


# The genomic landscapes of individual melanocytes from human skin

<https://doi.org/10.1038/s41586-020-2785-8>

Received: 7 April 2020

Accepted: 3 September 2020

Published online: 07 October 2020

 Check for updates

Jessica Tang<sup>1,2,5</sup>, Eleanor Fewings<sup>1,2,5</sup>, Darwin Chang<sup>1,2</sup>, Hanlin Zeng<sup>3</sup>, Shanshan Liu<sup>1,2</sup>, Aparna Jorapur<sup>1,2</sup>, Rachel L. Belote<sup>3,4</sup>, Andrew S. McNeal<sup>1,2</sup>, Tuyet M. Tan<sup>1,2</sup>, Iwei Yeh<sup>1,2</sup>, Sarah T. Arron<sup>1,2</sup>, Robert L. Judson-Torres<sup>3,4,6</sup>, Boris C. Bastian<sup>1,2,6</sup> & A. Hunter Shain<sup>1,2,6</sup>✉

Every cell in the human body has a unique set of somatic mutations, but it remains difficult to comprehensively genotype an individual cell<sup>1</sup>. Here we describe ways to overcome this obstacle in the context of normal human skin, thus offering a glimpse into the genomic landscapes of individual melanocytes from human skin. As expected, sun-shielded melanocytes had fewer mutations than sun-exposed melanocytes. However, melanocytes from chronically sun-exposed skin (for example, the face) had a lower mutation burden than melanocytes from intermittently sun-exposed skin (for example, the back). Melanocytes located adjacent to a skin cancer had higher mutation burdens than melanocytes from donors without skin cancer, implying that the mutation burden of normal skin can be used to measure cumulative sun damage and risk of skin cancer. Moreover, melanocytes from healthy skin commonly contained pathogenic mutations, although these mutations tended to be weakly oncogenic, probably explaining why they did not give rise to discernible lesions. Phylogenetic analyses identified groups of related melanocytes, suggesting that melanocytes spread throughout skin as fields of clonally related cells that are invisible to the naked eye. Overall, our results uncover the genomic landscapes of individual melanocytes, providing key insights into the causes and origins of melanoma.

Cutaneous melanomas are skin cancers that arise from melanocytes, the pigment-producing cells in the skin. Thousands of melanomas have been sequenced, and the results have revealed a high burden of somatic mutations with patterns that implicate sunlight as the major mutagen responsible for their formation. It is unclear when these mutations are acquired during the course of tumorigenesis and whether their rate of accumulation accelerates during neoplastic transformation.

In normal skin, melanocytes reside within the penetrable range of ultraviolet (UV)-A and UV-B radiation in the basilar epidermis. They make up a minor fraction of the cells in the epidermis, which is mainly comprised of keratinocytes. Keratinocytes have a p53-dependent program that triggers apoptosis after exposure to high doses of UV radiation, resulting in the sloughing off of epidermal sheets after sunburn<sup>2</sup>. As a result, clonal patches of *TP53*-mutant keratinocytes are prevalent in sun-exposed skin<sup>3,4</sup>, and these can eventually give rise to keratinocyte cancers.

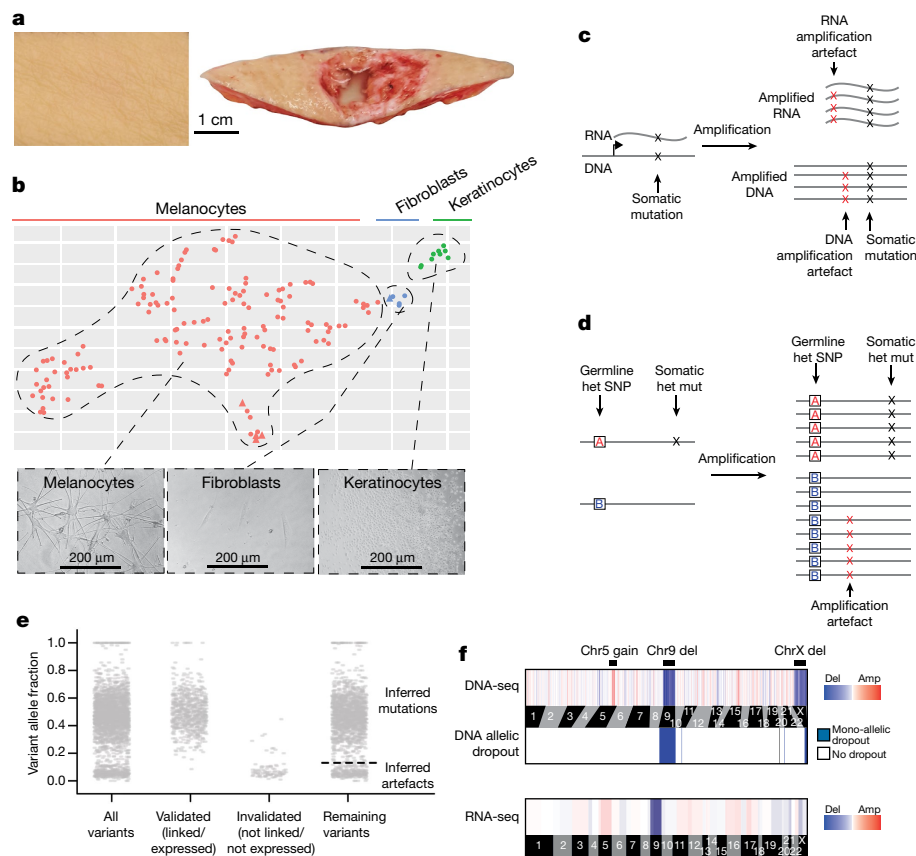
By contrast, the homeostatic mechanisms that govern melanocytes and the selective pressures that operate on these cells during early phases of transformation are less well understood. Although some melanomas arise from naevi (common moles), most arise in the absence of a precursor lesion. Understanding the mutational processes and kinetics of mutation acquisition in pre-malignant melanocytes of normal skin would provide important insights into the early phases

of transformation, before clinically visible neoplastic proliferations have formed.

Most DNA sequencing studies have been performed on bulk groups of cells, yielding an average signal from the complex mixture of cells that are sampled. Bulk-cell sequencing of normal blood<sup>5</sup>, skin<sup>4</sup>, oesophageal mucosa<sup>6</sup> and colonic crypts<sup>7</sup> has identified mutations in these tissues, including the presence of pathogenic mutations, and offered valuable insights into the earliest phases of carcinogenesis in these tissue types. However, bulk-cell sequencing is not suitable for detecting mutations in melanocytes because melanocytes are sparsely distributed in the skin<sup>4</sup>.

Genotyping studies at the resolution of individual cells are rare and, to our knowledge, none has been performed on melanocytes. It is difficult to genotype an individual cell because there is only one molecule of double-stranded DNA corresponding to each parental allele in a diploid cell. There are two primary strategies to overcome this bottleneck. First, an individual cell can be sequenced after its genomic DNA has been amplified *in vitro*<sup>8,9</sup>. Unfortunately, *in vitro* amplification regularly fails over large stretches of the genome, reducing the sensitivity for detecting mutations, and errors are frequently incorporated during amplification, which diminishes the specificity of subsequent mutation calls<sup>1</sup>. Alternatively, a cell can be clonally expanded in tissue culture, before sequencing, to increase the amount of genomic starting material<sup>10–13</sup>; however, only limited types of primary human cells can

<sup>1</sup>Department of Dermatology, University of California San Francisco, San Francisco, CA, USA. <sup>2</sup>Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA. <sup>3</sup>Department of Dermatology, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>4</sup>Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA. <sup>5</sup>These authors contributed equally: Jessica Tang, Eleanor Fewings. <sup>6</sup>These authors jointly supervised this work: Robert L. Judson-Torres, Boris C. Bastian. ✉e-mail: hunter.shain@ucsf.edu



**Fig. 1 | A workflow to genotype individual skin cells. a**, Examples of healthy skin from which we genotyped individual cells. Left, skin from the back of a cadaver; right, skin surrounding a basal cell carcinoma. **b**, Expression profiles classify the cells that we genotyped into their respective lineages. Each cell is depicted in a  $t$ -distributed stochastic neighbour ( $t$ -SNE) plot and coloured according to morphology. Five cells were engineered (Methods, depicted as triangles). See Extended Data Fig. 1b, c for further details on cell identity. **c, d**, Patterns to distinguish true mutations from amplification artefacts. **c**, Mutations in expressed genes are evident in both DNA and RNA sequencing data, whereas amplification artefacts are not. **d**, Germline polymorphisms (A and B alleles) are in linkage with somatic mutations but not amplification artefacts. Het, heterozygous; SNP, single nucleotide polymorphism; Mut, mutation. **e**, Variant

allele fractions (VAFs) from an example cell indicate how we inferred the mutational status of variants outside the expressed and phase-able portions of the genome. Variants that were validated as somatic mutations had VAFs around 1 or 0.5, and variants that were invalidated had lower VAFs; however, PCR biases sometimes skewed these allele fractions. Variants that could not be directly validated or invalidated were inferred by their VAF (Methods). The dotted line indicates the optimal VAF cutoff to distinguish somatic mutations from amplification artefacts for the variants of this particular cell (Extended Data Fig. 2b). **f**, Copy number was inferred from DNA and RNA sequencing (DNA-seq and RNA-seq) depth as well as from allelic imbalance. An example of a cell with a gain over chromosome 5q, loss of chromosome 9 and loss of the X chromosome is shown.

expand sufficiently in tissue culture, reducing the scope of this strategy. Here, we combine elements of each strategy to genotype melanocytes from normal skin at single-cell resolution.

### A workflow to genotype individual skin cells

We collected clinically normal skin from 19 sites across 6 donors. Skin biopsies were obtained from cadavers of individuals with no history of skin cancer or from the peritumoral tissue of donors with skin cancer (Fig. 1a). All donors had light skin tone and European ancestry (Extended Data Fig. 1a), and they ranged from 63 to 85 years in age.

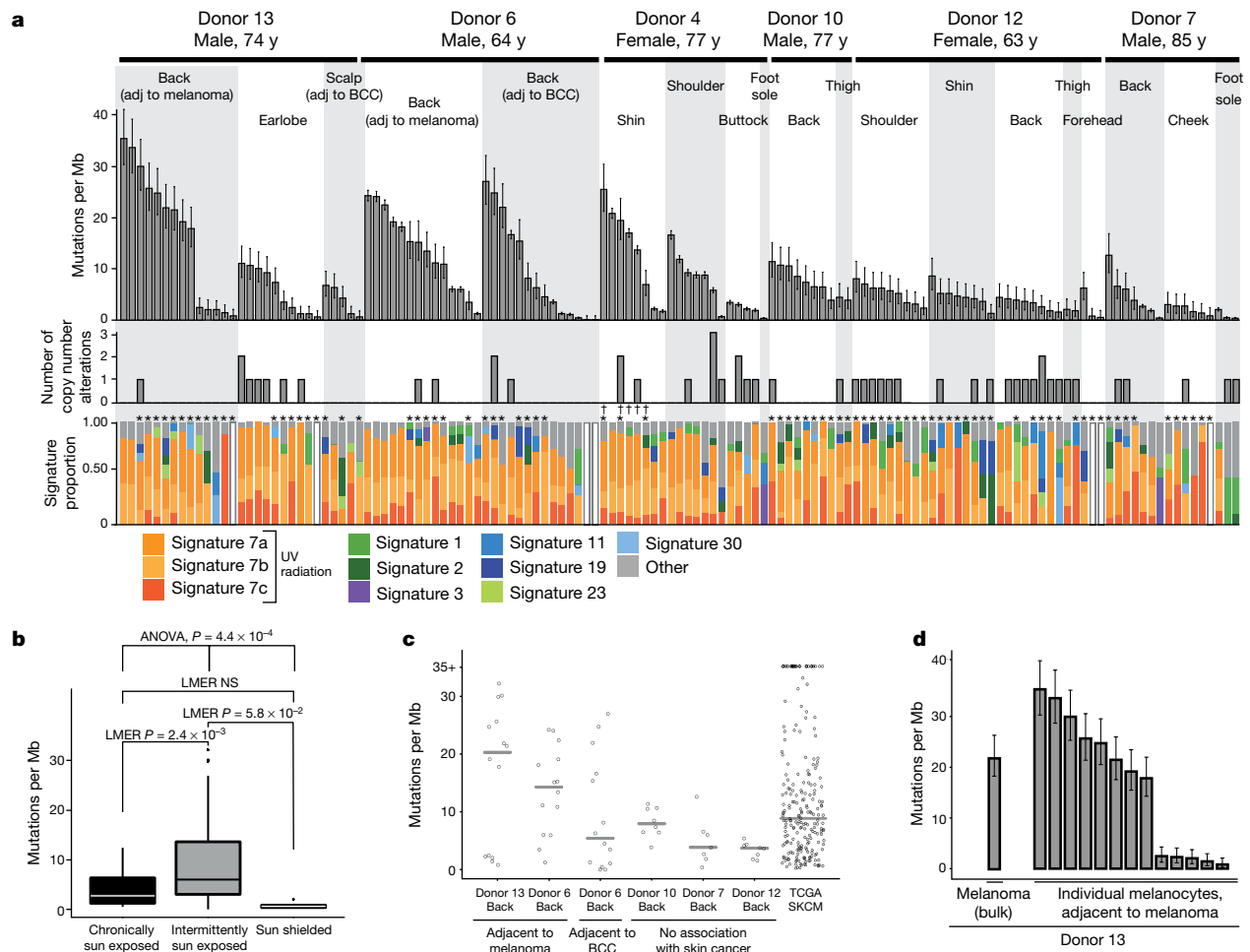
From each skin biopsy, epidermal cells were established in tissue culture for approximately two weeks and subsequently single-cell-sorted and clonally expanded. On average, 38% of flow-sorted melanocytes produced colonies, ranging from 2 to 3,000 cells (median 184 cells, Supplementary Table 1), indicating that we are studying a prevalent and representative population. Despite the small size of these colonies, there was sufficient starting material to achieve an allelic dropout rate of only 0.14% (Extended Data Fig. 2a).

Next, we extracted, amplified and sequenced both DNA and RNA from each clonal expansion, as previously described<sup>14,15</sup>. Our tissue

culture conditions were tailored for melanocyte growth, but some keratinocytes and fibroblasts also grew out. The RNA sequencing data confirmed the identity of each cell (Fig. 1b, Extended Data Fig. 1b, c). The matched DNA and RNA sequencing data also permitted comparisons of genotype and phenotype, as described below.

Polymerases often introduce errors during amplification, and these artefacts can be difficult to distinguish from somatic mutations. The matched DNA and RNA sequencing data improved the specificity of mutation calls because mutations in expressed genes could be cross-validated, whereas amplification artefacts arise independently during DNA and RNA amplifications and thus do not overlap (Fig. 1c). To further improve the specificity of mutation calls, we leveraged haplotype information to identify amplification artefacts. When reads are phased into their maternal and paternal haplotypes using heterozygous germline variants, neighbouring somatic mutations occur within all amplified copies of that haplotype, whereas amplification artefacts rarely display this pattern<sup>16,17</sup> (Fig. 1d).

We were able to distinguish true somatic mutations from amplification artefacts confidently in portions of the genome that were expressed and/or could be phased. Variants that fell outside these regions were classified as somatic mutations or artefacts on the basis



**Fig. 2 | The genomic landscape of individual melanocytes from physiologically normal human skin.** **a**, Top, mutation burden of melanocytes from physiologically normal skin of six donors across different anatomic sites (adj, adjacent; BCC, basal cell carcinoma). Middle, number of copy number alterations identified within each melanocyte. Bottom, the proportion of the mutations of each cell that are attributable to established mutational signatures. Each bar represents one cell ( $n=1$ ). Error bars, 95% confidence intervals (CIs) by two-sided Poisson test. White bars indicate that there were too few mutations for signature analysis. Asterisks, samples that underwent only targeted DNA-seq; crosses, *CDKN2A*-engineered cells. **b**, Comparisons between mutation burdens of chronically sun-exposed ( $n=24$ ), intermittently sun-exposed ( $n=105$ ) and sun-shielded sites ( $n=4$ ). An analysis of variance

(ANOVA), comparing the results of linear mixed-effect models both including and excluding sun exposure to account for repeated donor measurements, gave  $P = 4.43 \times 10^{-4}$ , showing that sun exposure has a significant effect on mutation burden. Pairwise  $P$  values from the linear mixed-effects model are also shown (LMER  $P$ ). Box plots show median and 25th and 75th percentiles; whiskers extend to the largest and smallest values no further than  $1.5 \times$  the interquartile range; dots show outliers. **c**, Mutation burdens of site-matched melanocytes adjacent to cancer compared with not adjacent to cancer. Melanoma mutation burdens from The Cancer Genome Atlas (TCGA) are shown as a reference. Grey lines, median. **d**, Mutation burdens of melanocytes as compared to an adjacent melanoma. Each bar represents one cell ( $n=1$ ). Error bars, 95% CIs by two-sided Poisson test.

of their variant allele frequencies. Heterozygous mutations should have allele frequencies of 50%, whereas amplification artefacts tend to have much lower allele frequencies. For each cell, we identified the variant-allele-frequency cutoff that would maximize the specificity and sensitivity of somatic mutation calls by comparing the variant allele frequencies of known somatic mutations and known amplification artefacts in the expressed and phase-able portions of the genome (Fig. 1e, Extended Data Fig. 2b–d).

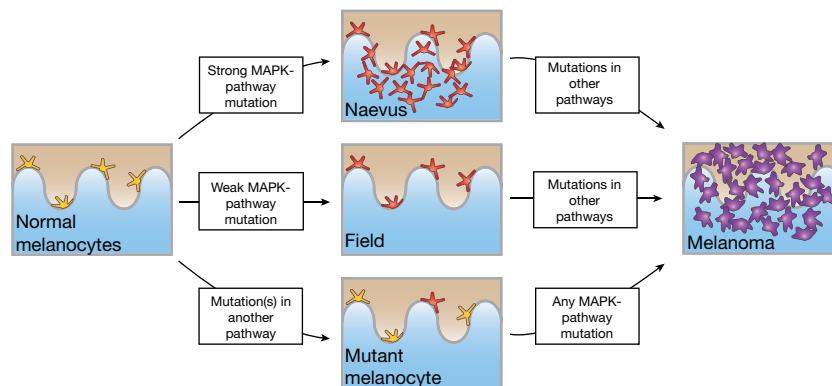
To assess the quality of mutation calls, we explored the genomic contexts of somatic mutations and amplification artefacts classified by each of the methods described above (Extended Data Fig. 3). Somatic mutations—whether ascertained by cross-referencing RNA sequencing data or from their haplotype distribution, or inferred by their allele frequency—showed a pattern similar to signature 7, which is known to be associated with exposure to UV radiation. By contrast, amplification artefacts were more similar to signatures scE and scF, which have been recently defined as likely artefacts resulting from multiple displacement amplification<sup>18</sup>.

Finally, we deduced copy number alterations from both the DNA and RNA sequencing data using the CNVkit software suite<sup>19,20</sup>. As an additional filter, we required that copy number alterations coincided with a concordant degree of allelic imbalance over the region affected (Fig. 1f).

In summary, we implemented a series of experimental and bioinformatic solutions to overcome the major obstacles associated with genotyping individual melanocytes. One hundred and thirty-three melanocytes passed our quality control metrics and were included in all subsequent analyses. Tissue pictures, cellular morphologies and genomic features are shown for each melanocyte in an extended dataset hosted by Figshare (<https://doi.org/10.6084/m9.figshare.11794296.v1>)<sup>21</sup>.

## Mutational landscape of melanocytes

For each clone, we performed RNA sequencing of the entire transcriptome and DNA sequencing on a panel of 509 cancer-relevant genes (Supplementary Table 2). For a subset of 48 cells we performed an



**Fig. 3 | Distinct trajectories of melanoma evolution.** On the basis of the data shown here and previously, we propose that melanomas can evolve via distinct trajectories depending upon the order in which mutations occur.

additional round of DNA sequencing over the entire exome, providing more power to measure the mutational signatures of those cells. The mean numbers of mutations per cell from targeted and whole-exome sequencing were 37 and 790, respectively.

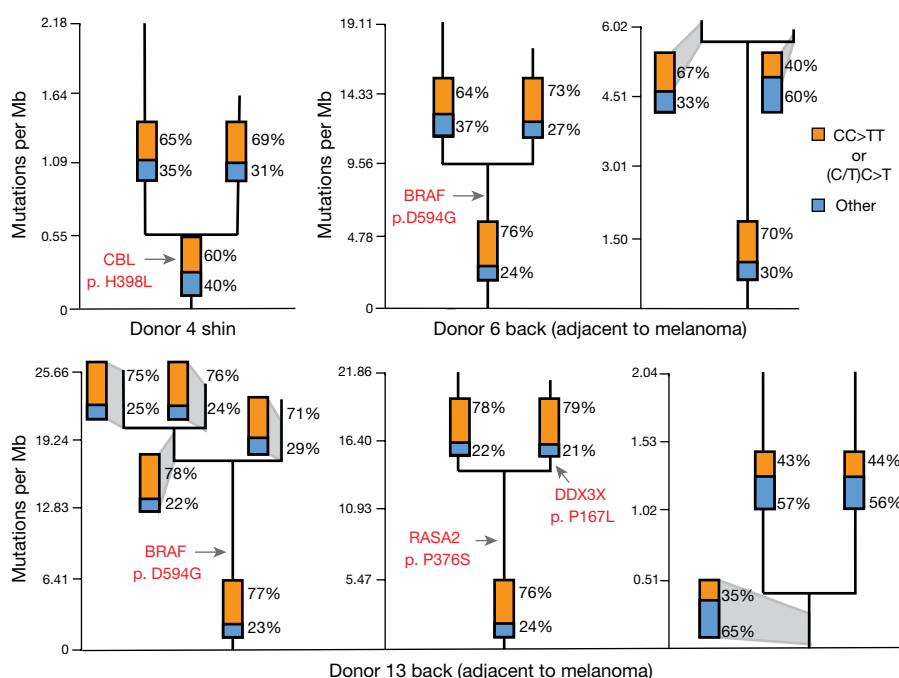
We observed an average mutation burden of 7.9 mutations per megabase (Mb); however, this ranged from less than 0.82 to 32.3 mutations per Mb, and depended on several factors (Supplementary Table 3). First, the mutation burdens of melanocytes varied within people by anatomic site. As expected, melanocytes from sun-shielded sites had fewer mutations than those on sun-exposed sites (Fig. 2a, b, Extended Data Fig. 4). Consistently, sun-shielded melanocytes had little evidence of UV-radiation-induced mutations, whereas this was the dominant mutational signature in melanocytes from sun-exposed skin (Fig. 2a).

Unexpectedly, among sun-exposed melanocytes in this dataset, cells from the back and limbs had more mutations than cells from the face (Fig. 2a, b, Extended Data Fig. 4). Typically, skin from the back and limbs is exposed only intermittently to sunlight and would be expected to accumulate lower levels of cumulative sun exposure than skin from the face, neck and bald scalp. The finding of lower mutation burdens in chronically sun-exposed sites deserves further study, as it indicates possible differences in mutation rate, DNA repair or turnover among melanocytes from these anatomic sites. However, our observations

are consistent with the fact that melanomas are disproportionately common, compared to other forms of skin cancer, on intermittently sun-exposed skin<sup>22,23</sup>.

The mutation burdens of melanocytes also varied between donors. For example, we sequenced melanocytes from the backs of five donors. Among these, the melanocytes from donors 6 and 13 contained the highest mutation burdens (Fig. 2c) with more than half of melanocytes exceeding the median mutation burden of melanoma—this was notable because these donors had skin cancer adjacent to the skin that we sequenced.

For several donors, we observed a wide range of mutation burdens among melanocytes taken from the same anatomic site. This was unexpected, as cells that originated from the same area of skin (about 3 cm<sup>2</sup>) would be expected to have similar levels of exposure to UV radiation and therefore comparable mutagenic profiles. To further understand the broad range of mutation burdens, we sought to identify genes with expression that correlated with mutation burden using differential expression analysis (Extended Data Fig. 5, Supplementary Table 4). Among genes for which the correlation was significant, *MDM2* was more highly expressed in melanocytes with elevated mutation burdens. *MDM2* promotes the rapid degradation of p53, raising the possibility that there is heterogeneity among melanocytes with respect to p53



**Fig. 4 | Fields of related melanocytes identified in normal human skin.** Phylogenetic trees in which each branch corresponds to an individual cell. Mutations that are shared between cells comprise the trunk of each tree and private mutations in each cell form the branches. Trunk and branch lengths are scaled equivalently within each tree but not across trees. The proportion of mutations that can be attributed to UV radiation (CC>TT or (C/T)C>T) is annotated in the bar charts on each tree trunk or branch. Pathogenic mutations and their locations on each tree are indicated in red text.

**Table 1 | Pathogenic mutations in melanocytes from normal human skin**

Pathway	Gene	Protein change	Donor	Site
MAPK	<i>BRAF</i>	G466R	6	Back (adjacent to a BCC)
	<i>BRAF</i>	G466R	6	Back (adjacent to a melanoma)
	<i>BRAF</i>	D594G	6	Back (adjacent to a melanoma)
	<i>BRAF</i>	D594G	6	Back (adjacent to a melanoma)
	<i>BRAF</i>	D594G	13	Back (adjacent to a melanoma)
	<i>BRAF</i>	D594G	13	Back (adjacent to a melanoma)
	<i>BRAF</i>	D594G	13	Back (adjacent to a melanoma)
	<i>CBL</i>	H398L	4	Shin
	<i>CBL</i>	H398L	4	Shin
	<i>MAP2K1</i>	E203K	4	Shoulder
	<i>MAP2K1</i>	E203K	10	Thigh
	<i>NF1</i>	W1314 <sup>a</sup>	6	Back (adjacent to a BCC)
	<i>NF1</i>	P1847L	13	Back (adjacent to a melanoma)
	<i>NF1</i>	Q2239 <sup>a</sup>	13	Back (adjacent to a melanoma)
	<i>NF1</i>	R1276 <sup>a</sup>	6	Back (adjacent to a melanoma)
	<i>NF1</i>	V2511fs	10	Back
	<i>RASA2</i>	L83I	6	Back (adjacent to a BCC)
P376S		13	Back (adjacent to a melanoma)	
P376S		13	Back (adjacent to a melanoma)	
P376S		13	Back (adjacent to a melanoma)	
<i>NRAS</i>	Q61L	13	Back (adjacent to a melanoma)	
Cell cycle	<i>CDKN2A</i>	V43M	6	Back (adjacent to a BCC)
	<i>PPP6C</i>	R264C	10	Back
Epigenetic	<i>ARID2</i>	E1670K	7	Cheek
	<i>ARID2</i>	Q1591 <sup>a</sup>	4	Buttock
	<i>ARID2</i>	A18V	6	Back (adjacent to a melanoma)
	<i>ARID2</i>	L202S	6	Back (adjacent to a melanoma)
	<i>ARID2</i>	P1392L	13	Ear
PI3K	<i>PTEN</i>	QYPFEDH87fs	13	Ear
RNA processing	<i>DDX3X</i>	P167L	13	Back (adjacent to a melanoma)

A curated list of pathogenic mutations in melanocytes found in this study (see Methods for details on how these were defined).

<sup>a</sup>Nonsense mutations.

activity, which could affect the ability of a cell to repair mutations or to undergo DNA-damage-induced cell death. Although MDM2 provides a convincing narrative to explain the heterogeneity in mutation burden, it is just one out of a number of significantly correlated genes that may contribute to the phenotype. Another possibility is that the melanocytes may have resided in the epidermis for different times. For instance, melanocytes with low mutation burdens may reside, or have resided for some portion of their life, in a privileged niche, such as the hair follicle, which would have protected them from UV radiation. Future studies will be needed to resolve why melanocytes from a single site can exhibit such a broad range of mutation burdens.

Melanocytes were collected near a site with melanoma in two patients, and tumour tissue was available from one of these donors. The mutation burden of the melanoma, determined by bulk sequencing, was comparable to that of individual melanocytes from its surrounding skin (Fig. 2d). There was no overlap between mutations in the melanoma and mutations in surrounding melanocytes, suggesting that few, if any, melanoma cells strayed beyond the excision margins into the normal skin. Although more cases need to be studied, our findings suggest that melanomas have mutation burdens similar to neighbouring normal cells. This would contrast with colorectal

cancers, which have higher mutation burdens than surrounding normal colorectal cells<sup>24</sup>.

Copy number alterations were relatively uncommon in melanocytes (Fig. 2a (middle), Extended Data Fig. 6), with the exception of recurrent losses of the Y chromosome and the inactive X chromosome (Supplementary Table 5). Mosaic loss of the Y chromosome and the inactive X chromosome has been reported in normal blood<sup>25,26</sup>, suggesting that this is a generalized feature of ageing. The rarity of autosomal copy number alterations in melanocytes from normal skin is consistent with previous reports that copy number instability is acquired during the later stages of melanoma evolution, and is therefore unlikely to be operative in pre-neoplastic melanocytes<sup>27,28</sup>.

## Pathogenic mutations in melanocytes

We next explored the mutations to identify those that had previously been defined as drivers of neoplasia. We identified 29 pathogenic mutations in 24 different cells (Table 1). In particular, numerous mutations were predicted to activate the mitogen-activated protein kinase (MAPK) pathway. These include loss-of-function mutations in genes that encode negative regulators of the MAPK pathway, affecting *NF1*, *CBL* and *RASA2*. There were also gain- or change-of-function mutations in *BRAF*, *NRAS* and *MAP2K1*; however, we did not detect *BRAF*<sup>V600E</sup> mutations—the most common mutation in the MAPK pathway to occur in melanocytic neoplasms<sup>29,30</sup>.

The World Health Organization (WHO) classification of melanoma distinguishes two major subtypes of cutaneous melanoma—the low cumulative sun damage (low CSD) and high cumulative sun damage (high CSD) subtypes. Low-CSD melanomas are driven by *BRAF*<sup>V600E</sup> mutations and often originate from naevi<sup>31</sup>. By contrast, high-CSD melanomas are driven by a more diverse set of MAPK-pathway mutations, similar to those identified here, and they arise de novo rather than from naevi<sup>31</sup>. Previous functional studies suggested that the MAPK-pathway mutations in our study are weak activators of the MAPK signalling pathway<sup>32–35</sup>, possibly explaining why they do not give rise to discernible neoplasms by themselves, but they could eventually progress to high-CSD melanomas should additional driver mutations arise (Fig. 3).

We also observed driver mutations in other signalling pathways, including mutations that disrupt chromatin remodelling factors and cell-cycle regulators (Table 1). These mutations are presumably not sufficient to induce a neoplasm, but are likely to accelerate progression if the cell that contains them acquires a MAPK-pathway mutation<sup>36</sup> (Fig. 3). This evolutionary trajectory may explain the evolution of nodular melanoma, a type of melanoma that occurs in the absence of a naevus and grows rapidly<sup>37</sup>.

Notably, we found no *TERT* promoter mutations, despite their prominence in melanoma<sup>38,39</sup>. This suggests that *TERT* promoter mutations confer little, if any, selective advantage to melanocytes outside the neoplastic context.

## Fields of related melanocytes in the skin

We found shared mutations between nine separate pairs or trios of melanocytes, suggesting that these cells stemmed from clonal fields of melanocytes in the skin (Fig. 4, Extended Data Fig. 7). We ruled out the possibility that these melanocytes emerged during our brief period of tissue culture by growing neonatal melanocytes for several months and measuring their mutation burdens over time (Extended Data Fig. 8). The number of private mutations in the related sets of melanocytes (Fig. 4) was many orders of magnitude higher than would be expected from two weeks in tissue culture. Moreover, the private mutations from sun-exposed melanocytes showed evidence of UV-radiation-induced DNA damage (Fig. 4 and Extended Data Fig. 7)—a mutational process that does not operate in tissue culture<sup>18</sup>.

Four of the sets of related melanocytes contained a pathogenic mutation in the trunk of their phylogenetic trees, implicating the mutation in the establishment of the field. It is possible that the remaining fields of melanocytes had a pathogenic mutation that we did not detect or appreciate, but we favour the explanation that fields of related melanocytes can also form naturally over time—for instance, as the body surface expands or as part of homeostasis.

## Discussion

A complex set of risk factors is associated with melanoma, including cumulative levels of sun exposure, peak doses and timings of exposure throughout life, skin complexion, tanning ability and DNA repair capacity<sup>40</sup>. It is nearly impossible to quantify and integrate the effects of each of these variables, but we have shown here that it is feasible to directly measure the mutational damage in individual melanocytes. Moving forward, the number and types of mutations in melanocytes warrant further exploration as biomarkers to measure cumulative sun damage and melanoma risk.

Our study also offers important insights into the origins of melanoma. Idealized progression models typically depict melanomas as passing through a series of precursor stages, but in reality, most melanomas appear suddenly, without association with a precursor lesion<sup>41</sup>. We show that human skin is peppered with individual melanocytes or fields of related melanocytes that contain pathogenic mutations that drive melanoma. These poised melanocytes are likely to give rise to melanomas that appear in the absence of a pre-existing naevus, once additional mutations are acquired.

Finally, our genomic studies are a resource for improving understanding of basic melanocyte biology. For example, we found that melanocytes from sun-damaged skin vary in their mutation burdens by multiple orders of magnitude. Of note, a similar pattern of variable mutation burdens was recently reported in bronchial epithelial cells from former smokers<sup>13</sup>. Melanocytes with few mutations are likely to be more efficient at DNA repair and/or to have occupied privileged niches, protected from the sun (such as in the hair follicle). Melanocyte stem cells in the hair follicle can contribute to the intraepidermal pool of melanocytes, as is evident in patients who have repigmenting areas<sup>42</sup>—a similar process may be operative in the general population to replenish sun-damaged melanocytes.

In summary, the genetic observations described here offer insights into the early phases of melanocytic neoplasia, melanocyte homeostasis and the consequences of UV radiation.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2785-8>.

- Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
- Ziegler, A. et al. Sunburn and p53 in the onset of skin cancer. *Nature* **372**, 773–776 (1994).
- Jonason, A. S. et al. Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc. Natl Acad. Sci. USA* **93**, 14025–14029 (1996).
- Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).

- Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
- Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
- Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
- Hou, Y. et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
- Xu, X. et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
- Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
- Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
- Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
- Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
- Macaulay, I. C. et al. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protocols* **11**, 2081–2103 (2016).
- Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
- Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
- Bohrson, C. L. et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat. Genet.* **51**, 749–754 (2019).
- Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
- Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
- Talevich, E. & Shain, A. H. CNVkit-RNA: copy number inference from RNA-sequencing data. Preprint at <https://www.biorxiv.org/content/10.1101/408534v1> (2018).
- Fewings, E., Tang, J., Chang, D. & Shain, A. H. Genomic landscape of 133 melanocytes from human skin. <https://doi.org/10.6084/m9.figshare.11794296.v1> (2020).
- Elwood, J. M. & Gallagher, R. P. Body site distribution of cutaneous malignant melanoma in relationship to patterns of sun exposure. *Int. J. Cancer* **78**, 276–280 (1998).
- Nehal, K. S. & Bichakjian, C. K. Update on keratinocyte carcinomas. *N. Engl. J. Med.* **379**, 363–374 (2018).
- Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
- Machiela, M. J. et al. Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7**, 11843 (2016).
- Forsberg, L. A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
- Shain, A. H. et al. Genomic and transcriptomic analysis reveals incremental disruption of key signaling pathways during melanoma evolution. *Cancer Cell* **34**, 45–55.e4 (2018).
- Bastian, B. C., Olshen, A. B., LeBoit, P. E. & Pinkel, D. Classifying melanocytic tumors based on DNA copy number changes. *Am. J. Pathol.* **163**, 1765–1770 (2003).
- Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
- Pollock, P. M. et al. High frequency of BRAF mutations in nevi. *Nat. Genet.* **33**, 19–20 (2003).
- Shain, A. H. & Bastian, B. C. From melanocytes to melanomas. *Nat. Rev. Cancer* **16**, 345–358 (2016).
- Yao, Z. et al. Tumours with class 3 BRAF mutants are sensitive to the inhibition of activated RAS. *Nature* **548**, 234–238 (2017).
- Krauthammer, M. et al. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat. Genet.* **47**, 996–1002 (2015).
- Grand, F. H. et al. Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood* **113**, 6182–6192 (2009).
- Arafah, R. et al. Recurrent inactivating RASA2 mutations in melanoma. *Nat. Genet.* **47**, 1408–1410 (2015).
- Shain, A. H. & Bastian, B. C. The genetic evolution of melanoma. *N. Engl. J. Med.* **374**, 995–996 (2016).
- Kelly, J. W., Chamberlain, A. J., Staples, M. P. & McAvoy, B. Nodular melanoma. No longer as simple as ABC. *Aust. Fam. Physician* **32**, 706–709 (2003).
- Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
- Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
- Schadendorf, D. et al. Melanoma. *Lancet* **392**, 971–984 (2018).
- Shitara, D. et al. Nevus-associated melanomas: clinicopathologic features. *Am. J. Clin. Pathol.* **142**, 485–491 (2014).
- Abu Tahir, M., Pramod, K., Ansari, S. H. & Ali, J. Current remedies for vitiligo. *Autoimmun. Rev.* **9**, 516–520 (2010).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

## Methods

### Skin tissue collection

Physiologically normal skin tissue was collected from cadavers (up to 8 days post-mortem) or from tissue discarded after surgery on living donors. Skin tissue from cadavers was collected from either the UCSF Autopsy program or the UCSF Willed Body Program. Family members consented to donate tissue from the UCSF Autopsy program, and Willed Body donors consented to donate their tissues for scientific research before their deaths. Surgical discard tissue was collected from donors undergoing dermatologic surgery at UCSF, and their consent was obtained at the time of surgery. Donors from the UCSF Willed Body Program have consented to have any data derived from the donation to be deidentified, stored and shared securely, and used for research, as required by the Federal Privacy Act of 1974, California Information Practices Act of 1977, and HIPAA (Health Insurance Portability and Accountability Act). Donors from clinical practice have consented to the release and sharing of deidentified clinical data and genetic testing information via HIPAA as guided by the NIH National Human Genome Research Institute. Specifically, we used tissue samples banked under the Pathogen Discovery in Cutaneous Neoplasia/Cutaneous Neoplasia Tissue Bank protocol (10-01451) at UCSF.

Here, we define physiologically and clinically normal skin as skin lacking palpable or visible lesions. High-resolution photos (Nikon D3300 fitted with AF-S DX Micro-NIKKOR 40 mm f/2.8G lens) of each skin sample are available at <https://doi.org/10.6084/m9.figshare.11794296.v1>. Skin tissue was stored at 4 °C and processed within 24 h of collection.

### Establishment of epidermal skin cells in tissue culture

Skin tissue was briefly sterilized with 70% ethanol and rinsed with Hank's balanced salt solution (Thermo, 14175095). Excess dermis was trimmed off and the remaining skin was cut into pieces (approximately 2 × 2 mm<sup>2</sup>) using surgical scalpel blades. Tissue was incubated in 10 mg/ml dispase II (Thermo, 17105-041) for 18 h at 4 °C. The epidermis was peeled away from the dermis, incubated in 0.5% trypsin-EDTA (Thermo, 15400-054) at 37 °C for 4 min, and neutralized with 0.5 mg/ml soybean trypsin inhibitor (Thermo, 17075-029). Epidermal cells were plated in Medium 254 (Thermo, M254500) supplemented with human melanocyte growth supplement-2 (HGMS-2, Thermo, S0165) and antibiotic-antimycotic (Thermo, 15240062). Cells were incubated at 37 °C in 5% CO<sub>2</sub> for 7–14 days.

### CRISPR engineering of a subset of cells

Initially, we presumed that it would be impossible to clonally expand single-cell-sorted melanocytes from adult human skin, so we engineered mutations into the *CDKN2A* locus, as previously described<sup>43</sup>. This decision was based on our previous success in engineering *CDKN2A* mutations into foreskin melanocytes and our ability to clonally expand these melanocytes, thereby producing isogenetic population of engineered melanocytes. However, during the course of these experiments, we recognized that control melanocytes, which were not engineered, could clonally expand under optimized tissue culture conditions, so we subsequently stopped engineering melanocytes. In total, five melanocytes were engineered before genotyping, as indicated in Supplementary Table 1. Removal of these cell does not affect any of the conclusions of this study.

### Flow cytometry and cell culture of individual cell clones

Establishing epidermal cells in tissue culture produced a heterogeneous mixture of cells, comprised primarily of melanocytes and keratinocytes with some fibroblasts present. Differential trypsinization was used to separate melanocytes from keratinocytes using 0.05% trypsin-EDTA (Thermo, 25300054) at 37 °C for 2 min and 10 min, respectively. Trypsin was neutralized with 0.5 mg/ml soybean trypsin inhibitor. Cells were centrifuged at 300 rpm for 5 min, resuspended

in 300 µl sorting buffer (1× PBS without Ca<sup>2+</sup> and Mg<sup>2+</sup> (Caisson Labs, PBL-01), 1 mM EDTA (Thermo, AM9262), 25 mM HEPES, pH 7.0 (Thermo, 15630130), and 1% bovine serum albumin (Thermo, BP67110)), strained using a test tube with a 35-µm cell strainer snap cap (Corning, 352235), and single-cell-sorted into 96-well plates filled with 100 µl complete Medium 254 using a Sony SH800S Cell Sorter. Cell sorting was performed using a 100-µm microfluidic sorting chip with the 488-nm excitation laser without fluorescent markers.

On the next day, cells were screened (Zeiss Axiovert microscope) to decipher their morphology and to confirm that each well contained only one cell. Individual melanocytes were grown in CnT-40 melanocyte medium (CELLnTEC, CnT-40) supplemented with antibiotic-antimycotic. A small number of cells had keratinocyte or fibroblast morphology. Keratinocytes were grown in 50:50 complete Medium 254 and keratinocyte-SFM medium (Thermo, 17005042), and fibroblasts were grown in complete Medium 254 for 10–14 days. After 10–21 days, clone sizes ranged from 2–3,000 cells (Supplementary Table 1) and ceased any further expansion, prompting us to collect these clones at their peak cell count. Approximately 37.5% of flow-sorted cells typically produced colonies, providing evidence that we are studying a prevalent and representative population.

### Extraction and amplification of DNA and RNA from each clone

Clones of 2–3,000 cells do not yield enough genomic material to directly sequence using conventional library preparation technologies. For this reason, we elected to isolate both DNA and RNA from each clone and pre-amplify the nucleic acids before sequencing. To do this, we used the G&T-Seq protocol<sup>14,15</sup>.

G&T-Seq was performed, as previously described<sup>14,15</sup>. In brief, clones of cells were lysed in 7.5 µl RLT Plus Buffer (Qiagen, 1053393). mRNA and genomic DNA were separated using a biotinylated oligo d(T)<sub>30</sub> VN mRNA capture primer (5'-biotin-triethyleneglycol-AAGCAGTGGTATCAACGCAGAGTACT30VN-3', where V is A, C or G, and N is any base; IDT) conjugated to Dynabeads MyOne Streptavidin C1 (Thermo, 65001). cDNA was synthesized using the Smart-Seq2 protocol using SuperScript II reverse transcriptase (Thermo, 18064014) and template-switching oligo (5'-AAGCAGTGGTATCAACGCAGAGTACrGrG+G-3', where r indicates a ribonucleic acid base and + indicates a locked nucleic acid base; Qiagen). cDNA was amplified using KAPA HiFi HotStart ReadyMix kit (Roche, KK2502) and purified in a 1:1 volumetric ratio of Agencourt AMPure XP beads (Thermo, A63880). The average yield of amplified cDNA was 305 ng. Genomic DNA was purified in a 0:0.72 volumetric ratio of Agencourt AMPure XP beads and amplified using multiple displacement amplification with the REPLI-g Single Cell Kit (Qiagen, 150345) to yield an average of 815 ng amplified genomic DNA per clone.

### Library preparation and next-generation sequencing of amplified DNA and RNA

We next prepared the amplified cDNA and amplified genomic DNA for sequencing. Library preparation was performed according to the Roche Nimblegen SeqCap EZ Library protocol. In brief, 250 ng DNA input was sheared to 200 bp using Covaris E220 in a Covaris microtube (Covaris, 520077). End repair, A-tailing, adaptor ligation (xGen Duel Index UMI adapters; IDT), and library amplification were performed using the KAPA HyperPrep kit (Roche, KK8504) and KAPA Pure Beads (Roche, KK8001). Library quantification was performed using the Qubit dsDNA High Sensitivity kit and quantitative PCR with the KAPA Quantification kit (Roche, KK4854) on a QuantStudio 5 real-time PCR system.

Target enrichment for next-generation sequencing was performed with the UCSF500 Cancer Gene Panel (developed by the UCSF Clinical Cancer Genomics Laboratory; Roche) or the SeqCap EZ Exome + UTR library probes (Roche, 06740294001). All cells initially underwent targeted sequencing, and if a cell had a low mutation burden, or if a cell was phylogenetically related to other cells, we sequenced it again

# Article

with exome baits. The exome sequencing data yielded more mutations, allowing us to infer mutational processes in low mutation burden cells and in distinct branches of phylogenetically related cells.

The hybridization reaction was performed using the SeqCap EZ Hybridization and Wash Kit (Roche, 05634253001). xGen Universal blocking oligos (IDT, 1075474), human COT 1 DNA (Thermo, 15-279-011), and custom xGen Lockdown probes targeting the telomerase reverse transcriptase (*TERT*) promoter (IDT) were additionally added to the hybridization reaction. After library wash and PCR amplification, the captured library was quantified by Qubit and analysed using the High Sensitivity DNA kit on Agilent's Bioanalyzer 2500.

*TERT* promoter spike-in baits were made with xGen Lockdown probe sequences (2× tiling): 1) /5Biosg/GGGCAGACGCCAGGACCGCGCTTCCCACGTGGCGGAGGGACTGGGGACCCGGGACCCCGTCCGTCCCTTACCTTCCAGCTCCGCCTCTCCGCGCGACCCCGCCCTCCAGCCCTCCCGGGTCCCCGGCCAGCCCCCTCCGGGCCCTCCAGCCCCCTCCCTTCTCTT; 2) /5Biosg/GTCCCGAC; 2) /5Biosg/CCGTCCTGCCCTTACCTTCCAGCTCCGCCTCTCCGCGCGGACCCCGCCCTCCAGCCCTCCCGGGTCCCCGGCCAGCCCCCTCCGGGCCCTCCAGCCCCCTCCCTTCTCTT; 3) /5Biosg/CGACCCCTCCCGGGTCCCCGGCCAGCCCCCTCCGGGCCCTCCAGCCCTCCCTTCTCTTCCGCGGCCCGCCCTCTCTCCGCGCGGAGTTTCAGGCAGCGCTGCGTCTCTGC; 4) /5Biosg/CTTCCGCGGCCCGCCCTCTCTCCGCGCGGAGTTTCAGGCAGCGCTGCGTCTCTGC; 5) /5Biosg/TGCGCAGCTGGGAAGCCCTGGCCCCGGCCAGCCCGCGATGCCGCGCGCTCCCGCTGCCGA; 5) /5Biosg/TGCGCAGCTGGGAAGCCCTGGCCCCGGCCAGCCCGCGATGCCGCGCGCTCCCGCTGCCGAGCCGTGCGCTCCTTCCGCGGACCCCGGAGGTGCTGCCGCTGCCGACGTTCCG.

Libraries were sequenced on an Illumina HiSeq 2500 or Novaseq (paired end 100 bp or 150 bp). On average, we achieved 489-fold unique coverage from targeted sequencing data, 86-fold unique coverage from exome-sequencing data, and 7.75 million reads per clone from RNA-sequencing data.

## Calling a preliminary set of variants

Variant call format files for each clone were generated as previously described<sup>27</sup>. In brief, Fastq files underwent quality checks using FastQC (v.2.4.1) and were subsequently aligned to the hg19 reference genome using the BWA-MEM algorithm (v0.7.13). BWA-aligned bam files were further groomed and deduplicated using Genome Analysis Toolkit (v2.8) and Picard (v.2.1.1). For each clone, variants were called using Mutect (v3.4.46) by comparing to bulk normal cells from a distant anatomic site. At this stage, the variants were composed primarily of amplification artefacts and somatic mutations. We leveraged the matched DNA and RNA sequencing data and haplotype information, detailed below, to distinguish between these entities.

## Using the matched DNA and RNA sequencing data to remove amplification artefacts

The DNA and RNA from each clone were separately amplified, and consequently, amplification artefacts were unlikely to affect the same genomic coordinates in both the DNA and RNA sequencing reads (Fig. 1c). By contrast, somatic mutations should always overlap, assuming there was coverage of the mutant allele in both the DNA and RNA sequencing data. We applied the following criteria to determine whether this assumption could be met.

To begin, we established rates of allelic dropout in our DNA and RNA sequencing data. From known heterozygous SNP sites, we empirically deduced that allelic dropout rates were less than 0.15% in our DNA sequencing data. We achieved low levels of allelic dropout because of our high sequencing coverage, relatively uniform levels of coverage, and low levels of PCR-bias during amplification. Coverage in the RNA sequencing data was more variable owing to differences in gene expression, but from known heterozygous SNP sites, we empirically deduced that 15× coverage was sufficient to sample both alleles at nearly all variant sites. There were a small number of exceptions for which this did not hold true. Truncating mutations (nonsense, splice-site, and frameshift)

are prone to nonsense-mediated decay and were commonly under-sampled in our RNA sequencing data relative to the wild-type allele. Also, mutations on the X chromosome from female donors tended to be in 100% or 0% of RNA sequencing reads, depending on whether they resided on the active or inactive X chromosome. Aside from these examples, allelic variation in expression was minimal, particularly for highly expressed genes, as was previously reported<sup>44</sup>.

On the basis of these observations, a variant was considered a somatic mutation if it was present in both the DNA and RNA sequencing data from the same clone. Conversely, a variant was considered an amplification artefact if the following conditions were met: the variant was present in the DNA sequencing data but not the RNA sequencing data, and there was at least 15× coverage in the RNA sequencing data, and the variant was not truncating or on the X chromosome. We declined to make a call in either direction for any variant that did not fulfil these conditions.

A limitation to this approach was that some variants did not reside in genes that were expressed. Nevertheless, 11.6% of variants could be classified as either a somatic mutation or amplification artefact by cross-validating the DNA and RNA sequencing data.

## Using haplotype information to remove amplification artefacts

We also used haplotype information to distinguish between somatic mutations and amplification artefacts. Somatic mutations occur in *cis* with nearby germline polymorphisms, and this pattern is preserved during amplification (Fig. 1d). By contrast, amplification artefacts do not occur in complete linkage with nearby germline polymorphisms for the reasons described below (Fig. 1d).

The germline polymorphisms operate similarly to unique molecular barcodes, designating which amplicons descended from each parental allele. The main reason why amplification artefacts are not in complete linkage with nearby polymorphisms is because there are multiple template molecules, associated with each parental allele, from which to amplify, and each template molecule can be amplified more than once—it is unlikely that the exact same mistakes will be made during each independent amplification reaction over an error-free template. For example, we sequenced clonal expansions of cells, so each cell provided one molecule of double-stranded DNA from each allele. Furthermore, both strands of DNA are subject to amplification, thereby doubling the number of template molecules relative to the starting cell number. Finally, a single strand of DNA is repeatedly amplified during multiple displacement amplification, further enhancing the number of times an error-free template is used during amplification. Amplification artefacts therefore reveal themselves in the sequencing data by not occurring in complete linkage with nearby polymorphisms.

There was an exception for which the pattern described above did not hold true. A copy number gain or copy-number-neutral loss-of-heterozygosity (LOH) results in two or more copies of a single parental allele. If a somatic mutation occurs after the allelic duplication, then the somatic mutation would not be in complete linkage with nearby polymorphisms. Consequently, we did not apply this methodology to identify amplification artefacts over regions of the genome for which there was an allelic duplication.

A limitation to this approach is that we used short-read sequencing technologies, so some variants were too far away from the nearest polymorphic sites to be phased. Nevertheless, 14.7% of variants could be classified as either a somatic mutation or amplification artefact, using the phasing approach.

## Inferring the mutational status of variants outside the expressed or phase-able portions of the genome

In total, 25.1% of variants could be classified as either a somatic mutation or an amplification artefact, using either the expression or the phasing approaches described above. The remaining variants did not reside in portions of the genome that were sufficiently expressed or



close enough to germline polymorphisms to permit phasing. For these variants, we inferred their mutational status from their VAF.

The majority of somatic mutations in our study were heterozygous, and these mutations, as expected, exhibited a normal distribution of mutant allele frequencies centred at 50% (Fig. 1e, Extended Data Fig. 2b). The standard deviation of mutant allele frequencies in a given clone was dictated primarily by the number of starting cells, indicating that allelic biases—introduced during amplification—were the primary drivers of ‘noise’ in our data.

By contrast, amplification artefacts exhibited a different distribution of allele frequencies. Most amplification artefacts occurred in later rounds of amplification, and therefore had extremely low VAFs. However, a small number of amplification artefacts occurred in relatively early rounds of amplification and were disproportionately amplified thereafter. As a result, amplification artefacts exhibited a distribution of allele frequencies with a low peak but a long tail, sometimes extending into the range of allele frequencies seen for somatic mutations (Fig. 1e, Extended Data Fig. 2b). As expected, the tail of this distribution was more extreme in clones with fewer starting cells because amplification biases were more exacerbated in these clones.

Owing to the distinct distributions of variant allele frequencies for somatic mutations and amplification artefacts, a VAF cutoff could distinguish the vast majority of somatic mutations from amplification artefacts. However, the sensitivity and specificity of somatic mutation calls, using this approach, varied for each clone, primarily based on the clone size for the reasons described above. We were able to precisely define the sensitivity and specificity of mutation calls, and we could optimize the VAF cutoff for each clone by studying the overlap in VAFs from known somatic mutations and known amplification artefacts.

For each clone, we had a set of known somatic mutations and known amplification artefacts, situated in the expressed and phase-able portions of the genome. We were therefore able to determine the proportion of false positives and false negatives under the assumption that all variants above a given VAF were somatic mutations. Here, a false positive is an amplification artefact that would have been called a somatic mutation, and a false negative is a somatic mutation that would have been called an amplification artefact. We plotted the sensitivity and specificity of mutation calls at different VAF cutoffs for each clone, and we chose the VAF cutoff that maximized these values. This value was then applied to the variants whose mutational status was unknown—that is, the variants outside the expressed and phase-able portions of the genome. For clones greater than five cells, we could typically infer somatic mutations at greater than 98% specificity and 98% sensitivity (Extended Data Fig. 2c, d). We indicate in Supplementary Table 3 whether each mutation was validated or inferred by this approach.

### Copy number analysis

Copy number alterations were inferred from both the DNA and the RNA sequencing data using CNVkit (v.0.9.5.3)<sup>19,20</sup>. We also integrated allelic frequencies from somatic mutations and germline heterozygous SNPs.

First, we inferred copy number alterations from the DNA-sequencing data. CNVkit can be run in reference or reference-free mode. We elected to run CNVkit in reference mode, and in doing so, we created several references, encompassing panels of clones without copy number alterations that were amplified and prepared for sequencing in similar batches. This approach consistently produced the least noisy copy number profiles, as compared to reference-free mode or a universal reference. All other parameters were run on their default settings.

Second, we inferred copy number alterations from the RNA sequencing data. In brief, CNVkit assumes that the expression of a gene correlates with its copy number status. Of course, the expression of a gene is dictated by several factors, including, but not limited to, copy number. As an input, CNVkit accepts correlation values from an independent

dataset between expression and copy number. Here, we included correlation values from the melanoma TCGA project. Given this input, CNVkit downweights genes with expression that does not correlate well with copy number.

Third, we calculated allelic imbalance over germline heterozygous germline SNPs. Copy number alterations are expected to induce imbalances over these sites. Additionally, we calculated the allelic frequencies of somatic mutations across the genome, as these too would be modulated by copy number alterations.

Finally, we manually reviewed the copy number and variant allele information to call copy number alterations that were supported by each approach.

### Establishing cell identity

We made morphologic predictions when screening single cell clones of melanocytes, fibroblasts, and keratinocytes to designate cell identity. Melanocytes have a cell body with stellar or dendritic projections, are darker owing to the presence of melanin, and tend to grow in tighter clusters than fibroblasts, albeit not as tight as keratinocytes. Keratinocytes have a polygonal cell shape with more regular dimensions and grow in a very tight cluster owing to the presence of desmosomes. Fibroblasts are flat, oblong or triangular cells that divide very quickly in a diffuse cluster as a characteristic meshwork. In addition to cell morphology, we inspected the gene expression of *MLANA*, *TYR*, *PMEL*, and *SIOOB*. The protein products of these genes are well-established markers of the melanocyte cell lineage and are commonly used in the clinical setting to distinguish melanocytes and tumours of melanocytic origin from other cell lineages and other tumour types. There was a clear separation of gene expression levels of these genes between the cells that we nominated as melanocytes as opposed to keratinocytes or fibroblasts (Extended Data Fig. 1c).

### An overview of the genetic landscape of each sequenced melanocyte

Individual summaries of the 133 sequenced clones which describe cellular morphology and tissue images, validation of VAF of raw calls, copy number alterations, and *CDKN2A* status (where applicable) are available at <https://doi.org/10.6084/m9.figshare.11794296.v1><sup>21</sup>.

### Admixture analysis

Related to Extended Data Fig. 1a. Bulk normal cells were analysed to identify germline variants present in each studied donor. Donor ethnicity was inferred via Admixture analysis using a Bayesian modelling approach employed by the tool STRUCTURE (v2.3.4)<sup>45</sup>. A set of 7,662 common variants (1000 Genomes population allele frequency >0.05) with a sequencing depth of greater than 10 across all donors and all 2,504 samples from the 1000 Genomes study<sup>46</sup> were selected. The burn-in period and analysis period were both completed with 10,000 repetitions as per the tool recommendations to achieve accurate estimations of admixture. To select an appropriate number of populations (*K*), the algorithm was run using *K* estimations of 5–9. A final *K* value of 8 was selected to appropriately cluster populations without overfitting. The data were plotted using the STRUCTURE GUI plotting tool. The ethnicity of donors within this study was inferred by their similarity to known populations within the 1000 Genomes set<sup>46</sup>.

### RNA gene expression analysis

Related to Fig. 1b and Extended Data Fig. 1b. RNA sequencing reads were aligned to the transcriptome as well as the hg19 reference genome using STAR alignment tool (v.2.5.1b)<sup>47</sup>. Transcripts were quantified using RNA-seq by expectation–maximization (RSEM) (v.1.2.0)<sup>48</sup> and filtered to remove those with fewer than 10 reads across all clones as recommended by DESeq2 R package documentation. A variance stabilizing transformation was applied to the data and a Barnes–Hut *t*-SNE algorithm was performed to cluster related cells on the expression of

# Article

the top 500 genes using the Rtsne R package (v.0.15) with a perplexity of 6 over 1,000 iterations.

Differential expression analysis was completed on the quantified transcript values using DESeq2 R package<sup>49</sup> (v.1.22.2). Three experimental designs were produced, selecting for differentially expressed genes that are overexpressed in fibroblasts, melanocytes, and keratinocytes independently. The data were  $\log_2$ -transformed and a heatmap was generated presenting the top 20 significantly differentially overexpressed genes per cell type.

Gene set enrichment analysis was performed across the significantly differentially overexpressed genes from each cell type using the Molecular Signatures Database (v.6.2) webtool. The top significantly enriched pathways were examined for their relation to the cell type of interest.

## Mutation burden and signature analysis

Related to Fig. 2. The mutation burdens reported in Fig. 2 correspond to the number of somatic mutations in a given clone divided by the genomic footprint for which mutations could be detected. Owing to differences in depth of coverage across bam files and the unevenness of coverage in a given bam file, mutations were not callable at every base present in the target region. Additionally, we used both a targeted and exome sequencing panel in this study, which produce two different sequencing footprint sizes. To account for these issues, we calculated callable sequencing footprints for each clone and corresponding reference. On-target bam files were created per clone and per bulk normal. The coverage of each on-target base was calculated using the bedtools (v.2.25.0) `genomecov` command, and the number of bases covered by more than five reads was counted in each bam file. The minimum value between a clone's bam and its reference bam was used as the footprint from which to calculate a mutation burden for each clone.

Linear mixed-effect models were generated using the `lmer` library in R to identify any association between sun exposure (as determined by the anatomic site from which the single cell was derived) and mutation burden while correcting for the donor of origin. *P* values of each pairwise comparison derived from this model with the `lmer` package are shown in Fig. 2b. To further account for the repeated measurements per donor, a model was created excluding the sun-exposure variable and an ANOVA was performed comparing the fit of the two models.

To perform mutational signature analysis, surrounding genomic contexts were applied to single nucleotide variants identified in each clone using the `Biostrings` hg19 human genome sequence package (BSgenome.Hsapiens.UCSC.hg19 v.1.4.0). Variant contexts were used to assess the proportion of each clone's mutational landscape that could be attributed to a mutagenic process using the `deconstructSigs` R package (v.1.8.0). A recently described set of 48 signatures<sup>18</sup> was analysed, with particular attention paid to the single-base-substitution signatures 7a, 7b, and 7c that are associated with UV light exposure.

## Identifying pathogenic mutations

Related to Table 1. We define a pathogenic mutation to be a mutation that is under positive selection in cancer.

To identify gain- or change-of-function mutations that affect oncogenes, we investigated whether the mutations in our study overlapped with previously defined mutational hotspots. First, we referenced the Catalogue of Somatic Mutations in Cancer (COSMIC) database (see column M 'COSMIC\_ID' of Supplementary Table 3). There are thousands of entries in the COSMIC database, so mutations could recur at low frequencies at certain positions by chance alone. Therefore, we curated these mutations to identify those with a previously published biological function. From this analysis, we identified hotspot mutations affecting *BRAF*, *NRAS*, *MAP2K1*, *CBL*, and *PPP6C* (Table 1). In parallel, we referenced `cancerhotspot.org`, a curated database of mutational hotspots. From this analysis, we corroborated the hotspot

mutations affecting *BRAF*, *NRAS*, *MAP2K1*, and *PPP6C*. In addition, we found an E548K substitution affecting *PTPRT*. Upon further review, we concluded that the *PTPRT* mutation was unlikely to be biologically active because the gene is not expressed in the melanocytic cell lineage and the mutations in this gene do not show evidence of positive selection in melanoma<sup>50</sup>, and therefore we elected not to highlight this gene in our analysis.

To identify loss-of-function mutations that affected defined tumour suppressor genes in our study, we referred to previous melanoma-related publications<sup>33,50</sup>. From this analysis, we identified mutations affecting *NF1*, *CBL*, *RASA2*, *CDKN2A*, *ARID2*, *PTEN*, and *DDX3X*. There were also mutations affecting genes that are likely to be tumour suppressors in melanoma but have yet to be unequivocally defined as such. We elected not to highlight these mutations in Table 1; however, we encourage readers to consult the full list of mutations in Supplementary Table 3, as the number of pathogenic mutations is likely to exceed the more conservative assessment shown in Table 1.

## Gene expression correlation with mutation burden

Related to Supplementary Table 4 and Extended Data Fig. 5. RNA data were used to explore the variability in mutation burdens, often observed over a single site. Sites with more than 3 s.d. of mutation burdens, demonstrating the presence of both high and low mutation burden clones, were selected for analysis. Mutation burdens were normalized to the median of each anatomic site. Differential expression analysis was then performed using DESeq2 R package<sup>49</sup> (v.1.22.2). Genes with expression changes significantly associated (adjusted *P* < 0.01) with a continuous change in mutation burden are highlighted in Supplementary Table 4 and Extended Data Fig. 5.

## Estimating mutation acquisition over time in tissue culture

Related to Extended Data Fig. 8. We established skin cells in tissue culture for 7–14 days before single-cell sorting and clonal expansion. Any mutation that arose after clonal expansion would be recognizable because it would be present in only a proportion of daughter cells, thus appearing subclonal. However, mutations that arose during the brief period of tissue culture preceding clonal expansion could be mistaken as a mutation that occurred while the cell was still situated in the skin. We therefore sought to establish the rate at which melanocytes accumulate *de novo* mutations in tissue culture to determine whether this was a meaningful contribution to the total mutation burden that we observed in our cells.

Towards this goal, we followed a framework that has recently been put forth<sup>18</sup>. In that study, the authors sequenced subclones of daughter cells from common cancer lines at different generational time points for up to 161 days, thereby revealing the mutational processes that were operating during their time in tissue culture. Here, we sequenced a bulk culture of normal human melanocytes derived from human foreskin to establish the germline variants and somatic mutations in the dominant clones. We continued to culture these cells, and at time points of 51, 63, 120, and 239 days, we single-cell sorted and clonally expanded individual cells. We genotyped each clonal expansion, following the same protocol that was applied to melanocytes. From these analyses, we estimate that mutations occur at a rate of 0.045 mutations per Mb per 7 days in tissue culture. To put this in perspective, the mutation burden of melanocytes from the bottom of the foot was 0.25 mutations per Mb. On the basis of these findings, we conclude that the number of mutations accumulated in tissue culture was negligible as compared to the number of mutations that pre-existed in melanocytes that were profiled for this study.

We also analysed publicly available data<sup>18</sup> to deduce the rate at which melanoma cell lines accumulate mutations in tissue culture. From these analyses we estimate that mutations occur at a rate of 0.043 mutations per Mb per 7 days, in line with our estimates for normal human melanocytes.

Thus, it is not surprising that the number of mutations collected after 7 days in tissue culture is negligible as compared to the number of mutations collected from decades in the skin.

### Phylogenetic tree construction

Related to Fig. 4 and Extended Data Fig. 7. Pairwise comparisons of melanocyte mutation calls were performed to identify sets of melanocytes with shared mutations, and when this occurred, phylogenetic trees were constructed from the shared and unshared mutations. In Fig. 4, trunk lengths correspond to the number of shared mutations, and branch lengths correspond to the number of unshared mutations. If there was an allelic deletion in one clone, we did not assign mutations in the clone lacking the deletion over the deletion area to the branch. Shared mutations were discarded if there was insufficient coverage in the reference to rule out the possibility that the mutation was a germline SNP. Unshared mutations were discarded if sequencing coverage was insufficient in one clone to definitively make a call. In practice, few mutations needed to be discarded by these filtering criteria because we achieved high sequencing coverage in our clones.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

Sequence data have been deposited in dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>) with the accession code phs001979.v1.p1. Individual sample summaries of every single cell clone are available at <https://doi.org/10.6084/m9.figshare.11794296.v1><sup>21</sup>. Source data are provided with this paper.

### Code availability

Scripts and resources to perform analyses downstream of variant calling are available at [https://github.com/elliefewings/Melanocytes\\_Tang2020](https://github.com/elliefewings/Melanocytes_Tang2020).

43. Zeng, H. et al. Bi-allelic loss of CDKN2A initiates melanoma invasion via BRN2 activation. *Cancer Cell* **34**, 56–68.e9 (2018).
44. Reinius, B. et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* **48**, 1430–1435 (2016).
45. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
46. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
51. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

**Acknowledgements** We acknowledge support from the following: National Cancer Institute K22 CA217997 (A.H.S.), Melanoma Research Alliance (A.H.S.), LEO Foundation (A.H.S.), George and Judy Marcus Precision Medicine Fund (A.H.S. and S.T.A.), National Center for Advancing Translational Sciences and the National Institutes of Health through UCSF-CTSI TL1-TRO01871 (J.T.), 1R35CA220481 (B.C.B.), Mt. Zion Health Research Fund (A.H.S.), Dermatology Foundation (A.H.S.), the American Federation of Aging Research (A.H.S.), and the NIH Director's Common Fund DP5 ODO19787 (R.L.J.). We thank the tissue donors, whose tissue was obtained through the UCSF Willied Body Program for medical education, and patients who consented to donate surgical discard tissue. Cell sorting was performed in the Laboratory for Cell Analysis of UCSF's Helen Diller Family Comprehensive Cancer Center which is supported by a National Cancer Institute Cancer Center Support Grant (P30 CA082103).

**Author contributions** Conception and design of the work: A.H.S. Data collection: J.T., D.C., S.L., E.F., H.Z., A.J., R.L.B., A.S.M., S.T.A. Data analysis and interpretation: E.F., J.T., D.C., T.M.T., R.L.J.-T., B.C.B., A.H.S. Drafting the Article: E.F., J.T., A.H.S. Critical revision of the Article: E.F., J.T., A.H.S., R.L.B., I.Y., S.T.A., R.L.J.-T., B.C.B.

**Competing interests** S.T.A. is an employee at Rakuten Medical and a consultant for Castle Biosciences and Enspectra Health.

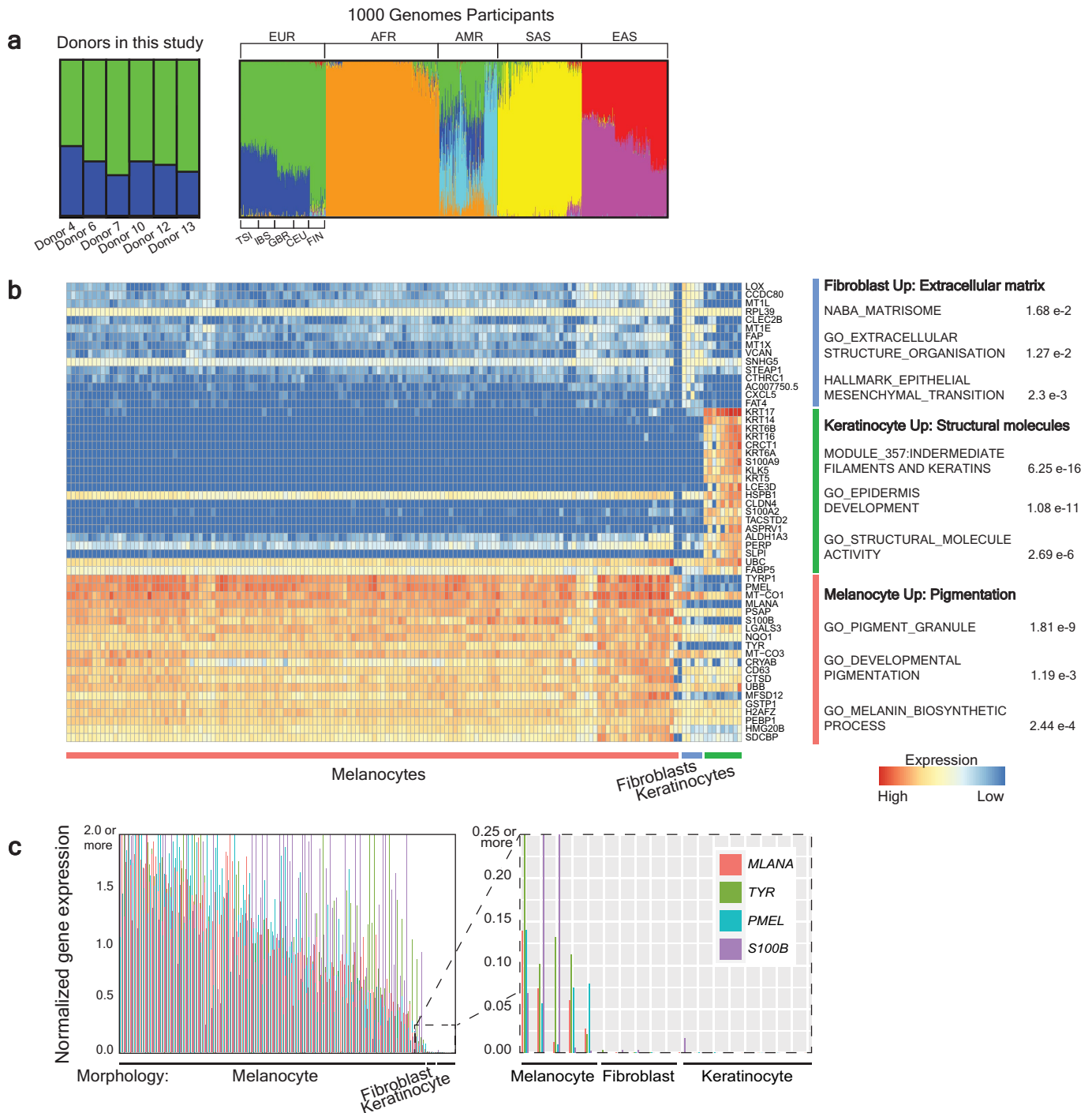
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-2785-8>.

**Correspondence and requests for materials** should be addressed to A.H.S.

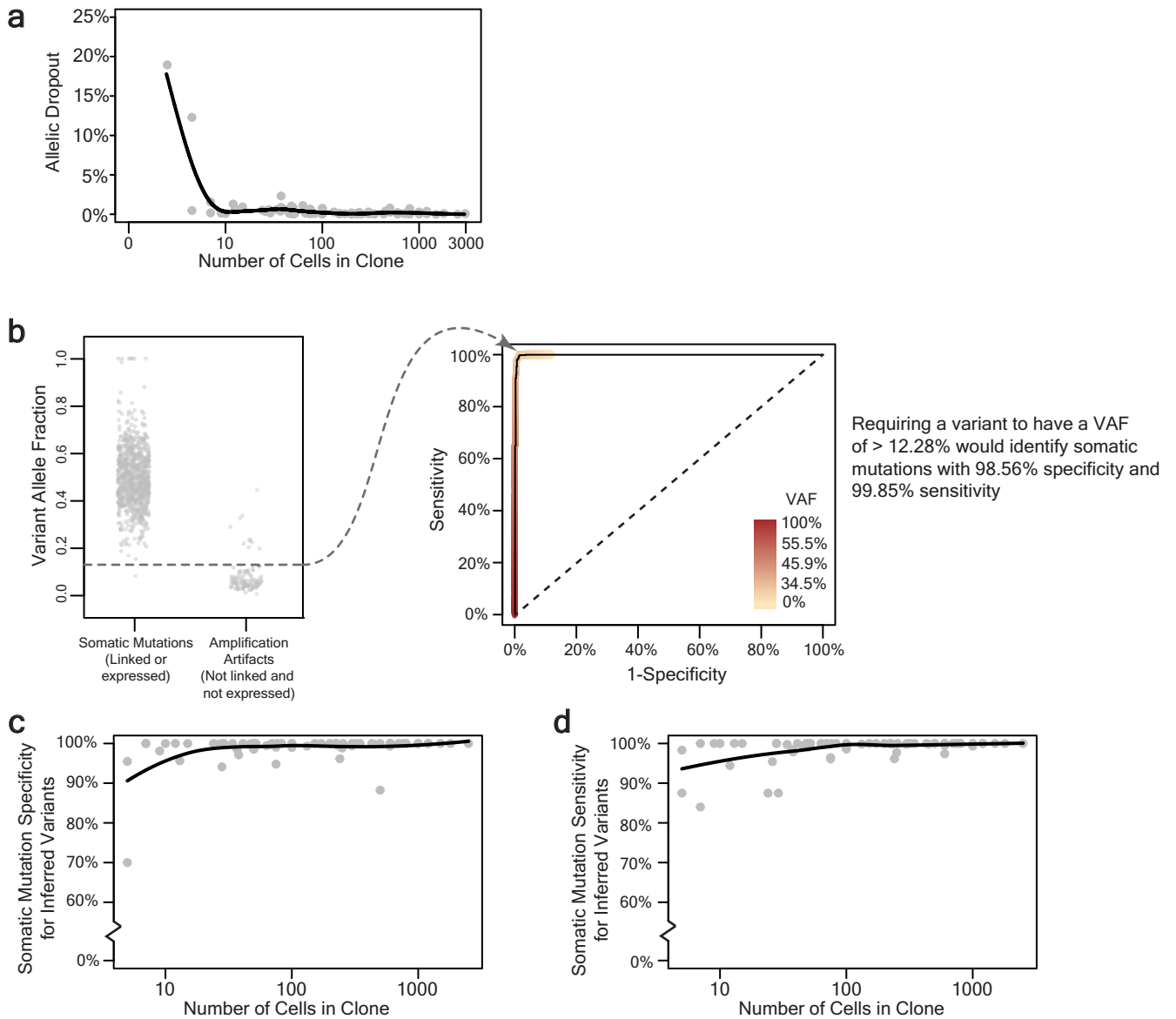
**Peer review information** *Nature* thanks Meenhard Herlyn, Inigo Martincorena and Göran Jönsson for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Establishing the ethnicity of donors and identity of cells in this study.** **a**, Admixture analysis of donors included in this study alongside participants from the 1000 Genomes Project. Donors in our study were genotypically most similar to European participants from the 1000 Genomes Project. EUR - European (TSI - Toscani in Italia, IBS - Iberian Population in Spain, GBR - British in England and Scotland, CEU - Utah Residents with Northern and Western European Ancestry, FIN - Finnish in Finland), AFR - African, AMR - Latin American, SAS - South Asian, and EAS - East Asian. **b**, Differential expression analysis comparing cells that were morphologically predicted to be keratinocytes, melanocytes, or fibroblasts (see Fig. 1b for more

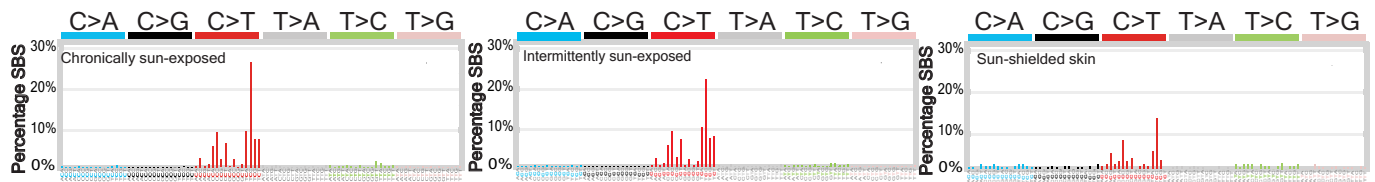
details). The top 20 differentially expressed genes for each group are shown along with gene ontology terms with significant overlap. **c**, Cells with melanocyte morphology express higher levels of known melanocyte markers. Bar plots showing gene expression levels of *MLANA*, *TYR*, *PMEL*, and *S100B*, colored as indicated. A value of 1 is equivalent to the medium FPKM value for that gene across cells. Each quartet of bars corresponds to an individual clone, and clones are rank ordered by their medium normalized gene expression values for these 4 genes. The zoomed inset portrays the 5 melanocyte clones with lowest expression levels of melanocyte markers adjacent to the fibroblast and keratinocyte clones.



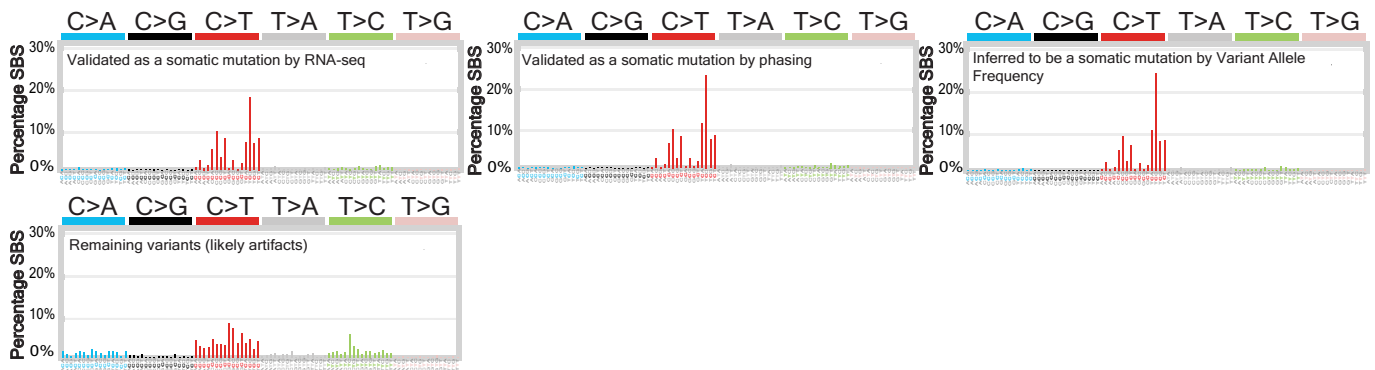
**Extended Data Fig. 2 | Detection of somatic mutations in small clones of skin cells with high specificity and sensitivity.** **a**, Allelic dropout declines rapidly as a function of clone size. Each data point represents the percent of germline SNP alleles that could not be detected for a given clone as a function of the number of cells within the clone. **b**, Establishing a VAF cutoff to infer somatic mutations within a clone. The left panel depicts the VAFs for known somatic mutations and known amplification artefacts from a single clone. The right panel depicts a ROC curve, showing the VAF at which sensitivity and specificity of somatic mutation calls would be maximized when inferring the

mutational status of variants based on VAF alone. Variants that fell within expressed or phase-able portions of the genome were classified as mutations or artefacts as described (Fig. 1c, d). The remaining variants were inferred based on the VAF cutoff, which maximized sensitivity and specificity of somatic mutation calls. **c**, **d**, The specificity (**c**), and sensitivity (**d**), of inferred somatic mutations as a function of clone size. The mean specificity and sensitivity of inferred somatic mutations was respectively 98.83% and 98.60% for all clones of at least 5 cells. All trendlines correspond to a moving average.

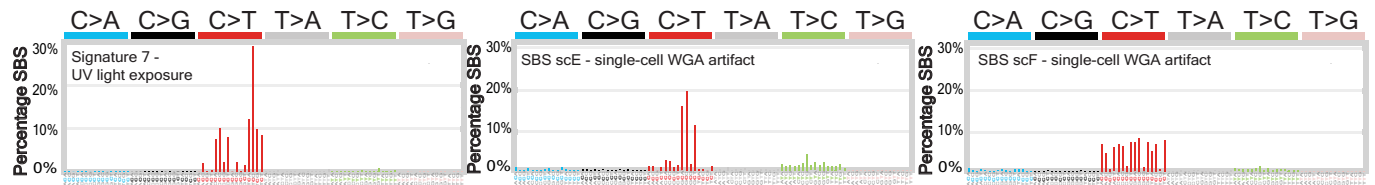
**a** Contexts of somatic mutations identified in skin of varying sun exposure



**b** Contexts of substitutions identified in sun-exposed skin, organised by validation status

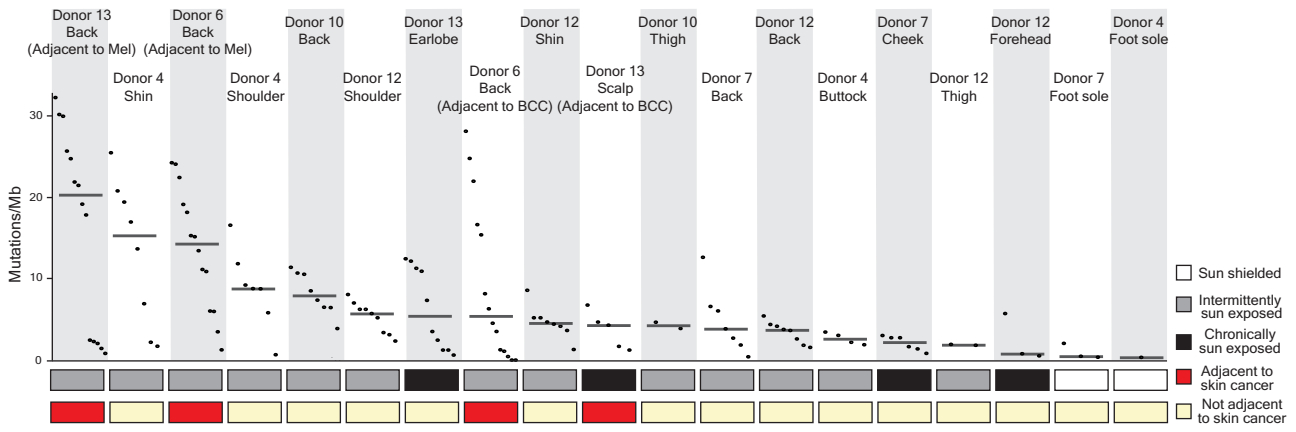


**c** Predefined mutation signatures



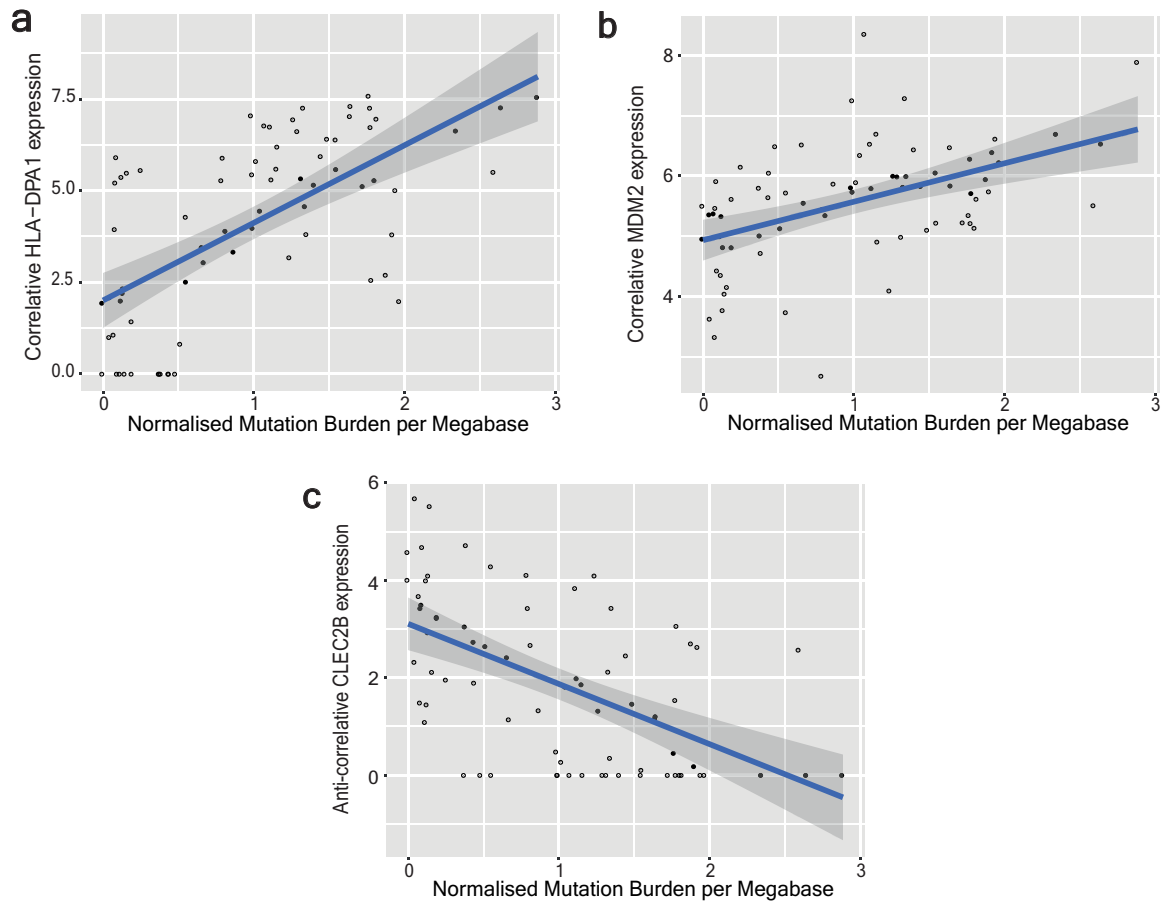
**Extended Data Fig. 3 | Contexts of single-base substitutions corroborate the quality of somatic mutation calls.** **a**, The proportion of somatic mutations identified in chronically sun-exposed, intermittently sun-exposed, and sun-shielded skin that belong to each of the 96 trinucleotide substitution contexts. Note the similarity to signature 7 (shown for reference in **c**), albeit to a lesser extent in sun-shielded skin cells. **b**, Tri-nucleotide contexts of variants from sun-exposed skin validated to be somatic mutations by RNA-seq or

phasing as well as variants inferred to be somatic mutations by their variant allele frequency (VAF). Note the similarity to signature 7. The tri-nucleotide contexts of remaining variants (assumed to be amplification artefacts) are also shown. **c**, Predefined mutation signatures shown for reference; Signature 7 (associated with UV-radiation-induced DNA damage)<sup>51</sup>, and SBS scE and SBS scF, which are associated with single-cell whole genome amplification artefacts<sup>18</sup>.



**Extended Data Fig. 4 | Median mutation burden of melanocytes from different anatomic sites.** Mutation burden of melanocytes from physiologically normal skin of six donors across different anatomic sites with

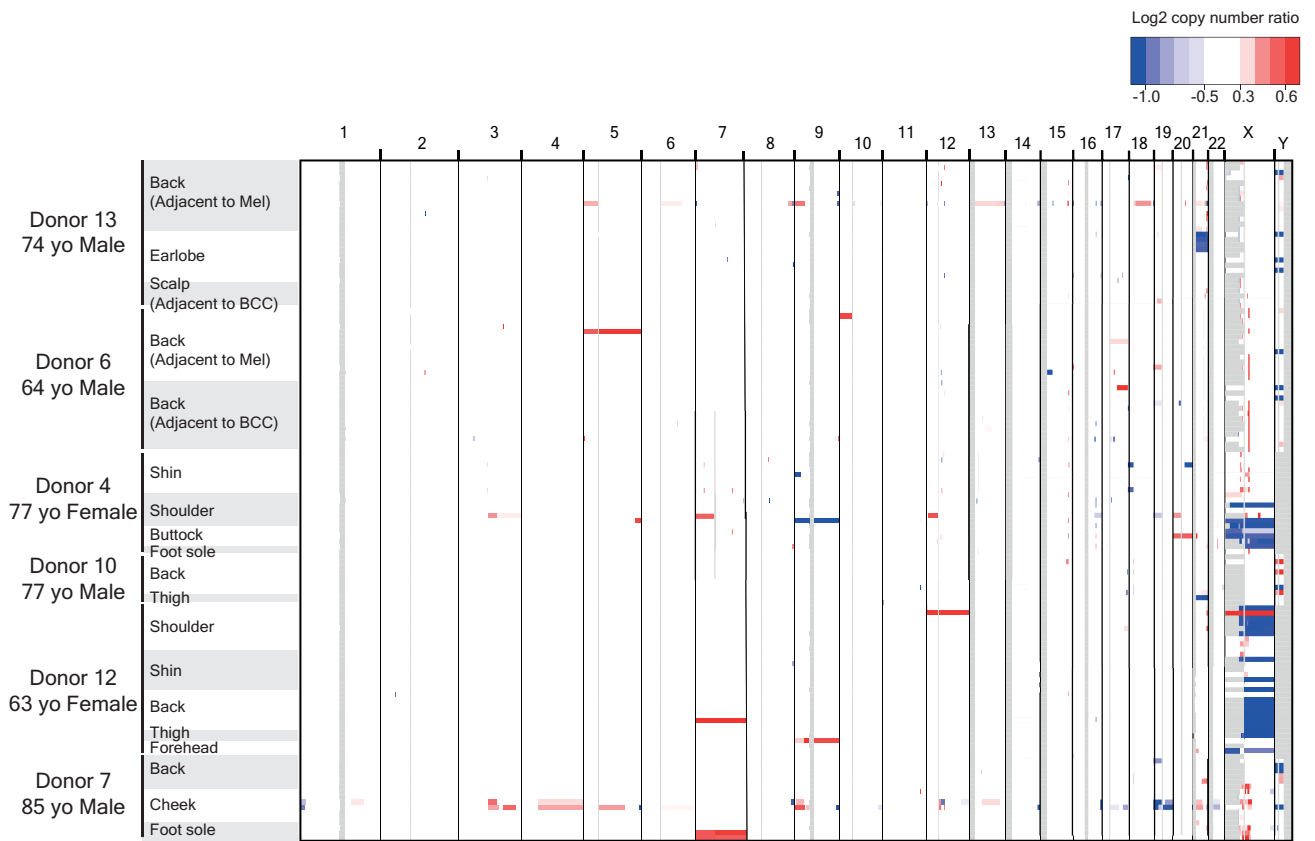
varied sun exposure that are rank ordered by median mutation burden (line) within each site. (BCC = Basal Cell Carcinoma, Mel = Melanoma).



**Extended Data Fig. 5 | Differential expression analysis revealing genes significantly correlating with mutation burden. a-c,** Gene expression versus normalized mutation burden is shown for two top correlative genes (*HLA-DPA1* and *MDM2*) and one (*CLEC2B*) anti-correlative gene of interest from Supplementary Table 4. Clones included in this analysis are from anatomic sites

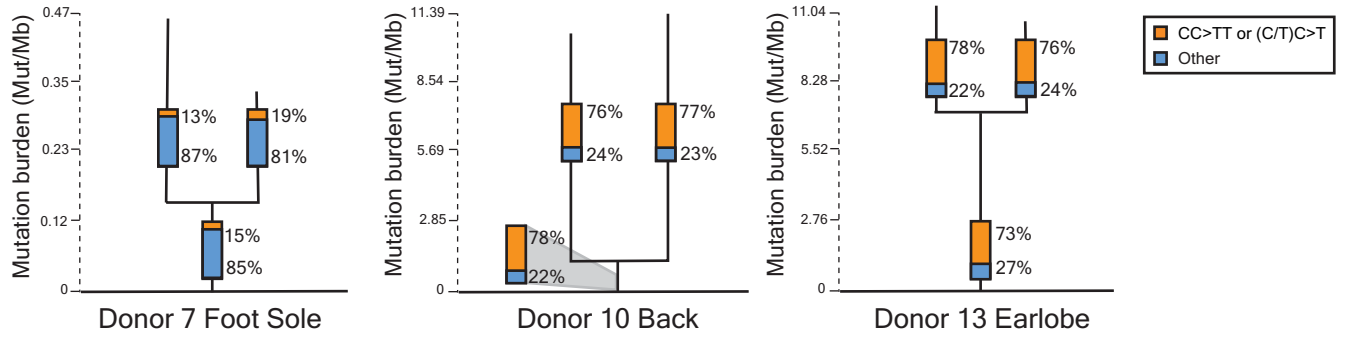
with greater than 3 standard deviations of mutation burdens among their cells, thus demonstrating a range of mutation burdens. The plotted blue line represents a linear model fit to the data with 95% confidence intervals for that model prediction shown in grey.





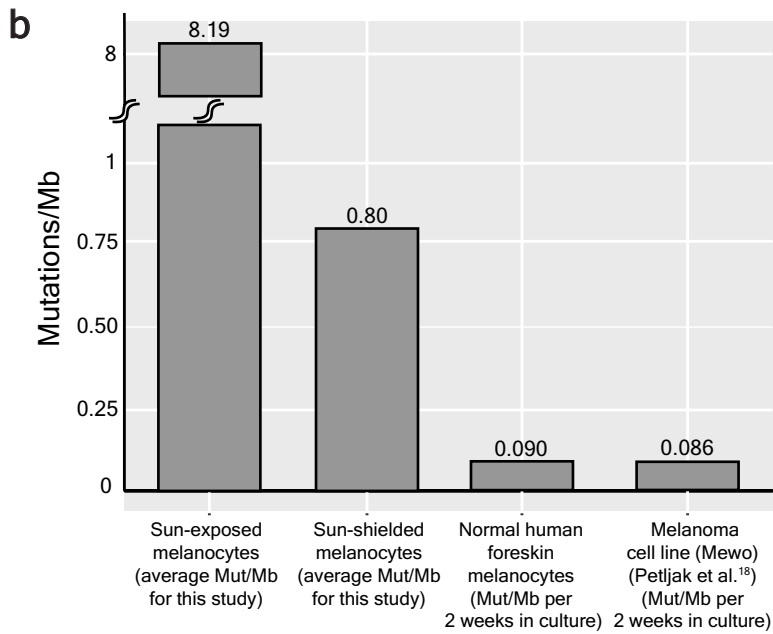
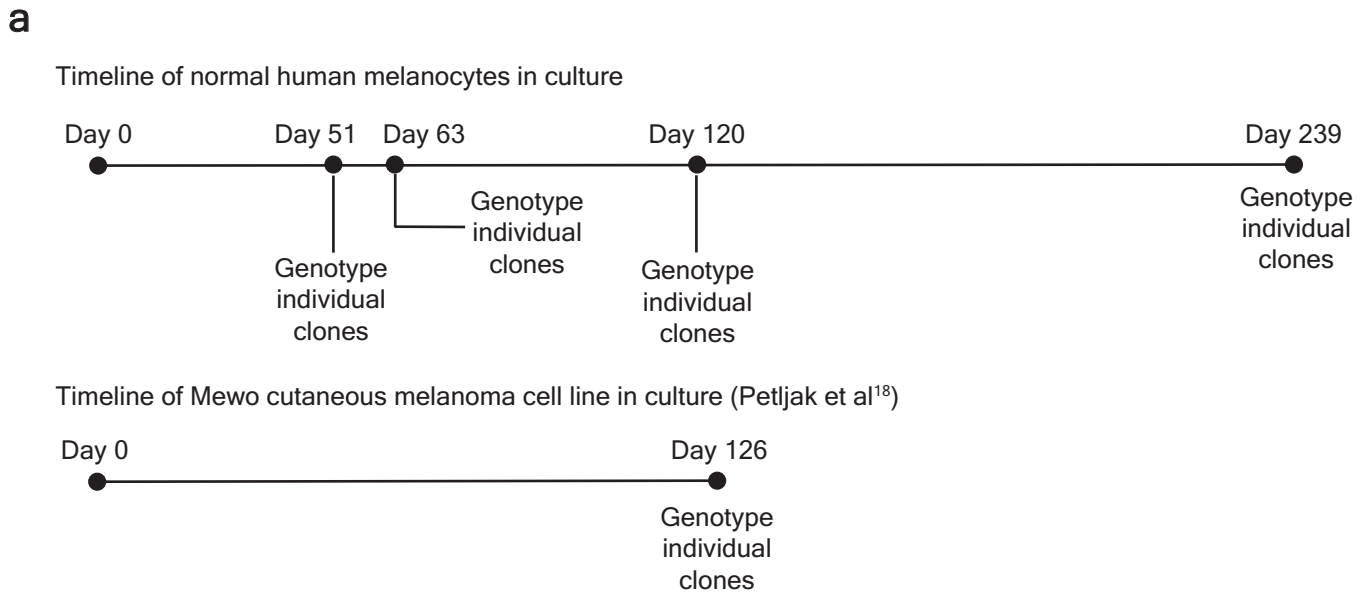
**Extended Data Fig. 6 | Copy number landscape of melanocytes from normal human skin.** Copy number was inferred, as described, and segments (regions of equal copy number) are depicted, here, denoting gains (red) and losses (blue) for each melanocyte (rows). Note that copy number alterations over

autosomes were rare, while the loss of one sex chromosome is a common occurrence. All X chromosome deletions in females affect the inactive X (see Supplementary Table 5).



**Extended Data Fig. 7 | Fields of related melanocytes exist within the skin.** Phylogenetic trees in which each branch corresponds to an individual cell. Mutations that are shared between cells comprise the trunk of each tree and private mutations in each cell form the branches. Trunk and branch lengths are

scaled equivalently within each tree but not across trees. The proportion of mutations that can be attributed to UV radiation (CC > TT or (C/T)C > T) is annotated in the bar charts on each tree trunk or branch.



**Extended Data Fig. 8 | Melanocytes accumulate few mutations in tissue culture.** **a**, We sequenced a bulk culture of neonatal melanocytes to establish the germline SNPs and somatic mutations in the dominant clones. We continued to passage the cell line for 239 days, genotyping individual clones at the time points indicated to establish the rate at which mutations were acquired in culture. In parallel, Petljak et al.<sup>18</sup> performed similar experiments on common cancer cell lines, and we analysed their data from a melanoma cell line (Mewo) included in their study. **b**, On average, the mutation burden of

neonatal melanocytes and Mewo cells respectively increased by 0.090 and 0.086 mutations/Mb for every 2 weeks in tissue culture (we typically cultured melanocytes 2 weeks or less in this study). To put these mutation burdens in perspective, the average mutation burdens of sun-exposed and sun-shielded melanocytes from this study are shown in comparison. Based on these results, we conclude that the brief period of tissue culture contributed little towards the mutation burdens observed in our study.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |  |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used during data collection.
Data analysis	Custom scripts generated to analyse data are available on GitHub ( <a href="https://github.com/elliefewings">https://github.com/elliefewings</a> ) and described in full in the manuscript. The following data analysis software were used: FastQC (v2.4.1), BWA-MEM algorithm (v0.7.13), Genome Analysis Toolkit (v2.8), Picard (v.2.1.1), Mutect (v3.4.46), CNVkit (v0.9.5.3), STRUCTURE (v.2.3.4), STAR alignment tool (v2.5.1b), RSEM (v1.2.0), DESeq2 R package (v1.22.2), Rtsne R package (v0.15), Molecular Signatures Database (v6.2) webtool, bedtools (v2.25.0), Biostrings hg19 human genome sequence package (v1.4.0), and deconstructSigs R package (v1.8.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data that support the findings of this study have been deposited in dbGaP with the accession code phs001979.v1.p1 and can be accessed here: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs001979.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001979.v1.p1). The following databases were also used for analyses, as detailed in the manuscript: Molecular Signatures Database (v6.2) (<https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>), COSMIC (<https://cancer.sanger.ac.uk/cosmic>), cancerhotspot.org (<https://www.cancerhotspots.org/#/home>), and cbiportal (<https://www.cbiportal.org/>).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	7 Donors were included in this study. Physiologically normal skin tissue was collected from cadavers (up to 8 days post-mortem) or from surgical discard tissue of living donors. Skin tissue from cadavers was collected from either the UCSF Autopsy program or the UCSF Willied Body Program. No statistical method or calculation was used to determine the sample size. Our goal was to collect data from a large number individual cells from several donors across multiple anatomic sites. Since we performed single cell genotyping from multiple donors, each cell represents a n=1 and we sequenced a total of 133 samples in this study, which we believe is a sufficient sample size for this descriptive study.
Data exclusions	No data were excluded from analysis.
Replication	In this study, we genotyped multiple cells per tissue from multiple donors. By genotyping multiple cells both within an across people, we were able to replicate the patterns described in this manuscript.
Randomization	All single cells were sequenced using the same protocol and no treatment was applied to any subset of our cells. Therefore, randomization was not applicable for this study design.
Blinding	No treatment or variable was applied to any of our samples as they were all processed equally. Therefore, blinding was not applicable for this study design.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Our study includes biospecimens that were anonymously donated to us as surgical discard tissue from living donors or from cadavers from the UCSF Autopsy program or UCSF Willied Body Program. All donors were of European ancestry and ranged from 63 to 85 years in age.
Recruitment	The biospecimens were obtained as they were made available from persons undergoing dermatologic surgery or they passed away. We sought to collect biospecimens from European-ancestry individuals.
Ethics oversight	The organizations that we procured deidentified biospecimens have obtained IRB approval, consent from donors, and ethical clearance to collect, store, share, and perform genetic testing of biospecimens that meet the standards set by the Federal Privacy Act of 1974, California Information Practices Act of 1977, HIPAA (Health Insurance Portability and Accountability Act), and the NIH National Human Genome Research Institute. Specifically, the study utilized tissue samples banked under the Pathogen Discovery in Cutaneous Neoplasia/Cutaneous Neoplasia Tissue Bank protocol (10-01451).

Note that full information on the approval of the study protocol must also be provided in the manuscript.