PLOS GENETICS

RESEARCH ARTICLE

UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin

Natalie Saini[1¤], Camille K. Giacobone[1], Leszek J. Klimczak[2], Brian N. Papas[2], Adam B. Burkholder[2], Jian-Liang Li[2], David C. Fargo[2], Re Bai[3], Kevin Gerrish[3], Cynthia L. Innes[4], Shepherd H. Schurman[4], Dmitry A. Gordenin[1]*

1 Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, US National Institutes of Health, Research Triangle Park, North Carolina, United States of America, **2** Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, US National Institutes of Health, Research Triangle Park, North Carolina, United States of America, **3** Molecular Genomics Core Laboratory, National Institute of Environmental Health Sciences, US National Institutes of Health, Research Triangle Park, North Carolina, United States of America, **4** Clinical Research Unit, National Institute of Environmental Health Sciences, US National Institutes of Health, Research Triangle Park, North Carolina, United States of America

¤ Current address: Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston, South Carolina, United States of America
* gordenin@niehs.nih.gov

OPEN ACCESS

Citation: Saini N, Giacobone CK, Klimczak LJ, Papas BN, Burkholder AB, Li J-L, et al. (2021) UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. PLoS Genet 17(1): e1009302. https://doi.org/10.1371/journal.pgen.1009302

Editor: Mitch McVey, Tufts University, UNITED STATES

Received: August 24, 2020

Accepted: December 7, 2020

Published: January 14, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009302

Data Availability Statement: All BAM and MAF files are available under controlled access from the dbGaP database (phs001182.v2.p1 https://www.

Abstract

Human skin is continuously exposed to environmental DNA damage leading to the accumulation of somatic mutations over the lifetime of an individual. Mutagenesis in human skin cells can be also caused by endogenous DNA damage and by DNA replication errors. The contributions of these processes to the somatic mutation load in the skin of healthy humans has so far not been accurately assessed because the low numbers of mutations from current sequencing methodologies preclude the distinction between sequencing errors and true somatic genome changes. In this work, we sequenced genomes of single cell-derived clonal lineages obtained from primary skin cells of a large cohort of healthy individuals across a wide range of ages. We report here the range of mutation load and a comprehensive view of the various somatic genome changes that accumulate in skin cells. We demonstrate that UV-induced base substitutions, insertions and deletions are prominent even in sun-shielded skin. In addition, we detect accumulation of mutations due to spontaneous deamination of methylated cytosines as well as insertions and deletions characteristic of DNA replication errors in these cells. The endogenously induced somatic mutations and indels also demonstrate a linear increase with age, while UV-induced mutation load is age-independent. Finally, we show that DNA replication stalling at common fragile sites are potent sources of gross chromosomal rearrangements in human cells. Thus, somatic mutations in skin of healthy individuals reflect the interplay of environmental and endogenous factors in facilitating genome instability and carcinogenesis.

ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001182.v2.p1). All other data including the underlying numerical data for all of graphs and summary statistics are in Supplementary Tables. The R-code for analysis of the trinucleotide-specific mutation signatures can be accessed via https://github.com/NIEHS/P-MACD".

Funding: This work was supported by the US National Institute of Health Intramural Research Program Project Z1AES103266 to D.A.G. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Skin forms the first barrier against a variety of environmental toxins and DNA damaging agents. Additionally, DNA of skin cells suffer from endogenous damage and errors during replication. Altogether, these lesions cause a variety of genome changes resulting in disease including cancer. However, the accurate measurement of the range and complete spectrum of genome changes in healthy skin was missing due to technical or biological limitations of prior studies. We present here accurate measurements of the various types of somatic genome changes that we found in skin fibroblasts and melanocytes from 21 donors ranging in ages from 25 to 79 years, which allowed to distinguish age related from age independent changes. Our cohort contains both White and African American donors, allowing an estimation of the impacts of skin color on mutagenesis. As a result, we revealed the complete spectrum and determined the range of somatic genome changes and their etiologies in healthy human skin fibroblasts and melanocytes and highlighted molecular mechanisms underlying these changes. Therefore, our study introduces a base line for defining disease levels of genome instability in skin.

Introduction

Cells within the human body encounter a vast variety of DNA damaging agents throughout an individual's lifetime. By some estimates, cells may receive 70,000 DNA lesions per day [1,2]. Erroneous repair or lack of repair of these lesions would lead to a variety of genome changes including somatic single base substitutions, insertions and deletions, rearrangements and copy number changes. Large-scale sequencing studies of single cells, clonally expanded single cells and bulk cells from healthy humans have demonstrated that healthy human tissues are genetically mosaic with thousands of somatic mutations [3–11]. Analysis of such accumulated somatic genome changes have enabled elucidation of the sources of the mutation-initiating lesions as well as the various DNA repair pathways that may be involved in error-prone repair of DNA damage in human cancers [12–15]. Since at least half of the somatic genome changes seen in cancers originate in healthy pre-cancerous cells [16], it is imperative to establish the sources of DNA damage and their impacts on genome stability in healthy cancer-free tissues.

Skin is the largest tissue in the human body and forms the first line of defense against environmental toxins and DNA damaging agents, with ultraviolet (UV) radiation being the most potent environmental mutagen in skin cells. In fact, melanoma genomes have the highest burdens of mutations with UV-induced mutation signatures predominating amongst the mutation signatures identified in this cancer type [12,13]. The pathogenic impact of UV-radiation in generating genome instability is multifaceted. UV-induced DNA lesions are a source of replicative polymerase stalling [17,18] and require translesion synthesis (TLS) over cyclobutane pyrimidine dimers (CPD) and pyrimidine 6–4 pyrimidone (6-4PP) [19–27]. Error-prone TLS over UV-induced lesions leads to C➔T changes in the yCn motif (y is any pyrimidine, n is any nucleotide, mutated base is capitalized). Cytosines within a CPD may also be deaminated to uracils and upon copying by the canonical DNA polymerases or by the TLS polymerase, Pol η, would be fixed as yCn➔yTn changes or to CC➔TT changes in the next round of replication [19,21]. Error-prone TLS across thymine CPDs can also lead to T➔C changes [19,21,28,29] preferring nTt➔nCt motif [8]. Altogether, these base-substitution motifs derived from experimental data constitute a significant part of mutation signature SBS7b extracted by non-negative matrix factorization analysis from mutation catalogs of thousands of whole-genome sequenced human cancers [13]. In the absence of TLS across UV-induced lesion it would not result in base substitutions but can lead to impediment of replication fork progression. Restart

of a stalled replication fork can result in the formation of single-stranded gaps in the sister DNA molecules and later convert to double strand breaks (DSBs) [30,31]. Inaccurate repair of such a DSB via homologous recombination (HR) or non-homologous end joining (NHEJ) can lead to a structural changes, copy number variation or generate a small insertion or a deletion.

In agreement with UV radiation being the major source of DNA damage in skin cells, various studies have demonstrated that C➔T changes in the yCn context is the most prevalent base substitution in skin fibroblasts, melanocytes, and keratinocytes. In addition, human skin cells also carry CC➔TT changes and T➔C in the nTt motifs [7,8,32,33]. Moreover, we previously demonstrated that fibroblasts obtained from sun-exposed body sites carry a higher mutation burden along with a higher contribution of a UV-mutation signature than fibroblasts obtained from sun-shielded sites [8]. Our findings were also supported in the study by Tang *et. al*. wherein they demonstrated higher mutation burden in melanocytes from sun-exposed body sites than sun-shielded body sites via either whole exome sequencing or targeted sequencing of 509 cancer-associated genes in single melanocytes [33]. In summary, numerous studies have established and verified the prominent mutagenic effects of the bypass of UV-induced lesions by translesion polymerases generating a characteristic base substitution signature in skin cells. However, the broad spectrum of somatic genome instability, including consequences of UV-induced DSBs in cells of healthy human skin have neither been established nor characterized by mutation signature analysis.

In addition to environmental DNA damage, cells may also accumulate somatic genome changes due to endogenous DNA damage or errors during DNA replication in the form of base substitutions, small insertions or deletions (indels), and gross chromosomal rearrangements. Somatic mutations in skin cells have been measured either by deep sequencing of bulk tissue [32] or whole-genome sequencing of single cell-derived induced pluripotent stem cells [5] or single cell-derived clonal lineages [7,8]. However, due to either small sample sizes or difficulties in accurately identifying somatic indels and chromosomal rearrangements using induced pluripotent stem cells or bulk cells, none of these studies have been able to adequately characterize the different sources of DNA damage and their mutagenic outcomes in skin cells from healthy donors [34].

Here, we present an integrated analysis of the various types of somatic genome changes that are found in skin fibroblasts and melanocytes from a total of 21 donors ranging in ages from 25 to 79 years. Unlike previous studies, our cohort contains both White and African American donors, allowing a better estimation of the impacts of skin color on mutagenesis in skin cells. Our work provides the normal range of the burden and types of somatic genome instability in human skin cells. We show here that in skin cells, endogenous DNA damage in the form of spontaneously deaminated cytosines at CpG motifs, oxidative DNA damage, as well as DNA replication errors, are a substantial source of somatic mutagenesis. Additionally, UV-induced DNA damage is prevalent even in sun-shielded skin cells and manifests as single base substitutions arising from DNA synthesis over lesions by TLS and by deletions of five or more nucleotides arising from end-joining repair of UV-induced DSBs. Our analysis also highlights the differences in the outcomes of UV-induced DSBs and DSBs induced by endogenous DNA damage in cancer-free skin cells. Overall, we provide a comprehensive analysis of the various UV-induced and endogenous genome de-stabilizing processes that operate in healthy skin cells.

Results

Study design

Based on our prior study [8], we performed whole-genome sequencing of hip skin cells as it would allow detection of versatile mutational processes, because it is mostly sun-shielded. UV-

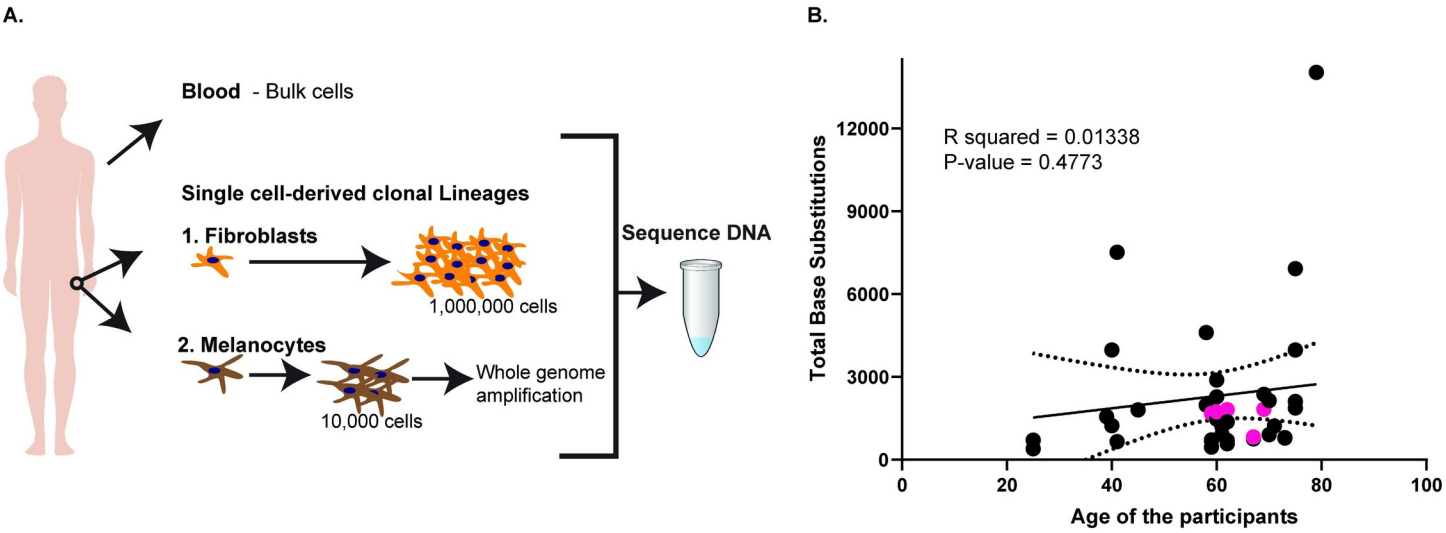

Fig 1. Schematics and total base substitutions identified per clonal lineage in this study. (A) Schematics of the study design. From each donor, we obtained blood for whole-genome sequencing. In addition, we obtained skin biopsies from the hips of the donors from which fibroblasts and melanocyte clonal lineages were obtained. Fibroblasts were grown up to a million cells and their DNA was directly used for whole-genome sequencing, while melanocytes grew up to 10,000 cells and the DNA was whole-genome amplified and thereafter sequenced. (B) The total base substitutions in each clonal lineage versus the age of the donors. The pink filled circles denote melanocyte clones. The x-axis denotes the ages of the donors, while the y-axis denotes the number of base substitutions. The solid black line is the linear regression line for the samples, while the dotted black curves are the 95% confidence intervals. The source data for this figure is in S1 and S2 Tables.

https://doi.org/10.1371/journal.pgen.1009302.g001

induced mutagenesis has lower contribution into overall mutation spectrum in this tissue which allows better detection of other types of mutagenesis. In this study, we analyzed somatic genome changes in single cell lineages from 34 fibroblasts obtained from skin biopsies taken from hips of a total of 21 donors, ages 25 to 79. Our dataset includes the hip fibroblasts from two donors sequenced previously [8]. In addition, we sequenced five genomes of clonal single melanocyte-derived lineages. Single skin fibroblasts were propagated in culture up to approximately 1,000,000 cells which provided sufficient high-quality genomic DNA for whole-genome sequencing and follow-up validation. The clonal single-melanocyte lineages were cultured in media up to 10,000 cells. DNA from these cells was whole-genome amplified and sequenced. In addition, we were able to grow one melanocyte clone up to 1,000,000 cells and performed whole-genome sequencing on this sample without whole-genome amplification. From each donor, we also sequenced whole blood DNA (Fig 1A).

The median sequencing depth for the samples was 78X with a minimum average coverage per site of 50X (S1 Table). The genome-wide changes detected in the clones were compared to blood samples from the same donors and only the variants unique to the clones were denoted as somatic changes in the clones. Stringent filtering criteria were applied to exclude changes that could have occurred during limited propagation of the clone. For this purpose, only base substitutions as well as indels calls within 45% and 55% (heterozygous alleles) or above 90% (homozygous alleles) allele frequencies were considered clonal and somatic in the initiating melanocyte or fibroblast cell. All other calls that did not conform to these allele frequencies were considered sub-clonal and were removed from the analysis as these most likely represented culture-induced artifacts. We also analyzed the allele frequencies of all somatic base substitutions in the clonal lineages. All fibroblast clonal lineages demonstrated a peak of mutation calls at the 45% to 55% allele frequencies, indicating that these samples were clonal (S1 Fig and S2 Table). We did not see such a peak in the whole-genome amplified melanocyte clones which could reflect uneven genome amplification and localized genome duplications during the whole-genome amplification step. Nonetheless, only analyzing heterozygous mutation

calls within the 45% and 55% allele frequencies and homozygous mutation calls at >90% allele frequencies allows us to estimate the minimum number of somatic mutations in the founder cells. Mutations that accrue during culture and/or polymerase errors during whole genome amplification are expected to not be clonal and have allele frequencies <45%. For structural changes, clonal calls with variant junction reads representing at least 30% of the total junction reads, and no reads representing the variants in the blood genomes were identified as clonal somatic rearrangements present in the initiating fibroblast. Multiple samples sequenced from the same donors allowed intra-individual comparisons of somatic genome instability in humans.

UV-induced base substitutions and C➔T changes due to spontaneous cytosine deamination are prevalent in skin cells

We detected 402 to 14029 base substitutions in each clonal lineage sequenced and the mutations did not increase with the age of the donors (Fig 1B). Analysis of the mutation spectrum revealed that the predominant mutation in many samples was C➔T base substitution (S2 Fig and S2 Table). The number and types of base substitutions in melanocyte clones were similar to those seen in the fibroblasts (Figs 1B and S2 and S2–S4 Tables). To identify the predominant mutation signatures in our samples, we determined the cosine similarities of the 96 tri-nucleotide motif mutation profiles in our samples versus all the published mutation signatures derived from analysis of thousands of mutation catalogs from human tumors ([13], and https://cancer.sanger.ac.uk/cosmic/signatures). This allowed us to agnostically determine the mutation signatures previously identified in cancers that were also overrepresented in our samples. We saw that the SBS7b signature was overrepresented in many samples. These signatures comprise of C➔T changes at cC or tC motifs (S2 and S3 Figs). In addition, SBS1 was only weakly represented in our samples. We also found mutation signatures SBS2 and SBS11 present in the samples which also carry a strong SBS7b mutation signature. These are likely due to the overlap between the SBS2 and SBS11 mutation signatures with UV-induced mutations (S3 Table). We also used non-negative matrix factorization (NMF)-based deconvolution of mutation signatures as a parallel approach to agnostically determine the predominant signatures in our samples for single base substitutions and dinucleotide substitutions. We were able to detect SBS1, SBS5, and prominent UV mutation signatures (SBS7b and DBS1) in our cohort (S3 Table, S2 and S3 Figs). In 28 samples we also detected either SBS4 or SBS18 which are indicative of oxidative damage in the cells leading to G➔T (C➔A) changes (S2 and S3 Figs and S3 Table).

We then sought to determine if the most prominent components of mutational signatures identified above were statistically enriched in our samples. For this purpose, we used our previously described knowledge-based trinucleotide-motif-centered pipeline [8,14,15,35]. This pipeline calculates enrichment with mutations within pre-defined trinucleotide motifs. It also calculates the sample-specific p-values for enrichments and minimum estimate of mutation load assigned to a motif-specific mutagenic process after stringent statistical filtering. nCg➔nTg changes likely arise upon spontaneous deamination of methylated cytosines [15]. These mutations constitute the major component of SBS1 in COSMIC [12,13]. SBS1-associated mutation load has been shown to increase with age in cancers [36] and in healthy individuals [7,10,37,38]. Analysis of the nCg➔nTg changes in our donors demonstrated that this mutation type is statistically enriched in all the samples and was also found to linearly increase with the ages of the participants with an average increase of 0.4 mutations per year (Fig 2 and S4 Table). We also detected statistically significant enrichment with UV-associated C➔T changes in the tC or cC context (yCn➔yTn, major component of COSMIC SBS7b) in many of

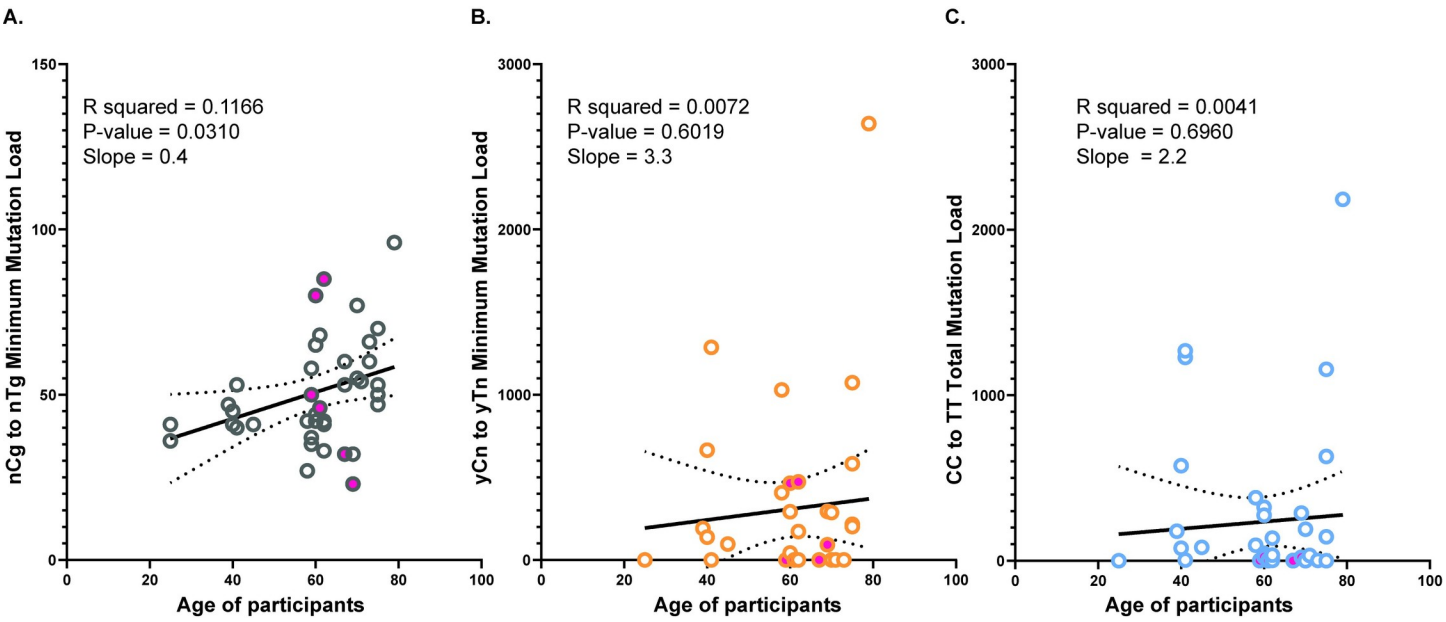

Fig 2. Analysis of the motif-specific mutation signatures in the genomes of skin cells. The minimum mutation load for the (A) nCg➔nTg mutation signature, (B) yCn➔yTn mutation signature and (C) the total mutation load for the CC➔TT dinucleotide changes are plotted against the ages of the participants. The solid pink circles denote the mutation load in melanocytes. The black solid line is the linear regression, and the dotted curves are the 95% confidence intervals for each dataset. The source data for this figure is S4 Table.

https://doi.org/10.1371/journal.pgen.1009302.g002

the sequenced samples as well as a prominent presence of CC➔TT changes (Fig 2 and S4 Table). The estimates of yCn➔yTn minimum mutation load correlate with direct counts of the less frequent CC➔TT and nTt➔nCt changes, which have previously been shown to be associated with UV-induced DNA damage in human cells [8,13] (S4 Fig). Altogether these three types of changes indicate the contribution of UV-induced changes in mutation load accumulated in skin. Interestingly, the UV-induced mutations did not correlate with the ages of the participants (Fig 2).

SBS1 mutations identified by SigProfilerExtractor (NMF-based deconvolution) and the minimum number of nCg➔nTg mutations analyzed by the knowledge-based pipeline correlate with each other. We saw a similar correlation between the mutations attributed to SBS7b by SigProfilerExtractor and the minimum number of yCn➔yTn mutations identified by the knowledge-based pipeline (S5 Fig). These data indicate that both methodologies perform similarly in the evaluation of mutation signatures.

Bulk exome sequencing reveals the presence of cancer drivers in the samples at less than 10% allele frequencies

Whole-exome sequencing up to 100X to 150X was also performed for bulk samples from which 14 fibroblast clones and four melanocyte clones, respectively (S1 Table). Analysis of single nucleotide variants (SNVs) in the bulk samples and comparisons with the clonal lineages derived from the bulk cells revealed the presence of overlapping SNVs in bulk and the corresponding clones. Interestingly, we did not see any overlapping SNVs between bulk samples and clonal lineages that were not derived from the same bulk sample, even if they were coming from the same donor (S5 Table). This observation validates our mutation calling pipeline and provides support for the presence of the mutations detected in the clonal lineages in the original skin biopsies.

The somatic mutations identified in the bulk samples were predominantly at or less than 10% allele frequencies (S5 Table and S6 Fig). This observation was also found to hold true for the allele frequencies of the overlapping SNVs in bulk samples and their corresponding clonal lineages (S7 Fig). The low allele frequency in the population demonstrates the large amount of heterogeneity in the dermal and the epidermal tissue.

We also annotated all the SNVs in all whole-genome sequenced clones and whole-exome sequenced bulk tissues for functional effects. All non-synonymous SNVs, stop gains, start or stop loss SNVs in the clones and bulk tissues were further analyzed using the Cancer Genome Interpreter [39] to determine if these were potentially cancer drivers. Of the 672 SNVs in the clones that had potentially functional impacts, 32 changes were in tumor driver genes and 13 changes were annotated as tumor drivers. One sample, DAG_H95, was found to have 3 tumor driver mutations, however the donor does not have any history of cancer (S2 Table). We also detected 3190 SNVs in the bulk tissues that could alter protein sequence, of which 390 were within tumor driver genes, and 62 of the mutations were annotated as driver mutations. Interestingly, these driver mutations were also present between 2 to 10% allele frequencies in the samples (S7 Fig). Overall, the results suggest that normal sun-shielded human skin carries a substantial proportion of cancer driver mutations, albeit at low allele frequencies.

Single base indels in homonucleotide repeats and deletions larger than 5 bases are ubiquitous in skin cells

We detected from 7 to 71 indels in the donors (Fig 3A and S6 Table). The insertions ranged from 1 base to 40 bases and deletions ranged from 1 to 171 bases (Fig 3B and S6 Table). The total number of indels per sample do not appear to increase statistically with the ages of the donors (Fig 3A). NMF-based deconvolution analysis of indel signatures or measuring the cosine similarities of indel patterns with the indel signatures currently annotated in cancers [13] demonstrated that two indel types were prevalent in our samples. The first was single base insertions or deletions in homopolymeric stretches associated in our samples with ID1, ID2 and ID7 (S8 Fig and S7 and S8 Tables). Since many samples had very low numbers of indels, it is possible that mathematical deconvolution of indel signatures may carry errors. Therefore, instead of the number of mutations within each signature, we used the total number of single base insertions or deletions in homopolymeric stretches for further downstream analyses. These types of indels were also found to increase linearly with the ages (0.22 mutations per year) of the donors consistent with the idea that they were associated with polymerase slippage at the homopolymeric repeats [40,41] during ongoing DNA replication in fibroblasts over the donors' lifetime (Fig 3C). The second class of indels were deletions spanning five nucleotides or more, many of which have microhomology of one or more bases at the deletion junction (S8 Fig and S8 Table). Based on cosine similarities, these indels were highly similar to ID8 (S8 Table), the indel signature associated with double-strand break repair via non-homologous end joining [13]. Consistently, NMF-based deconvolution of indel signatures applied to our samples identified these deletions of five or more bases as part of novel signature which is a composite signature made up of ID8-like indels as well as indels in homopolymeric repeats (Signature A in S8 Fig and S8 Table). Such deletions spanning five or more nucleotides were identified in almost all samples and did not demonstrate a statistically significant increase with the ages of the donors. We also did not see any differences between the indel load or signatures in melanocytes versus the fibroblasts indicating that the processes yielding indels in both cell types are likely the same (Fig 3C).

The number of deletions spanning five or more nucleotides were found to also correlate in our samples with the UV-associated trinucleotide-centered yCn to yTn mutation signature.

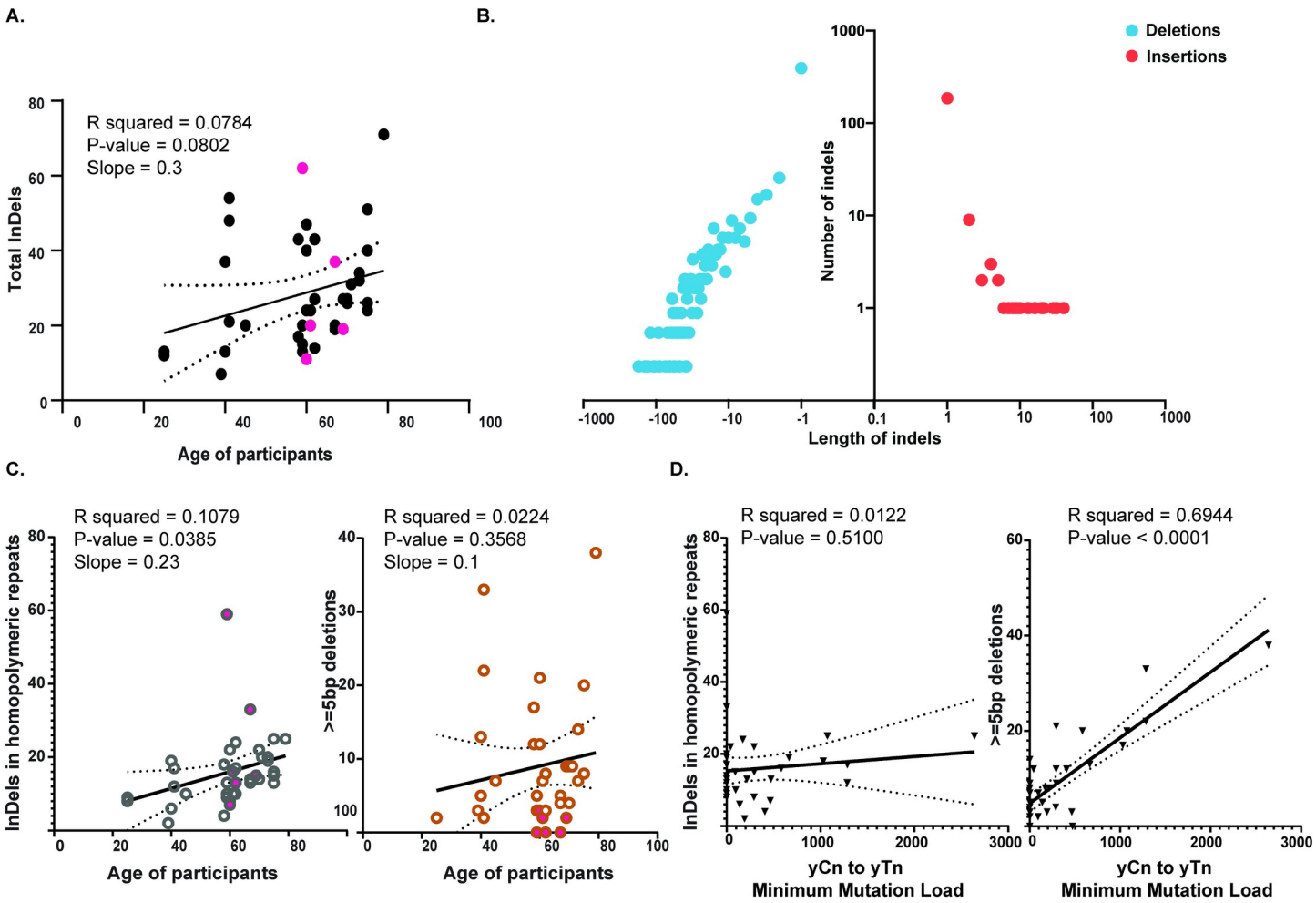

Fig 3. Analyses of indels in skin cells. (A) The total number of indels identified in each sample plotted against the ages of the donors. The black dots denote fibroblast clones, while the pink dots denote the melanocyte clones. (B) The distribution of the lengths of the insertions and deletions detected in the clonal lineages. The source data for panels A and B is in S6 Table. (C) The number of insertions and deletions in homopolymeric repeats and the deletions spanning 5 bases or more plotted against the ages of the donors. The open circles denote fibroblast clones, while the filled in circles denote melanocyte clones. The source data for this figure are in S7 Table. (D) The number of insertions and deletions in homopolymeric repeats and the deletions spanning five bases or more plotted against the yCn→yTn minimum mutation load in skin cells. The source data for this panel are in S4 and S6 Tables. In the graphs, the black solid line is the linear regression of the data, and the dotted black curves are the 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1009302.g003

We did not see a positive correlation between the UV-associated mutation signatures and single base indels in homopolymeric repeats (Fig 3D). These data indicate that unlike indels at homopolymeric repeats, UV-induced DNA double strand breaks are the underlying etiology for deletions of five or more bases in human skin cells.

The majority of the insertions in skin cells are templated

The predominant insertions detected in our clonal lineages were templated single-base insertions (i.e. copied from the neighboring bases). Of the 186 single base insertions, 163 of the insertions were copied from the adjacent base (S9 Table). Such insertions most likely represent polymerase slippage events within homopolymeric runs of bases or erroneous Okazaki fragment maturation [40,41] and constitute the ID1, ID2 and ID3 indel signatures as mentioned above [13].

We also detected 28 instances of insertions larger than two bases in length. 18 of these larger insertions were a duplication of the neighboring residues. Three of these templated insertions carried small mismatches likely due to errors during copying of the neighboring residues (S9 Table). Such templated insertions along with deletions spanning five bases or more have been shown to be characteristic of non-homologous or microhomology-mediated end-joining of double-strand breaks [13,42,43]. As such, it is likely that repair of UV-induced double strand breaks also leads to insertions of two bases and more. However, the low numbers of such events do not allow statistical verification of this hypothesis.

UV-induced mutation load varies by race and is not impacted by the sex of the donors

Our cohort included five African American or Black donors and 16 White donors, thus allowing us to also determine if the accumulation of somatic genome changes is different between the two races. The total base substitutions in samples from the White donors (median 1824), were higher than in the skin fibroblasts and melanocytes obtained from African American or Black donors (median 715, p-value = 0.00002193, calculated by two-tailed Mann Whitney test). We reasoned that this lower mutation load in the African American donors might reflect the protective effect of melanin in skin. Consistently, we did see a prominent presence of UV-associated yCn➔yTn changes in skin cells from White donors (median for White donors = 209). However, we did not detect statistically significant enrichment with this mutation type in skin cells from Black donors (minimum estimate of mutation load = 0, Fig 4A and S10 Table). The number of nCg➔nTg mutations, which are not associated with UV-lesions did not vary across the two categories of donors (median for White donors = 48, median for African American donors = 40). In addition, although we did not see any difference in the total number of indels or the number of indels in homopolymeric repeats in skin cells obtained from donors of either race, we found increased numbers of deletions spanning five bases or more in skin cells obtained from White donors (median = 9) as compared to the skin cells obtained from African American or Black donors (median = 2, P-value = 0.0002781, calculated by two tailed Mann Whitney test) (Fig 4B and S10 Table). In order to avoid skewing of the data due to differences in sequencing methodologies used for melanocytes and fibroblasts in our samples, we also calculated P-values for each of the cohorts using a two tailed Mann-Whitney test after excluding the data from the melanocytes. Even in this data set, we were clearly able to detect an increase in UV-induced mutations in White donors as compared to African American or Black donors (S10 Table). Overall, our data can be explained by melanin in skin providing strong protection against UV-associated somatic mutations in the form of both UV-signature base substitutions as well as deletions of five bases or more.

Our cohort also consists of eight men and 13 women. Analysis of mutation load based on sex did not demonstrate any differences between the skin cells obtained from the men or the women (S9 Fig and S10 Table).

Structural variant hotspots colocalize with common fragile sites

Structural variants were only analyzed for the fibroblast clonal lineages and the single melanocyte clonally grown lineage for which we were able to obtain sufficient cells for WGS without whole-genome amplification, since the genome amplification process can result in many false rearrangement calls. There were 120 structural variants in the 35 sequenced clonal lineages (from 1 to 14 in each isolate) (Fig 5A and S11 Table). The structural variants included deletions, duplications, inversions ranging in size from 225 bp to 39 Mbp, as well as translocations. No age-dependent increase in structural variants was evident.

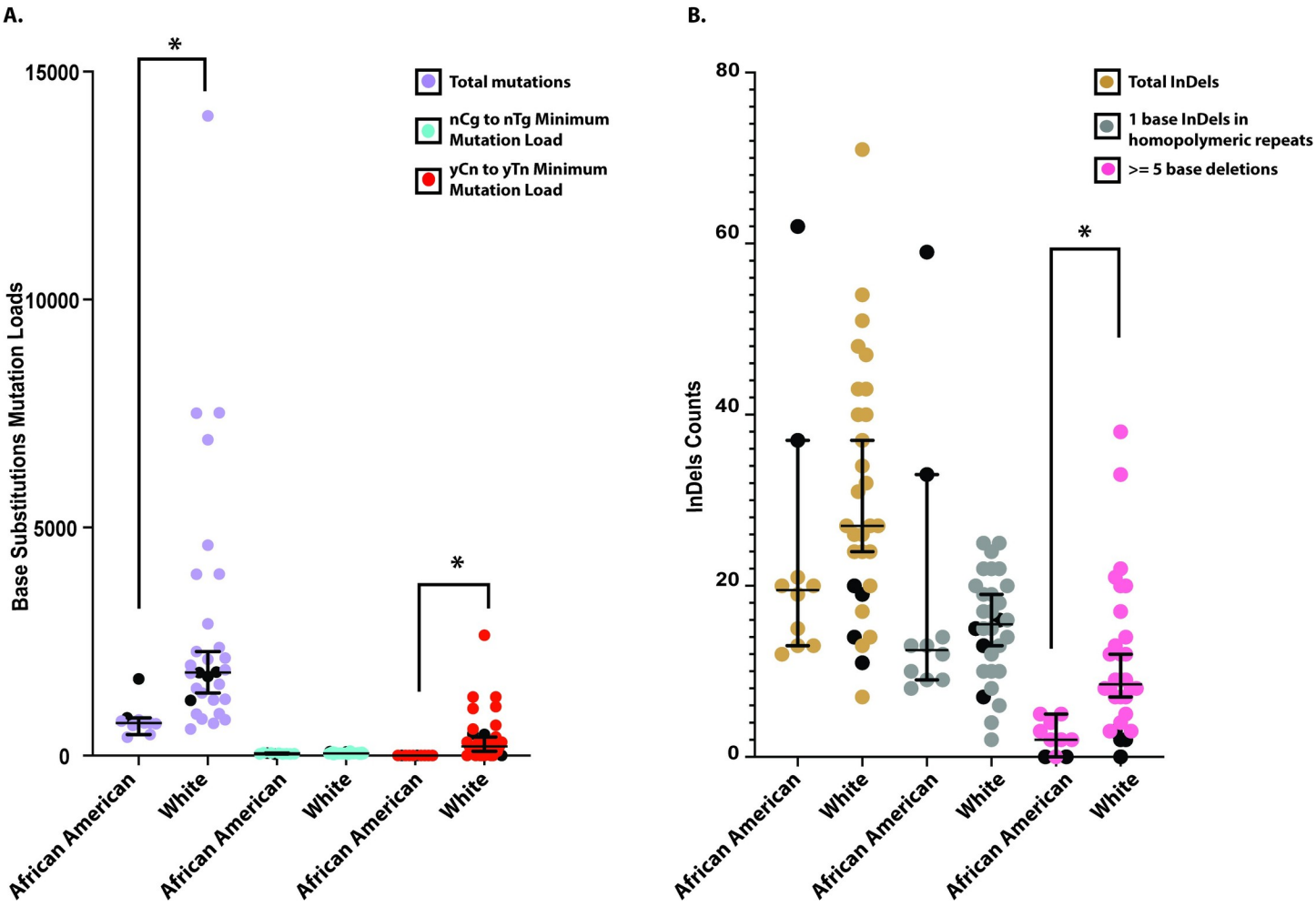

Fig 4. Base substitutions and indels in African American and White donors. (A) The total number of base substitutions, the nCg➔nTg minimum mutation load and the yCn➔yTn minimum mutation load in African American and White donors. Melanocytes are depicted as filled black circles. (B) The total number of indels, single nucleotide indels in homopolymeric repeats and deletions spanning five bases or more in the African American and White donors. Melanocytes are depicted as filled black circles. A two-sided Mann-Whitney U-test was performed to compare the mutation load across the two cohorts. * denotes a Bonferroni corrected P-value < 0.05. The source data for this figure is in S10 Table.

https://doi.org/10.1371/journal.pgen.1009302.g004

We identified genomic regions that are hotspots for chromosomal breakage and structural variation. Two or more rearrangements were denoted as part of a hotspot if they were less than or equal to 1Mbp apart and were present in different samples. Of the 120 structural variants identified, 55 rearrangements were within hotspots. Previously, we showed that structural variants identified in skin fibroblasts of two donors were often in the vicinities of common fragile sites (CFSs) [8]. To determine if the structural variants in this larger data set also often colocalize with CFSs, we identified those deletions, duplications and inversions that intersect common fragile sites within the HumCFS database [44]. We also identified those translocations as colocalizing with fragile sites, whose breakpoints were within 10kb of a CFS. 63 rearrangements were found to colocalize with CFSs. 18 of the rearrangements within CFSs were on chromosome 7, of which 14 rearrangements were within FRA7J, implying that this fragile site is expressed more prominently in fibroblasts than the other fragile sites, leading to higher levels of replication stalling and gross chromosomal breakage. Moreover, the majority of the rearrangements within hotspots also colocalized with CFSs, while the majority of

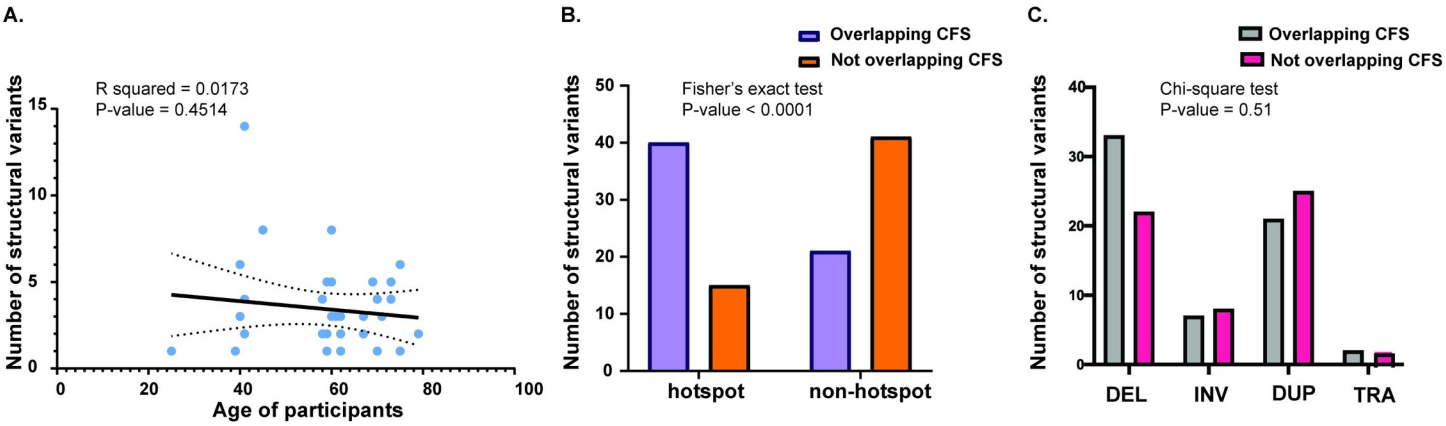

Fig 5. The structural variants identified in the genomes of skin cells. (A) The number of structural variants in each donor plotted against the ages of the donors. The black inclined line denotes the linear regression of the data, while the dotted curves denote the 95% confidence intervals. (B) The number of structural variants that were or were not within hotspots and common fragile sites. A Fisher's exact test was performed to determine if structural variants in hotspots were also preferentially present within common fragile sites. (C) The types of structural variants that overlap and do not overlap common fragile sites. A Chi-square test was performed to determine if the structural variant types within common fragile sites were different from those that did not overlap common fragile sites. The source data for this figure are in S11 Table.

https://doi.org/10.1371/journal.pgen.1009302.g005

rearrangements that were not within hotspots were scattered across the genome (Fig 5B and S11 Table). Interestingly, we did not see any difference in the types of structural variants that overlap and those that did not overlap CFSs (Fig 5C). Moreover, we determined the use of microhomology at the breakpoints to identify a role of microhomology mediated repair of DSBs at CFSs. Of the 120 rearrangements, only 15 rearrangements contained microhomology at the breakpoints (6 overlapping CFSs, and 9 not overlapping CFSs). These regions of microhomology were small and ranged from 2 to 3 bases. A Fisher's exact test demonstrated no significant bias in the use of microhomology between the variants that overlapped CFSs versus those that do not (P-value 0.41) (S11 Table).

Overall, we hypothesize that replication-associated difficulties at CFSs are responsible for the generation of rearrangement hotspots in healthy human cells.

Discussion

In this study, we revealed and accurately measured load of the major types of somatic genome changes in human skin. We grew single cell-clonal lineages derived from human skin fibroblasts and melanocytes. Whole genome sequencing from these samples allows us to detect somatic genome changes that are present in the original single cells with high accuracy. Moreover, the methodology provides sufficient DNA for orthogonal validation of the changes, allowing us to apply the most stringent criteria for identifying different kinds of genome changes without losing sensitivity.

Our work provides the range of normal somatic genome changes in human skin cells in donors across a wide range of ages and of different races. We demonstrate each skin cell carries from 402 to 14029 base substitutions, 7 to 71 indels and 1 to 14 structural variants per cell. The mutation burden in healthy skin cells was also similar to the median mutation load in cancers [45]. Interestingly, we identified various cancer driver mutations in the clones as well as in the bulk tissue samples, although these driver mutations were present at low allele frequencies in the bulk samples. This observation echoes previous findings where normal tissue often contains cells with driver mutations [11,32,37,38,46,47].

Analysis of mutation signatures in the clonal lineages allowed to differentiate between endogenous DNA damage-induced mutations, replication-associated errors as well as

environmental DNA damage-induced genome changes. C➔T changes at CpG motifs as well as single nucleotide insertions and deletions were found to increase with the ages of the donors and were indicative of endogenous mutational processes and replication errors, respectively. In addition, UV-induced base substitution signatures were prominent in many samples even though they were obtained from sun-shielded skin. UV-induced DNA damage can also lead to the formation of double strand breaks in the genome. We showed here that deletions spanning 5 or more nucleotides with or without microhomologies at the junctions strongly correlated with the UV-induced base substitution signature. Previously, this indel signature (ID8) has been identified in a wide variety of cancers, and likely represents repair of double strand breaks via non-homologous end joining (NHEJ) pathways [13]. As such, we hypothesize that ID8-like indels are characteristic of UV damage in human cells. Since we also detect deletions with limited microhomologies at the junctions, it is possible that in addition to canonical NHEJ, microhomology mediated end joining (MMEJ) or polymerase theta-mediated end joining (TMEJ) [48,49] may also participate in the repair of UV-induced DSBs in skin cells. In addition to deletions, we also detected a few instances of long insertions, often templated from the flanking sequences in our samples. Such locally templated insertions are highly characteristic of TMEJ and are likely formed by the Polθ-dependent synthesis wherein one resected DSB end uses the second resected DSB end for synthesis [42]. Overall, our data indicates that ID8-like indels along with a small number of templated insertion events accumulate in skin cells, due to UV-induced DNA damage and error-prone repair via NHEJ or TMEJ.

Interestingly, although non-UV mutations (nCg➔nTg and indels at homopolymeric repeats) increased with the ages of the participants, we did not see a similar correlation between UV-exposure-induced mutations and ages of the donors. Since we are measuring mutation load in sun-shielded skin cells, as such, even intermittent UV-exposure due to clothing and lifestyle choices of the participants during their lifetimes are likely to lead to the formation of UV-induced DNA damage and impact the lifetime accumulation of UV-induced mutation load in hip-derived cells. Thus, the absence of a correlation between age of the donors and the UV-induced mutation load might be due to differences in overall accumulation of UV-exposure across the lifetime of different donors.

DNA double strand breaks can be channeled into repair via two major pathways, HR or NHEJ. One major factor that determines the choice of the repair pathway in cells is the cell cycle stage. Cells in the S or early G2 phases of the cell cycle predominantly repair DSBs via HR, while NHEJ events peak in the G1 or late G2 phases [50,51]. Since HR is mostly error-free, we would not be able to detect HR activity that may have occurred in skin cells. Nonetheless, the prominent presence of UV-associated NHEJ or TMEJ-generated indels in human skin further indicates that the majority of UV-associated damage and mutagenesis accrues in quiescent non dividing cells.

We also demonstrated here that UV-associated mutation load is decreased in skin cells from African American donors as compared to White donors. We hypothesize that this effect may be due to the protective effects of melanin on UV-induced DNA damage. In agreement with the decreased mutation load, are the lower rates of skin cancer in African Americans. While skin cancer accounts for up to 35–45% of all cancers in Caucasians [52], it only accounts for 1–2% of the neoplasms in African Americans [53–55]. Moreover, the impact of UV-exposure as a risk factor for skin cancers is decreased in African Americans as compared to Caucasians [55,56]. These observations imply that lowered mutation burden due to UV-radiation is indicative of the lower risk of UV-induced skin cancer in African Americans.

In addition to small indels, we also detected large structural variant hotspots in our samples that often coincided with CFSs. For example, rearrangements in FRA7J were found in 14 different donors and was the most common hotspot in our samples (S11 Table). Recurrent

breakage at this fragile site has been implicated in the Williams-Beuren syndrome and and this region contains the genes LIMK1, EIF4H(WBSCR1), AUTS2 as well as the tumor suppressor gene FZD9 [57,58]. One explanation for the large number of rearrangements found at a single fragile site is that the genes within this fragile site are preferentially expressed in fibroblasts that may cause transcription-replication collisions often leading to breakage and rearrangements. Alternatively, the replication timing within the fragile locus may be delayed leading to unfinished replication and fragility. Tissue-specific expression and alteration in replication timing at fragile sites has been observed previously in cultured cells [59–61]. As such, we surmise that fibroblast-specific replication-associated difficulties at common fragile sites lead to the formation of rearrangement hotspots in normal skin.

Overall, our work provides an accurate and comprehensive catalog of the somatic genome changes attributable to different DNA damaging processes that act upon human skin cells over the lifetime of the individuals. Our analysis uniquely identifies and measures the impacts of endogenously operating DNA damage, DNA replication errors as well as environmental DNA damage on the somatic mutation load and profiles in each single cell-derived lineage. Finally, we provide the reference for the burden, types and etiologies underlying somatic genome instability in cells of healthy human skin which is required for defining disease level of somatic genome instability.

Materials and methods

Ethics statement

Written consent was obtained from all participants in the Environmental Polymorphisms Registry (registered with ClinicalTrials.gov, NCT00341237, and approved by the NIH Institutional Review Board, protocol 04-E-0053). Each participant provided their age, sex and self-identified race.

Sample collection and processing

4 mm punch skin biopsies were collected from donors' hips. Samples were collected from healthy cancer-free skin. After overnight incubation of the biopsies at 4˚C in 2.66 units/ml dispase (Roche) and 50μg/ml gentamycin (Sigma Aldrich), the epidermis and dermis were separated. The epidermis was emulsified and plated in a six-well cell culture dish in the DermaLife Ma Melanocyte Medium Complete Kit (Lifeline Cell Technology) supplemented with 100μg/ml primocin (Invitrogen). Melanocytes were identified based on their dendritic shape and ability to grow adhered to the dish in serum-free media. The dermis from each biopsy was divided into six to eight pieces which were then allowed to adhere to a six-well cell culture dish and were grown in Dulbecco's modified eagle's medium (Gibco) supplemented with 1X non-essential amino acids (Hyclone), 10% Cosmic Calf Serum (Hyclone), 10% AmnioMax C-100 supplement (Gibco) and 100μg/ml primocin. Fibroblasts were identified as adherent cells elongated in shape that grew from the dermis pieces. All cultures were incubated at 37˚C in a 5% carbon dioxide containing incubator. A portion of bulk cultures of both fibroblasts and melanocytes were harvested for genomic DNA, and another portion was diluted and plated to obtain single cell-derived clones. Fibroblast clones were expanded in culture for 5 to 6 additional passages (4–6 weeks) to obtain ~10^6 cells, and genomic DNA was extracted. Genomic DNA extraction from all samples was performed with DNeasy Blood and Tissue kit (Qiagen). Melanocyte clones were expanded in culture for 2 to 3 passages to obtain 10,000 cells and genomic DNA was extracted. 1 to 2.5 ng of the melanocyte genomic DNA was treated with USER (NEB) to remove deaminated cytosines from genomic DNA that are an artifact of DNA extraction [62]. The DNA was amplified using the REPLI-g Mini Kit (Qiagen). 12 to 14

different primer sets were used for PCR across random loci at different chromosomal positions, ranging from 100bp to 500bp, from the amplified genomic DNA. Samples with 10 or more reactions with the correct amplification product were subsequently purified and used for whole-genome sequencing. This quality check allows us to only sequence the genomic DNA with uniform amplification. Venous blood was collected in three to five 8.5mL PAXgene blood DNA tubes (PreAnalytiX/Qiagen) and DNA was isolated from whole blood samples. For 38 DNA samples DNA libraries were prepared using Truseq DNA PCR-free 350bp insert kit (Illumina), and were subsequently sequenced using Illumina HiseqX. For the 30 remaining samples, libraries were prepared using the Nextera DNA Flex library Prep kit (Illumina) and sequenced using the NovaSeq 6000 system. All samples were sequenced as 150 base-paired reads to a depth of 50X to 132X. For a subset of donors, additional skin biopsies were obtained for establishing a second clonal fibroblast lineage.

We also analyzed whole-genome sequenced clonal hip fibroblasts (D1-L-H, D1-R-H1, D1-R-H2, D2-L-H and D2-R-H) from 2 donors that were obtained in a previous study [8].

Calling somatic genome changes in sequenced clones

The FASTQ reads for each clone and blood sample were aligned to the hg19 genome using the GATK best practices pipeline [63]. Three base substitution callers, SomaticSniper [64], VarScan2 [65,66] and Mutect2 [67] were used to identify the clone-specific mutation calls that were not present in blood of the same donors. Only base substitutions detected by all three callers were analyzed further. Any somatic mutations that were also present in the dbSNP138 database, or any SNVs that overlapped SimpleRepeats tract in the UCSC Genome Browser were removed. The final mutation calls were filtered based on allele frequencies, such that only heterozygous mutations with allele frequencies between 45% and 55% or homozygous mutations with allele frequencies greater than 90% were kept. This methodology of using three independent mutation callers and stringent filtering criteria were used previously for accurate measurements of somatic mutations in human fibroblasts and has demonstrated very high accuracy by orthogonal validations [8]. For bulk whole exome sequencing, Mutect2 was used to call mutations. SNVs that overlapped SimpleRepeats tract or were in the dbSNP138 database were removed. The mutation calls for both whole exome and whole-genome sequencing have been organized as MAF files in the TCGA format and have been submitted to dbGAP study phs001182.v2.p1. Somatic structural variants within 1Mb of each other in different donors were marked as being within a "hotspot".

Delly was used to identify structural variants in the form of deletions, duplications, inversions and translocations [8,68]. Calls which were designated "LowQual" and/or "IMPRECISE" were removed. Clonality of structural variants was determined based on the allelic fraction of reads supporting the variants in the clone. Structural variants with 30% or more reads supporting the structural variant and the absence of any reads supporting the variants in blood were denoted as clonal somatic changes. Due to the low number of variants, we cannot rule out that some of the structural variants may have been generated due to a rearrangement in during the first few cell divisions of the founder cell in culture. However, we think this is unlikely as these cells were passaged less than 3 times before the generation of a clonal lineage. Moreover, the number and types of variants are similar to those detected in previous studies [69–71] indicating that the structural variants detected in our work were likely present in founder cells.

Indels were detected using the tool SV-ABA [72] and were filtered based on multiple criterion. Indel calls with a quality score less than 50 were removed and were only included if they occurred between 45%-55% allele frequency (heterozygous indel), or between 90%-100% allele frequency (homozygous indel). Indels that overlapped with the SimpleRepeats or the RepeatMasker tracts

were removed as these calls were often found to be erroneous. A subset of the indels were visually verified by inspection of the alignments using the Integrative Genomics Viewer [73]. 46 indels were orthogonally verified via PCR amplification and Sanger Sequencing.

Analyzing base substitution and indel signatures in clones

We used the SigProfilerMatrixGenerator [74] to identify the different types of indels in our samples. SigProfilerExtractor [13] was used to deconvolute the single base substitution, dinucleotide base substitution and indel signatures in our samples. 9 processes with 10 iterations were used within SigProfilerExtractor for extraction of indel and base substitution signatures. MutationalPatterns [75] was used to both identify the cosine similarities between the mutation patterns in samples in this study and the signatures identified in COSMIC and to also identify the contributions of COSMIC signatures in our samples. For base substitutions, the function fit_to_signatures() within MutationalPatterns was used to identify the contributions of the COSMIC signatures on the mutation profile of each sample. For indels, the function fit_to_-signatures() was modified to allow the matrix to have 83 rows instead of 96 so that the indels in our samples could be compared to the known ID signatures (83 channels) in COSMIC.

The enrichment of mutation signatures in each of the samples analyzed in this study was calculated as described in [8,14,35]. For this calculation, context is defined as the +/- 20 bases surrounding the mutated base. The mutated residue is capitalized in the annotation of the signature and the equation to calculate enrichment of a given mutation signature is provided below with the UV-mutation signature yCn➔yTn as an example.

$$Enrichment\ (yCn \rightarrow yTn) = \frac{[Mutations_{yCn \rightarrow yTn}] \times [Context_c]}{[Mutations_{C \rightarrow T}] \times [Context_{ycn}]}$$

For each motif, the reverse complement was also taken into account in the calculations. Mutations <10 bases apart, are excluded in this calculation as these are "complex" mutations that likely arise due to the activity of translesion polymerases and may confound the analysis of the mutation signatures. To determine if increased fold enrichments for the mutation signatures were statistically relevant a Fisher's Exact test was performed wherein the ratio of the number of mutations within the trinucleotide motif ($Mutations_{yCn \rightarrow yTn}$), and those that do not conform to the trinucleotide motif ($Mutations_{C \rightarrow T}$), were compared to the number of unmutated bases in the context that either were in the trinucleotide motif ($Context_{ycn}$) versus those that were not in the context ($Context_c$). Multiple hypothesis testing was further accounted for by the correction of the P-values via the Benjamini-Hochberg method. For samples where enrichment > 1 and the corrected P-value < 0.05, the Minimum Mutation load was calculated for the enriched signature. The equation for calculating this is provided below for the yCn➔yTn mutation signature.

$$Minimum\ Mut\ Load\ (yCn \rightarrow yTn) = \frac{[Mutations_{yCn \rightarrow yTn}] \times [Enrichment_{yCn \rightarrow yTn} - 1]}{[Enrichment_{yCn \rightarrow yTn}]}$$

Analysis of bulk DNA samples

We also sequenced the exomes of 15 fibroblast bulk samples and 4 melanocyte bulk samples directly cultured from the biopsies. Bulk samples are defined as cells that were not propagated clonally. Libraries were prepared using the Nextera Flex for Enrichment library prep kit, Illumina Exome Panel (Illumina) and IDT for Illumina UD Indexes (Illumina). All samples were sequenced using the NovaSeq 6000 system up to approximately ~150X depth. Somatic mutations were called in the samples by using Mutect2. Whole-genome-sequenced blood samples

from the donor corresponding to each bulk sample was used as a proxy for germline mutations. Only single nucleotide variants called by Mutect2 were further analyzed. Any mutations that were within the dbSNP138 database or in the SimpleRepeats tract were removed.

Annotation of SNVs

We used Annovar [76] to annotate SNVs for changes to protein sequence using the refGene track from UCSC Genome Browser. Nonsynonymous SNVs or SNVs affecting start or stop codons and splice sites were further annotated using the Cancer Genome Interpreter [39] as driver mutations or passenger mutations.

Supporting information

S1 Fig. The distribution of the allele frequencies of the whole genome sequenced clones in this study. The plots for melanocyte clones are in red. The source data for this figure is in S2 Table.
(TIF)

S2 Fig. (A)The mutation spectra in each sequenced clone in this study. The melanocyte clones are marked with an "M" in the X-axis. The source data for this figure are in S2 Table. (B) The NMF-derived mutation signature loads as determined by SigProfilerExtractor in each clone sequenced in this study. Samples form African American donors are annotated with an "A" and melanocyte clones are marked with an "M" in the X-axis. The source data for this figure are in S3 Table.
(TIF)

S3 Fig. The NMF-derived single base substitution and double base substitution signatures identified in human skin cells. Total mutations corresponding to the signature in the cohort as determined by SigProfilerExtractor are shown.
(TIF)

S4 Fig. The correlation of the different UV-specific mutation signatures in the samples in this study. The total nTt➔nCt mutations and the CC➔TT mutations are plotted against the yCn➔yTn total mutation load in the samples. The black inclined line denotes the linear regression of the data, and the dotted black lines denote the 95% confidence intervals. The source data for this figure are in S4 Table.
(TIF)

S5 Fig. Comparison of the NMF-derived mutation signatures and the trinucleotide-specific mutation signatures in this study. The nCg➔nTg minimum mutation load in each sample is plotted against SBS1-associated mutations as determined by SigProfilerExtractor, and the yCn➔yTn minimum mutation load in each sample is plotted against SBS7b-associated mutations. The linear regression of the data is shown, and the dotted lines denote the 95% confidence intervals. The source data for this figure are in S3 and S4 Tables.
(TIF)

S6 Fig. The distribution of the allele frequencies of the whole exome sequenced bulk samples in this study. The source data for this figure is in S5 Table.
(TIF)

S7 Fig. The distribution of the allele frequencies of consensus alleles and cancer drivers. (A) The allele frequencies of the consensus SNVs identified in the bulk and the corresponding clones are shown. (B) The allele frequency distribution of the cancer driver mutations

identified in the exome of the bulk samples. The source data for this figure is in S5 Table.
(TIF)

S8 Fig. The NMF-derived indel mutation signatures in this study. Total mutations corresponding to the signature as determined by SigProfilerExtractor are shown.
(TIF)

S9 Fig. The analyses of the impact of sex on mutation and indel load in the samples. The total base substitutions and the total indels in the clonal lineages derived from males and females in this study are shown. A Mann-Whitney U-test was used to determine if the distribution of mutation and indel load were statistically different between the two cohorts. The P-values for the base substitutions was 0.4041, while the P-value for the indels was 0.9401. The source data for this figure is in S10 Table.
(TIF)

S1 Table. Coverage statistics and donor characteristics for all samples sequenced in this study.
(XLSX)

S2 Table. The somatic base substitutions in the whole genome sequenced single skin cell clonal lineages. a) The fraction of all SNVs prior to filtering that correspond to each allele frequency bin. SNVs that corresponded to allele frequencies between 45% and 55% or above 90% were considered clonal. b) The exonic somatic base substitutions in the samples. c) The mutation spectra in the samples. The reverse complements are considered in the mutation spectra analyses.
(XLSX)

S3 Table. Agnostic base substitution signature analyses. a) The contributions of previously determined mutation signatures using MutationalPatterns. b) The number of mutations corresponding to each signature identified by SigProfilerExtractor
(XLSX)

S4 Table. Motif-specific mutation signature analyses. a) nCg➔nTg mutation signature analysis b)yCn➔yTn mutation signature analysis. c) nTt➔nCt mutation signature analysis. d) CC➔TT total dinucleotide substitutions.
(XLSX)

S5 Table. The somatic mutation list for whole exome sequenced bulk tissues.
(XLSX)

S6 Table. Somatic indels identified in the samples
(XLSX)

S7 Table. The distribution of the types of indels in each sample as identified by SigProfilerMatrixGenerator.
(XLSX)

S8 Table. Agnostic indel signature analysis. a) The contributions of previously determined mutation signatures using MutationalPatterns. b) The number of mutations corresponding to each signature identified by SigProfilerExtractor
(XLSX)

S9 Table. The somatic templated insertions identified in human skin fibroblasts and melanocytes.
(XLSX)

S10 Table. The somatic genome changes in the samples along with the cell type, and the sex and race of the donors. a) The total number of somatic genome changes in the samples along with the cell type, and the sex and race of the donors. b) The median values and P-values for Mann-Whitney tests for pairwise comparisons of the somatic genome changes between the sets of clones obtained from White and African American donors.
(XLSX)

S11 Table. The somatic structural variants identified in the donors. a) All the somatic structural variants annotated for hotspots, common fragile sites and microhomologies. b) A Fisher's exact test for the use of microhomologies in the SVs that are present within or out of common fragile sites.
(XLSX)

Acknowledgments

We are thankful to Drs Natasha Degtyareva, Kathleen Hudson, Scott Lujan and Sriram Vijayraghavan for critically reading this manuscript and providing their feedback.

Author Contributions

Conceptualization: Natalie Saini, Dmitry A. Gordenin.

Data curation: Natalie Saini, Leszek J. Klimczak, Brian N. Papas, Adam B. Burkholder, Jian-Liang Li, David C. Fargo.

Formal analysis: Natalie Saini, Camille K. Giacobone, Leszek J. Klimczak, Dmitry A. Gordenin.

Funding acquisition: Dmitry A. Gordenin.

Investigation: Natalie Saini, Camille K. Giacobone, Re Bai, Kevin Gerrish, Cynthia L. Innes, Shepherd H. Schurman.

Methodology: Natalie Saini, Shepherd H. Schurman, Dmitry A. Gordenin.

Project administration: Jian-Liang Li, David C. Fargo, Shepherd H. Schurman, Dmitry A. Gordenin.

Resources: David C. Fargo, Kevin Gerrish, Dmitry A. Gordenin.

Software: Leszek J. Klimczak.

Supervision: Natalie Saini, Dmitry A. Gordenin.

Validation: Natalie Saini, Re Bai, Kevin Gerrish, Cynthia L. Innes, Shepherd H. Schurman.

Visualization: Natalie Saini, Dmitry A. Gordenin.

Writing – original draft: Natalie Saini, Dmitry A. Gordenin.

Writing – review & editing: Natalie Saini, Kevin Gerrish, Shepherd H. Schurman, Dmitry A. Gordenin.

References

1. Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell. 2017; 168(4). https://doi.org/10.1016/j.cell.2017.01.002 WOS:000396277600015. PMID: 28187286

2. Lindahl T, Barnes DE. Repair of endogenous DNA damage. Cold Spring Harb Sym. 2000; 65:127–33. https://doi.org/10.1101/sqb.2000.65.127 WOS:000169676800014. PMID: 12760027

3. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science. 2015; 350(6256):94–8. Epub 2015/10/03. https://doi.org/10.1126/science.aab1785 PMID: 26430121; PubMed Central PMCID: PMC4664477.

4. Rouhani FJ, Nik-Zainal S, Wuster A, Li Y, Conte N, Koike-Yusa H, et al. Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. PLoS Genet. 2016; 12(4):e1005932. https://doi.org/10.1371/journal.pgen.1005932 PMID: 27054363; PubMed Central PMCID: PMC4824386.

5. Abyzov A, Tomasini L, Zhou B, Vasmatzis N, Coppola G, Amenduni M, et al. One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. 2017; 27(4):512–23. Epub 2017/02/24. https://doi.org/10.1101/gr.215517.116 PMID: 28235832.

6. D'Antonio M, Benaglio P, Jakubosky D, Greenwald WW, Matsui H, Donovan MKR, et al. Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. Cell Rep. 2018; 24(4):883–94. Epub 2018/07/26. https://doi.org/10.1016/j.celrep.2018.06.091 PMID: 30044985; PubMed Central PMCID: PMC6467479.

7. Franco I, Helgadottir HT, Moggio A, Larsson M, Vrtačnik P, Johansson A, et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. Genome Biol. 2019; 20(1):285–. https://doi.org/10.1186/s13059-019-1892-z PMID: 31849330.

8. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLoS genetics. 2016; 12 (10):e1006385–e. https://doi.org/10.1371/journal.pgen.1006385 PMID: 27788131.

9. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science. 2018; 359(6375):550–5. Epub 2017/12/09. https://doi.org/10.1126/science.aan8690 PMID: 29217587; PubMed Central PMCID: PMC6311130.

10. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016; 538(7624):260–4. Epub 2016/10/05. https://doi.org/10.1038/nature19768 PMID: 27698416; PubMed Central PMCID: PMC5536223.

11. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019; 574(7779):538–42. https://doi.org/10.1038/s41586-019-1670-9 PMID: 31645727

12. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500(7463):415–21. https://doi.org/10.1038/nature12477 PMID: 23945592; PubMed Central PMCID: PMC3776390.

13. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020; 578(7793):94–101. Epub 2020/02/07. https://doi.org/10.1038/s41586-020-1943-3 PMID: 32025018; PubMed Central PMCID: PMC7054213.

14. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45(9):970–6. Epub 2013/07/16. https://doi.org/10.1038/ng.2702 PMID: 23852170; PubMed Central PMCID: PMC3789062.

15. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nature reviews Cancer. 2014; 14(12):786–800. Epub 2014/11/25. https://doi.org/10.1038/nrc3816 PMID: 25568919.

16. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A. 2013; 110(6):1999–2004. https://doi.org/10.1073/pnas.1221068110 PMID: 23345422; PubMed Central PMCID: PMC3568331.

17. Kaufmann WK, Cleaver JE. Mechanisms of inhibition of DNA replication by ultraviolet light in normal human and xeroderma pigmentosum fibroblasts. J Mol Biol. 1981; 149(2):171–87. Epub 1981/06/25. https://doi.org/10.1016/0022-2836(81)90297-7 PMID: 7310880.

18. Rudolph CJ, Upton AL, Lloyd RG. Replication fork stalling and cell cycle arrest in UV-irradiated Escherichia coli. Genes Dev. 2007; 21(6):668–81. Epub 2007/03/21. https://doi.org/10.1101/gad.417607 PMID: 17369400; PubMed Central PMCID: PMC1820941.

19. Sugiyama T, Chen Y. Biochemical reconstitution of UV-induced mutational processes. Nucleic Acids Res. 2019; 47(13):6769–82. Epub 2019/05/06. https://doi.org/10.1093/nar/gkz335 PMID: 31053851; PubMed Central PMCID: PMC6648339.

20. Yu SL, Johnson RE, Prakash S, Prakash L. Requirement of DNA polymerase eta for error-free bypass of UV-induced CC and TC photoproducts. Mol Cell Biol. 2001; 21(1):185–8. Epub 2000/12/13. https://doi.org/10.1128/MCB.21.1.185-188.2001 PMID: 11113193; PubMed Central PMCID: PMC88792.

21. Stary A, Kannouche P, Lehmann AR, Sarasin A. Role of DNA polymerase eta in the UV mutation spectrum in human cells. J Biol Chem. 2003; 278(21):18767–75. Epub 2003/03/20. https://doi.org/10.1074/jbc.M211838200 PMID: 12644471.

22. Johnson RE, Prakash S, Prakash L. Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Poleta. Science. 1999; 283(5404):1001–4. Epub 1999/02/12. https://doi.org/10.1126/science.283.5404.1001 PMID: 9974380.

23. Masutani C, Kusumoto R, Iwai S, Hanaoka F. Mechanisms of accurate translesion synthesis by human DNA polymerase eta. EMBO J. 2000; 19(12):3100–9. Epub 2000/06/17. https://doi.org/10.1093/emboj/19.12.3100 PMID: 10856253; PubMed Central PMCID: PMC203367.

24. Washington MT, Prakash L, Prakash S. Mechanism of nucleotide incorporation opposite a thymine-thymine dimer by yeast DNA polymerase eta. Proc Natl Acad Sci U S A. 2003; 100(21):12093–8. Epub 2003/10/07. https://doi.org/10.1073/pnas.2134223100 PMID: 14527996; PubMed Central PMCID: PMC218718.

25. Dumstorf CA, Clark AB, Lin Q, Kissling GE, Yuan T, Kucherlapati R, et al. Participation of mouse DNA polymerase iota in strand-biased mutagenic bypass of UV photoproducts and suppression of skin cancer. Proc Natl Acad Sci U S A. 2006; 103(48):18083–8. Epub 2006/11/23. https://doi.org/10.1073/pnas.0605247103 PMID: 17114294; PubMed Central PMCID: PMC1838710.

26. Ziv O, Geacintov N, Nakajima S, Yasui A, Livneh Z. DNA polymerase zeta cooperates with polymerases kappa and iota in translesion DNA synthesis across pyrimidine photodimers in cells from XPV patients. Proc Natl Acad Sci U S A. 2009; 106(28):11552–7. Epub 2009/07/01. https://doi.org/10.1073/pnas.0812548106 PMID: 19564618; PubMed Central PMCID: PMC2710681.

27. Yoon JH, Prakash L, Prakash S. Highly error-free role of DNA polymerase eta in the replicative bypass of UV-induced pyrimidine dimers in mouse and human cells. Proc Natl Acad Sci U S A. 2009; 106(43):18219–24. Epub 2009/10/14. https://doi.org/10.1073/pnas.0910121106 PMID: 19822754; PubMed Central PMCID: PMC2775276.

28. Zhang H, Siede W. UV-induced T—>C transition at a TT photoproduct site is dependent on Saccharomyces cerevisiae polymerase eta in vivo. Nucleic Acids Res. 2002; 30(5):1262–7. https://doi.org/10.1093/nar/30.5.1262 PMID: 11861920; PubMed Central PMCID: PMC101249.

29. McCulloch SD, Kokoska RJ, Masutani C, Iwai S, Hanaoka F, Kunkel TA. Preferential cis-syn thymine dimer bypass by DNA polymerase eta occurs with biased fidelity. Nature. 2004; 428(6978):97–100. https://doi.org/10.1038/nature02352 PMID: 14999287.

30. Lopes M, Foiani M, Sogo JM. Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. Mol Cell. 2006; 21(1):15–27. Epub 2006/01/03. https://doi.org/10.1016/j.molcel.2005.11.015 PMID: 16387650.

31. Elvers I, Johansson F, Groth P, Erixon K, Helleday T. UV stalled replication forks restart by re-priming in human fibroblasts. Nucleic Acids Res. 2011; 39(16):7049–57. Epub 2011/06/08. https://doi.org/10.1093/nar/gkr420 PMID: 21646340; PubMed Central PMCID: PMC3167624.

32. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015; 348(6237):880–6. https://doi.org/10.1126/science.aaa6806 PMID: 25999502; PubMed Central PMCID: PMC4471149.

33. Tang J, Fewings E, Chang D, Zeng H, Liu S, Jorapur A, et al. The genomic landscapes of individual melanocytes from human skin. Nature. 2020; 586(7830):600–5. Epub 2020/10/09. https://doi.org/10.1038/s41586-020-2785-8 PMID: 33029006; PubMed Central PMCID: PMC7581540.

34. Saini N, Gordenin DA. Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. Environ Mol Mutagen. 2018; 59(8):672–86. Epub 2018/08/29. https://doi.org/10.1002/em.22215 PMID: 30152078; PubMed Central PMCID: PMC6188803.

35. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015; 47(9):1067–72. Epub 2015/08/11. https://doi.org/10.1038/ng.3378 PMID: 26258849; PubMed Central PMCID: PMC4594173.

36. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015; 47(12):1402–7. https://doi.org/10.1038/ng.3441 PMID: 26551669.

37. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020; 580(7805):640–6. Epub 2020/05/01. https://doi.org/10.1038/s41586-020-2214-z PMID: 32350471.

38. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019; 574(7779):532–7. https://doi.org/10.1038/s41586-019-1672-7 PMID: 31645730

39. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018; 10(1):25. Epub 2018/03/30. https://doi.org/10.1186/s13073-018-0531-8 PMID: 29592813; PubMed Central PMCID: PMC5875005.

40. Gordenin DA, Resnick MA. Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. Mutat Res. 1998; 400(1–2):45–58. Epub 1998/08/01. https://doi.org/10.1016/s0027-5107(98)00047-5 PMID: 9685581.

41. Sia EA, Jinks-Robertson S, Petes TD. Genetic control of microsatellite stability. Mutat Res. 1997; 383 (1):61–70. Epub 1997/01/31. https://doi.org/10.1016/s0921-8777(96)00046-8 PMID: 9042420.

42. Carvajal-Garcia J, Cho JE, Carvajal-Garcia P, Feng W, Wood RD, Sekelsky J, et al. Mechanistic basis for microhomology identification and genome scarring by polymerase theta. Proc Natl Acad Sci U S A. 2020; 117(15):8476–85. Epub 2020/04/03. https://doi.org/10.1073/pnas.1921791117 PMID: 32234782; PubMed Central PMCID: PMC7165422.

43. Yu AM, McVey M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. Nucleic Acids Res. 2010; 38(17):5706–17. Epub 2010/05/13. https://doi.org/10.1093/nar/gkq379 PMID: 20460465; PubMed Central PMCID: PMC2943611.

44. Kumar R, Nagpal G, Kumar V, Usmani SS, Agrawal P, Raghava GPS. HumCFS: a database of fragile sites in human chromosomes. BMC Genomics. 2019; 19(Suppl 9):985. Epub 2019/04/20. https://doi.org/10.1186/s12864-018-5330-5 PMID: 30999860.

45. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–8. Epub 2013/06/19. https://doi.org/10.1038/nature12213 PMID: 23770567; PubMed Central PMCID: PMC3919509.

46. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018; 362(6417):911–7. Epub 2018/10/20. https://doi.org/10.1126/science.aau3879 PMID: 30337457; PubMed Central PMCID: PMC6298579.

47. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020; 578(7794):266–72. Epub 2020/01/31. https://doi.org/10.1038/s41586-020-1961-1 PMID: 31996850; PubMed Central PMCID: PMC7021511.

48. Wang H, Xu X. Microhomology-mediated end joining: new players join the team. Cell Biosci. 2017; 7:6. Epub 2017/01/20. https://doi.org/10.1186/s13578-017-0136-8 PMID: 28101326; PubMed Central PMCID: PMC5237343.

49. Sfeir A, Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends Biochem Sci. 2015; 40(11):701–14. Epub 2015/10/07. https://doi.org/10.1016/j.tibs.2015.08.006 PMID: 26439531; PubMed Central PMCID: PMC4638128.

50. Karanam K, Kafri R, Loewer A, Lahav G. Quantitative live cell imaging reveals a gradual shift between DNA repair mechanisms and a maximal use of HR in mid S phase. Mol Cell. 2012; 47(2):320–9. Epub 2012/07/31. https://doi.org/10.1016/j.molcel.2012.05.052 PMID: 22841003; PubMed Central PMCID: PMC3494418.

51. Delacote F, Lopez BS. Importance of the cell cycle phase for the choice of the appropriate DSB repair pathway, for genome stability maintenance: the trans-S double-strand break repair model. Cell Cycle. 2008; 7(1):33–8. Epub 2008/01/17. https://doi.org/10.4161/cc.7.1.5149 PMID: 18196958.

52. Ridky TW. Nonmelanoma skin cancer. J Am Acad Dermatol. 2007; 57(3):484–501. Epub 2007/05/22. https://doi.org/10.1016/j.jaad.2007.01.033 PMID: 17512631.

53. Halder RM, Bridgeman-Shah S. Skin cancer in African Americans. Cancer. 1995; 75(2 Suppl):667–73. Epub 1995/01/15. https://doi.org/10.1002/1097-0142(19950115)75:2+<667::aid-cncr2820751409>3.0.co;2-i PMID: 7804993.

54. Gloster HM Jr., Neal K. Skin cancer in skin of color. J Am Acad Dermatol. 2006; 55(5):741–60; quiz 61–4. Epub 2006/10/21. https://doi.org/10.1016/j.jaad.2005.08.063 PMID: 17052479.

55. Bradford PT. Skin cancer in skin of color. Dermatol Nurs. 2009; 21(4):170–7, 206; quiz 178. Epub 2009/08/21. PMID: 19691228; PubMed Central PMCID: PMC2757062.

56. Halder RM, Bang KM. Skin cancer in blacks in the United States. Dermatol Clin. 1988; 6(3):397–405. Epub 1988/07/01. PMID: 3048822.

57. Plaja A, Castells N, Cueto-Gonzalez AM, del Campo M, Vendrell T, Lloveras E, et al. A Novel Recurrent Breakpoint Responsible for Rearrangements in the Williams-Beuren Region. Cytogenet Genome Res. 2015; 146(3):181–6. Epub 2015/09/19. https://doi.org/10.1159/000439463 PMID: 26382598.

58. Georgakilas AG, Tsantoulis P, Kotsinas A, Michalopoulos I, Townsend P, Gorgoulis VG. Are common fragile sites merely structural domains or highly organized "functional" units susceptible to oncogenic stress? Cell Mol Life Sci. 2014; 71(23):4519–44. Epub 2014/09/23. https://doi.org/10.1007/s00018-014-1717-x PMID: 25238782; PubMed Central PMCID: PMC4232749.

59. Le Tallec B, Dutrillaux B, Lachages AM, Millot GA, Brison O, Debatisse M. Molecular profiling of common fragile sites in human fibroblasts. Nat Struct Mol Biol. 2011; 18(12):1421–3. Epub 2011/11/08. https://doi.org/10.1038/nsmb.2155 PMID: 22056772.

60. Murano I, Kuwano A, Kajii T. Fibroblast-specific common fragile sites induced by aphidicolin. Hum Genet. 1989; 83(1):45–8. Epub 1989/08/01. https://doi.org/10.1007/BF00274145 PMID: 2504659.

61. Maccaroni K, Balzano E, Mirimao F, Giunta S, Pelliccia F. Impaired Replication Timing Promotes Tissue-Specific Expression of Common Fragile Sites. Genes (Basel). 2020; 11(3). Epub 2020/03/25. https://doi.org/10.3390/genes11030326 PMID: 32204553; PubMed Central PMCID: PMC7140878.

62. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017; 356(6334):189–94. Epub 2017/04/15. https://doi.org/10.1126/science.aak9787 PMID: 28408603; PubMed Central PMCID: PMC5538131.

63. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43(5):491–8. Epub 2011/04/12. https://doi.org/10.1038/ng.806 PMID: 21478889; PubMed Central PMCID: PMC3083463.

64. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28(3):311–7. Epub 2011/12/14. https://doi.org/10.1093/bioinformatics/btr665 PMID: 22155872; PubMed Central PMCID: PMC3268238.

65. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr Protoc Bioinformatics. 2013; 44:15 4 1–7. Epub 2015/01/02. https://doi.org/10.1002/0471250953.bi1504s44 PMID: 25553206; PubMed Central PMCID: PMC4278659.

66. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22(3):568–76. Epub 2012/02/04. https://doi.org/10.1101/gr.129684.111 PMID: 22300766; PubMed Central PMCID: PMC3290792.

67. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31(3):213–9. Epub 2013/02/12. https://doi.org/10.1038/nbt.2514 PMID: 23396013; PubMed Central PMCID: PMC3833702.

68. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28(18):i333–i9. Epub 2012/09/11. https://doi.org/10.1093/bioinformatics/bts378 PMID: 22962449; PubMed Central PMCID: PMC3436805.

69. Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, Tomasini L, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature. 2012; 492(7429):438–42. Epub 2012/11/20. https://doi.org/10.1038/nature11629 PMID: 23160490; PubMed Central PMCID: PMC3532053.

70. Knouse KA, Wu J, Whittaker CA, Amon A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proc Natl Acad Sci U S A. 2014; 111(37):13409–14. Epub 2014/09/10. https://doi.org/10.1073/pnas.1415287111 PMID: 25197050; PubMed Central PMCID: PMC4169915.

71. Knouse KA, Wu J, Amon A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res. 2016; 26(3):376–84. Epub 2016/01/17. https://doi.org/10.1101/gr.198937.115 PMID: 26772196; PubMed Central PMCID: PMC4772019.

72. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018; 28(4):581–91. Epub 2018/03/15. https://doi.org/10.1101/gr.221028.117 PMID: 29535149; PubMed Central PMCID: PMC5880247.

73. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14(2):178–92. Epub 2012/04/21. https://doi.org/10.1093/bib/bbs017 PMID: 22517427; PubMed Central PMCID: PMC3603213.

74. Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 2019; 20

(1):685. Epub 2019/09/01. https://doi.org/10.1186/s12864-019-6041-2 PMID: 31470794; PubMed Central PMCID: PMC6717374.

75. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018; 10(1):33. Epub 2018/04/27. https://doi.org/10.1186/s13073-018-0539-0 PMID: 29695279; PubMed Central PMCID: PMC5922316.

76. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. Epub 2010/07/06. https://doi.org/10.1093/nar/gkq603 PMID: 20601685; PubMed Central PMCID: PMC2938201.