Effects of psoriasis and psoralen exposure on the somatic mutation landscape of the skin

Sigurgeir Olafsson, Eike Rodriguez, Andrew R.J. Lawson, Federico Abascal, Philip H. Jones, Sascha Gerdes, Iñigo Martincorena, Stephan Weidinger, Peter J. Campbell, Carl A. Anderson

1. Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK
2. Department of Dermatology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.

Abstract

Somatic mutations are hypothesised to play a role in many non-neoplastic diseases. These diseases may also shape the somatic mutation landscape of affected tissues after onset. We performed whole-exome sequencing of 1182 microbiopsies dissected from lesional and non-lesional epidermis from 111 patients with psoriasis, a chronic inflammatory disease of the skin, to search for evidence that somatic mutations in keratinocytes may influence the disease process and to characterise the effects of the disease on the mutation landscape of the epidermis. We show that psoriasis is associated with increased mutation burden of the cell-intrinsic signatures SBS1 and SBS5 but not of UV-light, which remains the dominant mutagen in psoriatic skin. Despite the hyperproliferation of keratinocytes that characterises psoriasis, lesional skin remains highly polyclonal, showing no evidence of spread of clones carrying potentially pathogenic mutations. We find that the selection forces operating in the epidermis remain mostly unchanged in psoriasis and the mutational landscape continues to be dominated by clones carrying mutations in genes recurrently mutated in normal squamous epithelia. There is evidence of positive selection in previously reported driver genes, NOTCH1, NOTCH2, TP53, FAT1 and PPM1D and we also identify four driver genes (GXYLT1, CHEK2, ZFP36L2 and EEF1A1), that have not been previously described in studies of normal skin but which we hypothesise are selected for in squamous epithelium.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
irrespective of disease status. We describe the mutagenic effects of psoralens, a class of chemicals previously found in some sunscreens and which remain a part of a common photochemotherapy treatment for psoriasis (psoralens and UV-A, PUVA). Psoralens leave a distinct mutational signature in the genomes of exposed cells that is tightly linked with transcription, showing evidence of both transcription-coupled repair and transcription-coupled damage. These results suggest that somatic mutations in keratinocytes are unlikely to influence the pathogenesis of psoriasis and that while psoriasis has only modest effect on the mutation landscape of the skin, PUVA treatment has the potential to exert a unique and larger effects.

Main text - introduction

Somatic mutations have been extensively studied in cancers and have also been recognized for decades to cause mosaic forms of Mendelian disease. Somatic mutations occur in every cell of the body throughout life and technological advances have enabled significant progress in our understanding of the somatic mutation landscapes of phenotypically normal tissues, including blood, oesophagus, colon, liver, urothelium, and more. These studies have described how diverse mutagenic processes create the substrate for a ceaseless Darwinian competition between clones populating normal tissues. Clones acquiring mutations that provide a competitive advantage over their neighbours will expand, and may come to dominate. While there appears to be significant overlap between cancer driver mutations and those that drive clonal expansions in normal tissues, mutant clone expansions do not necessarily drive tumour formation. However, widespread replacement of wildtype cells with mutant clones can have functional consequences for the tissue, potentially contributing to common complex disease risk or influencing the disease progression or response to treatment.

Exposure to external factors, such as drugs or environmental insults, may influence somatic mutation landscapes and these exposures are sometimes associated with disease. For example, azathioprine, which is used to treat multiple immune-related conditions, leaves a characteristic mutational signature on the genomes of exposed cells in both the skin and colon. Disease itself can be a powerful external factor. The effects of non-neoplastic disease on the somatic mutation landscapes of affected tissues has been most comprehensively demonstrated in studies of liver biopsies from patients with non-alcoholic fatty liver disease and of colonic mucosa from patients suffering from inflammatory bowel disease (IBD). In both conditions, disease results in increased mutagenesis and large-
scale replacement of wildtype cells with clones carrying somatic mutations in metabolic and
immune related pathways, respectively. In IBD, chronic inflammation drives the expansion of
clones carrying somatic mutations affecting the toll-like receptor and interleukin-17 signalling
pathways within the colonic mucosa\textsuperscript{23–25}.

Here, we further characterise the somatic mutation landscape of diseased tissues by
studying chronically inflamed psoriatic skin. We focus on the epidermis, a relatively flat
structure punctured by hair follicles and sweat ducts and folded into extensions called rete
ridges which project into the underlying dermis, alternating with dermal projections known as
dermal papillae at most body sites. Targeted deep sequencing of cancer gene panels has
revealed that normal epidermis is a mosaic of clones and that genes commonly mutated in
squamous cell carcinomas and basal cell carcinomas, including \textit{TP53}, \textit{NOTCH1}, \textit{NOTCH2},
\textit{FAT1} are under strong positive selection in the epidermis under normal conditions\textsuperscript{22,26–28}.

The most common subtype of psoriasis, psoriasis vulgaris, accounts for about 90% of cases and manifests as well-defined plaques of thickened skin with an overlying silvery scale, most often on the knees, elbows and scalp\textsuperscript{29,30}. Histologically, psoriatic skin has a
distinct appearance characterised by epidermal hyperplasia, where hyperproliferation of
keratinocytes results in thickening of the epidermis and elongation of epidermal rete into the
dermis below. The disease can be managed using a variety of systemic therapies, including
interleukin-17 inhibitors, but also locally applied topical treatments and phototherapy. Here
we use whole-exome sequencing of hundreds of microbiopsies of epidermis isolated from
lesional and non-lesional skin of psoriasis patients to explore the extent to which somatic
mutations affect the disease process in psoriasis. We also describe the mutagenic
processes that are active in psoriatic skin, including an in-depth exploration of the mutagenic
effects of psoralen exposure, which is a part of commonly used phototreatment for psoriasis.

Lesional skin of psoriasis patients remains highly polyclonal

We recruited 111 psoriasis vulgaris patients between ages 18 and 88 at the
Department of Dermatology, University Hospital Schleswig-Holstein, Kiel, Germany. Patients
donated paired punch biopsies from lesional (n=111) and adjacent non-lesional (n=106) skin
(Figure 1a, Methods, Supplementary Table 1). Laser capture microdissection (LCM) was
used to isolate 1182 microbiopsies of epidermis from this material (946 lesional - 80%; 236
non-lesional - 20%) (Figure 1, Supplementary Tables 2 and 3, Methods). The median microbiopsy lengths along the basal membrane were 72 and 91 micrometers for lesional and non-lesional biopsies, respectively. Most microbiopsies correspond to <0.01 mm² of skin surface (Supplementary Figure 1) and commonly consist of individual rete of inflamed epidermis. DNA extracted from each of the microbiopsies was whole-exome sequencing to a median on-target depth of 56X and somatic substitutions, indels and structural variants were called (Supplementary Table 4, Methods).

We sequenced 18 near-replicate samples, where the same histological features visible on adjacent histological slides were dissected twice and independently sequenced. A median of 89% of substitutions and indels were called in both samples (Supplementary Figure 1d, Methods), providing a quantitative estimate of the sensitivity of the mutation calling pipeline (albeit excluding mutations with very low variant allele fractions in both samples).

Most somatic mutations are heterozygous within a cell and the clonal composition of a microbiopsy can be determined from the fraction of reads that report the alternate allele at the mutant site (the variant allele fraction - VAF). Perfectly clonal microbiopsies, where all cells derive from a relatively recent common ancestor cell, have VAF distributions centred around 50%. VAF estimates are shifted towards zero with increasing polyclonality. We found that even our small microbiopsies of epidermis rarely comprised fully clonal populations of cells, with most microbiopsies containing a mixture of clones. The VAFs of microbiopsies dissected from lesional and non-lesional skin were near identical (MedianVAFlesional = 0.26, MedianVAFNon-lesional = 0.25; Figure 1d) indicating that the hyperproliferation of keratinocytes associated with flare ups of psoriasis has minimal effect on the vertical spread of mutant clones. Some clones extended a few hundred micrometers across more than one microbiopsy and contributed a different cell fraction in each. The size of the clones appears similar to previous reports from people without psoriasis\textsuperscript{22,26}, although differences in methodologies prevent a formal quantitative comparison.
Figure 1: Overview of the sampling strategy and clonal structure of the tissue. a) Anatomical locations of paired biopsies from lesional (L) and non-lesional (NL) skin. b) An overview of the sampling pipeline. Punch biopsies from lesional and non-lesional skin were taken under local anaesthesia. These were histologically sectioned and stained. Laser capture microdissection was used to isolate multiple microbiopsies of epidermis from each skin biopsy. c) Histology image of a lesional biopsy (haematoxylin and eosin staining). Examples of microbiopsies of epidermis are highlighted. The black scale bar shows 500 micrometers. d) A histogram showing the median VAFs of mutations called in microbiopsies from lesional and non-lesional skin. The dashed lines show the medians across each tissue type which are 0.26 for lesional skin and 0.25 for non-lesional skin.
Supplementary Figure 1: Characteristics of the microbiopsies of epidermis. a) Histogram showing the volume of the biopsies by type. b) Histogram showing the surface area of the skin each microbiopsy covers. The vertical dashed lines show the medians of the distributions. c) Median on-target coverage for each microbiopsy. The medians for each type are 58X and 46X for the lesional and non-lesional microbiopsies, respectively. d) Sensitivity estimates derived by sequencing 18 near-replicate samples of epidermis from adjacent histological sections. The same histological features or rete tend to be visible across multiple histological sections. Near-replicate samples were obtained by repeated cutting of the same features into different wells. Refer to the methods for explanation of the “braid sampling”.

Mutagenic processes in the psoriatic skin

The oligoclonal composition of the microbiopsies means their mutation counts are not representative of the per-cell mutation counts in the skin. There is also a risk of counting the same mutation event multiple times when a clone extends across multiple microbiopsies. We performed Bayesian clustering of mutations based on their VAFs in different samples and used the statistical pigeonhole principle to derive a phylogenetic relationship between the clusters (Methods). Each tip of the phylogenetic trees represents a clone and we performed all subsequent analyses at the level of clones rather than the level of microbiopsies.
To determine which mutational processes are active in psoriatic skin, we used a Bayesian hierarchical Dirichlet process (Methods) to extract mutational signatures for each mutation clone and compared these to the COSMIC reference signatures\(^\text{31}\) (Figure 2a, Supplementary Figures 2 and 3, Supplementary Table 3). Unsurprisingly, the most abundant signature is SBS7b, which has been attributed to UV-exposure. This signature accounts for 80% of the mutations in the dataset. We also found the UV-related signature SBS7c, but it only accounts for 0.14% of the mutations. In agreement with previous studies\(^\text{16,22}\), we observed large variation in UV-associated mutation burden between cells 1-2mm apart in the tissue.

The second most prevalent signature in the dataset is not listed in COSMIC (v3.2). It is characterised by T>A, T>C and T>G mutations at TpA sites (Supplementary Table 5). This is consistent with the known mutagenic effects of a psoralens\(^\text{32,33}\), a class of chemicals that form a part of PUVA (psoralens and UV-A) phototreatment, which is used to treat psoriasis and other diseases of the skin. The signature is present both in lesional and non-lesional skin and accounts for 10.7% of the mutations in the whole-exome dataset, and for as much as 94% of the mutations in some individual clones (Figure 2a). As with UV-light, proximal cell clones sometimes show large variation in the number of psoralen-related mutations, with some showing extremely high mutation burdens. In the most extreme case, over 12,000 exonic mutations attributed to this signature were found in two related cell clones from patient 34, who has a history of extensive PUVA treatment. These clones were clear outliers in terms of the substitution burden and substitution mutational signature profiles (Figure 2b) but did not show a higher-than-expected burden of double-base substitutions or indels. Across the dataset in general, the indel mutation spectrum is identical (cosine similarity >0.99) between clones with high and low burden of the substitution psoralen signature (Supplementary Figure 5). The mutation spectra of double-base substitutions in both groups are also dominated by the UV-related CC>TT mutations (DBS1), suggesting that the mutagenic effects of psoralens are limited to single-base substitutions. The psoralen signature is further characterised below.

The remaining prevalent signatures identified are the result of cell-intrinsic mutational processes. One is a mix of the clock-like SBS1 and SBS5 signatures, that are found ubiquitously in normal tissues, including proliferative and post-mitotic tissues\(^\text{17,34}\). These signatures tend to be correlated within a tissue and could not be separated by our model. We also found a small number of samples exhibiting the mutational signatures of APOBEC activation (SBS2 and SBS13), which have been occasionally found to be present in a range of normal tissues\(^\text{10,34,35}\). While these signatures account for only 0.5% of the mutations in the
dataset as a whole, 22 clones had over 50 exonic mutations attributed SBS2/SBS13, accounting for up to 28% of mutations in the exposed clones.

We used a linear mixed effects model to estimate the effects of age and disease duration on the mutation burden after subtracting the mutations assigned to the psoralen signature and controlling for the anatomical location of the sample (Methods and Supplementary analysis 1). The mutation burden of the keratinocyte clones increased linearly with age, with each clone accumulating 14.6 mutations per exome per year (14.6 (10.1-19.0, 95% CI, $P=4.3 \times 10^{-9}$, linear mixed-effects models, Supplementary analysis 1). This is a much higher burden than that of other normal tissues, which range around 0.2-0.4 mutations per exome per year.$^{1,2,5,8,10,34}$

Lacking information about the frequency or severity of psoriasis flare-ups, we next tested if disease duration, which we have previously used as a proxy for inflammation exposure in IBD$^{23}$, is associated with mutation burden of keratinocyte clones (Methods). We did not find a significant effect of disease duration on the total mutation burden (-0.60 mutations per year (-3.4-2.2, 95% CI, $P=0.67$, likelihood ratio test of linear mixed-effects models)). However, the large variation in the burden of the UV-related SBS7b reduces statistical power to detect a disease effect on the total mutation burden. In our previous work on IBD-affected colonic mucosa, we found that the disease was associated with accelerated mutagenesis by the cell-intrinsic signatures SBS1 and SBS5$^{23}$. We fitted a linear mixed effects model using only the mutation burden attributed to SBS1/5 and found that psoriasis increases the mutation burden of these signatures by 0.16 mutations per exome per year of disease duration (0.038-0.29, 95% CI, $P=0.012$ - Likelihood ratio of linear mixed effects models, Supplementary analysis 1). We estimate the age effect of SBS1/5 to be 0.65 mutations per year (0.49-0.81, 95% CI, $P=2.5 \times 10^{-12}$, linear mixed effects model, Supplementary analysis 1). Psoriasis is thus associated with ~25% increase in the rate of these clock-like signatures in the skin.
Figure 2: Mutational signatures and mutation burden in lesional and non-lesional skin. a) The mutation burden of each clone and the relative contribution of each mutational signature. Clones are grouped by patient and patients are ordered by mutation burden of the clone with the most mutations. For clarity, only samples taken from the leg are shown. See Supplementary Figure 3 for other anatomical sites. b) The total burden of single base substitutions (SBS) as a function of age of the donor at the time of sampling. c) The burden of single base substitutions attributable to UV-light (SBS7b and SBS7d) as a function of age. d) The mutation burden of the clock-like SBS1/5 as a function of age. The y-axes in b) to d) refer to the exonic mutation burden.
Supplementary Figure 2: 96-class substitution profiles of the mutational signature components extracted using a hierarchical Dirichlet process. Apart from the Unassigned, components are ordered by the fraction of mutations assigned to each component in the dataset. Three unknown components were extracted that together account for 2.6% of the mutations. These have been merged with the Unassigned component in Main Figure 2 (see Methods).
Supplementary Figure 3: Mutational signatures across all cell clones by anatomical location of the skin microbiopsies.
Structural variants lead to loss of heterozygosity of driver mutations

We used the ASCAT software\cite{36,37} to call large (>1Mb) deletions, duplications and copy number neutral loss-of-heterozygosity (LOH) events in 1104 microbiopsies where ASCAT goodness-of-fit exceeded 90%. After filtering variants (Methods), 95 duplications, 294 deletions and 46 LOH events were called, with most events affecting only small segments of chromosomes (Supplementary Figure 4). Events affecting whole chromosomes were only detected in two of the 1104 microbiopsies. A single gain of chromosome 12 was observed in one microbiopsy and a loss of chromosome 19 in another. We observed nine and three LOH events affecting NOTCH1 and FAT1, respectively. Of these, eight and two events made putative driver mutations in these genes (see below) homozygous. LOH events that render mutations in NOTCH1 and FAT1 homozygous have been previously reported both in the normal skin and in the normal esophagus\cite{2,22,26}. 
Supplementary Figure 4: Structural variants identified in microbiopsies from each patient. The intensity of the colours increases with the fraction of microbiopsies from each patient in which the variant is called.
Mutational signature of psoralen exposure

Psoralens are a class of linear furocoumarins, polycyclic aromatic compounds naturally found in many crops, including citrus fruits and figs. They increase the photosensitivity of the skin and their synthetic forms are administered as part of phototherapy, (PUVA treatment, psoralens + UV-A) of psoriasis as well as other severe skin diseases such as eczema, vitiligo, and graft-versus-host disease. The patient is administered psoralens either orally, by applying a topical cream or by soaking in a bath containing a solution of psoralens. This is followed by high-dose exposure to UV-A light either administered locally with a UVA-lamp or body-wide in a UV-chamber. Psoralens were also included as tanning activators in some tanning lotions and sunscreens until a limit on their use was imposed by European safety regulations in 1996.

Psoralens are used as mutagens in molecular biology research and both PUVA treatment and the use of psoralen-containing sunscreens are known to increase the risk of skin cancers\textsuperscript{38–41}. Upon irradiation, the psoralen molecule binds to thymine forming a monoadduct. Further exposure to light can then cause the molecule to form an interstrand cross-link between thymines on opposite strands at TpA sites.

In our whole-exome data, we found a signature characterised by T>A, T>C and T>G mutations at TpA sites, consistent with the known mutagenic effects of psoralens\textsuperscript{32,33,42}. Out of the 111 donors, 21 patients had at least one microbiopsy with at least 50 exonic mutations attributed to this signature. Of these patients, 12 have a history of PUVA treatment in the clinical metadata. The 9 patients without documented PUVA treatment have disease durations ranging from 17-43 years prior to sampling and may have had undocumented PUVA treatment, but their ages and biopsy locations are also consistent with the possible use of psoralen containing lotions/sunscreens in the past.

To further characterise the mutagenic effects of psoralen exposure \textit{in vivo}, we performed whole-genome sequencing (WGS) of 16 microbiopsies from three patients who showed evidence of clear exposure in the whole-exome data and two of which have histories of PUVA treatment (Figure 3, Supplementary Table 2). The median sequencing depths for samples from each patient were 23.3X, 25.5X and 22.9X (Methods). We found an effect of sequence context that extended beyond the trinucleotide context, with sites that contain the double palindrome ApTpApT being most frequently mutated (Figure 3b).

We found a strong effect of transcription on psoralen-related mutagenesis in the WGS data, with 1.72 times more mutations occurring on the untranscribed strand than on the transcribed (1.86, 1.71 and 1.57 for T>A, T>C and T>G mutations, respectively, Figure 3a). This is partially due to the effects of transcription coupled repair (TCR), which removes
mutations from the transcribed strand, but the asymmetry is further compounded by transcription coupled damage (TCD) to the complementary non-transcribed strand (Figure 3c, Supplementary analysis 2). Although both TCD and TCR occur, the impact of TCR is about twice that of TCD, with the result that the mutation rate at TpA sites drops with increasing gene expression (Figure 3d).

In addition to the effects of transcription, there was an effect of replication on psoralen-related mutagenesis. The leading strand accumulates 9% more mutations compared to the lagging strand (Supplementary Figure 5b, Supplementary analysis 2). We divided the genome up into four bins in order of ascending replication timing and found that late replicating regions consistently showed higher mutation rates than early replicating regions (Supplementary Figure 5c).

In addition to its effect on mutagenesis, high burden of the psoralen signature is associated with clonal expansions of keratinocytes in our dataset. Microbiopsies derived from skin biopsies where at least one clone had >100 psoralen-related mutations have higher VAFs than microbiopsies dissected from unexposed skin (P=2.9 x 10^{-9}, Mann Whitney U test, Figure 3e). There was also a greater level of clonal spread between different microbiopsies, as indicated by the fraction of mutations shared between them. For each pair of microbiopsies, we calculated the fraction of mutations shared between the pair and plotted this as a function of the distance between the pair (Figure 3f). Microbiopsy pairs from psoralen-exposed skin account for 14.4% of all pairs in the dataset but for 35.8% of pairs that are more than 0.5mm apart and share more than 10% of their mutations (P=5.4x10^{-10}, Chi-squared test). This may be the result of psoralen exposure being correlated with age or severity of psoriasis, but may also be a direct consequence of the chemical itself. Psoralens are cytotoxic and may enable clonal expansions through the elimination of competitor clones. Exposure may also result in a higher number of driver mutations and higher fitness.

A number of canonical hotspot mutations in known skin cancer genes are caused by T>A, T>C or T>G mutations at TpA sites (Supplementary Table 6). There is evidence that UV-exposure affects the clonal dynamics in the skin and the spread of TP53 mutant clones in particular. We did not find evidence of stronger selection of mutations in any particular gene among clones which had >100 mutations attributed to the psoralen signature (Supplementary analysis 2).

Compared with a uniform mutational process, the psoralen-signature is 1.37 times more likely to result in splicing mutations in genes reported to be recurrently mutated in squamous cell carcinomas and normal skin. It would be especially likely to disrupt AT-AC introns, which have 5’-AT and AC-3’ boundaries, in contrast to the typical 5’-GT and AG-3’ splice sites.
Figure 3: Characterization of the mutational signature of psoralen exposure. a) A 96-class mutation profile of one of the outlier samples from patient 34 (see main text) who has a history of PUVA treatment and shows extremely high burden of the psoralen signature. The lighter shades represent mutated pyrimidine bases on the transcribed strand and the darker shades the untranscribed strand. b) A heatmap showing the effect of sequence context on psoralen-mutagenesis extends beyond the trinucleotide model, with bases 2bp either side of the mutated base affecting the mutation probability. c) Mutational densities in the vicinity of

Psoralen exposed overall: 13.8%
Poralen exposed in square: 35.8%
P = 5.4 \times 10^{-10}
the transcriptional start site of genes by expression level quintiles in sun-exposed skin in the GTEx dataset. Mutation rate on the transcribed strand drops in the transcribed region of genes, reflecting transcription coupled repair (TCR). In contrast, the mutation rate on the untranscribed strand is increased in the transcribed region, compared with intergenic regions, reflecting transcription coupled damage (TCD) of this strand. The effects become more pronounced with increasing expression. d) The mutation rate in ten bins of ascending gene expression. Higher expressed genes have lower mutation rates. e) The variant allele fractions (VAFs) of microbiopsies dissected from psoralen-exposed and unexposed skin. f) Pairwise comparison of microbiopsies. For any pair from the same skin biopsy, the plot shows the fraction of mutations found in both microbiopsies as a function of the distance between them (measured centre-to-centre). Psoralen exposed microbiopsies are enriched among pairs that are distant but nevertheless share many mutations in common (grey square).

Supplementary Figure 5: Further characterization of the psoralen signature. a) A comparison of the indel mutation spectra between samples with high and low burden of the psoralen substitution signature. The spectra are identical (cosine similarity >0.99), indicating that psoralens don’t cause indels. n is the number of mutation events. b) The effect of replication direction of psoralen mutagenesis in whole-genome sequenced samples. More mutations are observed on the leading versus lagging strand. c) The effect of replication timing on psoralen mutagenesis. Across all samples that underwent whole-genome sequencing the mutation rate is greater in late-replicating regions.
Positive selection in lesional and non-lesional skin

Previous studies have shown that a large fraction of cells in the normal skin carry mutations in genes that are recurrently mutated in keratinocyte cancers\textsuperscript{22,26}. The recurrent cycles of inflammation, hyperproliferation and remission that characterise psoriasis could create an environment where different selection forces drive clonal expansions, for example by propagating mutations that render cells resistant to cytotoxic effects of inflammation\textsuperscript{19,20,24,44}. We used the dNdScv software\textsuperscript{45} to assess the ratio of non-synonymous mutations to synonymous mutations after accounting for sequence context and regional differences in mutation rate across the genome (Supplementary Table 7, Supplementary analysis 3). We first carried out an unbiased, exome-wide screen for recurrently mutated genes in all microbiopsies and found nine genes which passed correction for multiple testing (Figure 4a). Mutations in five of these (\textit{NOTCH1}, \textit{FAT1}, \textit{PPM1D}, \textit{TP53} and \textit{NOTCH2}) have been previously shown to be under positive selection in normal skin in studies that have used deep sequencing of targeted gene panels\textsuperscript{22,26}. Four additional genes (\textit{GXYLT1}, \textit{CHEK2}, \textit{ZFP36L2} and \textit{EEF1A1}) that were not a part of the targeted panels or have not been previously reported to harbour mutations under positive selection in the normal skin reached significance in the current study. To explore if mutations in these four genes are uniquely under positive selection in psoriatic skin we repeated the selection analysis using only samples from lesional skin or non-lesional skin. No additional genes reached significance in these restricted analyses but \textit{ZFP36L2} and \textit{GXYLT1} did not reach significance when non-lesional microbiopsies were excluded (Supplementary analysis 3). \textit{ZFP36L2} and \textit{GXYLT1} showed nominal significance (P<0.05) in the non-lesional skin, suggesting that mutations in these genes are not specifically selected for under conditions of inflammation and that our failure detect significant evidence of positive selection is likely driven by lower power due to the reduced sample size.

To look for evidence of positive selection in additional genes, we performed restricted hypothesis testing of 27 genes in which mutations have been reported to be under selection in either the normal skin or normal oesophagus\textsuperscript{2,3,22,26}. Even with relaxed correction for multiple testing, none of those genes reached significance in this dataset (Supplementary analysis 3). A previous report found evidence of negative selection of mutations in \textit{PIK3CA}, \textit{DICER1}, \textit{CUL3}, \textit{NSD1}, and \textit{NOTCH4} in the skin\textsuperscript{22}, but this could not be replicated in the current analysis, possibly due to lack of power.

We next carried out a pathway-level dN/dS analysis, searching for enrichment of missense and truncating variants across 11 gene sets that were defined \textit{a priori} because of their relevance in either keratinocyte cancers or psoriasis pathology (Figure 4c, Methods).
We observed a strong enrichment of mutations in genes previously reported to be recurrently mutated in normal skin or in squamous cell carcinomas. Genes reported to be recurrently mutated in basal cell carcinomas showed a much weaker enrichment (after excluding \textit{TP53}, \textit{NOTCH1} and \textit{NOTCH2}). No evidence was seen for enrichment of either missense or nonsense mutations in any of the other pathways implicated in psoriasis pathogenesis. Together with the gene-level analysis, this suggests that somatic mutations in keratinocytes are unlikely to play a role in the pathogenesis of psoriasis.

Based on the VAFs of individual mutations and the volume of the microbiopsies in which they were detected, we estimated for each patient the fraction of cells that carry a mutation in each gene (Methods, Figure 4d). We tested if the fraction of mutant cells was different between lesional and non-lesional skin using a Wilcoxon test for paired measurements but found no differences that were significant after correction for multiple testing. This further supports the hypothesis that the selection forces operating in psoriasis are the same as those operating in the skin of people without psoriasis.

Figure 4: Positive selection in psoriatic skin. a) The number of mutations in genes found to carry a significant excess of mutations by functional annotation. b) Observed-to-expected

Supplementary Figure 6: Genes showing a significant enrichment of somatic mutations in psoriatic skin. a) Histogram showing the fraction of cells carrying one or more mutations in each gene for each patient. b) Lollipop plots showing the location and annotation of the identified mutations in each gene.

Discussion

We have used whole-exome sequencing of microbiopsies of epidermis derived from 111 patients with psoriasis vulgaris to study the effects of this chronic disease on the somatic mutation landscape of the skin. Psoriasis is associated with modestly increased risk of keratinocyte cancers⁴⁸,⁴⁹ but our results support the view that this increase in risk may be predominantly due to the effects of treatment rather than features of the disease per se.
Disease duration of psoriasis, our best proxy for inflammation exposure, is associated with increased mutation burden of the clock-like signatures SBS1 and SBS5. We estimate that psoriasis disease duration is associated with an additional 0.16 mutations per exome per year, which is approximately a quarter of the age effect of these signatures. While the P-value for this effect is modest, the increased burden of SBS1 and SBS5 in the context of inflammation is in line with our previous finding of increased mutation burden of SBS1 and SBS5 in colonic mucosa affected by inflammatory bowel disease\textsuperscript{23}. However, the increase in mutation burden associated with psoriasis appears to be smaller than in IBD, where the disease duration effect is almost equal in size to the age effect\textsuperscript{23}. The relative importance of this increase is also different in psoriasis versus IBD. The mutation burden of colonic mucosa is dominated by cell intrinsic processes and SBS1 and SBS5 contribute a large proportion of the total mutations\textsuperscript{5,23}. In contrast, the increased mutation burden in psoriasis tends to be drowned out by the mutagenic effects of UV-light (SBS7), leading to a much smaller relative increase of the mutation burden. In the 16 samples that underwent whole-genome sequencing, around 1% of the mutations overlap the bait-capture region used for the WES. From this we may roughly extrapolate that the effect of psoriasis disease duration is 16 mutations per genome per year (compared with 36 mutations/year for IBD).

We discovered that mutations in four genes that have not been reported before in studies of normal epidermis are under positive selection in our dataset. We hypothesise that non-synonymous mutations in these genes are under positive selection in squamous epithelia in general and are not specific to psoriasis or selected as a consequence of the inflammatory environment. Two of the genes, \textit{GXYLT1} and \textit{ZFP36L2} plausibly assert their effects through perturbation of the Notch signalling pathway. \textit{GXYLT1} encodes a xylosyltransferase that adds xylose to O-glucose-modified residues on epidermal growth factor-like repeats, which are found on the extracellular domains of all Notch proteins and are where missense mutations cluster in \textit{NOTCH1} (Supplementary Figure 6). We are not aware that \textit{GXYLT1} has been found to be recurrently mutated in keratinocyte cancers and it may be positively selected in normal skin without promoting cancer formation. \textit{ZFP36L2} is an RNA binding protein of the tristetraprolin family of post-transcriptional regulators. It targets the 3' untranslated regions of multiple mRNAs, including those of \textit{NOTCH1}\textsuperscript{46,47}. Consistent with a general role in squamous epithelia, mutations in \textit{ZFP36L2} have been reported to be under positive selection in normal oesophagus\textsuperscript{3}, a squamous epithelial tissue which has a similar selection landscape to the skin. \textit{CHEK2} encodes checkpoint kinase 2, a cell cycle checkpoint regulator and a well-established tumour suppressor. Like \textit{ZFP36L2}, it has been reported to be recurrently mutated in normal oesophagus\textsuperscript{3} and selection of mutations in \textit{CHEK2} seems unlikely to be a specific feature of psoriasis. Finally, \textit{EEF1A1}
encodes the alpha subunit of the elongation factor-1 complex. Unlike \textit{GXYLT1}, \textit{ZFP36L2}
and \textit{CHEK2}, it was included in an extended panel of genes tested for evidence of selection
in a subset of samples in a previous study\textsuperscript{22}, but did not reach significance. \textit{EEF1A1} did
reach significance in our previous analysis of normal urothelium from bladder cancer
patients\textsuperscript{10}, suggesting it too may be generally selected for in squamous epithelia. Mutations
across all nine genes that reached significance in the study appeared to be present at a
similar frequency across lesional and non-lesional skin from psoriasis patients.

Complex diseases vary greatly in their effects on the clonal competition within affected
tissues. In contrast to inflammatory bowel disease and chronic liver disease, which show
mechanisms of positive selection distinct from normal colon and liver\textsuperscript{9,23–25}, psoriasis seems
to have minimal impact on the selection landscape of the skin. Lesional skin remains highly
polyclonal and the selection landscape is dominated by the same genes which are found to
be recurrently mutated in squamous epithelial tissues of people without psoriasis. IBD and
liver disease are both associated with significant stem cell death at affected sites which
selects for cells that are able to resist the cytotoxicity, enabling their bottleneck expansion as
the tissue is regenerated. In contrast, psoriasis is associated with hyperproliferation of
keratinocytes but limited cell death in the basal stem-cell layer. Mutations that enable cells to
outcompete their neighbours in the basal layer are favoured, and those appear to be the
same mutations as are selected for in the skin of people without psoriasis.

Genome-wide association studies have revealed both disease specific and
widespread sharing of disease mechanisms between immune-mediated diseases. Many
pathways and disease mechanisms are shared across multiple autoimmune conditions but
the extent to which these relationships manifest in the somatic mutation landscape of
affected tissues is not clear. Comparison with our earlier work on IBD shows that although
they are both Th17 mediated chronic inflammatory diseases of epithelial tissues, and share
several GWAS loci in common\textsuperscript{50–52}, IBD and psoriasis have very different effects on the
somatic mutation landscapes of affected tissues. In IBD, mutations in immune-related genes
are under positive selection and these may enable cells to escape the cytotoxic effects of IL-17.
Similar mutations were not observed in psoriasis, and thus we conclude that although
somatic mutations may offer insights into the cellular dynamics that accompany disease,
somatic mutations in the epidermis are unlikely to play a role in the pathogenesis of
psoriasis.

We have given a detailed description of a mutational signature that we believe to be
the result of psoralen exposure. While psoralens have been known to be mutagenic for
decades, we add considerable detail to our understanding of this mutagen in humans \textit{in vivo}.
We show that the mutational signature has a strong transcriptional strand bias which results
both from transcription coupled repair and transcription coupled damage. Transcription
coupled damage has been most comprehensively described for T>C mutations in COSMIC
signature SBS16, which is found almost exclusively in the liver and may result from alcohol
consumption. SBS16 is thought to result from the chemical modification of adenine
residues which likely occurs more frequently when the DNA is in unwound form during
transcription. Our data suggest that psoralen molecules similarly preferentially react with
thymine in unwound DNA. Psoralens are used in experimental systems to study
mutagenesis and the repair of interstrand cross-link lesions. It has been demonstrated that
the repair of psoralen-related interstrand cross links does not depend on the Fanconi
anaemia pathway but rather depends on the DNA glycosylase NEIL3, which facilitates the
removal of the lesion without the formation of a double-strand break. In agreement with
this, we find no effect of psoralens on the burden of indels or structural variants, as might be
expected if the cross-links required double strand breaks to resolve.

One patient with extremely high burden of the psoralen signature also has history of
extensive PUVA-therapy, suggesting that PUVA therapy can result in a large number of
somatic mutations and may exert a much larger effect on the somatic mutation burden of the
skin than the disease itself. It is also associated with clonal expansions of clones in exposed
skin. This may occur through changes in the selection environment, although we were not
able to detect any gene-level differences on selection in the current analysis. However, a
limited number of rounds of PUVA treatment are generally considered safe and establishing
a dose-response curve for the PUVA process is an important direction of future work.

While psoralen exposure in the context of psoriasis is particularly likely to occur
during PUVA-therapy, other means of exposure include the use of sunscreen or tanning
lotions containing psoralens or even the consumption of furocoumarin-rich foods. It has been
hypothesised that orally ingested furocoumarins increase risk of skin cancers, although
further validation and functional work is required to confirm the relationship. We propose that
testing for a relationship between furocoumarin consumption and the psoralen mutational
signature in sun exposed skin offers a mechanistic way to test this hypothesis.

In summary, our results suggest that chronic inflammation of the skin is associated with
increased mutation burden of cell-intrinsic signatures but has little effect on the clonal
competition within the epidermis. PUVA treatment, and likely psoralen exposure in general,
leaves a distinct mutational signature on the genomes of exposed cells and can result in a
very large number of somatic mutations without resulting in malignant transformation. In
recent years, our understanding of somatic evolution in normal tissues has been
transformed. We are now in a position to carry out comparative studies of non-malignant
disease, treatments and other powerful environmental exposures to understand the effects these have on cancer risk and the biology of affected cells.

Acknowledgements

We thank the staff of the Wellcome Sanger Institute’s Sample Management, Sequencing, and Informatics teams for their contribution to the study. Special thanks go out to Yvette Hooks, for histological processing and sectioning of the skin biopsies and to Deborah Plowman and Sophie Leggett for practical assistance in sample and data management. We also thank the German psoriasis patients who donated tissue samples for the study. This research was funded by the Wellcome Trust [Grant numbers 206194 and 108413/A/15/D].

Author contributions

SO, PJC and CAA designed the study with contributions from ER, SW, PHJ and IM. ER, SG and SW consented patients for the study, collected clinical metadata and took and fixed the tissue biopsies. SO stained and imaged the tissue sections and performed laser capture microdissection of microbiopsies of epidermis. SO performed the calling and quality control of substitutions and indels. ARJL performed the calling of structural variants. SO performed all statistical and bioinformatics analyses with contributions from ARJL, FA, IM, PJC and CAA. SO wrote the initial draft of the manuscript and all authors contributed to the interpretation of the findings and the final version of the manuscript.

Competing interests

S.W. is co-principal investigator of the German Atopic Eczema Registry TREATgermany; has received institutional research grants from Sanofi Deutschland GmbH, LEO Pharma, and La Roche Posay; has performed consultancies for Sanofi-Genzyme, Regeneron, LEO Pharma, AbbVie, Pfizer, Eli Lilly, Kymab, and Novartis; has lectured at educational events sponsored by Sanofi-Genzyme, Regeneron, LEO Pharma, AbbVie, Novartis, and Galderma; and is involved in performing clinical trials with many pharmaceutical companies that manufacture drugs used for the treatment of psoriasis and atopic eczema. C.A.A. has received consultancy or lecture fees from Genomics plc, BridgeBio and GlaxoSmithKline.
Data and materials availability

Raw sequencing data are available in the European Genome-phenome Archive (EGA) under accession number EGAS00001004882. Intermediary and supporting files, including mutation calls, mutational cluster assignments, phylogenetic trees, histological images, spatial relationship matrices and more are available in a Mendeley data repository (DOI:10.17632/rfcy88sb9s.1). Code supporting the main analyses of the manuscript is provided in supplementary analysis files. Custom scripts documenting the mutation filtering and clustering, signature extraction and more are publicly available at https://github.com/Solafsson/somaticPsoriasis.

References

rates and distinctive mutational signatures in normal human cells. *bioRxiv*


Effects of psoriasis and phototreatment on the somatic mutation landscape of the skin - Methods

Sigurgeir Olafsson\textsuperscript{1}, Eike Rodriguez\textsuperscript{2}, Andrew R.J. Lawson\textsuperscript{1}, Federico Abascal\textsuperscript{1}, Philip H. Jones\textsuperscript{1}, Iñigo Martincoren\textsuperscript{a,1}, Stephan Weidinger\textsuperscript{2}, Peter J. Campbell\textsuperscript{1}, Carl A. Anderson\textsuperscript{1}

1. Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK
2. Department of Dermatology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.

1. Human tissue attainment and processing
   1.1 Tissue donation and fixation
   1.2 Laser-capture microdissection of epidermis

2. DNA sequencing
   2.1 Whole exome sequencing
   2.2 Whole genome sequencing

3. Somatic mutation calling
   3.1 Mutation calling in whole-exomes
      3.1.1 Substitution calling with CaVEMan
      3.1.2 Indel calling with cgpPindel
      3.1.3 Binomial filtering of somatic mutation calls
      3.1.4 Sensitivity of the substitution and indel calls from whole-exome data
      3.1.4 Structural variant calling in the exome data with ASCAT
   3.2 Mutation calling in whole-genomes

4. Identification of SNV clusters by hierarchical Dirichlet processing

5. Inference of phylogenetic trees

6. Mutational signatures
   6.1 Extraction of mutational signatures by hierarchical Dirichlet processing
   6.2 Characterization of the PUVA signature

7. Mutation burden estimation
   7.1 Adjusting the burden for coverage and VAF
   7.2 Linear mixed effect models to test for an effect of disease duration

8. Selection and driver analyses
1. Human tissue attainment and processing

1.1 Tissue donation and fixation

Round punch biopsies, 4 mm in diameter, from lesional and adjacent non-lesional skin were
donated by psoriasis patients presenting to the Department of Dermatology, UKSH Kiel
between 2017 and 2019. Biopsies were taken under sterile conditions under local
anesthesia and the emerging holes were closed with seam stitching and bandaged. All
donors gave informed consent for genetic research of the material and the study was
approved by the research ethics committee of Christian-Albrechts University in Kiel
(A100/12), the National Health Service (NHS) Research Ethics Committee (Yorkshire & The
Humber - South Yorkshire Research Ethics Committee, REC ID 20/YH/0244, IRAS ID
286843) and by the Wellcome Trust Sanger Institute Human Materials and Data
Management Committee (approval number 20/0085).

Biopsies were fixed in RNAlater (AM7021, ThermoFisher) upon collection following the
manufacturer’s instructions and frozen. One half of each biopsy was used in this study and
one half retained for use in future projects.

1.2 Laser-capture microdissection of epidermis

Skin biopsies were processed using a Tissue Tek VIP 6 AI tissue processor (Sakura Finetek,
Leiden, Netherlands) and embedded in paraffin. The blocks were sectioned into 10 μm thick
sections using an Accu-Cut SRM 200 microtome (Sakura Finetek). Sections were fixed to 4
μm PEN membrane slides (11600288, Leica). Slides were stained with haematoxylin and
eosin and imaged using a NanoZoomer 2.0-HT slide scanner (Hamamatsu Photonics,
Hamamatsu, Japan). Samples of epidermis were dissected from this material using laser
capture microdissection microscopy (LMD7000, Leica). Cells were lysed using Arcturus
PicoPure DNA extraction kit (Applied Biosystems) according to the manufacturer’s
instructions.

The volume of the microbiopsies was determined by adding together the size estimates of
the cuts from the LCM software and multiplying this by the thickness of the sections (10 μm).
The same histological features from serial sections were often cut into the same well to
increase the DNA-yield of the samples (a practice referred to as z-stacking). The surface
area of the samples was determined by measuring the width of the samples along the basal membrane and multiplying this with the section thickness and the number of sections (z-stacks) separating the first and last sections dissected into the same well. When microdissecting, occasionally, individual microdissections do not drop to the bottom of wells but either fall outside a well or get stuck to the side of a well. Those dissections do not contribute any DNA to the sample but are nevertheless part of the volume and surface area estimation. The size estimates should therefore be considered upper bound estimates.

To assess the sensitivity of the mutation calling, we cut 18 technical duplicate samples (see methods section 3.1.4). These were cut either from histological sections directly adjacent to the original samples (where the same histological features were clearly visible) or using a “braid sampling” strategy, where alternate z-stacks were cut into different wells of the sample collection plate.

2. DNA sequencing

2.1 Whole-exome sequencing

1207 microbiopsies (Supplementary Table 2) from 111 individuals were whole-exome sequenced on Illumina NovaSeq 6000® machines using 150bp paired-end reads and the Agilent SureSelect Human All Exon V5 bait set (S04380110). Paired-end reads were aligned to the human reference genome (build hg38) using BWA-MEM1. PCR duplicates were marked using biobambam2 and duplicate statistics were calculated using Picard (v.1.131) (http://broadinstitute.github.io/picard/). Sample contamination estimates were calculated using VerifyBamID3 and the on-target coverage was calculated using samtools (v.1.11) depth command, considering only reads with base quality and mapping quality greater than 30. The median on target coverage for all samples was 56X (range: 18-190) for the dataset as a whole but 58X and 46X for microbiopsies from lesional and non-lesional skin, respectively. The difference in coverage is due to the greater volume of the microbiopsies from lesional skin (see Supplementary Figure 1), which results in greater library complexity and translates to lower PCR duplicate rate and higher coverage. We cut columns of cells and the thicker epidermis of lesional skin results in a greater sample volume for the same surface area covered.

2.2 Whole-genome sequencing

Whole-genome sequencing was performed on 16 microbiopsies from 3 donors (patients 18, 21 and 34) which showed the signature of PUVA exposure in the WES data (see main text). The sequencing was carried out in two pools of 8. Samples from patients 18 and 21 were multiplexed together in one run on an Illumina NovaSeq 6000® machine and samples from patient 34 in a second run. The median coverages were 23.3X, 25.5X and 22.9X for microbiopsies from patients 18, 21 and 34, respectively. The specific microbiopsies were selected for whole-genome sequencing on the basis of relatively high levels of clonality and...
limited relatedness on the phylogenetic trees constructed from the WES data. Higher clonality results in a greater sensitivity for calling mutations when sequencing to 20-30X coverage. The limited sharing of mutations allows more independent mutation events to be identified.

3. Somatic mutation calling

3.1 Mutation calling in whole-exomes

Substitution and indel calling in the whole-exomes was carried out in three steps: discovery, initial filtering of the call set and genotyping of all mutations called in an individual in all samples from that sample donor. Further filtering was applied after genotyping somatic mutations as described below.

3.1.1 Substitution calling with CaVEMan

Substitutions were called using CaVEMan (v.1.15.1) (Cancer Variants through Expectation Maximization) (https://cancerit.github.io/CaVEMan/). Mutations were called against an unmatched normal with the copy number options manually set to 10 and 2 for the mutant and wild-type copy numbers, respectively. The samples were compared against a normal panel consisting of 75 unrelated normal samples to remove common SNPs. Mutations were further filtered if the reads reporting the mutations had a median alignment score lower than 140 or if >50% of the reads were clipped.

All mutations passing these filters in any sample from a donor were next genotyped in all samples from that donor. Mutations could thus be called in a microbiopsy if they failed the criteria above assuming they had passed the filtering criteria in another microbiopsy from the same individual (for example, a mutation in a clone that dominates in one microbiopsy and extends into an adjacent microbiopsy but at lower levels could still be called in the latter sample). We used the bam2R() function of the deepSNV package (v. 1.40.0) in R to generate pileups of all sites mutated in any sample using only mapped reads that had base quality and mapping quality greater than 30 and which were mapped in a proper pair, were not PCR duplicates, were the primary alignment and which passed platform quality check (sam flags 3847 see https://broadinstitute.github.io/picard/explain-flags.html). After these filters, we required a coverage of at least 4X at the site and at least three reads reporting the alternate allele in at least one sample from a donor to call a mutation.

Adjacent substitutions called in the same sample were merged into a double-base substitution call if the number of reads reporting the reference and alternative alleles were not significantly different (Fisher test).
3.1.2 Indel calling with cgpPindel

Indels were called using cgpPindel (v.3.5.0) (https://github.com/cancerit/cgpPindel)\(^6\) using the same unmatched normal sample that was used to call substitutions. We generated pileups of the indel calls in the same way as described above for substitutions and required a coverage of at least 4X and at least three reads reporting the alternate allele in at least one sample from a donor to call a mutation as before.

3.1.3 Binomial filtering of somatic mutation calls

To filter rare germline variants not removed by the comparison with the normal panel we applied an exact binomial test of the number of reads reporting each mutation, as previously described\(^7\). Heterozygous germline variants are expected to be present at a VAF of 0.5 in every sample from a patient. For each mutation, we compared the number of reads reporting the reference and alternate alleles across all samples from that patient. We tested the hypothesis that the read counts for the variants were drawn from a binomial distribution with a probability of success of 0.5, or 0.95 for mutations on the sex chromosomes in men. We applied Benjamini-Hochberg correction for multiple testing and excluded mutations with \(q>10^{-3}\). We also used binomial filtering to remove erroneous mutation calls. Recurrent sequencing artefacts will be randomly distributed across samples and can be modelled as being drawn from a binomial distribution. In contrast, true somatic mutations will have a high VAF in some samples whilst being completely absent from others. The latter are best represented by a beta-binomial with a high overdispersion. For every mutation call, we calculated the maximum likelihood overdispersion parameter (\(\rho\)) in a grid-based way (ranging the value of \(\rho\) from \(10^{-6}\) to \(10^{-0.05}\)), like previously described\(^7\). Calls with \(\rho<0.1\) were filtered as likely artifactual.

3.1.4 Sensitivity of the substitution and indel calls from whole-exome data

To estimate the sensitivity of the mutation calls we sequenced 18 technical duplicate samples and compared the mutations called in the original samples to those in the technical duplicates. The technical duplicates constitute the same histological features (rete ridges) as the original samples. Assuming the same sensitivity in both samples, a maximum likelihood estimate for the sensitivity when mutations not present in either sample go unobserved is:

\[
s = \frac{2 \times n_2}{n_1 + 2 \times n_2}
\]

Where \(n_2\) is the number of mutations called in both samples and \(n_1\) is the sum of mutations called in only one sample. Using this equation, we estimated a median sensitivity of 89% in the technical duplicates. This estimate should be considered to be a lower bound because:

1. The sensitivity depends on the coverage, which is uneven for the members of a pair.
2. Although the technical duplicates were cut from adjacent histological sections to the original samples, they may comprise slightly different combinations of keratinocyte clones compared with the original samples. Some mutations may therefore be truly unique to each sample. This does not apply to samples where alternate z-stacks were cut into different wells of the sample collection plate (referred to as “braid sampling” in Supplementary Figure 1D).

3.1.4 Structural variant calling in the exome data with ASCAT

The B-allele fractions and coverages were calculated at all common SNP sites (as defined by dbSNP build 150 for GRCh38p7) located within the exome panel using ConstructASCATFiles (https://github.com/MathijsSanders/ConstructASCATFiles). Normalised LogR and BAF values were generated by carrying out quality control and PCA analysis of the coverage using PREASCAT (https://github.com/MathijsSanders/PREASCAT). Nominated matched normal samples in Supplementary Table 1 were used as germline controls. Structural variants were called using Allele-Specific Copy Number Analysis of Tumors (ASCAT8; https://github.com/VanLoo-lab/ascat). A penalty score of 100 was used for segmentation. 101 microbiopsies were excluded from this analysis either due to the inability of ASCAT to find an optimal solution or if the goodness-of-fit for the optimal solution was <90%. We removed from the call list all segments smaller than 1 Mb in size, segments on chromosome X, segments within the extended MHC region (Mb 20-38 on chr6) and recurrent artifactual calls on Mb19 of chr14. We further removed segments that were also called in the matched normal of the sample or were called in microbiopsies from both lesional and non-lesional skin (which never showed a clonal relationship in the SNV data) or were called in all samples except 0, 1 or 2 from the same individual (these being likely germline in origin). Finally, we removed microbiopsies for which over 10 structural variants were called and which also had purity of less than 0.4, as we considered these likely problematic.

3.2 Mutation calling in whole-genomes

For the 16 microbiopsies that were whole-genome sequenced we used CaVEMan and cgpPindel to call and filter small variants in the same way as described above. However, in this case calling was performed against a matched normal sample chosen because it was phylogenetically unrelated to the index sample. This was possible because samples were whole-genome sequenced only after whole-exome sequencing and was preferable because the low number of samples per patient (and the lower aggregate coverage) reduces the accuracy of the binomial filtering of germline variants described above.

We used ascatNgs (internal version.4.5.0, https://github.com/cancerit/ascatNgs)9 to estimate the copy number and purity of the samples. Those estimates were used by CaVEMan, rather than the copy number options being manually pre-set as described for the whole-exomes above.
4. Identification of SNV clusters by hierarchical Dirichlet processing

We implemented a nonparametric Bayesian hierarchical Dirichlet process (HDP) to cluster autosomal single base substitutions with similar variant allele fractions (VAFs). The full mathematical and implementation details of the model are described in a previous publication\(^\text{10}\). Briefly, clones of cells are present across different microbiopsies and this manifests as clusters of mutations that are found at similar VAFs. For every mutation, we have two vectors, one containing the number of reads reporting the alternate allele and another containing the total sequencing depth at each microbiopsy. We assume that each mutation can be assigned to exactly one cluster but the number of clusters is unknown. We aim to estimate the number of clusters present across all the microbiopsies dissected from a patient, the location of each cluster in the n-dimensional VAF hypercube and the allocation of mutations to each cluster.

We model the data using an N-dimensional Dirichlet process (NDP) clustering model, where the distribution of clone sizes and numbers follows a Dirichlet process. This has the advantage that there is no need to pre-specify the number of clusters present. Instead, mutations are moved around the clusters and in each sampling iteration, there is a defined probability that a mutation will initiate a new cluster which was not present in previous iterations. Clusters can also cease to exist if all member mutations are assigned to other clusters. Thus, the number of clusters varies throughout the sampling chain.

We ran the Gibbs sampler for 25,000 iterations, dropping the first 15,000 as burn-in. We used the ECR algorithm\(^\text{11}\), implemented in the R package label.switching, to resolve the label-switching problem associated with mixture models. To avoid overly complex solutions, we imposed an upper limit of 100 clusters per patient. We kept for downstream analysis only those clusters that were present at a minimum VAF of 0.05 in at least one microbiopsy and had a minimum of 10 unique mutations allocated to them.

5. Inference of phylogenetic trees

Each cluster of single-base-substitutions identified by the NDP algorithm represents a branch of the phylogenetic tree for that patient. We applied the statistical pigeonhole principle to infer phylogenetic relationships between clusters. Given clusters A and B, if the combined mutant cell fraction (CF) of both is >100% (VAF > 0.5) within the same microdissections and B consistently shows a lower CF than A, then that is strong evidence that B is nested within A, that is mutation cluster B represents a sub-clone of clone A. If the combined mutant cell fraction is ≤100%, only weak evidence of nesting exists. If B is found at a higher VAF than A in some microdissections but at lower VAF in others, the clusters are interpreted as being independent clones without nesting.

We treated each tip of the phylogenetic tree for each patient as a clone. The length of the branches from the root (germline) was used in the mutation burden calculations, as described below.
6. Mutational signatures

6.1 Extraction of mutational signatures by hierarchical Dirichlet processing

To extract mutational signatures and estimate the exposure of each signature, we used a second hierarchical Dirichlet process\(^\text{12}\), as implemented in the hdp R package (https://github.com/nicolaroberts/hdp). HDP performs comparably to signature extraction methods that use non-negative matrix factorization but has the advantage of being able to model the relationships between samples and to simultaneously discover new signatures and quantifying the exposure to known signatures. Note that this is a separate implementation from the model used to cluster single base substitutions described above. In this case, we sought to identify clusters of mutations corresponding to mutational signatures. The data was organised into a tree structure where the root contained all the mutations in the dataset. This node had as children one node that represented the most-recent common ancestor of all the patients and additionally frozen pseudo-count nodes for signatures that are to be used as priors in the model. The pseudo-nodes contained 10,000 pseudocounts each. We used signatures 1, 5, 2, 13, 7a, 7b, 7c, 7d, 17a, 17b, 18 and 38 as priors. During the Dirichlet process, mutations from the dataset may join the pseudocount clusters but the pseudocounts are frozen such that they are unable to leave the initial cluster. The patient ancestor node had as children one node for each of the patients and each patient node had as children one node for each branch of the phylogenetic tree (see the R-code accompanying this manuscript).

The hyperparameters for the \(\alpha\) clustering parameter (\(\alpha\) and \(\beta\)) were both set to 1. The model was initialised with 13 data clusters (number of priors plus one, parameter ‘initcc’). After the initiation, an MCMC algorithm (Gibbs sampler) probabilistically moves each individual mutation to a cluster with a high proportion of mutations in the same category and/or a high proportion of mutations in that branch and/or parent node. The first 100,000 iterations of the Gibbs sampler were not collected (parameter ‘burnin’) but after that we sampled the posterior 200 times (parameter ‘n’) every 2000 Gibbs iterations (parameter ‘space’). We sampled the concentration parameter three times (parameter ‘cpiter’) after each Gibbs sampling. We ran 20 chains using different random seeds and combined the results for signature extraction. Any clusters with a cosine similarity greater than 0.9 were merged while clusters with no significant data categories or sample exposures were combined into a null signature, representing the fraction of the data that is unexplained by the model.

The model extracted nine signature components in addition to the Unassigned component (Supplementary Figure 2). Among these were components corresponding to COSMIC signatures SBS7b, SBS1/5, SBS2, SBS7c and SBS13, all of which were included as priors in the model. A novel component characterised by mutations at TpA sites accounted for just under 11% of the mutations in the dataset. This is the signature we attribute to psoralen exposure, which in the context of psoriasis is likely to occur during treatment with psoralens and high-dose UV-A (PUVA), see the main text. Finally, three additional components were extracted, Unknown components N1-3 in Supplementary Figure 2. These together accounted for 2.6% of the mutations in the dataset. They may represent individual variation in repair of UV-damage or they may be artefacts of the signature extraction model. We do not have sufficient confidence in these components to draw...
conclusions from them and have added them to the Unassigned component for subsequent analyses. The model did not extract the UV-related SBS7a and SBS7d. The Unassigned component has a dominant T>C peak that likely corresponds to the prominent peak from SBS7d (Supplementary Figure 2). The SBS7b component may be a composite component consisting of a mixture of SBS7b and SBS7a which could not be separated by the model.

6.2 Characterisation of the psoralen signature

From the whole-exome data, we identified a number of samples which showed a large number of mutations at TpA sites, consistent with the known mutagenic effects of psoralens\(^{13-15}\). To enable further characterization of this signature, we selected 16 microbiopsies from patients showing clear evidence of psoralen exposure for whole-genome sequencing.

To visualise the trinucleotide and pentanucleotide spectrums associated with psoralen exposure, we used the R-package MutationalPatterns\(^{16}\) (v. 3.4.0) together with BSgenome (v 1.60.0). To calculate the transcriptional strand bias, we used the gene definitions from the R-package TxDb.Hsapiens.UCSC.hg38.knownGene (v 3.13.0) and the strand_occurrences() and strand_bias_test() functions from MutationalPatterns.

To test for transcription coupled damage, we carried out a similar analysis to that originally used to describe transcription coupled damage in liver cancers\(^{17}\). We divided protein coding genes into quintiles by ascending expression in sun-exposed skin from the GTEx dataset\(^{18}\) (v8). We extracted the transcriptional start site (TSS) and the strand of each gene from Gencode (v27) and defined ten 1kb bins upstream and downstream of the TSSs. We pooled T>[ACG] and A>[CGT] mutations at TpA or ApT sites from all whole-genome sequenced samples. If the gene is on the (-) strand, the transcribed strand is the reference and we counted the number of T>[ACG] mutations overlapping each 1kb bin. If the gene is on the (+) strand, the transcribed strand is the complement of the reference and we counted the number of A>[CGT] mutations. This was reversed for the untranscribed strand. We observed a drop in the mutation rate on the transcribed strand upstream of the TSSs, indicating transcription coupled repair. However, we also found an increased mutation burden on the untranscribed strand, indicative of transcription coupled damage of this strand. To test the statistical significance of the increased mutation burden, we fit two linear models with and without a parameter indicating if each position was upstream or downstream of the TSSs and used a likelihood ratio test to test if the fit of the model was improved. Figure 3c shows the mutation rate in each 1kb bin relative to the -10kb bin, the intergenic bin furthest from the TSS.

To test the effect of gene expression levels on PUVA mutagenesis we again used expression data from sun-exposed skin from the GTEx dataset\(^{18}\) (v8). We split protein coding genes into 10 equally sized bins by ascending levels of expression. We used the Bedtools\(^{19}\) intersect function (v. 2.18) to count the number of mutations overlapping genes in each bin. Figure 3d shows the relative mutation rate in each bin compared to the lowest expression bin.
The replication strand is dependent on the replication origins, which for the human genome are only partially known and likely vary to some extent between cell types. An analysis of replication strand bias requires replication timing data and must assume that the position of the replication origins are at least partially conserved between the cell type of interest and the cell type in which the timing of replication was established. We used two external sources of replication timing data. The results reported in Supplementary Figure 4 use data from a previous analysis of replication timing\(^{17}\) that focused on "timing transition regions", which are highly conserved across cell types and species of eukaryotes\(^{20,21}\). As a second independent validation, we used replication timing data from the ENCODE project\(^{22}\) that is provided with the MutationalPatterns package and obtained similar results.

To assess the effect of replication timing on psoralen-related mutagenesis, we used the same replication timing data as for the replication strand analysis described above. The genome was split into four bins in ascending order of replication timing. For the regions in each bin, we estimated the psoralen mutation rate by counting the number of T\(>\)ACG/A\(>\)CGT mutations overlapping the regions and dividing this by the number of TpA/ApT sites in each bin. The mutation overlap was determined using the Bedtools\(^{19}\) intersect function (v. 2.18).

To see if psoralen exposure was associated with increased levels of clonal spread, we defined patients as psoralen exposed if at least one microbiopsy from the patient had over 100 mutations attributed to the psoralen signature. We then compared the VAFs between all microbiopsies dissected from exposed patients and microbiopsies dissected from non-exposed patients using a Mann Whitney U test. To also test for a greater expansion of clones across microbiopsies, we generated spatial relationship matrices for lesional and non-lesional biopsies from which more than one microbiopsy was sequenced. From the histological images, we manually measured the distance between the centres of the microdissections. This data is made available through the Mendeley repository referred to in the main text. For all pairs of microbiopsies dissected from the same skin biopsy, we plotted the fraction of mutations the two microbiopsies shared in common as a function of the distance between the pair (Figure 3f). We tested if microbiopsy-pairs from psoralen-exposed patients were enriched among the set of pairs that were both distant (separated by >500 micrometres) and which shared a high fraction of their mutations (>10%) using a Chi-square test. The cut-offs described in this paragraph were arbitrarily chosen but the results of both of these tests hold true across a range of threshold values.

7. Mutation burden estimation

7.1 Adjusting the burden for coverage and VAF

The number of mutations detected depends partially on both the sequencing coverage and clonal composition of the microbiopsy. In this study, we required a minimum coverage of 4X of a site and at least 3 reads supporting the alternative allele to call a mutation. Higher coverage enables the detection of more sub-clonal variants while greater polyclonal
composition means the available reads are split between more clones. Any comparison of
the mutation rate must take this technical variability into account.

To adjust the mutation burden of each cluster for coverage and VAF, we estimated
the sensitivity for calling a true mutation belonging to the cluster in each microbiopsy, $S_{bi}$.
The total sensitivity of the cluster, $S_{c}$, is one minus the sensitivity of the cluster in each
microbiopsy, $S_{bi}$, or formally:

$$S_{c} = 1-(1-S_{b1})*(1-S_{b2})*...*(1-S_{bi})*...*(1-S_{bn})$$

The microbiopsies in which a cluster is present were sequenced to varying depths and the
clusters are present at different cell fractions across different microbiopsies. To estimate the
sensitivity for mutations in a given cluster in a microbiopsy, we fitted a truncated binomial
distribution to the VAF distribution of the mutations assigned to the cluster in each
microbiopsy (this is different from the estimated median VAF, as low VAF mutations may be
missed at low coverage). We next simulated 100,000 mutation call attempts by drawing the
coverage of each call from a Poisson distribution with the lambda parameter set as the
median on-target coverage of that microbiopsy and multiplying that with the VAF estimate
from the truncated binomial. The product represents the number of reads that report the
alternative allele in the simulation, and the estimated sensitivity is the fraction of simulations
in which this value was greater than four.

7.2 Linear mixed effect models to test for an effect of disease
duration

We wish to know if psoriasis affects the mutation burden of keratinocytes. The ideal predictor
variable would be some measure of ‘inflammation exposure’ which might comprise both the
number of flare ups and the severity of each flare up. We lack this information for our
patients, many of which have had psoriasis for decades. We therefore use disease duration
as a proxy for disease exposure (after setting the disease duration of non-lesional biopsies
to zero).

To test if disease duration is associated with mutation burden we used linear mixed
effect models (see the supplementary code accompanying this manuscript). The model
describing the mutation rate must take into account that observations are not independent.
We included in the model random effects for patients and for biopsy, which were nested
within patients. We further include fixed effects for age and the anatomical site of the biopsy.
This model was compared against a second model which additionally included a fixed effect
for disease duration. We tested if the inclusion of disease duration improved the fit of the
model using likelihood ratio test (see Supplementary analysis 1, an R-markdown file that
accompanies this manuscript).

8. Selection and driver analyses

We used the dNdScv software\(^23\) (https://github.com/im3sanger/dndscv) to identify genes
enriched in non-synonymous mutations, indicative of positive selection. The ratio of non-
synonymous (dN) over synonymous (dS) mutation rate (dN/dS) is a widely used measure of
evolutionary pressures. dNdScv is a maximum-likelihood implementation of this method,
specifically developed for somatic mutation data. dN/dS ratios for each gene (or groups of genes together) are calculated using a substitution model that takes into account the sequence composition of the genes, the trinucleotide context of the observed mutations, and transcriptional strand biases. Mutation densities vary substantially depending on chromatin state, replication timing and expression levels of genes. This is modelled by dNdScv using epigenetic covariates calculated from the ENCODE\textsuperscript{24} project and the Cancer Genome Atlas (TCGA). We refer to the original dNdScv publication for an extensive description of the method.

Double-base substitutions (DBS) represent a small problem to dndscv, because the same mutation event can affect more than one codon and give rise to two annotation types of mutations (for example a missense and a synonymous change). DBSs are by default combined with indels in a negative-binomial model which estimates the mutation frequency across the genome without taking the sequence context into account. This typically has only a small effect on the model, as the number of DBSs is small. In the current dataset however, UV-light exposure has resulted in a large number of DBS mutations. We therefore fitted separate negative binomials for indels and for DBSs and combined the P-values with those obtained from analysing single-base substitutions using Fisher’s method (see Supplementary analysis 3, an R-markdown file that accompanies this manuscript).

### 8.1 Exome-wide driver discovery

We first calculated dN/dS ratios across all coding genes using mutations identified in either lesional or non-lesional skin. Mutations shared between microbiopsies dissected from the same patient were collapsed into single mutation events to avoid counting them multiple times. This assumes that mutations are much more likely to be observed across multiple samples from an individual because the samples share a cellular origin than the same (hotspot) site being independently mutated multiple times. For the purposes of identifying selection, the former explanation is the more conservative.

As described in the main text, this revealed positive selection of mutations in nine genes. While not all genes have been identified in studies of normal skin, they all are likely all under selection in normal squamous epithelial tissues and not specific to psoriasis. We confirmed that the results are robust against the use of covariates, inclusion of hypermutator samples (which add many passengers and comparatively few drivers) and the inclusion of microbiopsies from non-lesional skin (Supplementary analysis 3).

We performed restricted hypothesis testing of 27 genes in which mutations have been reported to be under selection in either the normal skin or normal oesophagus\textsuperscript{25–28} but identified no additional genes under selection. We also used the sitedsnds() function of the dndscv package to look for recurrently mutated sites and mutation hotspots but this did not reveal selection in any additional genes.

### 8.2 Pathway-level selection

To search for an enrichment of mutations affecting particular gene sets or pathways relevant to psoriasis we defined eleven gene sets \textit{a priori} as follows.
A list of genes previously found to be recurrently mutated either in normal skin or in squamous cell carcinomas was compiled by taking the union of all genes found to be under positive selection in Martincorena et al\textsuperscript{25} or Fowler et al\textsuperscript{28}. To this list, we added \textit{CDKN2A}, \textit{MLL2}, and \textit{CASP8}, which are recurrently mutated in cutaneous squamous cell carcinomas but have not been reported to be under positive selection in the normal skin\textsuperscript{25}. A second list of genes reported to be recurrently mutated in basal cell carcinomas was obtained from Bonilla et al\textsuperscript{29}. From this list, we excluded \textit{TP53}, \textit{NOTCH1} and \textit{NOTCH2} so mutations in these genes would not drive enrichment of the pathway, which otherwise contains many genes involved in Sonic Hedgehog signalling. The third list consisted of genes implicated in psoriasis pathogenesis through GWAS\textsuperscript{30}. Assigning genes to GWAS signals is an on-going problem within the field of human genetics. In this work, we used the Open Targets Genetics portal\textsuperscript{31} to assign a gene to each GWAS locus. For loci where any gene achieved an L2G score of >0.5, we used that gene but otherwise used the gene closest to the variant with the lowest P-value in the GWAS analysis.

The fourth list contained genes identified as under positive selection in any of three recent papers describing somatic evolution in IBD affected colonic mucosa\textsuperscript{7,32,33}. The three papers identify distinct patterns of selection in the IBD-affected colon compared with the normal colon, characterised by selection of immune-related genes, including \textit{PIGR}, \textit{ZC3H12A}, \textit{NFKBIZ} and other genes in the IL-17 and toll-like receptor pathways. We included this gene set to test the hypothesis that similar selection forces might be operating across different types of epithelia under conditions of chronic inflammation. \textit{PIGR} was excluded from the list, as this gene is not expressed in the skin.

Lists five through eleven are TNF signalling, IFNg signalling, IL12/23 signalling, IL36/MyD88 signalling, TLR signalling, IL-17 signalling and class I MHC mediated antigen processing and presentation. These pathways have all been strongly implicated in psoriasis pathogenesis\textsuperscript{34}. We used the Reactome database\textsuperscript{35} to decide which genes to include in the definition of each pathway.

As described in the main text, the mutational signatures of UV-light and psoralen exposure dominate the dataset. Both have an effect of sequence context which extends beyond the trinucleotide model used by dNdScv. Using the trinucleotide model has previously been shown to result in arbitrarily low estimates of the number of driver mutations in melanomas\textsuperscript{23}. To account for this, we implemented a pentanucleotide model which takes into account two base pairs on either side of the mutated base. For each of the gene sets described above, and using this model, we extracted the global dN/dS ratios for missense and nonsense variants separately. We corrected for 22 statistical tests (11 gene sets, missense and nonsense) using Benjamini-Hochberg correction. We believe the effects of the pentanucleotide model are especially important when aggregating mutation counts across multiple genes and that implementing this on a per-gene level would be unlikely to affect our results.

### 8.3 Fraction of mutated cells

We compared the fraction of cells that carry mutations in any of the nine genes that showed a significant enrichment of mutations between lesional and non-lesional skin. The fraction of cells carrying a mutation can be estimated from the fraction of reads that report the mutation
in each sample (the variant allele fraction, VAF). We calculated the fraction of mutated cells separately in lesional and non-lesional biopsies from each individual by multiplying twice the VAF of each mutation in each microbiopsy by the volume of the microbiopsy and dividing that by the total volume of microbiopsies dissected from that skin biopsy. For clones that carry more than one mutation in the same gene, we only counted the mutation with higher VAF. When the second mutation was present in a sub-clone, we used only the mutation present in the parent clone.

Methods references


1365

1366