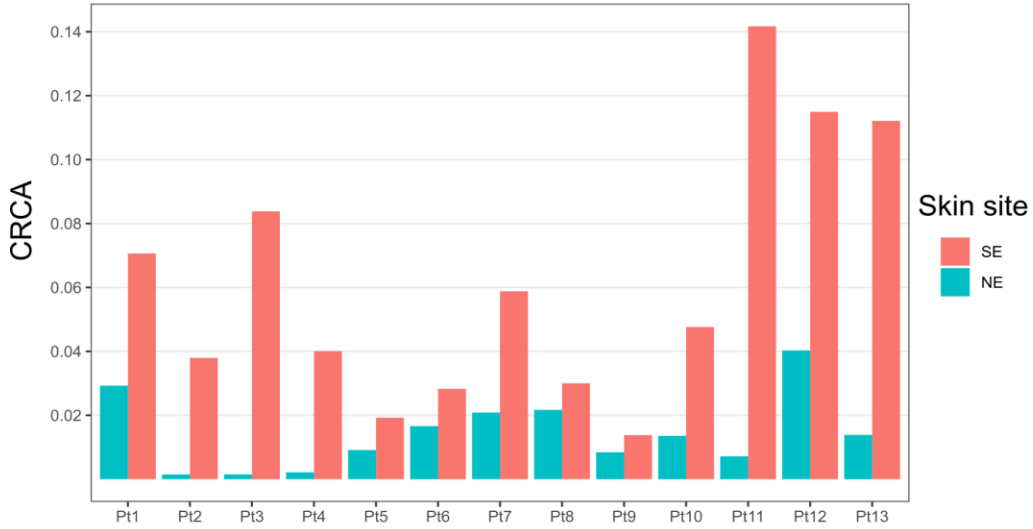


Figure 5. Optimization of punch size for detecting USMs

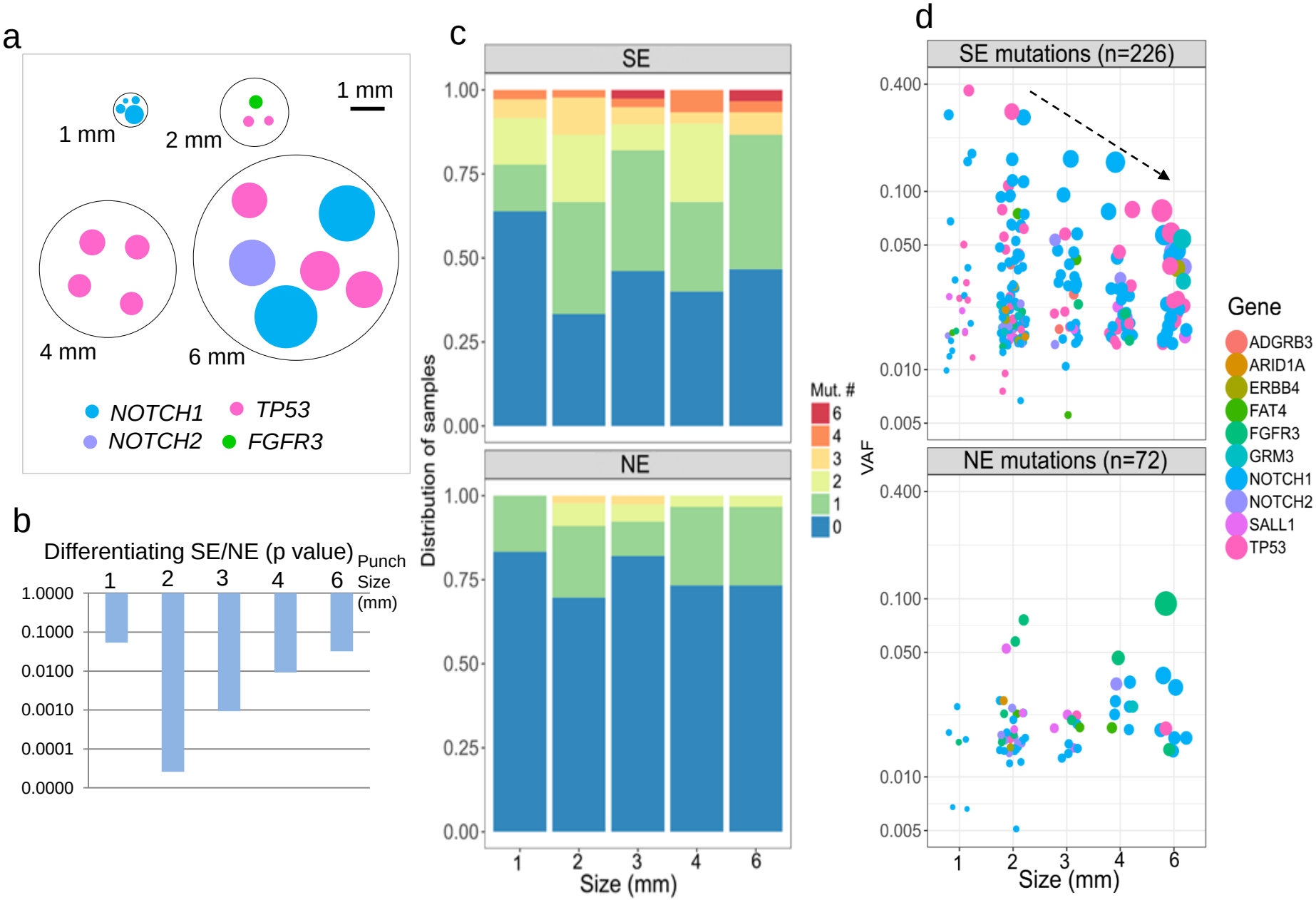
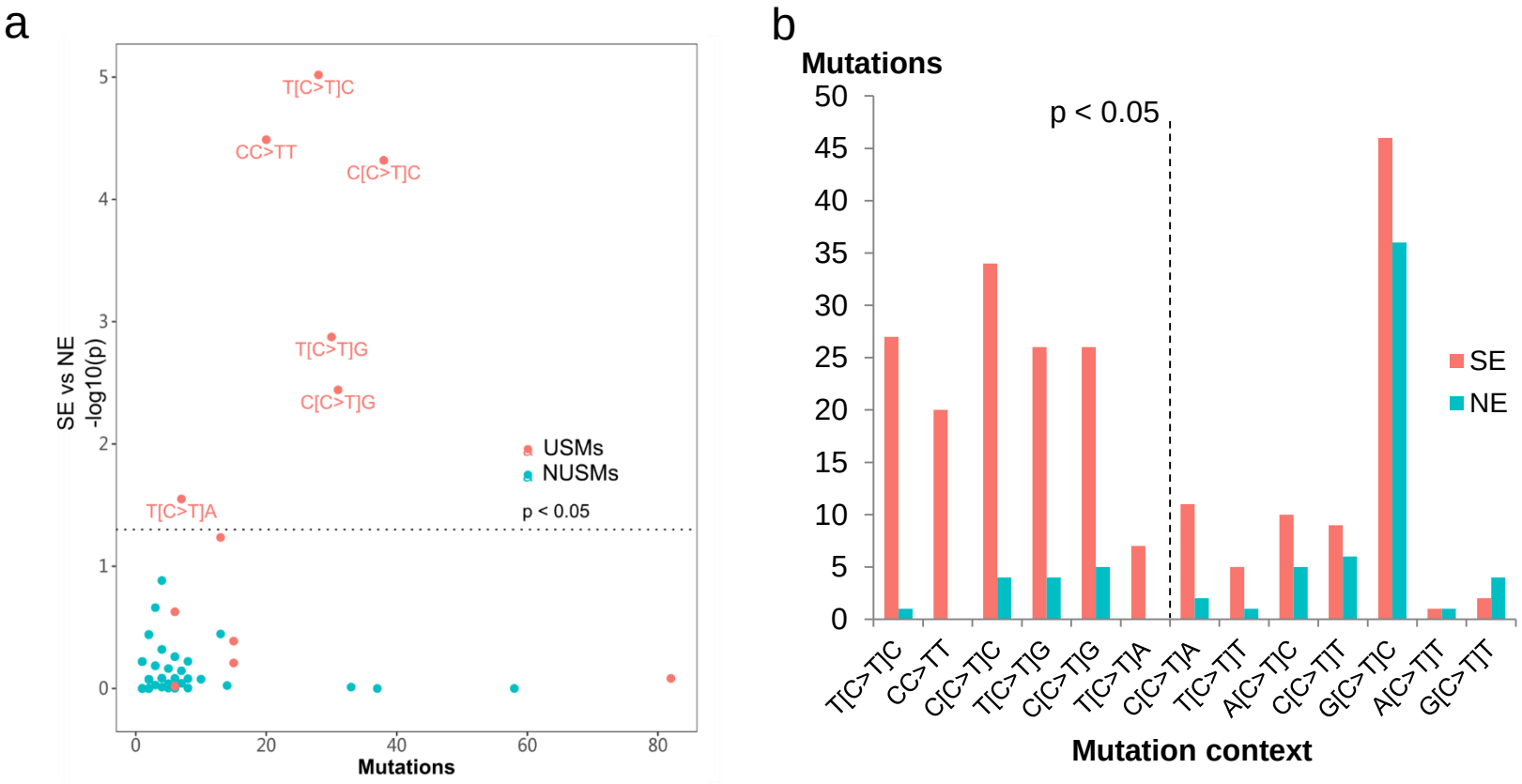



Figure 6. Mutational contexts associated with UV-exposure