SUPPLEMENTARY MATERIALS

Content

Other online-available Supplementary Material for this work includes:

Supplementary Datasets S1 and S2

SUPPLEMENTARY METHODS

Sample collection

A total of 123 cancer-free volunteers donated a normal skin sample obtained from the margin of skin excision biopsies undertaken to remove a cutaneous benign lesion (Table S1). Seventy-nine samples were collected from skin areas with intermittent sun exposure (back, chest, legs and upper arms), and 44 samples were obtained from chronically sun-exposed skin areas (neck, face and hands). Samples were recruited at the Department of Dermatology of two hospitals from Castellon Province, Spain (Castellon University General Hospital and La Plana University Hospital). Only one sample was recruited per donor due to ethical reasons. All participants provided a written informed consent. The study was approved by the Ethics Committee of the Jaume I University of Castellon (Spain).

An elliptical excision was performed for removing each skin lesion. Excision design followed the standard length-to-width ratio of 3:1, with apical angles of less than 30 degrees and surgical margins of 1 to 3 mm. In all excisions performed, the surgical margin was the minimum possible to perform the primary closure achieving a cosmetically acceptable scar. One of the two sharp ends of the biopsies (benign area) was collected for this study.

Immediately after resection, tissue samples were submerged in RNA*later* Tissue Collection Solution (Thermo Fisher Scientific, Walham, MA, USA) and stored at 4ºC overnight. Then, the epidermis was separated from the dermis by incubating the tissue sample in 3.8% ammonium thiocyanate (Sigma-Aldrich, St Louis, MO, USA) in PBS (pH 7.4) at room temperature for 3 hours. Subsequently, the epidermis was immersed in RNA*later* solution and stored at -20ºC until sample processing.

Genomic DNA was isolated from fresh-frozen normal epidermal samples with the QIAamp DNA Mini Kit, (Qiagen, Hilden, Germany). DNA was stored at -20ºC until use.

Phenotypic data collection

Each participant completed a standardised questionnaire to collect information on sex, age, pigmentation traits (skin, hair and eye colour), skin sensitivity to sunlight (tanning ability *versus* tendency to burn), freckling degree, history of childhood sunburns, and sun exposure habits. Pigmentation- and sun sensitivity-related traits were used to group individuals according to Fitzpatrick's skin type classification. Detailed information related to signs of sun damage in the skin area biopsied (pigmented spots, blotches, and wrinkles) was also recorded. To avoid misclassification, each participant completed the questionnaire under the supervision of a professional.

Sequencing of *MC1R* coding region

The coding sequence of the *MC1R* gene was directly sequenced in all samples, as previously described [1]. Non-synonymous *MC1R* mutations were then defined as 'R', 'r' or 'p' (pseudoallele) alleles according to their impact on protein function, following criteria previously described [2].

Ultra-deep targeting sequencing

A panel of 46 genes was chosen to perform ultra-deep targeted sequencing (Table S3). These genes have been found to be often involved in skin cancer development [3–6] and/or have been shown to be frequently mutated in normal skin samples [7]. A custom bait capture was designed using NimbleGen SeqCap EZ (Roche, Basel, Switzerland) in order to target the exonic regions of the selected genes. The total size of the targeted regions was 0.32 Mb.

The 46 genes selected for sequencing are: *ADAM29, ADAMTS18, ARID1A, ARID2, BAI3, BRAF, CDKN2A, CRNKL1, EPHA2, EZH2, FAT1, FAT2, FGFR3, GRIN2A, GRM3, HRAS, IL7R, KMT2B, KRAS, MECOM, NF1, NOTCH1, NOTCH2, NOTCH3, NRAS, PIK3CA, PLCB1, PPP1R3A, PPP6C, PREX2, PTCH1, PTEN, PTPRB, PTPRK, RAC1, RB1, RBM10, SALL1, SCN1A, SF3B1, SPHKAP, STAT5B, TERT, TP53,* and *ZNF750*.

Sequencing of paired-end 100bp reads was performed on an Illumina HiSeq 2000 machine. The average on-target coverage across samples was 923.44x, ranging from 377.96x to 1657.37x. The variation in coverage across genes and samples is displayed in Figure S11. Note that the mutation burden found per gene, as well as per sample, was not strongly influenced by differences in coverage (Figure S11C-D).

Paired-end reads were aligned to the reference human genome (GRCh37d5) using the BWA-MEM algorithm with default parameters [8]. Alignment files (BAM format) containing only properly paired, uniquely mapping reads were processed using Picard tools version 1.110 (http://broadinstitute.github.io/picard/) to add read groups and remove PCR duplicates. Local realignments and base-quality recalibrations were conducted using GATK (v.3.2.2) [9].

Variant calling

Processed BAM files were analysed to identify single-nucleotide variants (SNVs) and small insertions and deletions (indels). Somatic mutations are normally called by detecting mismatches present in a tissue sample that are absent in a matched control sample (normal tissue or blood from the same patient). Due to the absence of matched normal sample from each individual, processed BAM files were used to perform somatic variant calling by applying Mutect2 in tumour-only mode (version 4.0.8.1). Following Broad Institute recommendations for variant calling, putative artefacts and germline variants were removed with FilterMutectCalls and FilterByOrientationBias. We provided FilterMutectCalls the set of human variants from gnomAD (https://gnomad.broadinstitute.org). Functional annotations were added to the resulting list of variants using SnpEff [10], with the gene annotation based on Ensembl data release 75. Variants were then annotated using SnpSift [11], with population frequencies, conservation scores and deleteriousness predictions obtained from dbNSFP [12]. Each variant was also annotated using gnomAD, COSMIC, ExAC, and ClinVar.

Then, a number of post-processing filters were applied. Firstly, we focused on identifying and removing germline variants. The variant caller Platypus was used to identify germline variants, which were filtered out from the list of somatic mutations [13]. The tool was run using the human variant set from dbSNP as the reference, instead of using a matched normal sample. Mutations detected in each sample were additionally called against the aggregate variants from a panel of normal samples of 200 Spanish individuals sequenced in the facilities of the CNAG-CRG (Barcelona, Spain) in order to remove common single-nucleotide polymorphisms (SNPs) and frequent technical artefacts. Indeed, variants were filtered out if they appeared in any of the ExAC, 1000 Genomes Project and dbSNP databases. As our filtering strategy seems to be quite rigorous, we decided to not remove those mutations that are included in the catalogue of somatic mutations found in human cancers (according to COSMIC and DoCM databases) for downstream analyses. These 75 putative driver mutations had a low VAF (mean = 0.021, max = 0.093) and prevalence (mean = 1.42%, max = 2.74%) in our cohort. To reduce false positive calls, variants were also filtered out based on their allele frequency. Our study was designed to detect mutations present in a small fraction of the skin cells of the biopsy. Variant allele frequencies for somatic mutations in normal samples are more likely to have values below 50%, as shown previously in normal skin samples from eyelids [7]. Therefore, we filtered out a variant when the 95% confidence intervals (CIs) of its VAF (determined by the binomial distribution taking into account the depth of coverage) reached values greater than 50%. To increase the sensitivity of our analyses, we also opted to remove all variants present with VAF values two standard deviations away from the mean per sample. As we were working with a small cohort of unrelated patients, mutations with a prevalence higher than 25% in our cohort were further excluded. Next, we also excluded variants that have not been found to have a clinical relevance in human cancers (according to DoCM database) with a prevalence two standard deviations away from the mean of our cohort. That is because spontaneously-arising neutral mutations are extremely unlikely to affect samples collected from different patients. Finally, sites with very low coverage (n<50 reads) were also excluded to avoid testing sites with limited power to detect variants.

To check if the majority of mutations removed were germline variants or technical artefacts, we studied the context-specific mutation spectra for each set of mutations removed at each filtering step (Figure S2C). Note that nearly all substitutions removed were not related to UV damage (C>T mutations at dipyrimidine sites). Indeed, having a global dN/dS ratio << 1 may denote that the pre-filtering dataset of variants is contaminated with germline SNPs (Figure S2D).

Validation of germline filtering procedure

The efficiency of post-processing filters in removing germline variants from the final list of putative somatic mutations was tested by using two external datasets.

Firstly, we called somatic mutations from the genome in a bottle set (NA12878, https://www.nist.gov/programs-projects/genome-bottle) using our variant calling pipeline. We use only the genomic regions that were captured in our experiments. The efficiency of the filtering procedure was assessed by comparing the predicted set of somatic mutations before and after applying the post-processing filters with the golden set of germline variants. Four out of 257 germline variants (1.55%) remained in the pre-filtering list of somatic mutations, but they were filtered out after applying the filtering procedure (false positive rate of 0%).

In addition, we used sequencing data from melanoma and adjacent non-malignant FFPE samples from six patients. The same panel of 46 genes was sequenced. Germline variants were called using HaplotypeCaller. Putative somatic mutations were predicted in all FFPE samples using our variant calling pipeline and filtering procedure. Then, we evaluated the proportion of germline variants included in the predicted set of somatic mutations before and after applying the filtering procedure. A significant number of germline variants were filtered out from the pre-filtering set of somatic mutations after applying the post-processing filtering steps (mean rate of false positives decreased from 24.07% to 0.04%; Figure S2A).

Evaluation of somatic mutations missed

Since we applied a very stringent filtering procedure, we also evaluated the percentage of real somatic mutations likely missed in our dataset by two different ways.

Firstly, after calling somatic mutations from the genome in a bottle set (NA12878, https://www.nist.gov/programs-projects/genome-bottle) with Mutect2 tumour-only mode, we removed the known set of germline variants. Then, all post-processing filters were subsequently applied to the list of putative somatic mutations, except for the one that removes sites with low coverage (<50 reads). One out of 10 somatic mutations was lost from the pre-filtering set (false negative rate of 10%).

Additionally, we called somatic mutations accumulated in six melanoma FFPE samples using Mutect2 paired mode, since a non-malignant FFPE sample from each patient was also sequenced. After applying our filtering procedure, an average of 21.63% of putative somatic mutations were lost. That is, since many of variants in FFPE samples may be technical artefacts, we expected to have a maximum false negative rate of 21.63%.

Prediction of mutational burden

After evaluating different types of model (Supplementary Text and Figure S3), a log-linear model was applied to correlate the number of mutations detected per sample with the sun exposure pattern of the skin sample and the individual's age, including different phenotypic traits as covariates. The covariates included in the model were sex (female *vs.* male), skin phototypes (I *vs.* II, III or IV), sun damage in the tissue (absence *vs.* presence), history of sunlight exposure (frequently *vs*. occasionally), and *MC1R* genotype (wild-type *vs.* r carriers or R carriers). The R package 'relaimpo' was used to assess the relative importance of the different variables included in the model. Non-photoexposed skin samples were excluded in this analysis due to the reduced sample size of this group (n=4). Nonparametric bootstraps (1000 runs) were conducted to estimate the 95% confidence intervals (CI95) of the age effect in each skin phototype subgroup.

Additionally, the correlation of the different intrinsic and extrinsic risk factors with the number of mutations accumulated per sample was also tested taking into account the mean clone's size per sample (an indicator of mutation detectability threshold).

In order to double-check that the majority of mutations were real somatic variants, we applied the log-linear model using the pre-filtering dataset of variants (Figure S2E). Note that the total variance of mutational burden explained by the model was very low (adjusted-R^2 = 6.08%), and the major predictors of mutational burden were completely different from those obtained when the filtering mutation dataset was used.

Coverage down-sampling analysis

To evaluate the effect of sequencing depth on mutational burden variation across samples, bam files were randomly down-sampled to different coverage levels (1000x, 800x, 600x and 400x) using the function DownsampleSam in Picard tools. Once the down-sampled datasets were generated, the variant calling and filtering steps were repeated to obtain a list of somatic mutations from each down-sampled dataset. The number of samples included in each down-sampled dataset varied according to the original coverage of the sample. A total of 54 samples were included in the 1000x dataset, 95 samples in the 800x dataset, 113 samples in the 600x dataset, and 122 samples in the 400x dataset.

Analysis of local mutational context and extraction of mutational signatures

Mutational spectrum and signatures analyses were performed by using the deconstructSigs R package [14]. Due to the limited number of mutations found in some samples (less than 50 mutations), we decided to group the samples by (a) age of individuals, and (b) pattern of sunlight exposure of skin tissue biopsied. Firstly, the proportion of each distinct single base substitution, as well as of each dinucleotide mutation, occurring within a given trinucleotide context was determined per each sample and per group. Hierarchical clustering of samples based on trinucleotide context of mutations was performed by applying the Ward's criterion. Samples were mainly divided into the different clusters by age and body site exposure, confirming that the vast majority of mutations included in our final list are real somatic mutations (Figure S5A).

Then, we evaluated the transcriptional strand bias for the mutations that are located within exons. A Poisson test was applied to assess whether the mutations occurred more often in the transcribed or untranscribed strand, or vice versa. Exon definitions for human reference genome were retrieved from BiomaRt by loading a TxDb annotation package from Bioconductor [15].

The limited number of variants hampers the discovery of new mutational signatures. Therefore, we ran deconstructSigs including only the mutational signatures related to aging and/or previously observed in the different skin cancers subtypes that contribute at least 6% of all of the observed mutations across the 127 samples (SBS2, SBS6, SBS7a, SBS7b, SBS7d, and SBS17a). Figure S5B shows the weights assigned to all of these mutational signatures for the combined set of mutations within cohort (Total) and per age group. Due to the limited number of samples and the low number of mutations, non-photoexposed skin samples were excluded.

Prevalence of non-synonymous mutations and selection analyses

Selection across the normal skin samples was quantified by using the dNdScv R package [16], which adapts the traditional implementation of dN/dS ratio by using trinucleotide context-dependent substitution models to avoid common mutation biases affecting dN/dS. Selection tests were performed on different subsets of mutations by grouping samples per age. Briefly, global and gene-level dN/dS ratios were quantified for missense and truncating (nonsense and essential splicing) mutations, as well as for indels, and then were used to compare the selection intensities between skin samples biopsied from elderly and young individuals. Global dN/dS ratios were performed for cancer and non-cancer genes independently. Again, the fact that our results reveal an excess of non-synonymous mutations in genes previously associated with cancer development (dN/dS > 1), which is not seen in non-cancer genes (dN/dS ~ 1), suggests that our list of mutations is not contaminated by germline variants or technical artefacts (Figure 3B).

The fraction of non-synonymous mutations fixed by positive selection (and thus may be driver mutations) was calculated from the estimated dN/dS ratios for missense, truncating and essential splicing substitutions [16].

The database of curated mutations (DoCM, docm.genome.wustl.edu) was used to identify canonical hotspot mutations with characterized functional or clinical evidence in cancer.

To further evaluate evidence for drift and selection, a log-linear model was applied to explore the clonal expansion of non-synonymous and synonymous mutations (mean VAF per sample of both mutation types) with age.

Detection of copy number aberrations

Copy number aberrations were identified applying ExomeDepth [17], a R package that uses read depth data to call CNVs from exome/targeted sequencing datasets. Briefly, the BAM file from each sample was compared to a reference BAM file, which was constructed by combining sequencing data from the most compatible samples of the dataset. Thus, the reference BAM file was optimized for each sample. By applying this method, we assumed that the CNV of interest was absent from the aggregate reference set (recurrent CNVs may be missed), and that the coverage was equal in all genome regions sequenced in all samples (the same sequencing procedure was applied in all samples). This analysis was performed per gene.

Putative copy number aberrations were confirmed by detecting allelic imbalances. After identifying heterozygous polymorphisms, a proportion test was used for testing if the fraction of minor SNP-allele reads was similar to the minor allele frequency expected for a heterozygous SNP. To be more prudent, the observed proportion was compared to two theoretical proportions, which were calculated by two different ways: (1) per sample (mean of min($BAF_{s,i}$, 1-$BAF_{s,i}$)), and (2) per all SNPs (mean of min(BAF_i, 1-BAF_i). A biallelic fraction was considered as statistically different to the theoretical fraction when the two-sided *P*-values of both comparisons were lower than 0.001. Only genes that had at least 50% of their heterozygous SNPs with a statistically different proportion were considered to have a significant allelic imbalance.

SUPPLEMENTARY TEXT

Model selection for explaining mutational burden

Although there was a theoretical reason for modelling mutational burden increase as a log-linear function of age, other possible model forms (including linear, log-linear, quadratic, cubic and non-linear) were explored to make sure that we were choosing the best model for fitting our data.

Model selection was based on the Akaike information criterion (AIC). Among all models explored, the log-linear model presented the lowest AIC value and thus was the optimal model (Figure S3). The log-linear model also presented the lowest significant value, supporting that this statistical model was the most compatible with the data.

Effect of mutation detectability on mutational burden

The number of detectable mutations may be sensitive to sequencing depth. In this study, the mutational burden found per gene was not influenced by differences in coverage (Figure S11C). Besides, the variation in the number of mutations across samples seemed to not be affected by coverage – even when age and skin phototype were considered (Figure S11D).

To further evaluate the effect of sequencing depth on mutational burden variability across samples, bam files were down-sampled to different depths in order to compare mutational burden from down-sampled datasets with the original mutational burden values. A high correlation was observed between the original and each down-sampled dataset (Figure S4A), even though the impact of sequencing depth on mutational burden estimates was higher with decreasing coverage (the 400x down-sampled dataset presented the highest β value). This observation was also confirmed by estimating the ratio of difference (expressed as fold change) in mutational burden estimates for each increase in coverage metrics (Figure S4B). In fact, fold change values started to plateau with increased sequencing depth.

However, general conclusions of our work remained the same in all down-sampled datasets. Mutational burden variability across samples was mainly explained by age and skin phototype after down-sampling sequencing reads (Figure S4C). Local mutational context also remained similar when deconstructSigs was run on down-sampled datasets.

The sensitivity of mutation detection may also be affected by the number of reads supporting the presence of the variant. The mutational burden detected in each sample was not affected by the average frequency of variant reads per sample (Figure S4D). We also observed similar correlations of mutational burden with each risk factor evaluated when clone size in each sample (measured by the average VAF per sample) was considered, confirming that the effect of risk factors on mutational burden was not dependent on a hypothetical mutation detectability threshold (Figure S4E).

Copy number events in normal skin

As previously shown in normal cells from different tissues [7, 18, 19], our results also suggests that genomic instability is rare in normal skin (only 4 samples with putative copy number changes) and thus structural changes may be a key evolutionary event in carcinogenesis (Figure S8A-B). Copy number alterations across genes were explored using ExomeDepth [17], and confirmed using a method for identifying allelic imbalances of germline heterozygous polymorphisms (evident deviation from the expected fraction of reads supporting one of the two alleles). Note that the ability to detect allelic imbalances was variable across samples and genes because of differences in the number of heterozygous polymorphisms in the region.

NOTCH2 was the gene most frequently subject to copy number aberrations, with five samples having enough statistical power to confirm the duplication or deletion of this gene. Four of these samples also carried a missense or nonsense mutation in *NOTCH2*. We could not confirm the other copy number alterations identified by ExomeDepth.

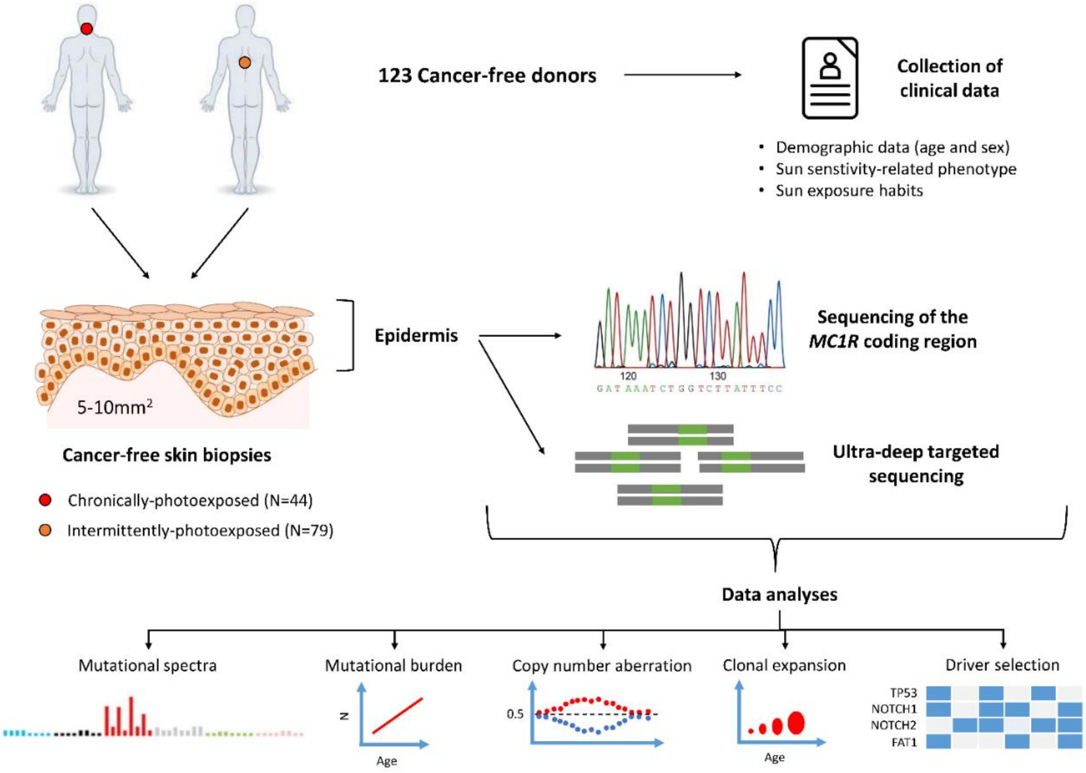
123 Cancer-free donors
Collection of clinical data
• Demographic data (age and sex)
• Sun senstivity-related phenotype
• Sun exposure habits
5-10mm^2
Cancer-free skin biopsies
● Chronically-photoexposed (N=44)
● Intermittently-photoexposed (N=79)
Epidermis
Sequencing of the *MC1R* coding region
120
130
GATAAATCTGGTCTTATTTCC
Ultra-deep targeted sequencing
Data analyses
Mutational spectra
Mutational burden
N
Age
Copy number aberration
0.5
Clonal expansion
Age
Driver selection
TP53
NOTCH1
NOTCH2
FAT1

Figure S1. Schematic overview of the experimental design.

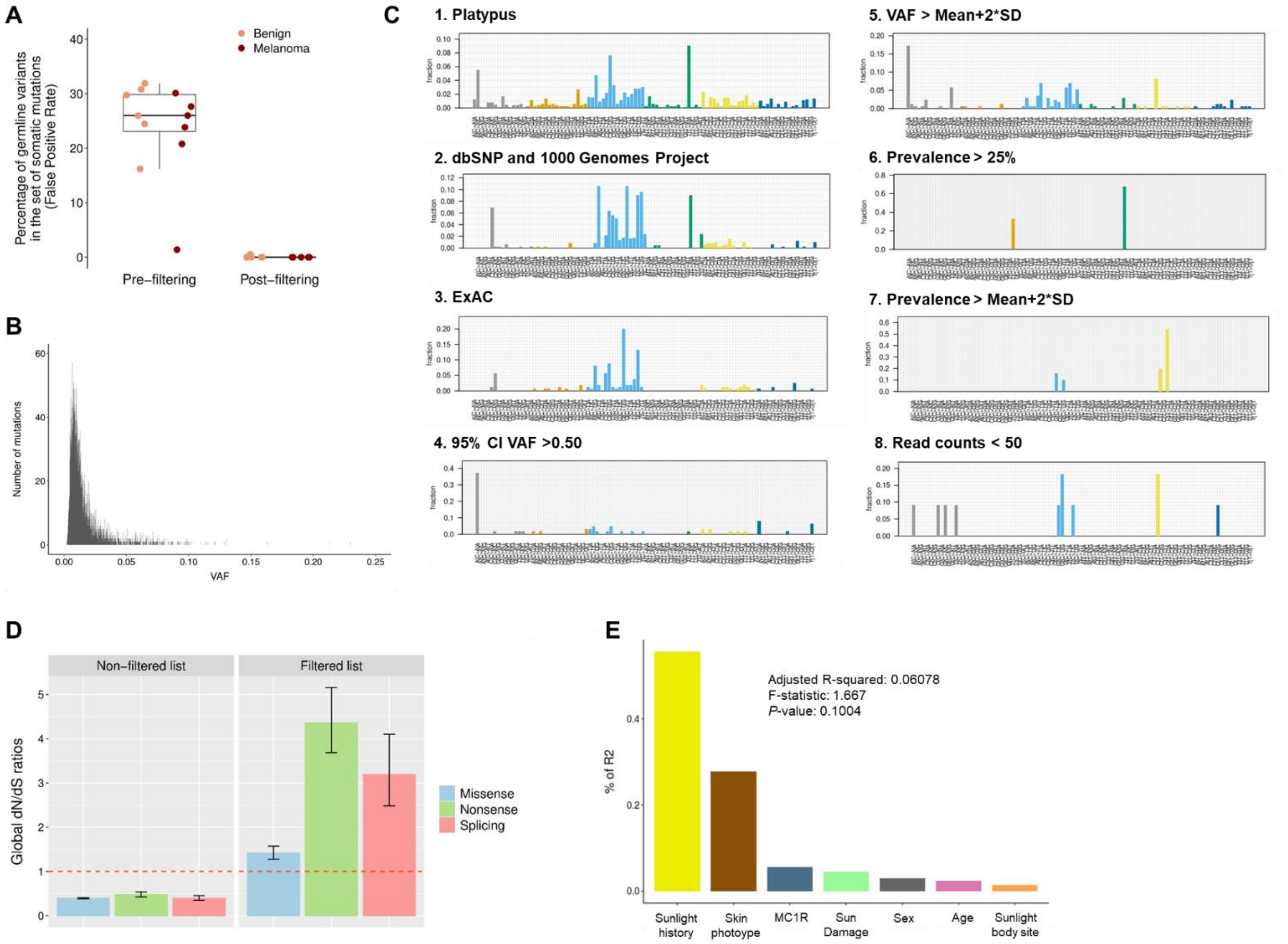

Figure S2. Evaluation of variant calling and filtering. (A) Validation of filtering procedure efficiency in an independent dataset comprising tumour and adjacent benign FFPE samples collected from six melanoma patients. A noteworthy decrease of false positive rates (proportion of germline variants in the set of somatic mutations) was denoted after applying the procedure for filtering out germline variants. **(B)** Histogram of somatic mutations identified by VAF. Most somatic mutations remain in a subclonal state with low VAFs (VAF << 5%). **(C)** Spectra of mutation sets removed after applying a specific filtering step. All mutational spectra are very different from the typical UV-related mutational spectrum, indicating that the filtered variants are unlikely to be real somatic mutations. **(D)** Global dN/dS ratios estimated before and after mutation filtering called with Mutect2 tumour-only mode. The global dN/dS << 1 denotes contamination of germline variants and/or technical artefacts in the non-filtered dataset of somatic mutations. This problem seems to be solved after applying the different filtering steps (dN/dS > 1). Error bars denote 95% confidence interval. **(E)** Results of applying a log-linear regression model in the non-filtered mutation dataset for predicting the number of mutations per sample. The low variance explained by the model (adjusted-R^2 = 6.08%) denotes that the non-filtered list of mutations includes a large number of likely false positive calls.

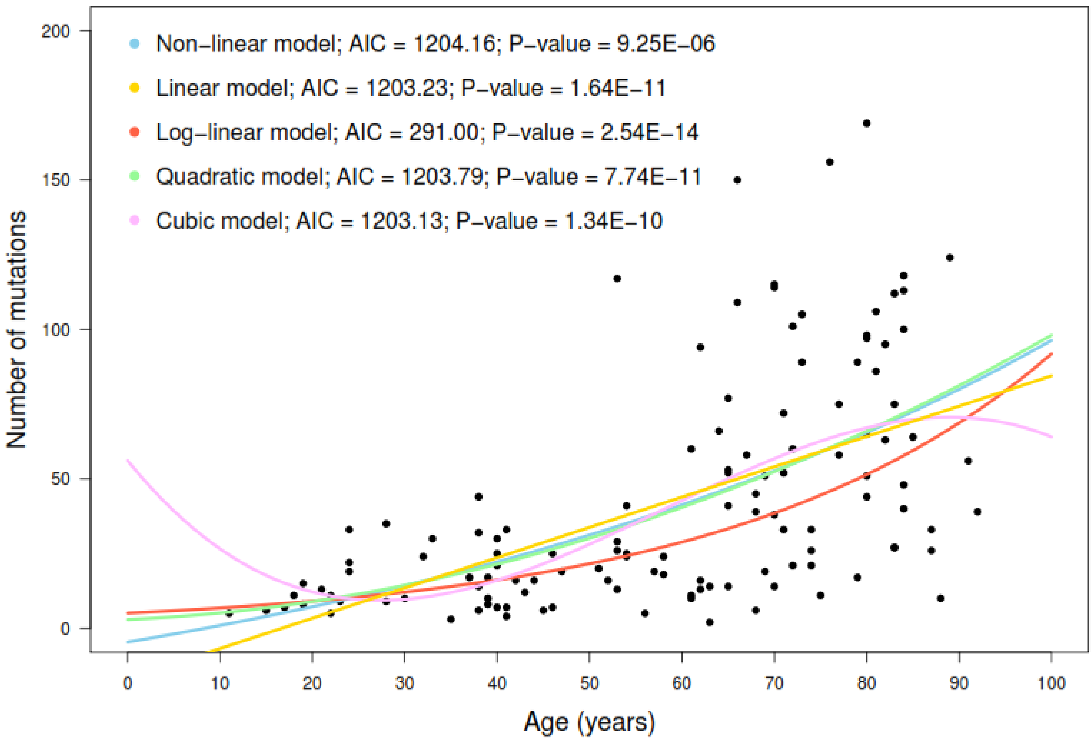

Figure S3. Selection of the best model explaining the age-dependent increase of somatic mutations in normal skin. Model selection was performed using the Akaike information criterion (AIC).

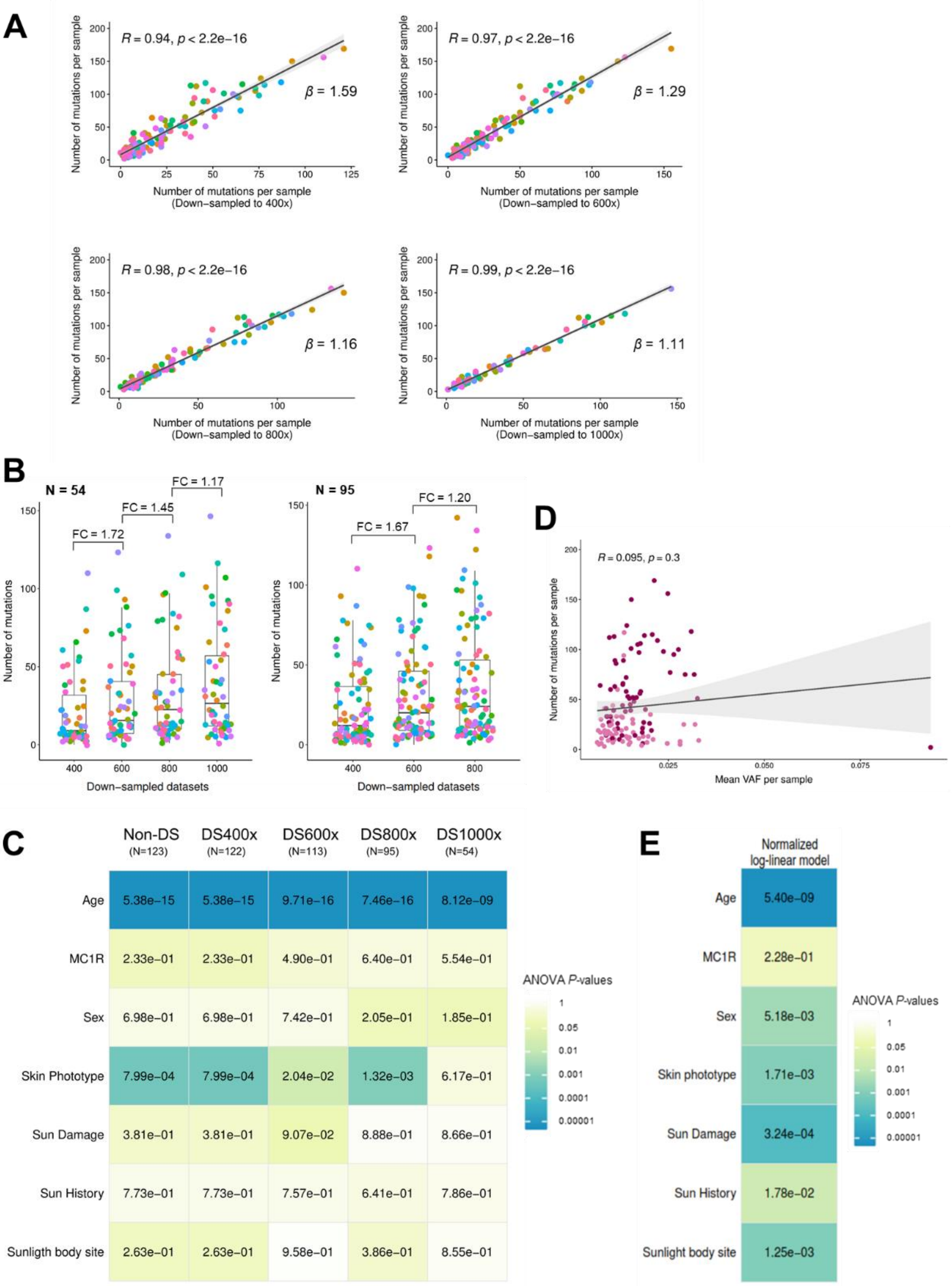

Figure S4. Evaluation the impact of mutation detectability on mutational burden variability across samples. (A) Scatter plots showing a high correlation between the number of mutations predicted from the original dataset and from each down-sampled dataset. *β* values denote the gradient of impact of coverage metric on mutational burden estimates. **(B)** Box plots showing the ratio of increase, expressed as fold change (FC), in mutational burden estimates for each increase in coverage metrics. **(C)** Heatmap showing the analyses-of-variance (ANOVA) *P*-values of multivariate log-linear model coefficients of each dataset. DS, down-sampled. N, sample size of the dataset. **(D)** Scatter plot showing the mean VAF and the number of all mutations found per sample. **(E)** Heatmap showing the *P*-values of univariate log-linear model coefficients from the ANOVA tables. The normalized mutational burden of each sample was calculated by dividing the number of mutations per sample by the mean VAF of all mutations found in the sample. All these plots show that mutation detectability did not significantly influence the number of mutations found across samples.

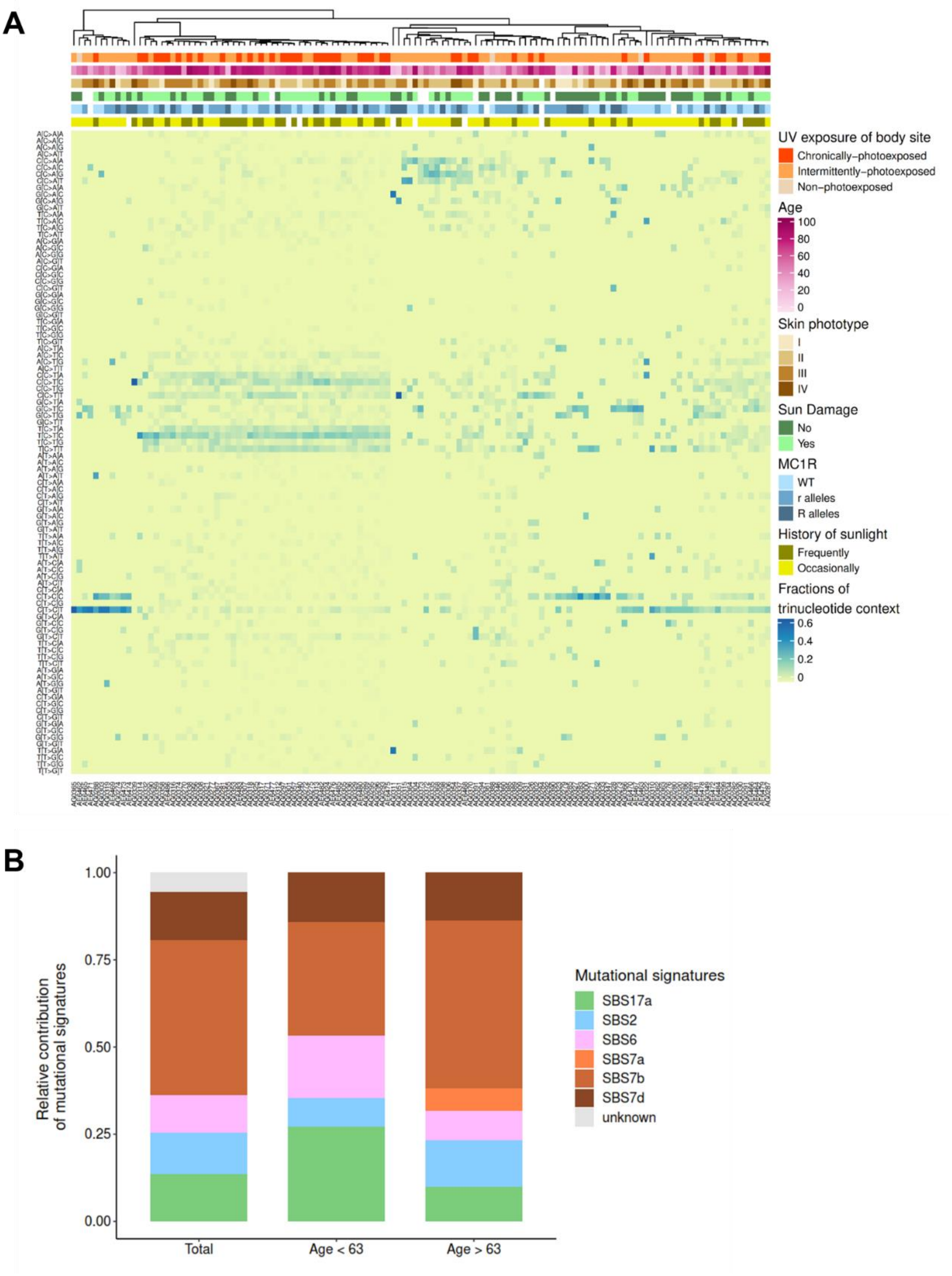

Figure S5. Mutational spectra in normal skin. (A) Heatmap showing the fraction of each 96-mutation type per sample. Clinical and demographic characteristics are presented above each sample. **(B)** Percentage of substitutions attributed to each one of the six mutational signatures for all mutations from all 127 samples together (Total), as well as for all mutations included in each age subgroup.

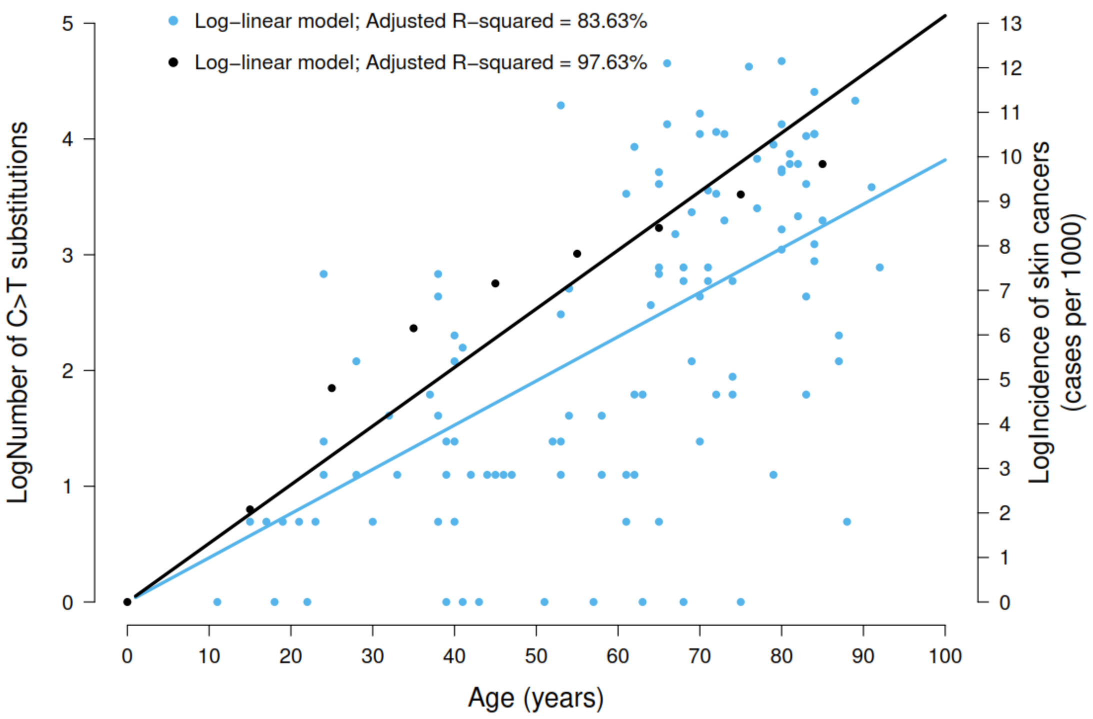

Figure S6. Linearization of the exponential increase of both UV-mutation accumulation and skin cancer incidence with age. Logarithmic transformation of data displayed in Figure 2D. The relatively high R-squared values denote that a high proportion of the total variance in UV-mutation accumulation (blue dots) and in skin cancer incidence (black dots) is explained by the respective log-linear model.

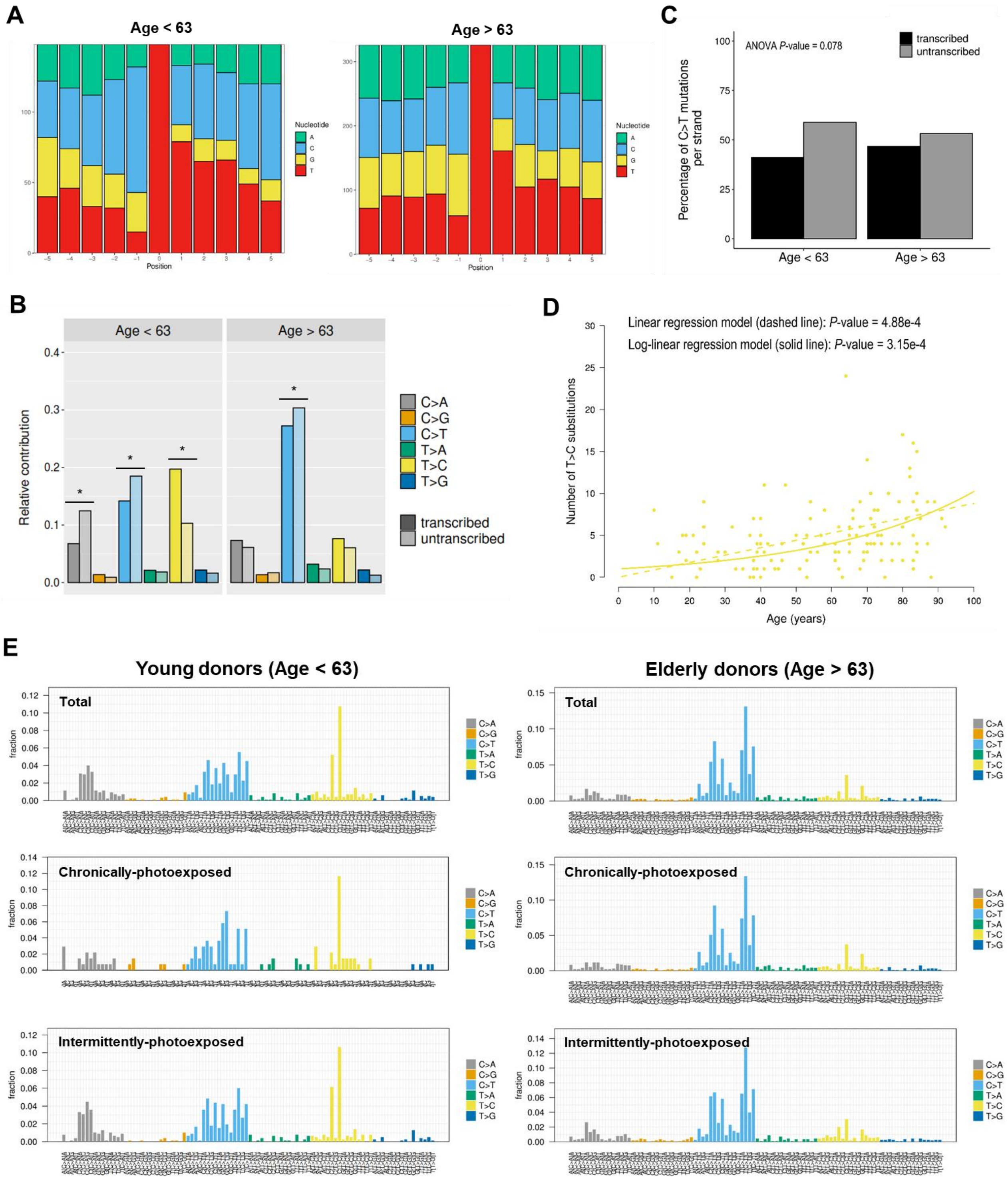

Figure S7. Age-related mutational spectra in normal skin. (A) Local mutational context of T>C substitutions in samples biopsied from young and elderly donors. **(B)** Relative number of each substitution type present on the transcribed (dark shading) and untranscribed strand (light shading) in samples biopsied from young and elderly donors. Asterisks indicate significant transcriptional strand asymmetries (Poisson test). **(C)** Percentage of C>T mutations per strand in young and elderly individuals. ANOVA test used for comparing the ratio of non-coding/coding C>T mutations between age groups. **(D)** Age-dependent increase of T>C substitutions. **(E)** 96-barplot depicting the number of mutations observed at each trinucleotide context taking together all samples biopsied from young and elderly individuals (Total), as well as splitting samples of each age group by the body site pattern of sun exposure (Chronically- and Intermittently-photoexposed).

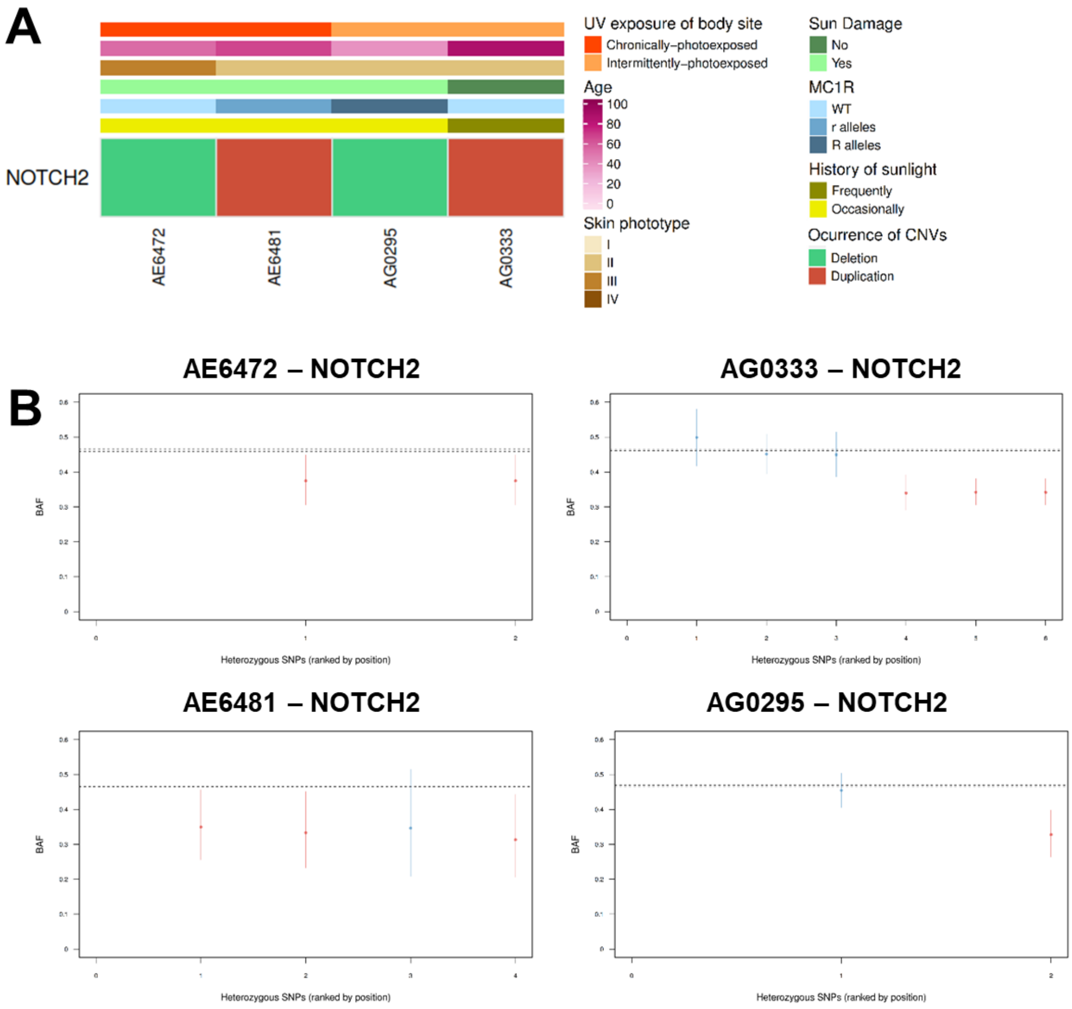

Figure S8. Occurrences of copy number alterations in the 46 cancer genes across samples. (A) Heatmap showing the significant copy number events detected in our cohort. **(B)** Scatter plots of four samples showing allelic imbalances in *NOTCH2*. The b-allele fraction (BAF) and 95% confidence interval of each germline heterozygous polymorphism in *NOTCH2* is shown. Red dots denote a deviation of the observed fraction of reads supporting the minor allele from the expected fraction (dashed lines), which is calculated by averaging the BAFs of all germline heterozygous SNPs in each sample and in all samples.

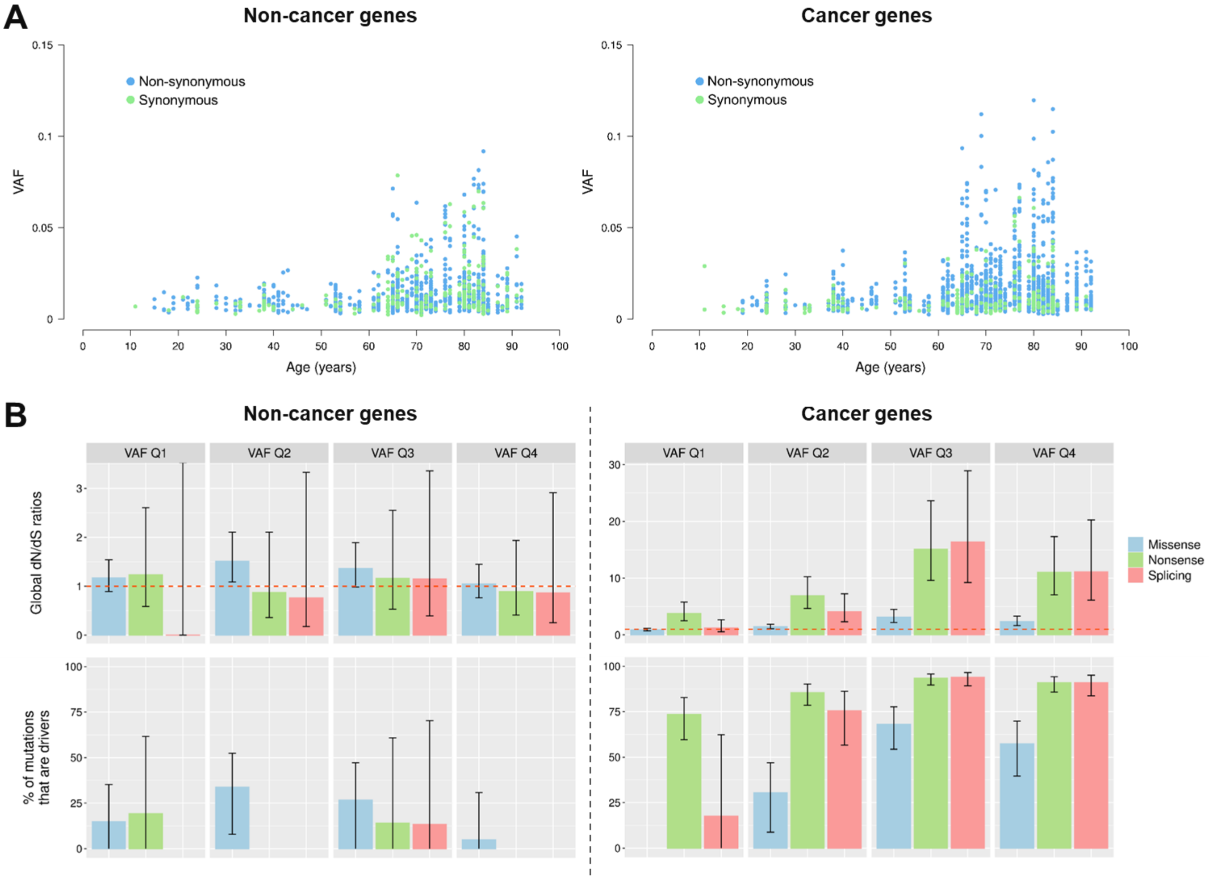

Figure S9. Mutation effect in cell fitness, selection and clonal expansion. (A) Age-related VAF spectra of non-synonymous and synonymous mutations in both cancer and non-cancer genes. **(B)** Global dN/dS values (top) and frequency of driver mutations (bottom) estimated in cancer and non-cancer genes according to mutation frequencies. Percentage of driver mutations was only calculated when dN/dS ratios denoted positive selection (dN/dS > 1). Mutations were divided into four equal parts according to their VAF. VAF Q1, mutations with VAF values below the first quartile. VAF Q2, mutations with VAF values between the first and second quartiles. VAF Q3, mutations with VAF values between the second and third quartiles. VAF Q4, mutations with VAF values above the third quartile.

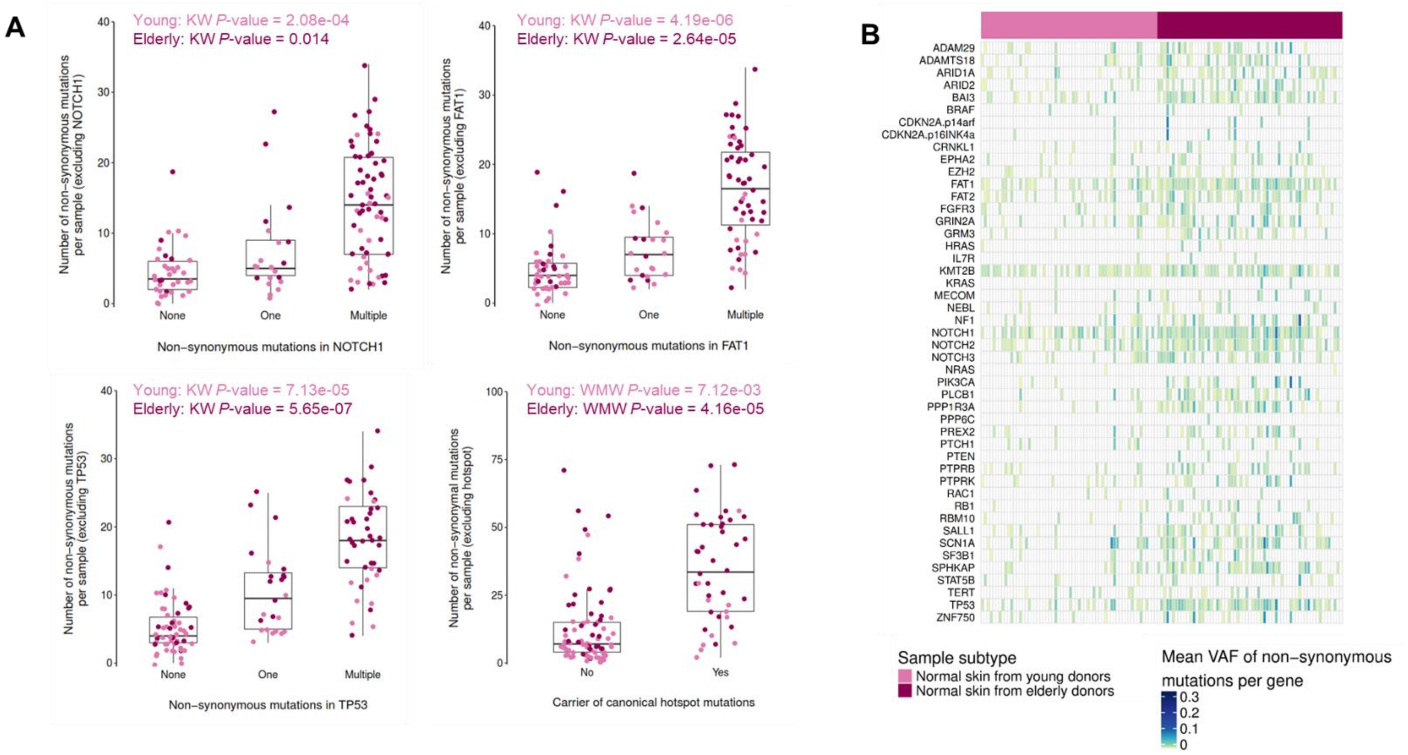

Figure S10. Clonal expansion of clones with oncogenic mutations. (A) Number of non-synonymous mutations per sample in normal skin samples non-carriers or carriers of one or multiple non-synonymous mutations in *NOTCH1*, *TP53* and *FAT1*, as well as in normal skin without or with canonical hotspot mutations. Each dot represents a sample and is coloured according to the donor's age. For avoiding the confounding effects of age, samples were stratified according to donor's age for statistical analyses. In panels comparing more than two groups, a Kruskall-Wallis (KW) test is used for testing differences among groups. In panels comparing two groups, a Wilcoxon-Mann-Whitney (WMW) test is used for testing differences among groups. **(B)** Heatmap showing the mean VAF of all non-synonymous mutations found per gene across normal samples collected from young and elderly individuals.

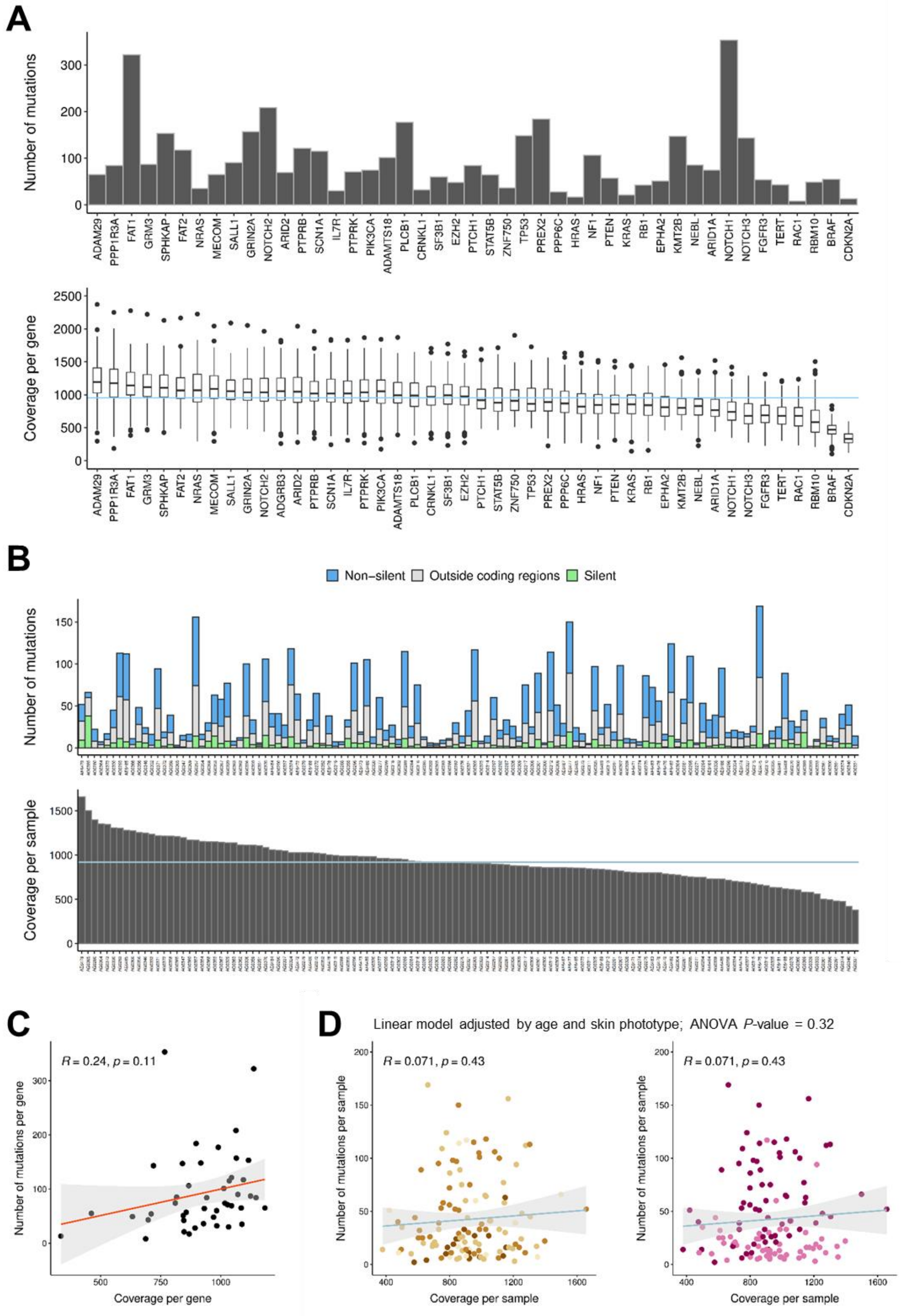

Figure S11. Coverage and mutational burden across genes and samples. (A) Plot showing the number of mutations per gene across all samples (bar plot, top) and the mean coverage per gene and sample (box plot, bottom). Genes in the x-axis sorted by mean coverage across samples. Blue line indicates the mean coverage across all samples. **(B)** Plot showing the number of mutations per sample (bar plot, top) and the mean coverage per sample (bar plot, bottom). Samples in the x-axis sorted by mean coverage across all sequenced regions. Blue line indicates the mean coverage across all samples. **(C)** Scatter plot showing the coverage and number of mutations per gene. **(D)** Scatter plot showing the coverage and number of mutations per sample coloured by skin phototype (left) and per age group (right). These plots show that coverage did not significantly influence the number of mutations found across genes and/or across samples.

Table S1. Demographic and clinical data of all Spanish donors.

		Pattern of sunlight exposure of normal skin samples					
		Chronically (N = 44)		Intermittently (N = 79)		Total (N = 123)	
		Mean	SD	Mean	SD	Mean	SD
Age (years)		69.86	15.99	52.16	20.91	58.50	21.03
		N	%	N	%	N	%
Sex	Females	12	27.27	47	59.49	59	47.97
	Males	32	72.73	32	40.51	64	52.03
	Unknown	0	0.00	0	0.00	0	0.00
Fitzpatrick skin type ‡	I	3	6.82	11	13.92	14	11.38
	II	13	29.55	34	43.04	47	38.21
	III	21	47.73	25	31.65	46	37.40
	IV	6	13.64	9	11.39	15	12.20
	Unknown	1	2.27	0	0.00	1	0.79
***MC1R* genotype**	Wild-type	18	40.91	28	35.44	46	37.40
	r carrier	18	40.91	26	32.91	44	35.77
	R carrier	7	15.91	22	27.85	29	23.58
	Unknown	1	2.27	3	3.80	4	3.25
History of sun exposure ¥	Occasional	18	40.91	54	68.35	72	58.54
	Frequent	22	50.00	21	26.58	43	34.96
	Unknown	4	9.09	4	5.06	8	6.50
Sun damage in the skin area	No	6	13.64	46	58.23	52	42.28
	Yes	35	79.55	27	34.18	62	50.41
	Unknown	3	6.82	6	7.59	9	7.32

N, number of individuals; %, percentage of individuals per group among the total

‡ Fitzpatrick’s skin type classification is based on pigmentation traits (skin, hair and eye color) and sun sensitivity-related traits (ability to tan versus tendency to burn, and freckling degree)

¥ History of sun exposure is based on occupancy, outdoor sport activity, and sunbed use

Table S2. Log-linear modelling of the accumulation of somatic mutations in normal skin.

Variable	Categories	β	SE	P-value ‡	R^2 (%)
Age		0.028	0.004	**1.46E-09**	55.17
Sex	Female	reference	-	-	3.89
	Male	-0.109	0.165	0.51	
Sunlight exposure of body site	Chronic	reference	-	-	7.83
	Intermittent	-0.171	0.179	0.34	
Sun damage	No	reference	-	-	7.95
	Yes	0.187	0.168	0.27	
Skin phototype	I	reference	-	-	17.93
	II	-0.387	0.266	0.11	
	III	-0.414	0.253	0.15	
	IV	-1.220	0.312	**1.74E-04**	
MC1R genotype	wild-type	reference	-	-	2.00
	r carrier	0.104	0.180	0.56	
	R carrier	-0.226	0.191	0.24	
History of sunlight exposure	Frequent	reference	-	-	5.23
	Occasional	-0.189	0.161	0.24	

β, coefficients; SE, standard error; R^2, percentage of relative contribution of each predictor to the total variance

‡ *P*-value for the multivariate log-linear model

Table S3. Information from literature about the function and the role in carcinogenesis of the list of genes sequenced in this study

Gene name	Description	Evidences from function ‡	Gene classification ¥
ADAM29	ADAM Metallopeptidase Domain 29	Membrane-anchored proteins implicated in a variety of biological processes involving cell-cell and cell-matrix interactions	Non-cancer
ADAMTS18	ADAM Metallopeptidase with Thrombospondin Type 1 Motif 18	Metalloproteinase with thrombospondin motifs that regulates hemostatic balance and functions as a tumor suppressor	Non-cancer
ARID1A	AT-Rich Interaction Domain 1A	Involved in transcriptional activation and repression of select genes by chromatin remodeling. Component of SWI/SNF chromatin remodeling complexes with key enzymatic activities.	Cancer
ARID2	AT-Rich Interaction Domain 2	Involved in transcriptional activation and repression of select genes by chromatin remodeling (alteration of DNA-nucleosome topology). Required for the stability of the SWI/SNF chromatin remodeling complex	Cancer
BAI3/ADGRB3	Brain-Specific Angiogenesis Inhibitor 3	A p53-target gene that encodes for an angiogenesis inhibitor	Non-cancer
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase	Oncogene that encodes for a serine/threonine protein kinase. Involved in the regulation of the MAP kinase/ERK signaling pathway affecting cell division, differentiation and secretion.	Cancer
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A	Tumor suppressor gene encoding p16 and p14. Involved in cell cycle arrest in G1 and G2 phases	Cancer
CRNKL1	Crooked Neck Pre-MRNA Splicing Factor 1	Involved in pre-mRNA splicing process	Non-cancer
EPHA2	Ephrin Receptor A2	Receptor tyrosine kinase which regulates cell adhesion and differentiation through DSG1/desmoglein-1 and inhibition of the ERK1/ERK2 signaling pathway. May also participate in UV radiation-induced apoptosis.	Cancer
EZH2	Enhancer Of Zeste 2 Polycomb Repressive Complex 2 Subunit	Catalytic subunit of the PRC2/EED-EZH2 complex. Involved in maintaining the transcriptional repressive state of genes via histone H3 methylation	Cancer
FAT1	FAT Atypical Cadherin 1	Tumor suppressor that plays an essential role for cellular polarization, directed cell migration and modulating cell-cell contact	Cancer
FAT2	FAT Atypical Cadherin 2	Functions as a cell adhesion molecule, controlling cell proliferation and playing an important role in cerebellum development	Cancer
FGFR3	Fibroblast Growth Factor Receptor 3	Tyrosine-protein kinase that acts as cell-surface receptor for fibroblast growth factors and plays an essential role in the regulation of cell proliferation, differentiation and apoptosis	Cancer
GRIN2A	Glutamate Ionotropic Receptor NMDA Type Subunit 2A	Component of NMDA receptor complexes that function as heterotetrameric, ligand-gated ion channels with high calcium permeability and voltage-dependent sensitivity to magnesium.	Non-cancer
GRM3	Glutamate Metabotropic Receptor 3	G-protein coupled receptor for glutamate. Signaling inhibits adenylate cyclase activity.	Non-cancer
HRAS	HRas Proto-Oncogene, GTPase	Involved in the activation of Ras protein signal transduction	Cancer
IL7R	Interleukin 7 Receptor	Receptor for interleukin-7	Non-cancer
KMT2B	Lysine Methyltransferase 2B	Histone methyltransferase that methylates 'Lys-4' of histone H3, a specific tag for epigenetic transcriptional activation.	Non-cancer
KRAS	KRAS Proto-Oncogene, GTPase	Possess intrinsic GTPase activity and plays an important role in the positive regulation of cell proliferation	Cancer
MECOM	MDS1 And EVI1 Complex Locus	Transcriptional regulator and oncoprotein that may be involved in hematopoiesis, apoptosis, development, and cell differentiation and proliferation.	Non-cancer
NF1	Neurofibromin 1	Negative regulator of the Ras signal transduction pathway.	Cancer
NEBL	Nebulette	May functionally link sarcomeric actin to the desmin intermediate filaments in the heart muscle sarcomeres	Non-cancer
NOTCH1	Notch Receptor 1	Involved in the Notch signaling pathway, a evolutionarily conserved intercellular signaling pathway that regulates cell fate decision and affects the implementation of differentiation, proliferation and apoptotic programs.	Cancer

NOTCH2	Notch Receptor 2	Member of the Notch family that plays a role in a variety of developmental processes by controlling cell fate decisions. Involved in immune system function, tissue repair and bone remodeling.	Cancer
NOTCH3	Notch Receptor 3	Member of the Notch family that plays a key role in the function and survival of vascular smooth muscle cells and in neural development	Non-cancer
NRAS	NRAS Proto-Oncogene, GTPase	Possess intrinsic GTPase activity and plays an important role in the positive regulation of cell proliferation	Cancer
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha	Oncogene encoding a phosphoinositide-3-kinase that activates signaling cascades involved in cell growth, survival, proliferation, motility and morphology.	Cancer
PLCB1	Phospholipase C Beta 1	Catalyzes the formation of inositol 1,4,5-trisphosphate and diacylglycerol from phosphatidylinositol 4,5-bisphosphate	Cancer
PPP1R3A	Protein Phosphatase 1 Regulatory Subunit 3A	Participates in the regulation of glycogen metabolism, muscle contractility and protein synthesis	Non-cancer
PPP6C	Protein Phosphatase 6 Catalytic Subunit	Component of a signaling pathway regulating cell cycle progression in response to IL2 receptor stimulation	Cancer
PREX2	Phosphatidylinositol-3,4,5-Trisphosphate Dependent Rac Exchange Factor 2	Functions as a RAC1 guanine nucleotide exchange factor (GEF), activating Rac proteins by exchanging bound GDP for free GTP	Cancer
PTCH1	Patched 1	Component of the hedgehog signaling pathway involved in embryonic development and tumorigenesis	Cancer
PTEN	Phosphatase And Tensin Homolog	Tumor suppressor that antagonizes the PI3K-AKT/PKB signaling pathway modulating cell cycle progression and cell survival	Cancer
PTPRB	Protein Tyrosine Phosphatase Receptor Type B	Plays an important role in blood vessel remodeling and angiogenesis	Non-cancer
PTPRK	Protein Tyrosine Phosphatase Receptor Type K	Regulation of processes involving cell contact and adhesion such as growth control, tumor invasion, and metastasis.	Non-cancer
RAC1	Rac Family Small GTPase 1	GTPase that binds to a variety of effector proteins to regulate cellular responses such as secretory processes, phagocytosis of apoptotic cells, epithelial cell polarization, neurons adhesion, migration and differentiation	Cancer
RB1	RB Transcriptional Corepressor 1	Key regulator of the cell cycle that acts as a tumor suppressor. Hypophosphorylated form of the protein binds transcription factor E2F1, leading to cell cycle arrest	Cancer
RBM10	RNA Binding Motif Protein 10	Nuclear protein that may be involved in post-transcriptional processing, most probably in mRNA splicing	Cancer
SALL1	Spalt Like Transcription Factor 1	Zinc finger transcriptional repressor involved in organogenesis	Non-cancer
SCN1A	Sodium Voltage-Gated Channel Alpha Subunit 1	Mediates the voltage-dependent sodium ion permeability of excitable membranes	Non-cancer
SF3B1	Splicing Factor 3b Subunit 1	Involved in pre-mRNA splicing	Cancer
SPHKAP	SPHK1 Interactor, AKAP Domain Containing	Anchoring protein that may act as a converging factor linking cAMP and sphingosine signaling pathways	Cancer
STAT5B	Signal Transducer And Activator Of Transcription 5B	After being phosphorilated in response to cytokines and growth factos, this STAT family member tanslocates to the nucleous and acts as transcription factor	Non-cancer
TERT	Telomerase Reverse Transcriptase	Ribonucleoprotein polymerase that maintains telomere ends by addition of the telomere repeat TTAGGG	Non-cancer
TP53	Tumor Protein P53	The encoded tumor suppressor protein responds to diverse cellular stresses to induce cell cycle arrest, apoptosis, senescence, DNA repair or changes in metabolism	Cancer
ZNF750	Zinc Finger Protein 750	Transcription factor involved in epidermis differentiation.	Cancer

‡ Information taken from GeneCards (http://www.genecards.org)

¥ Classification based on www.cancer-genes.org

SUPLEMENTARY REFERENCES

1. Martínez-Cadenas C, López S, Ribas G et al. Simultaneous purifying selection on the ancestral MC1R allele and positive selection on the melanoma-risk allele V60L in south Europeans. Mol. Biol. Evol. 2013; 30(12):2654–2665.

2. Hernando B, Ibañez MV, Deserio-Cuesta JA et al. Genetic determinants of freckle occurrence in the Spanish population: Towards ephelides prediction from human DNA samples. Forensic Sci. Int. Genet. 2018; 33:38–47.

3. Hayward NK, Wilmott JS, Waddell N et al. Whole-genome landscapes of major melanoma subtypes. Nature 2017; 545(7653):175–180.

4. Inman GJ, Wang J, Nagano A et al. The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. Nat. Commun. 2018; 9(1):3667.

5. Iranzo J, Martincorena I, Koonin EV. Cancer-mutation network and the number and specificity of driver mutations. Proc. Natl. Acad. Sci. U. S. A. 2018; 115(26):E6010–E6019.

6. Jayaraman SS, Rayhan DJ, Hazany S, Kolodney MS. Mutational landscape of basal cell carcinomas by whole-exome sequencing. J. Invest. Dermatol. 2014; 134(1):213–220.

7. Martincorena I, Roshan A, Gerstung M et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 2015; 348(6237):880–886.

8. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009; 25(14):1754–1760.

9. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–1303.

10. Cingolani P, Platts A, Wang LL et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012; 6(2):80–92.

11. Cingolani P, Patel VM, Coon M et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front. Genet. 2012; 3:35.

12. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. 2013; 34(9):E2393-2402.

13. Rimmer A, Phan H, Mathieson I et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 2014; 46(8):912–918.

14. Rosenthal R, McGranahan N, Herrero J et al. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016; 17:31.

15. Durinck S, Moreau Y, Kasprzyk A et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma. Oxf. Engl. 2005; 21(16):3439–3440.

16. Martincorena I, Raine KM, Gerstung M et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017; 171(5):1029-1041.e21.

17. Plagnol V, Curtis J, Epstein M et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinforma. Oxf. Engl. 2012; 28(21):2747–2754.

18. Lee-Six H, Olafsson S, Ellis P et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 2019; 574(7779):532–537.

19. Martincorena I, Fowler JC, Wabik A et al. Somatic mutant clones colonize the human esophagus with age. Science 2018; 362(6417):911–917.