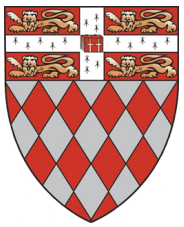
Factors Influencing the Somatic Mutational Landscape of Ageing Squamous Epithelium

Charlotte Rosemary King



Fitzwilliam College
University of Cambridge

March 2022

This thesis is submitted for the degree of Doctor of Philosophy
Supervised by Professor Philip Jones

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the Faculty of Biology Degree Committee.

Factors Influencing the Somatic Mutational Landscape of Ageing Squamous Epithelium

Charlotte Rosemary King

The incidences of many cancers vary substantially across the world, reflecting genetic differences between populations and exposure to environmental carcinogens. This is illustrated by keratinocyte skin cancers and oesophageal squamous cell carcinoma, which both develop from squamous epithelium, yet are remodelled during ageing by very different mutagenic processes and environmental exposures. In this thesis, I investigate the influence of cancer risk factors on the somatic mutations present in normal aged skin and oesophageal epithelium using a range of sequencing methods.

I find sun-exposed facial skin from donors of the UK to have a 4-fold increased mutation burden and 10-fold increase in copy number aberrant clones compared to donors of Singapore, a country with a 17-fold lower incidence of keratinocyte skin cancer. The majority of these mutations in the UK are due to ultraviolet radiation (UV) but, in Singapore, age-related signatures predominate. Mutations in *TP53* are more strongly selected in epidermis of the UK, whilst those in *NOTCH1* and *NOTCH2* are preferentially selected in Singapore, reflecting differences in the level of competition within the tissue. A survey of mutations in UK skin across body sites reveals differences in UV signature and selection between sites. In aged oesophageal epithelium from UK donors, I observe an increase in mutations with an alcohol-associated signature with reported alcohol consumption. Furthermore, mutation burden increases with smoking, without a detectable change in the mutational signature, consistent with tobacco smoke increasing oesophageal cancer risk independent of its mutagenic effects. Finally, in donors over the age of 60, mutations in *TP53* and *FAT1* are more strongly selected, whilst those in *NOTCH3* more weakly selected, suggesting changes to levels of competition within the tissue with age.

I conclude that the mutational landscapes of normal oesophagus and skin are shaped by age and environmental exposures and that this, in turn, may alter the risk of keratinocyte cancers.

Acknowledgements

The work within this thesis was reliant on the help and contribution of a large number of people. Primarily, I would like to thank the donors and their families who have consented for the use of their tissue and, without which, my thesis would not exist.

I am grateful to the help of collaborators Muly Tham, Jingxiang Juang, Birgit Lane, Amer Durrani, Kate Fife, Edward Rytina, Doreen Milne, Amit Roshan, Ben Hall and Kourosh Saeb-Parsy. In addition, I would like to thank the Wellcome Trust for funding my PhD and the work of Wellcome Sanger Sequencing Pipelines. Within CASM, special thanks go to CASM IT, Sarah Moody, Tim Coorens, Tim Butler and Inigo Martincorena for providing code and advice on computational analysis.

I want to express my appreciation for the time and valuable direction provided by Doug Winton and Mortiz Gerstung as advisors on my thesis committee.

Finally, I would like to thank the entire Jones group for providing warmth, collaboration, inspiration and humour throughout my PhD, in particular, to Tomeu Colom, David Fernandez-Antoran, Gabriel Piedrafita-Fernandez, Esther Choolun, Swee Hoe Ong, Albert Herms, Emilie Abby, Michael Hall, Roshan Sood, Chris Bryant, Kasumi Murai, Tom Metcalf and Irina Abnizova. Extended thanks go to Jo Fowler and my supervisor, Phil Jones, for their tireless support, continued guidance and unparalleled kindness - I am extremely grateful for everything they have helped me to achieve.

Contents

Chapter 1: Introduction

Somatic mutations have been shown to accumulate in tissues with age (Gerstung et al. 2020) at a rate of one to ten mutations per cell division (Martincorena and Campbell 2015). Although the majority of these mutations will not have a noticeable effect on a cell's physiology, some will alter the fitness of a cell. As in species evolution, this variation between cells of a tissue leads to competition and cells with a fitness advantage will increase in prevalence over time (Cairns 1975), forming a clone. It is this accumulation of somatic mutations within a clone that persists within normal tissue that leads to cancer, with the average tumour estimated to harbour approximately four coding mutations under positive selection (Martincorena, Raine, et al. 2018; Consortium and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020) and there is evidence to suggest that the mutant clones that give rise to cancer are resident in normal tissues for years to decades (Gerstung et al. 2020).

In this thesis, I analyse sequencing data from two squamous epithelia (normal skin and middle third of the oesophagus) to study the somatic mutations and evolution of clones. The results are considered in the context of the keratinocyte cancers that develop in these tissues: basal cell carcinoma and squamous cell carcinoma of the skin and oesophageal squamous carcinoma. The incidence of these cancers varies greatly across human populations. In this first chapter, I begin by reviewing the current epidemiological data to give insight into environmental factors that may increase an individual's risk. Epidemiological studies, however, give limited information on the cellular processes leading to carcinogenesis and I next discuss the current sequencing data available for keratinocyte cancers to characterise the genomic changes present at this tumour end point of somatic evolution. Recurrently mutated genes and changes to chromosomal copy number across DNA sequencing of tumour samples can give insight into the genome of the cell that founded the cancer and consequently the events required for carcinogenesis in the environment of that tissue. It has been shown that some of the endogenous and environmental processes that lead to somatic mutation can be identified by the specific trinucleotide context of the mutated base, known as a mutational signature (L. B. Alexandrov and Stratton 2014). I summarise our current knowledge of the mutagenic processes and signatures present in keratinocyte cancers and how these link to epidemiological risk factors. Finally, I review the studies conducted to date describing the somatic mutations detectable in normal skin and oesophageal epithelium. A comparison of the genomic changes recurrent in normal ageing tissue with respect to tumour data can give insight into those genes that provide a

proliferative advantage to clones in normal tissue, those which are essential for carcinogenesis and others which may in fact confer protection against the development of cancer.

Epidemiology of Cutaneous Keratinocyte Cancers

Keratinocyte skin cancers are the most common human cancer worldwide and develop from keratinocytes in the epidermis (**Fig. 1.1**). The deepest layer of the epidermis, the stratum basale, is composed of stem cells and is separated from the dermis by the basement membrane. These stem cells produce keratinocyte daughters, 50% of these are stem cells and 50% differentiating cells which commit to terminal differentiation, exit the basal layer and migrate through approximately 11-18 nucleated cell layers, before reaching the stratum corneum (Yousef, Alhajj, and Sharma 2021). The stratum corneum varies in thickness across body sites and acts as a protective barrier made up of keratin and dead, anucleate squamous cells, which are eventually shed from the tissue. Cells are constantly shed from the epidermal surface and stem proliferation must continue throughout life to maintain the tissue.

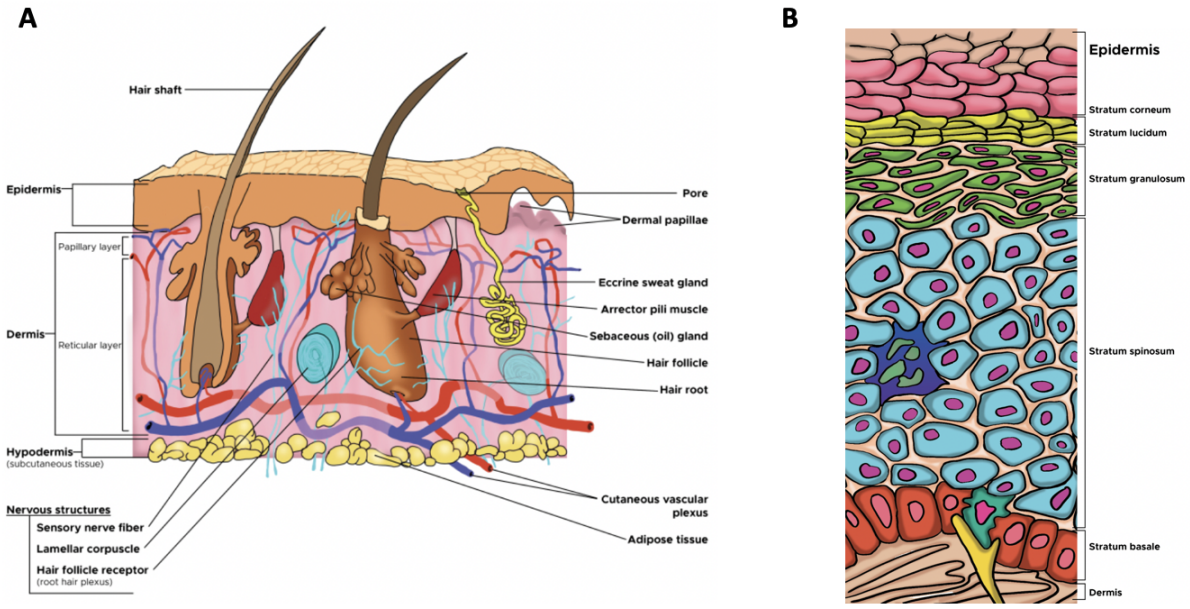


Figure 1.1: Structure of human skin, taken from (Yousef, Alhajj, and Sharma 2021). **A** Cross-section of human skin. The epidermis forms the uppermost layer of the skin and is composed of keratinocytes. **B** Layers comprising the epidermis. From top to base: stratum corneum, stratum lucidum, stratum granulosum, stratum spinosum and stratum basale.

The two major histological sub-groups of keratinocyte skin cancers are cutaneous squamous cell carcinoma (cSCC) and basal cell carcinoma (BCC). Whilst BCC is the more common, cSCC is more likely to metastasise and is consequently more fatal (Perry, Barton, and Alberg 2017). There are multiple lines of epidemiological evidence that suggest both BCC and cSCC are caused by exposure to solar UV radiation, with cancer risk increasing with both duration and intensity of this exposure (Gallagher et al. 2010). The majority of epidemiological risk factors identified for keratinocyte cancers are consistent with an increase in an individual's cumulative UV exposure and include: age (Muzic et al. 2017), outdoor work (Schmitt et al. 2011), use of tanning beds (Veierød et al. 2014; Ferrucci et al. 2012) and phenotypes such as red hair, blue eyes, freckles and lightly pigmented skin (Armstrong and Kricker 2001). In fact, in human populations with lower levels of pigmentation, the incidence of keratinocyte cancer exceeds that of all other cancers combined (Sung et al. 2021). Numerous genetic loci have been identified as being associated with inter-individual variation in levels of skin pigmentation, with genes having diverse functions from the biology of melanocytes (which produce the photoprotective pigment melanin) to the DNA repair of UV damage (Batai et al. 2021; Crawford et al. 2017). Within an individual, levels of melanin have been shown to increase in response to UV radiation (Coelho et al. 2009). In populations with high levels of skin pigmentation, cSCCs are more common than BCCs and are more frequently observed in body sites of chronic ulceration, after tissue injury or burn and in individuals that are immunocompromised (Wright et al. 2020).

A few studies have identified risk factors for keratinocyte cancers that are not associated with UV light, for example, exposure to ionising radiation (Leisenring et al. 2006; Watt et al. 2012) and the carcinogen arsenic (Maloney 1996). Chronic immune suppression, for example in organ transplant donors and the treatment of inflammatory diseases, has been shown to increase risk of developing both cSCC and BCC (Hartevelt et al. 1990; Long et al. 2012). There is little or inconclusive evidence to suggest that smoking, human papilloma viruses, alcohol consumption or diet affect risk of developing keratinocyte cancer (Verkouteren et al. 2017).

Keratinocyte cancer incidence rates are increasing globally, particularly in North America, Europe and Australia and it is proposed this increased incidence is due partly to increased awareness and detection of skin cancers by the general public and clinicians, but also due to an ageing population and increased UV exposure in younger generations, for example, through increasing accessibility of global travel and use of tanning beds (Verkouteren et al. 2017).

Epidemiology of Oesophageal Squamous Cell Carcinoma

Oesophageal squamous cell carcinoma (ESCC) develops from the squamous cells which form the epithelial layer lining the lumen (**Fig. 1.2**). Patterns in global incidence and risk factors associated with ESCC are very different to those for keratinocyte cancers of the skin. Like the epidermis, the oesophageal epithelium comprises a basal layer of stem cells beneath squamous cells which progressively stratify as they progress towards the lumen. However, unlike skin, these upper most cells are not keratinised, making them vulnerable to abrasive, thermal and pH injury (Barbera et al. 2015).

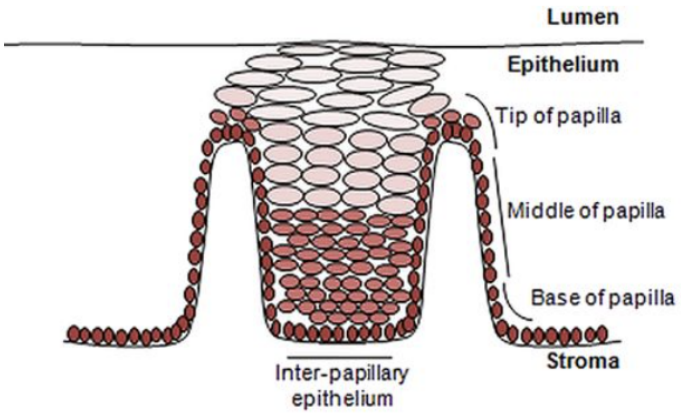


Figure 1.2: Structure of oesophageal tissue in humans, from (Barbera et al. 2015). Stem cells reside in layers at the base, nearest the stroma. As cells differentiate and stratify, they progressively move towards the lumen and are shed from the tissue.

Oesophageal cancer is the sixth leading cause of cancer death in humans (Bray et al. 2018) and is classified into two histological types: squamous cell carcinoma (ESCC), which occurs in the upper and mid parts of the oesophagus, and adenocarcinoma (EAD), which develops at the junction of the oesophagus and stomach. Around 90% of oesophageal cancer cases globally are ESCC and around 10% are EAD (Abnet, Arnold, and Wei 2018).

Countries among those with the highest ESCC incidence include those in the south and east of Africa, Brazil and China (**Fig. 1.3**). In countries where ESCC risk is relatively low, such as in the UK, USA and Australia, incidence rates can be four times higher for men than for women (Abnet, Arnold, and Wei 2018). It is possible this sex-bias is explained by increased historical exposure of males to tobacco smoke and alcohol, two of the main epidemiological risk factors, in addition to age, for ESCC in the west (Abnet, Arnold, and Wei 2018). Furthermore, tobacco smoke and alcohol appear to act synergistically to increase ESCC risk (Murphy et al. 2017; Yang et al. 2017). In high incidence countries, the ratio between sexes

nears 1, suggesting the presence of a stronger, sex-independent risk factor (Abnet, Arnold, and Wei 2018). Finally, ESCC incidence is found to vary greatly even within populations of the same country. For example, black Americans have a five-fold higher risk of ESCC than white Americans (Kgomo et al. 2017). In some regions of the North Central Taihang mountain range in China, ESCC is the leading cause of death (Abnet, Arnold, and Wei 2018).

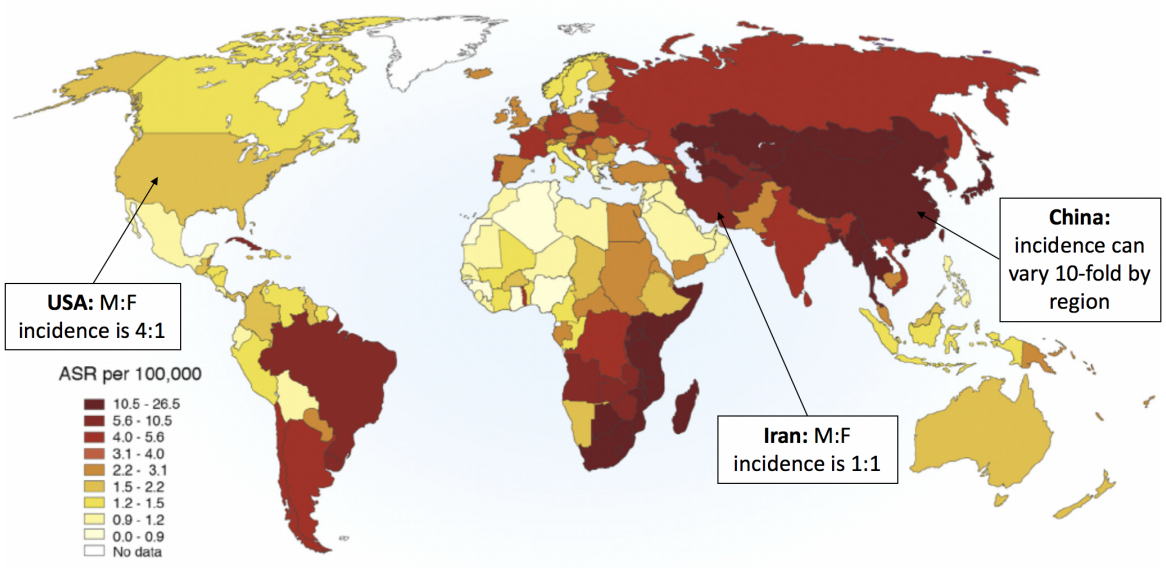


Figure 1.3: Incidence of ESCC in men worldwide (2012). Adapted from (Abnet, Arnold, and Wei 2018; Arnold et al. 2015). ESCC incidence varies more than ten-fold worldwide. In mid to low incidence countries, incidence is greater in males than females (Abnet, Arnold, and Wei 2018). In high incidence countries, there is no differential incidence with sex, suggesting the action of a stronger sex-independent risk factor (Abnet, Arnold, and Wei 2018). ASR, age standardised rate; F, female; M, male.

Globally, the cause of ESCC is unknown, with epidemiological studies reporting a large number of potential risk factors, in addition to tobacco smoking and alcohol consumption, that include tobacco chewing (Dar et al. 2012), exposure to polycyclic aromatic hydrocarbons (carcinogens produced during the incomplete combustion of organic material), drinking hot tea, red meat consumption, poor oral hygiene, low intake of fruit and vegetables and low socioeconomic status (Asombang et al. 2019; Abnet, Arnold, and Wei 2018; Murphy et al. 2017).

The dramatic geographical variations in ESCC incidence suggest a strong role for environmental risk factors (Lao-Sirieix, Caldas, and Fitzgerald 2010). Some studies have also identified genetic variants which may contribute to population differences in risk,

although most have not produced reproducible results (Abnet, Arnold, and Wei 2018). Polymorphisms in *ADH1B* and *ALDH2*, both involved in alcohol metabolism, are associated with increased ESCC risk with alcohol consumption (Druesne-Pecollo et al. 2009). In addition, a study in a Japanese population found a synergistic effect between these variants with both increased alcohol consumption and smoking (Lao-Sirieix, Caldas, and Fitzgerald 2010; Cui et al. 2009). An *ADH1B* polymorphism, associated with increased production of acetaldehyde, is present in up to 90% of some South Asian populations, with prevalence much lower in African (15%) and European (10%) populations (Druesne-Pecollo et al. 2009). It has been suggested that individuals with this polymorphism are more likely to avoid alcohol due to an unpleasant associated flushing reaction (Brooks et al. 2009). However, those that do drink alcohol can increase their risk of ESCC by as much as 43-fold and 73-fold with moderate or heavy drinking respectively (Abnet, Arnold, and Wei 2018).

Genomic Characterisation of Keratinocyte Cancers

Many of the risk factors consistently identified across epidemiological studies, such as low socio-economic status, are complex and multi-dimensional, giving limited insight into the biological processes leading to carcinogenesis. Understanding the biological mechanisms that lead to cancer progression is important as it allows the identification of interventions that may prevent cancer. With the rise of DNA sequencing technology in the 2000s, so came a push to sequence the coding regions of tumours. The International Cancer Genome Consortium (ICGC) was established to detail somatic genomic abnormalities across the main different tumour types and identify ‘driver’ mutations which contribute to cancer development (International Cancer Genome Consortium et al. 2010). In addition, The Cancer Genome Atlas (TCGA) pan-cancer project was set up to profile and collate DNA, RNA, protein and epigenetic data across numerous cancer types (Weinstein et al. 2013). However, much of this data was collected at a time when sequencing technology was in its infancy, the methods used are often inconsistent and quality metrics poor. More recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium was established to aggregate raw sequencing data across both ICGC and TCGA in order to deliver a high quality set of somatic mutations from whole-genome sequencing only for downstream analyses (Consortium and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Unfortunately, there is a sparsity of whole-genome sequencing data of keratinocyte cancers and consequently no data on these cancer types were included in the PCAWG pan-cancer analysis.

Our knowledge of cSCC genomic characterisation is mainly limited to a few, small studies of whole-exome sequencing. These reveal an average tumour mutation burden of 50 mutations/Mb, with recurrent inactivating mutations in *NOTCH1*, *NOTCH2*, *TP53* and *CDKN2A* and, less frequently, activating point mutations in the oncogenes *HRAS*, *NRAS* and *KRAS* (Inman et al. 2018; van der Schroeff et al. 1990; South et al. 2014). Copy number aberration (CNA) in cSCC is common (**Fig. 1.4**), with recurrent gains at 3q, 7q and 9q and losses at 3p and 9p (Inman et al. 2018).

To date, the largest study of BCC analysed 126 whole-exomes and estimated a mutation burden of 65 mutations/Mb, higher than in any other cancer studied to date (Bonilla et al. 2016). The most frequent alterations observed are at the tumour suppressor genes *PTCH1* and *TP53*. *PTCH1* represses the Hedgehog pathway and its loss leads to constitutive activation Hedgehog signalling that is the critical driver of BCC (Von Hoff et al. 2009). *PTCH1* is located on chromosome 9q in proximity to *NOTCH1*. Loss of heterozygosity (LOH) at 9q is the most frequent (~64%) CNA observed in BCC (van der Riet et al. 1994). Inactivating mutations in *TP53* are observed in ~55% of BCCs but LOH of its locus on 17p is rare (van der Riet et al. 1994). Other recurrently mutated genes include the Sonic Hedgehog pathway genes *SMO* (20%) and *SUFU* (8%), in addition to *MYCN* (30%), *PTPN14* (23%), *PPP6C* (15%), *STK19* (10%), *LATS1* (8%), *RB1* (8%), *FBXW7* (5%), *ERBB2* (4%), *PIK3CA* (2%) and *NRAS*, *KRAS* or *HRAS* (2%), suggesting BCC has a distinct biology of carcinogenesis to cSCC (Bonilla et al. 2016).

The majority of studies genomically characterising ESCC have involved the analysis of whole-exome sequencing, particularly across Chinese and Japanese populations (Lin et al. 2014; Y. Song et al. 2014; Urabe et al. 2019). The most frequent alteration observed in ESCC is inactivation of *TP53* through either mutation (91% of tumours) and/or copy number loss (Chen et al. 2017). Other commonly inactivated genes include *NOTCH1* and *NOTCH2*, *ADAM29*, *FAM135B*, the cell cycle regulators *RB1*, *CCND1*, *CDKN2A*, and *NFE2L2* and histone modifiers *KMT2D* and *EP300*. Activating point mutations in the oncogene *PIK3CA* are also common. Recurrent chromosome gains include 3q, 5p and 8q as well as loss at 3p (**Fig. 1.4**). Regions of oesophageal squamous dysplasia have a high mutation and copy number burden across genes commonly affected in ESCCs (Chen et al. 2017). It has been suggested that ESCC forms from such regions of dysplasia and comparative sequencing between matched oesophageal dysplasia and ESCC samples suggests a ‘two-hit’ event to inactivate *TP53* is required for full progression to cancer (Chen et al. 2017). Recently, a large study of over 500 whole-genomes of ESCC calculated a median tumour burden of 10,000 mutations/genome (Moody et al. 2021).

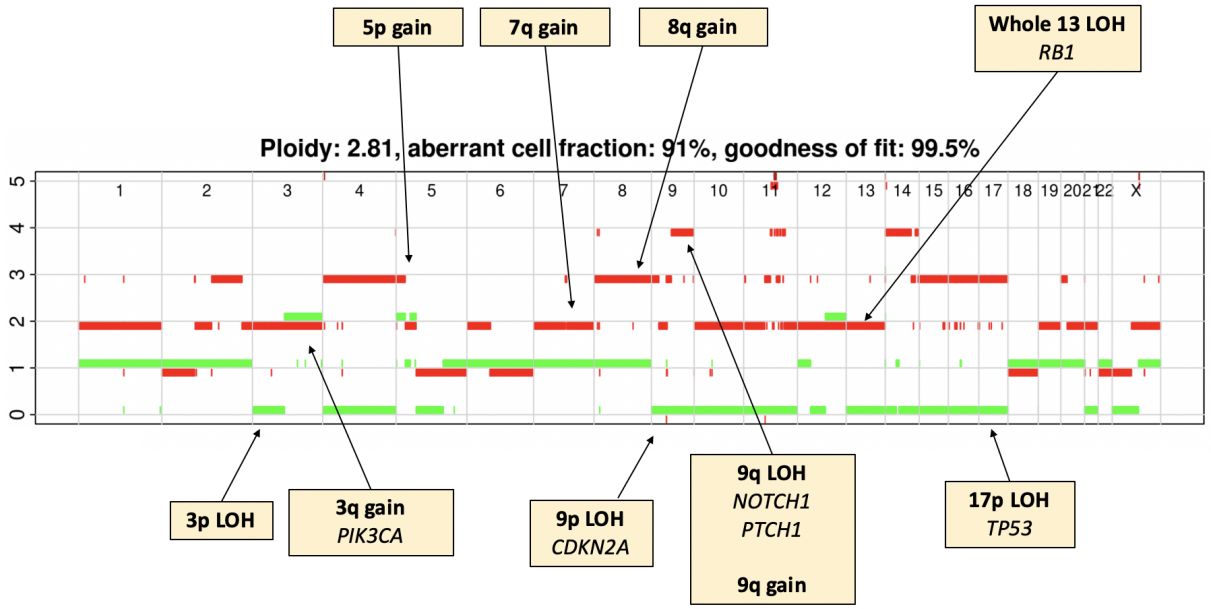


Figure 1.4: An example ASCAT (Raine et al. 2016) plot for an ESCC genome showing estimated frequency of the minor (green) and major (red) allele across each chromosome. Copy number aberrations common to squamous carcinoma are labelled with putative driver genes.

Mutational Signatures of Keratinocyte Cancers

Exposure to environmental mutagens can lead to DNA damage and, if not repaired, this results in somatic mutations that become fixed in the genomes of cells. In some cases, it has been shown that such exogenous processes can be identified by the specific trinucleotide context of the mutations caused, known as a mutational signature (L. B. Alexandrov and Stratton 2014). Consequently, mutational signatures can provide a more direct insight into the biological mechanisms that lead to carcinogenesis than epidemiological study. Across both BCC and cSCC, mutations are predominantly UV-induced C>T single- or double-base substitutions at pyrimidines. The reference mutational signatures SBS7a-d were defined in melanoma samples and characterise the trinucleotide context of single-base substitutions (SBS) indicative of UV damage (L. Alexandrov et al. 2018). The high mutation burden observed in keratinocyte cancers can be explained by these SBS7 signatures, however, SBS32 has also been identified in cSCCs from immunocompromised patients taking azathioprine (Inman et al. 2018).

A large study of over 500 ESCC genomes across eight countries of varying ESCC incidence found little variation in mutational signature between patients (Moody et al. 2021). The reference mutational signatures SBS2 and SBS13, both associated with apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) activity, were present in 88% of

ESCCs, accounting for 25% of the mutation burden, suggesting that activation of APOBEC is a common step present in, but not necessary for, ESCC carcinogenesis (Moody et al. 2021). APOBEC associated signatures are common in samples of oesophageal squamous dysplasia (Liu et al. 2017; Yokoyama et al. 2019) and clonal substitutions in an APOBEC context have been observed in ESCCs, suggesting that APOBEC mutagenesis occurs before the development of ESCC (Moody et al. 2021).

The majority of non-APOBEC associated substitutions in ESCCs have been attributed to the reference mutational signatures associated with endogenous ageing processes (SBS1, SBS5 and SBS40) and SBS18, a C>A signature thought to be caused by reactive oxygen species (Moody et al. 2021). SBS16, associated with alcohol consumption, was identified in samples mainly from Japan and Brazil, two lower incidence countries (Moody et al. 2021). This study found no mutational signature associated with smoking and no evidence of an unknown endogenous process to explain the large variation in ESCC incidence by country. Another study of over 500 Asian ESCCs identified SBS16 in samples from heavy drinkers and/or smokers, but did not compare mutational signature by risk factor separately (Yokoyama et al. 2019). ESCCs with no evidence of APOBEC activity have often achieved a high mutation burden through defects in DNA mismatch repair (Moody et al. 2021).

Whilst one environmental ESCC risk factor has been associated with a specific mutational signature (SBS16 and alcohol consumption), for the remainder, it is currently unknown whether they act directly by generating their own mutational signature, indirectly by influencing mutation rates of endogenous processes or through non-mutational mechanisms. A study of tumour genomes from mice chronically exposed to known or suspected human carcinogens revealed that most agents do not generate a distinct mutational signature or increase mutation burden (Riva et al. 2020). This argues that many factors that increase the risk of cancer may not be associated with a distinct mutational signature.

Calling Somatic Mutations in Normal Tissue

A combination of endogenous and environmental mutational processes can lead to the accumulation of somatic mutations in tissues over the course of a lifetime (Martincorena and Campbell 2015). If a mutation gives a cell a competitive advantage, a mutant clone is likely to persist, expand and the proportion of cells with this mutation will increase within the overall population (Frede et al. 2016). It is thought that cancers arise from these mutant clones, due to the high mutational burden observed in many tumours and the increase in cancer incidence with age (Bray et al. 2018). However, the genomes of cells at this end-point

of disease are so abnormal, it is difficult to determine from the sequencing of cancer genomes alone, which genes are true drivers of tumour progression and which are merely passengers. Furthermore, drivers of tumour initiation and carcinogenesis are likely to be different from events which drive tumour proliferation, survival and metastasis. Consequently, the study of clonal expansions in normal tissues can give unparalleled insight into both cancer development and the ageing process.

Since cancers develop from a small number of cells, tumour samples are largely clonal and therefore these clonal mutations, in addition to large sub-clones, can be detected through sequencing at a modest coverage (~20X). However, several techniques have been developed to overcome the challenge of detecting mutations in non-cancerous, poly-clonal tissues. One approach is to take a single cell from the tissue, for example blood or epithelium, and grow that cell into a clonal culture (Lee-Six et al. 2018; Yoshida et al. 2020; Yokoyama et al. 2019). This serves as *in vitro* amplification of the DNA of this single cell and, after sequencing, mutations present in the original cell will be clonal, whilst sub-clonal mutations are likely to have been acquired during culturing or through sequencing error and can be filtered out. The disadvantage of this method is that growing clones in culture serves as a form of selection and so it is difficult to obtain an unbiased overview of all clones present in the tissue since particular cells may not survive the cell culture process.

In tissues that form histologically distinct subunits, such as crypts in the colon, laser capture microdissection (LCM) can be used to carefully dissect clonal (or oligoclonal) areas of tissue before DNA sequencing (Lee-Six et al. 2019). By sampling an area of tissue that is histologically defined, this acts as *in vivo* amplification of the DNA of the stem cells that serve this niche (Ellis et al. 2021). In addition, detailed spatial information can be obtained for each clone that is sequenced (Grossmann et al. 2021). However, one disadvantage of this approach is that it is time and labour intensive, limiting the number of donors and area of tissue that can be sampled at one time.

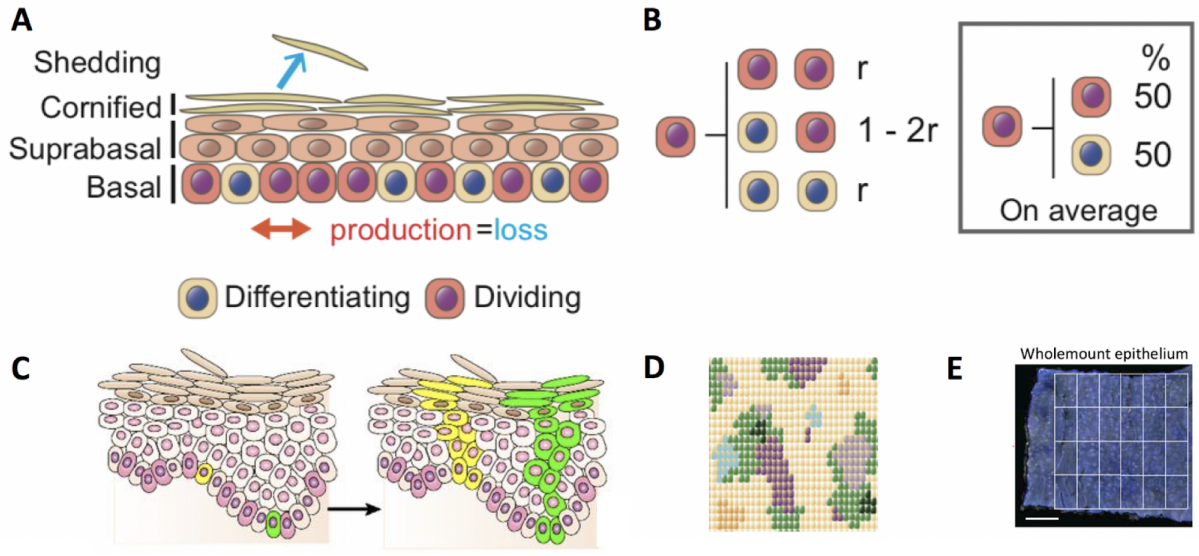


Figure 1.5: Clonal dynamics of murine epidermis. Adapted from (Murai et al. 2018; Simons 2016). **A** Structure of murine epidermis. Progenitor cells (pink) are confined to the basal layer. Differentiating cells exit the cell cycle and migrate through the suprabasal layer until they are shed, where the rate of shedding is equal to the rate of proliferation (Frede et al. 2016). **B** In mouse models, each progenitor produces one proliferating (pink) and one differentiating (white) daughter with each division. Occasionally, a progenitor will produce two proliferating or two differentiating daughters, but the likelihood (r) of each fate is the same so that, on average, the number of proliferating and differentiating cells produced across the progenitor population is equal (Murai et al. 2018; Frede et al. 2016). **C** Adapted from (Simons 2016). A somatic mutation which causes a progenitor (green) to shift its cell fate towards producing more proliferating daughters will cause the number of cells with this mutation in the basal layer to increase over time. **D** Normal squamous epithelium is a patchwork of mutant clones, some of which are large enough to detect with deep sequencing. **E** Squamous epithelium of the skin and oesophagus can be peeled from the underlying muscle and cut into adjacent 2-mm^2 grids.

Finally, a method of targeted sequencing at high depth (>600X) can be used to survey a large area of tissue comprising many samples in genes of specific interest. The physiology and structure of squamous epithelium makes it most appropriate for this approach (Martincorena et al. 2015; Martincorena, Fowler, et al. 2018). Transgenic mouse studies of the epidermis have found proliferating progenitor cells confined to the basal layer with each, on average, producing an equal number of either proliferating or differentiating daughters (**Fig. 1.5A-B**; (Murai et al. 2018)). This balance in cell fate means most neutral mutations are eventually lost through differentiation. During wounding, progenitors are able to shift their cell fate in favour of producing more proliferating daughters, until the wound is healed (Frede et al. 2016): a potential vulnerability which can be exploited by some somatic mutations (**Fig. 1.5C**; (Murai et al. 2018)). Over time, the number of cells in the basal layer with an

advantageous mutation will increase as the mutant clone expands laterally, displacing wild type basal cells (Murai et al. 2018). The lack of features or boundaries in squamous epithelium of the skin and oesophagus allows unrestricted growth of these mutant clones, some of which can grow up to centimetres in area and many of which are large enough to be detected with deep sequencing (**Fig. 1.5D**). Epithelium of the skin and oesophagus can be peeled from the underlying muscle as a single 2-dimensional layer of tissue. This has the advantage of ensuring very little contamination from stroma, as well as enabling the sampling of tissue as a contiguous grid, allowing the mapping of clones which span multiple samples (**Fig. 1.5E**).

The Mutational Landscape of Normal Skin Epidermis

The first study to describe the mutational landscape of normal epithelium took samples of eyelid epidermis of varying sizes (ranging from 0.8 to 4.7 mm^2) across four donors, aged from 55 to 73 years old (Martincorena et al. 2015). Each sample was sequenced at high depth for 74 genes commonly mutated in cancers and mutations were called using the *ShearwaterML* algorithm (Gerstung, Papaemmanuil, and Campbell 2014). *ShearwaterML* allows the detection of somatic mutations present in only a small fraction of a sample that has been sequenced at high depth. The method uses a large panel of lowly mutated samples, sequenced using the same method as the sample of interest, to generate an error model for each site of the genome covered by the targeted sequencing. Samples in this panel are assumed to have an absence of somatic mutations in the genes of interest and can therefore be composed of bulk samples from a different available tissue (for example muscle or fat) or samples that are known to have a low somatic burden (for example from young donors). All observed base changes at a site in this panel are therefore attributed to be technical sequencing error, allowing *ShearwaterML* to build a site-specific error model, which can be used to call somatic mutations at low allele frequencies in the sample of interest.

Once somatic mutations have been called across each gene, the ratio of substitution rates at non-synonymous (N) and synonymous (S) sites (dN/dS) can be used as a method for identifying genes under evolutionary selection in the tissue (Martincorena et al. 2017; Davoli et al. 2013). The basis of this method, originating from population genetics, is used to determine if mutations in a particular gene are likely to have become fixed through neutral drift or positive selection (Kryazhimskiy and Plotkin 2008). If the ratio of non-synonymous to synonymous mutations in a gene is equal (dN/dS = 1), there is no evidence to suggest this gene is under selection. If there is a greater number of non-synonymous to synonymous

mutations than expected by chance (dN/dS > 1), this suggests this gene is under positive selection. Recent versions of this model take account of the gene sequence and the mutational spectrum to estimate the expected frequency for each possible nucleotide substitution in the gene (Martincorena et al. 2017).

This study of deep sequencing of normal eyelid epidermis from the UK revealed a mutation burden similar to that seen in many cancers, with many key drivers of cSCCs under strong positive selection (Martincorena et al. 2015). *NOTCH1* was the most frequently mutated gene and was found to be under the greatest positive selection (Martincorena et al. 2015). However, one of the four donors was found to have a high proportion of non-synonymous mutations in *NOTCH2*, suggesting that in this donor *NOTCH2* is under greater positive selection than *NOTCH1 (Martincorena et al. 2015)*. This donor was of South Asian ancestry whilst the other three donors were of European descent. This led to a hypothesis that the genetic background of Asian individuals leads to stronger selection of *NOTCH2* mutations (Martincorena et al. 2015). An alternative hypothesis is that *NOTCH2* is a better competitor than *NOTCH1* in a lowly mutated environment. This South Asian donor had more pigmentation than the other donors. Individuals with darker skin are able to filter more UV radiation than those with light skin due to increased epidermal melanin, increased melanocyte activity and larger, more dispersed melanosomes (Que, Zwald, and Schmults 2018; Kim, Del Rosso, and Bellew 2009). However, because samples of differing sizes were used to call mutations in each donor, with a different *ShearwaterML* panel, it is not possible to accurately compare mutation burden across donors in this study and reliable conclusions cannot be drawn from a single individual.

The main finding of this study, however, is that 1 cm^2 of skin can harbour up to 45 detectable clones carrying mutations in genes associated with cancer (**Fig. 1.6**). Genes commonly mutated in cSCC, *NOTCH1*, *TP53*, *NOTCH2*, *FAT1* and *NOTCH3* were all found to be under positive selection within normal skin, questioning our understanding of cancer ‘drivers’. If a gene is already commonly mutated in normal tissue and is found to be mutated at a similar prevalence in cancer, it may not play a role in the additional steps required to transform a physiologically normal tissue into a tumour.

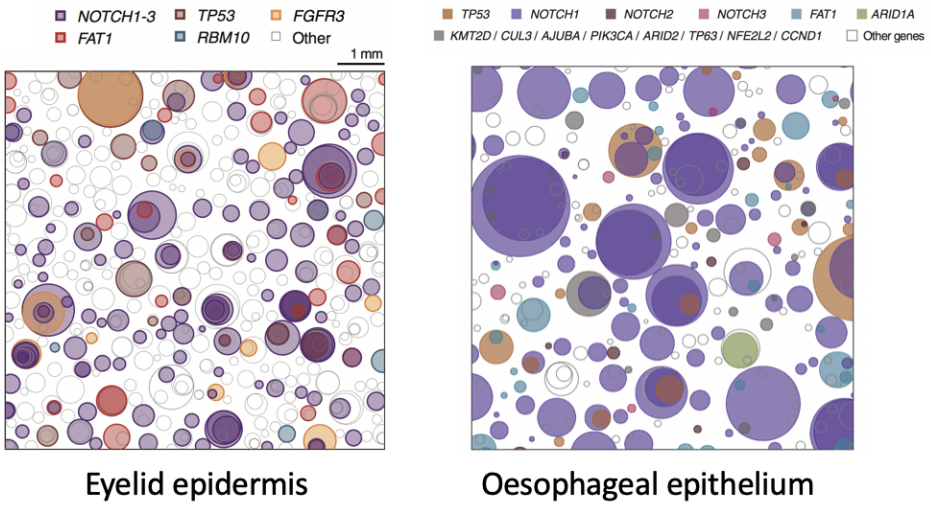


Figure 1.6: A representation of somatic clones present in 1 cm^2 of normal eyelid epidermis and oesophageal epithelium. Figures taken from (Martincorena et al. 2015) and (Martincorena, Fowler, et al. 2018), respectively. Non-synonymous mutations are displayed as circles, coloured by positively selected gene and randomly distributed over the space. Sequencing data, including copy number, was used to infer the size and number of clones and, where possible, the nesting of sub-clones. Otherwise, sub-clones are nested randomly.

The Mutational Landscape of Normal Oesophageal Epithelium

The first work describing the mutational landscape of the oesophageal epithelium (Martincorena, Fowler, et al. 2018) used a sampling method of taking multiple adjacent samples per donor (**Fig. 1.5E**), each of a consistent size (2 mm^2). This method is preferred as it enables the merging of mutations belonging to large clones which span multiple samples of a donor. This allows a more accurate estimate of both clone size and mutation burden and therefore a more reliable comparison of the mutational landscape across donors. Both mutation burden and the size of mutant clones were found to increase with the age of the donor. Furthermore, the oldest donor of this study, was found to have the greatest proportion of his tissue mutant for *TP53*, giving insight into a possible mechanism that explains the increase in ESCC risk with age.

In this study, the updated method of *dNdScv* was used to identify genes under selection (Martincorena, Raine, et al. 2018). *dNdScv* is a more appropriate measure of selection in somatic tissues because it adjusts the *d*N/*d*S ratio for trinucleotide mutational signatures, sequence composition and the variable mutation rates observed across genes (Martincorena, Fowler, et al. 2018; Martincorena et al. 2017). This study identified *NOTCH1,*

TP53, NOTCH2, FAT1, NOTCH3, ARID1A, KMT2D, CUL3, AJUBA, PIK3CA, ARID2, TP63, NFE2L2 and CCND1 as being under positive selection in normal oesophageal epithelium (**Fig. 1.6**). Many of these genes are recurrently mutated in ESCC, again questioning our understanding of cancer 'drivers' and the differences between normal epithelium and cancer. Furthermore, *NOTCH1* was found to be disproportionately mutated in normal tissue compared to the frequency observed in ESCC, suggesting that some mutant genes that colonise normal tissues may in fact protect against carcinogenesis. This is a valuable avenue to explore as understanding the mechanisms by which such mutations inhibit transformation could lead to the development of preventative interventions in high risk groups.

Thesis Summary

In Chapter 2, I expand on the study of eyelid epidermis by Martincorena *et al.* (2015) by sequencing 237 samples of epidermis from either the eyelid or brow across five donors from Singapore and six donors from the UK. The UK has a 17-fold higher age-adjusted incidence of keratinocyte cancer than Singapore which motivated me to compare the mutational landscape of sun-exposed skin from the two countries. In Chapter 3, I analyse a dataset of somatic mutations from normal epidermis from different body sites across over 30 donors of the UK. The incidences of keratinocyte cancers vary by body site in humans with low levels of skin pigmentation (Subramaniam et al. 2017). In Chapter 4, I expand on the work of Martincorena, Fowler *et al.* (2018) which identified inter-individual variation in clone size, mutation burden and genes under selection in normal oesophageal epithelium that could not be completely explained by age. I analyse somatic mutations from targeted sequencing of a further 30 cm^2 of aged oesophageal epithelium taken from 20 donors over the age of 60 years. In each chapter, epithelium is sampled as 2-mm^2 adjacent grids, allowing the merging of mutations spanning multiple samples of a donor (Martincorena, Fowler, et al. 2018).

I use mutational signature analysis to link the observed somatic mutations and their trinucleotide context with the mutational processes, either environmental or endogenous, that caused them. Mutational spectra and signatures are described using the notation employed by the PCAWG Mutational Signatures working group (L. B. Alexandrov et al. 2020a). I use non-negative matrix factorisation to estimate the contribution of 49 single-base-substitution and 11 doublet-base-substitution reference signatures (originally characterised in human cancers) to the mutational spectra (L. B. Alexandrov et al. 2020a). It is particularly interesting to study the clonal landscape of skin and oesophagus in the context of their environments. Cutaneous epithelium is exposed to UV light throughout a person's

lifetime, leading to a high mutation burden in this tissue, a highly competitive clonal landscape and very many small clones. In contrast, oesophageal epithelium has no exposure to UV light and a mutational burden about a tenth of that observed in sun-exposed skin (Martincorena, Fowler, et al. 2018; Martincorena et al. 2015). This lower mutational burden means competition between mutant clones is lower and they are able to grow very large over time.

Finally, in Chapters 3 and 4, I map normal skin epidermis and oesophageal epithelium at higher spatial resolution by calling somatic mutations from sequencing of punch biopsies that measure 0.05 mm^2. This smaller sample area provides a higher clonality and greater insight into events that may have happened more recently in the tissue. I use whole-genome sequencing of these punch samples to obtain more accurate estimates of mutation burden, mutational signatures and CNA and I use shared mutations between multiple samples of a donor to construct phylogenetic trees to give insight into the timing of these events in the tissue.

On a molecular level, the causes of cSCC, BCC and ESCC are poorly understood with drivers of carcinogenesis and disease progression needing to be clarified. The World Health Organisation aims to reduce premature mortality from cancer through the acceleration of research on the causes of cancer and mechanisms of carcinogenesis. The aim of this thesis is to further characterise the mutational landscape of normal squamous epithelium and explore the effect that epidemiological cancer risk factors have on that landscape.

Chapter 2: The Mutational Landscape of Aged Normal Skin from Two Countries of Contrasting Skin Cancer Risk

Introduction

Keratinocyte cancers are the most common malignancy to occur in fair-skinned populations worldwide (Ciążyńska et al. 2021) and chronic exposure to UV radiation in sunlight is the principal risk factor (Trakatelli et al. 2007). Keratinocyte cancer incidence is increasing and predicted to worsen over the next thirty years with Europe's ageing population (Trakatelli et al. 2007; Venables et al. 2019). UV index is a linearly-scaled measurement of the strength of sunburn-inducing UV radiation reaching the Earth's surface at a given location and time (Fioletov, Kerr, and Fergusson 2010). The UK is situated 55 degrees north of the equator and experiences an average daily maximum UV Index of 3 (defra.gov.uk). In contrast, Singapore lies just 1 degree north of the equator and experiences an average daily maximum UV Index of 8 (nea.gov.sg). Despite this, age-adjusted incidence of keratinocyte cancers is 17-fold lower in Singapore than in the UK (**Fig. 2.1**; (Koh et al. 2003)).

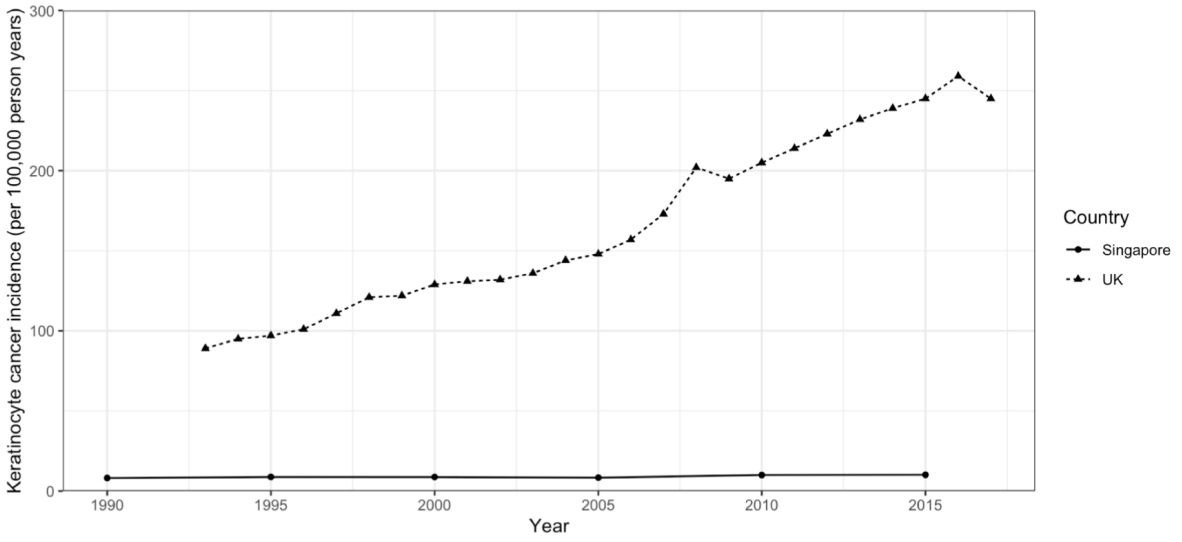


Figure 2.1: Age standardised incidence rates per 100,000 person years for non-melanoma skin cancers in UK and Singapore. Data collated from (1) and the Singapore Cancer Registry (Koh et al. 2003).

High, year-round humidity and temperature in Singapore means many prefer to stay indoors (Sng et al. 2009). In contrast, Western desirability of tanned skin and the rise of affordable air travel means many in the UK receive more intense and greater cumulative sun exposure

than previous generations (Trakatelli et al. 2007). However, it is unlikely that the large difference in skin cancer incidence observed between countries is explained completely by behavioural differences; genetics also plays a role. A study of Singaporean genomes identified 20 loci under positive evolutionary selection (Wu et al. 2019). One such gene is *OCA2*, which encodes a precursor of melanin (an absorbent of UV light), implying a direct link to lowered skin cancer risk. Another gene found to be under positive selection is *HYAL2*, a gene induced in response to UV-B in keratinocytes (Wu et al. 2019). *HYAL2* is situated on a region of chromosome 3p21, which also includes *HYAL1*, *HYAL3* and a cluster of tumour suppressor genes. Neanderthal introgression of this region has been found by Ding *et al.* (2014) to be under positive selection and present at high frequency in East Asians (Southern China: 65%) but near absent in Europeans (UK: 0.56%). This suggests there may be cellular mechanisms, unrelated to pigmentation, which act to reduce skin cancer risk in some populations.

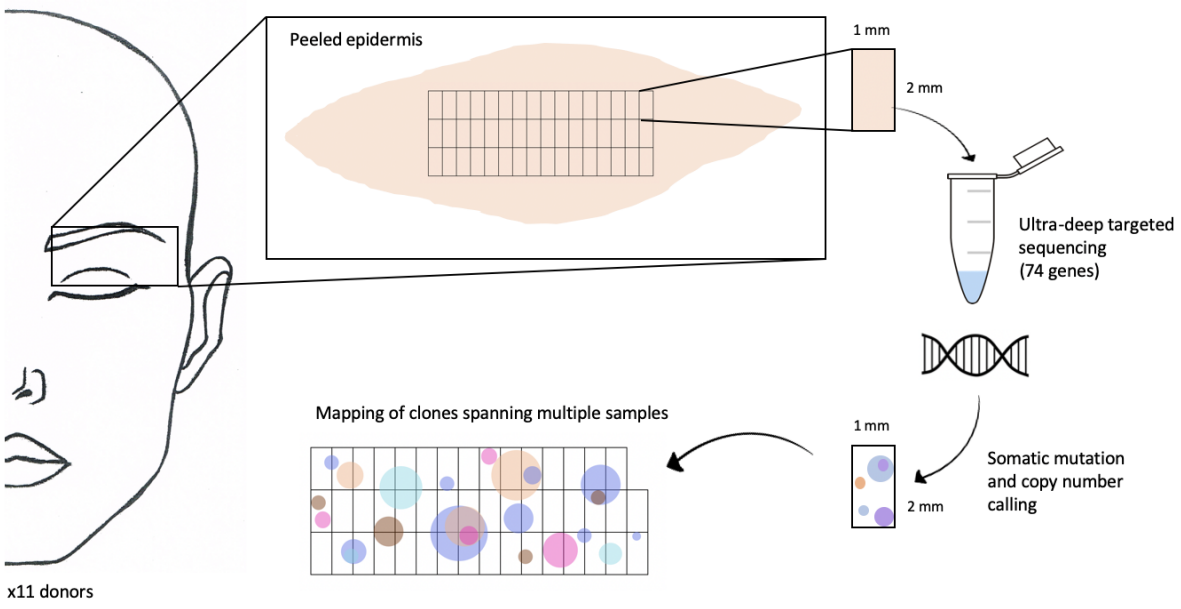


Figure 2.2: Sampling method to allow mapping of clones which span multiple samples of healthy skin.

The rationale for this chapter has come from previous work describing the somatic mutations detectable in biopsies of normal eyelid skin, taken from four donors (Martincorena et al. 2015). This study found somatic positive selection (Martincorena, Raine, et al. 2018) (i.e. the enrichment of protein-altering mutations compared to the expected background rate) in multiple genes commonly mutated in keratinocyte cancers, including *NOTCH1*, *NOTCH2*, *FAT1* and *TP53*. One donor, of South Asian ancestry, was found to have a disproportionately high number of *NOTCH2* mutations compared to three other donors, who are of European

ancestry (Martincorena et al. 2015). This study was first to hypothesise that the genetic background of an individual could lead to differences in selection in the somatic mutational landscape of skin epidermis. In this chapter, I expand on this work, sequencing 191 samples of skin epidermis from five Singaporean donors to compare with published sequencing data of 237 skin samples from six UK donors (Fowler et al. 2021). The same method is employed across both countries (**Fig. 2.2**), taking histologically normal eyelid or eyebrow skin from individuals undergoing blepharoplasty or brow surgery. There is a balance of donors of both sexes from each country and a comparable mean donor age (Singapore = 62 years, UK = 68 years, **Table 2.1**).

Donor	Site	Age	Sex	Occupation	Fitzpatrick score	Smoker	Total area sampled (mm^2)
SG1	Eyelid	70	F	Homemaker	3	No	68
SG2	Eyelid	30	F	Homemaker	4	No	78
SG3	Eyelid	74	M	Retired from military	4	No	76
SG4	Eyelid	77	M	Retired	3	No	80
SG5	Eyelid	59	M	Hawker centre worker	4	Ex	80
UK1	Eyelid	62	M	Indoor worker	2	Yes	88
UK2	Eyelid	77	F	Indoor worker	2	Unknown	90
UK3	Brow	67	F	Indoor worker	2	Ex	78
UK4	Brow	79	M	Journalist	Unknown	Unknown	78
UK5	Brow	73	M	Outdoor geologist	2	No	72
UK6	Brow	51	F	Researcher	Unknown	Yes	68

Table 2.1: Donor demographics.

Materials & Methods

Sample Collection - completed by J Fowler

Sample collection and DNA sequencing of Singaporean skin was carried out as described for published UK skin samples here (Fowler et al. 2021). Eyelid skin was collected from patients undergoing blepharoplasty surgery in Singapore. Informed consent was obtained in all cases under ethically approved protocols (Singapore: NHG DRSB study 2016/00659-AMD0001). Underlying fat and dermis was removed from the skin and the remaining tissue cut into approximately 0.25 cm^2 pieces. Each piece was incubated in 20 mmol/L EDTA for two hours at 37°C. The epidermis was peeled from the dermis using fine forceps under a dissecting microscope and fixed for 30 minutes with 4% paraformaldehyde (PFA; FD Neurotechnologies) before being washed three times in 1x PBS. The fixed epidermis was then cut into a contiguous array of approximately 40 samples per donor, each measuring 2 x 1 mm (**Table 2.1**). DNA was extracted from each sample using the QIAamp micro DNA extraction kit (Qiagen) by digesting overnight and following the manufacturer's

instructions. DNA was eluted using pre-warmed AE buffer where the first eluent was passed through the column twice more.

DNA Sequencing - completed by Wellcome Sanger Institute Sequencing Pipelines

Deep (~700x), targeted sequencing was performed across 74 genes commonly mutated in cutaneous squamous cell carcinomas and other cancers. This custom bait capture, first described by Martincorena *et al.* in 2015, targets the exonic regions of these 74 genes, in addition to 1,734 SNPs across the genome to aid with copy number analysis. The targeted regions cover 0.67 Mb of the genome, with 0.33 Mb being exonic. Samples were multiplexed and sequenced on HiSeq 2000 (Illumina) with version 4 chemistry to generate 75 bp paired-end reads. BAM files were mapped to the GRCh37d5 reference using BWA-MEM (Li and Durbin 2009). Duplicate reads were marked using Biobambam2 (https://gitlab.com/german.tischler/biobambam2).

Indel Realignment & Coverage

I realigned reads around indels using GATK IndelRealigner and calculated depth of coverage for targeted regions per sample using SAMtools. After removing off-target reads, duplicates and those with mapping quality of 25 or less and base quality of 30 or less, I calculated the mean quality sequencing coverage over all 428 epithelial samples (Singapore and UK) to be 749.0x.

Copy Number Analysis

I estimated the allele frequency for each gene in each sample by statistically phasing heterozygous SNPs, as described by (Martincorena et al. 2015). All samples of a patient were used as a panel in order to identify heterozygous SNPs, at sites with at least 1000x total coverage. Due to the variation in read depth across targeted regions, only copy number alterations which lead to an allelic imbalance, including loss of heterozygosity and gains, are detectable via this method. Using the results from published work (Fowler et al. 2021), copy number aberration through SNP phasing was only ever reliably called in clones with a variant allele fraction equivalent to at least 5.1% of cells in a sample. In each case, I regarded the estimation of allele number as reliable when it is matched by a non-synonymous mutation present at that locus in a corresponding proportion of cells, as described by (Martincorena et al. 2015). After excluding samples below this 5.1% detection limit, I still found CNA to be more frequent in UK samples (15.2%) than in Singapore samples (1.6%).

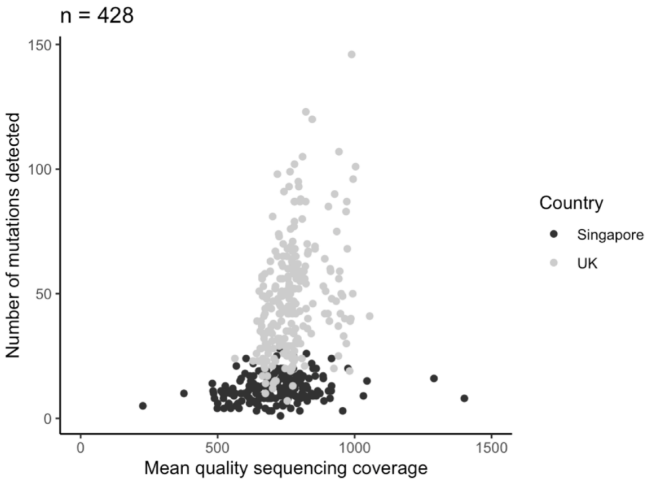


Figure 2.3: Poor correlation between mean quality sequencing coverage per sample and the number of mutations detected (Pearson's r = 0.33). Mean coverage was calculated after removing off-target reads, duplicates and those with mapping quality of 25 or less and base quality of 30 or less.

Mutation Calling

I called mutations using ShearwaterML (Gerstung, Papaemmanuil, and Campbell 2014), an algorithm designed to reliably detect mutations present in a small proportion of cells in a sample and described in detail in (Martincorena, Fowler, et al. 2018). The algorithm uses a deeply sequenced panel of normal, sparsely mutated samples to determine a base-specific error model for each site in the targeted region. Mutations in each sample are then called by comparing the observed mutation rate against the background model using a likelihood-ratio test. I used 51 samples of muscle or fat, sequenced using the same method as the epithelial samples, to create a reference panel with a mean coverage of 42,611x over the targeted regions. After variant calling, filters were applied as described by (Martincorena, Fowler, et al. 2018). Mutations were assumed to be germline and removed if present in more than 10% of all reads across all samples of a single patient. Across all samples, 13,850 mutations were detected, down to a minimum variant allele fraction of 0.0021 (median variant allele fraction = 0.015). I found no evidence of correlation between the mean quality sequencing coverage per sample and the number of mutations detected (Pearson's r = 0.33, **Fig. 2.3**). Mutations were annotated using VAGrENT (Menzies et al. 2015).

Spatial Mapping of Clones

Sampling the tissue in a grid of adjacent samples allows the mapping of large clones that spread over multiple samples. For all downstream analysis, identical mutations called in separate samples of the same donor were merged if the samples were known to have been within 10 mm of each other in the original tissue, rationale explained in (Martincorena,

Fowler, et al. 2018). For donors UK2-UK6, samples were not collected in a single contiguous grid (**Fig. 2.4**).

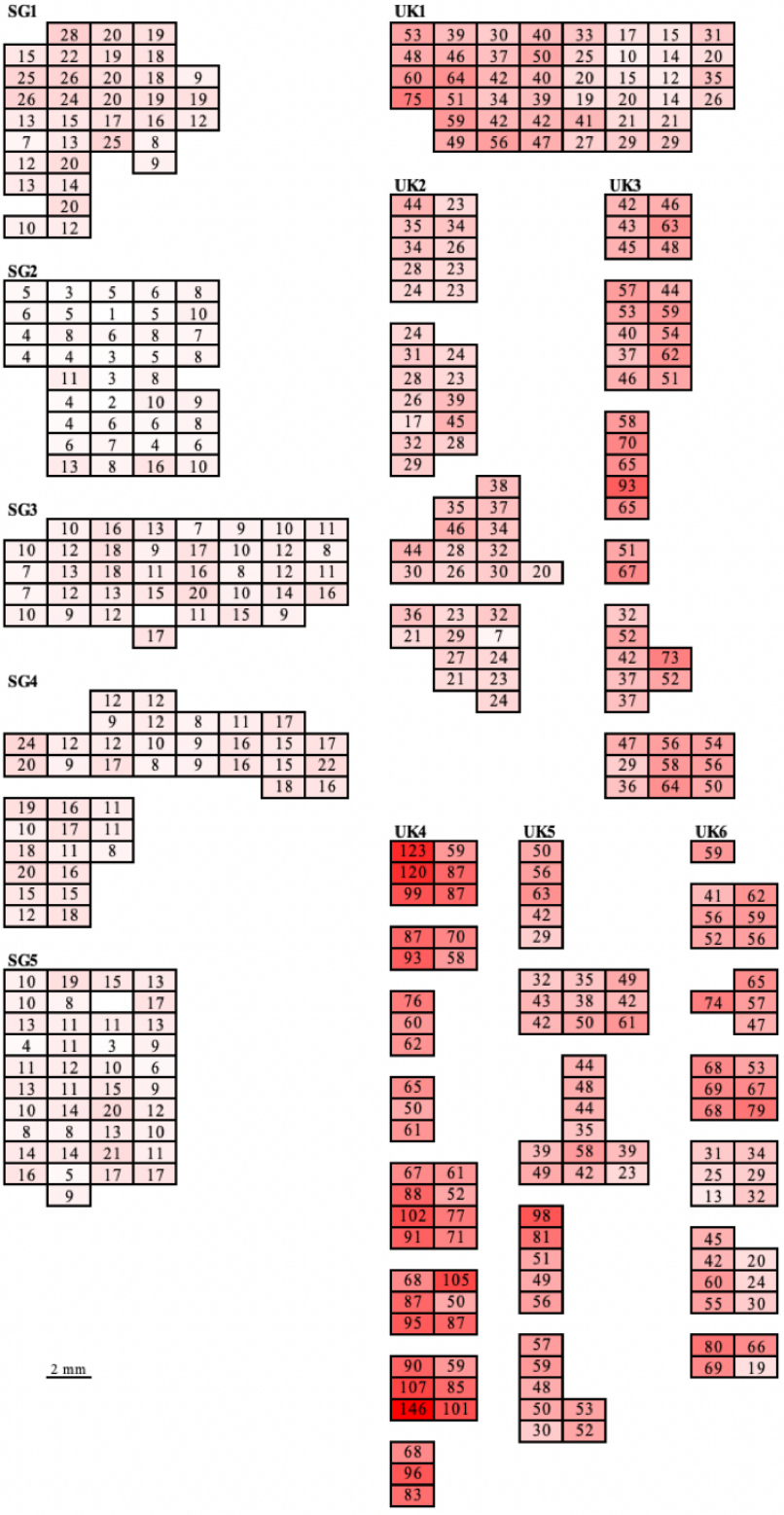


Figure 2.4: Maps showing spatial arrangement of epidermis sampling in each donor. Samples are coloured and labelled by the number of mutations called across targeted regions.

The 193 samples from these donors were taken from 30 distinct pieces of epidermis. In contrast, the 235 samples taken from the remaining six donors were taken from a total of 7 pieces of epidermis. Mutations called in separate pieces of epidermis cut from the same individual were not merged as the distance between these samples cannot be accurately known. However, re-running the analysis with equivalently-sized pieces of epidermis per donor (an artificial maximum of 7 adjacent grids) confirmed that the number of samples per piece of epidermis did not confound estimates of mutation burden or clone size.

Estimates of Mutation Burden and Percentage Mutant Tissue

In the absence of CNA, the proportion of cells in a sample carrying a mutation can be estimated as double the variant allele fraction (the proportion of sequencing reads with a corresponding base change at that position). In this chapter, the genome regions targeted cover genes commonly found to be mutated in cancers and consequently the mutation density observed is not likely to be representative of that genome-wide. I therefore used the method described in (Martincorena et al. 2015; Martincorena, Fowler, et al. 2018), to estimate mutation burden per cell per megabase exclusively from synonymous sites in the bait region. In this estimation I excluded the 32 samples where CNA was detected. The percentage of tissue covered by a mutation and patchwork plots were plotted as described in (Martincorena, Fowler, et al. 2018).

Mutational Signatures

The trinucleotide context of each single-base substitution was determined and the contribution of 49 reference mutational signatures (characterised across multiple cancers as part of the PCAWG study (L. B. Alexandrov et al. 2020b)) to this distribution was estimated using non-negative matrix factorisation with SigProfiler. Spectra for double-base substitutions and indels are also shown, however, the low numbers of each precludes formal signature decomposition.

Selection Analysis

Genes under selection were estimated using dNdScv (Martincorena, Raine, et al. 2018).

Results

Copy number aberration and mutation burden was lower in Singaporean donors

A total of 13,850 mutations were detected across the 428 samples of epidermis (**Fig. 2.5**). Across all donors, there was large variation in sample burden, ranging from 0.5 to 73 mutations/mm^2 (mean = 16.2 mutations/mm^2).

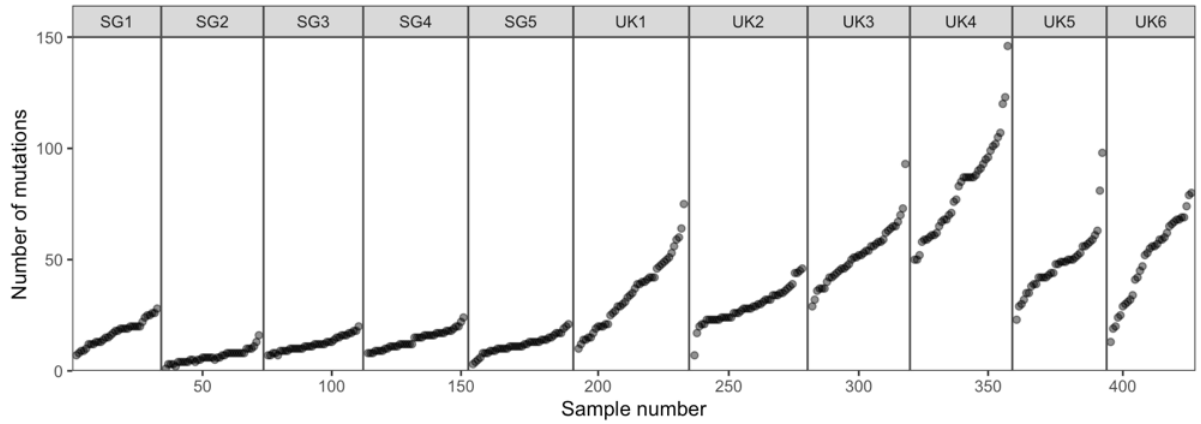


Figure 2.5: Distribution of the number of mutations detected per 2 mm^2 sample of epidermis (n = 428) across 74 genes per donor.

I estimated genome-wide burden per donor using independent synonymous sites only and found the mean burden in UK donors (6.3 mutations/Mb) to be 4-fold higher than that of Singaporean donors (1.6 mutations/Mb, Student's t-test: p = 8.8 $x10^{-3}$, **Fig. 2.6A**). I found no evidence to suggest a difference in mean mutation burden between the eyebrows and eyelids of donors (Welch's t-test: p = 0.14). I detected CNA in 32 of the 428 samples, of which 27 are likely to be independent events (**Fig. 2.6B**). The majority of aberrations detected (78%) were loss of heterozygosity (LOH) at the *NOTCH1* locus on 9q, consistent with mutant *NOTCH1* being the most common driver of clonal expansion in normal skin (Martincorena et al. 2015; Fowler et al. 2021). Other CNA included two *TP53* LOH events, one *NOTCH4* duplication and an LOH event at each of *NOTCH4*, *FGFR2* and *RB1*. I detected CNA in 12.7% of UK samples and in 1.0% of Singapore samples.

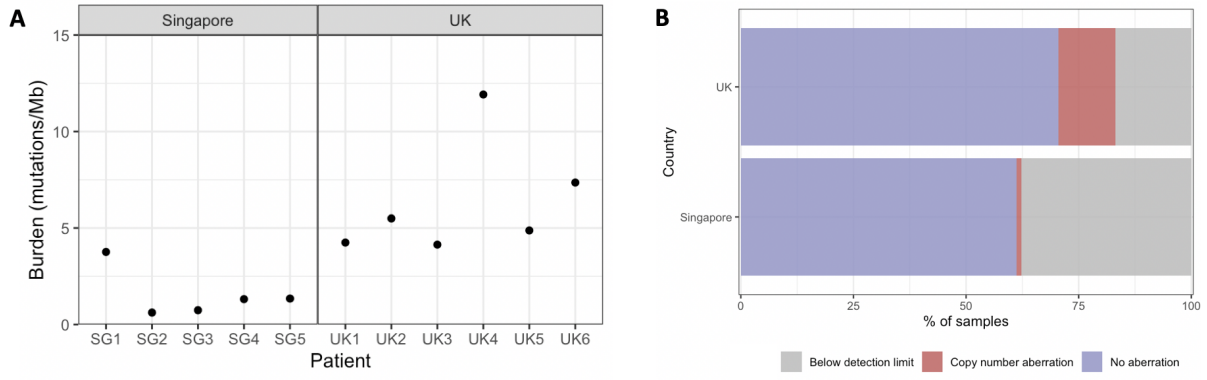


Figure 2.6: A Mutation burden was higher in skin from UK donors (Student's t-test: p = 8.8x10^{-3}). Genome-wide burden was estimated using independent events at synonymous sites only. **B** CNA detected by heterozygous SNP phasing in UK (n = 237) and Singapore (n = 191) samples.

Clones were larger in Singaporean donors

The sensitivity of mutation calling is base-dependent, however in some cases, mutations were detected down to a variant allele fraction (VAF or proportion of mutant reads) of 0.002 (**Fig. 2.7**).

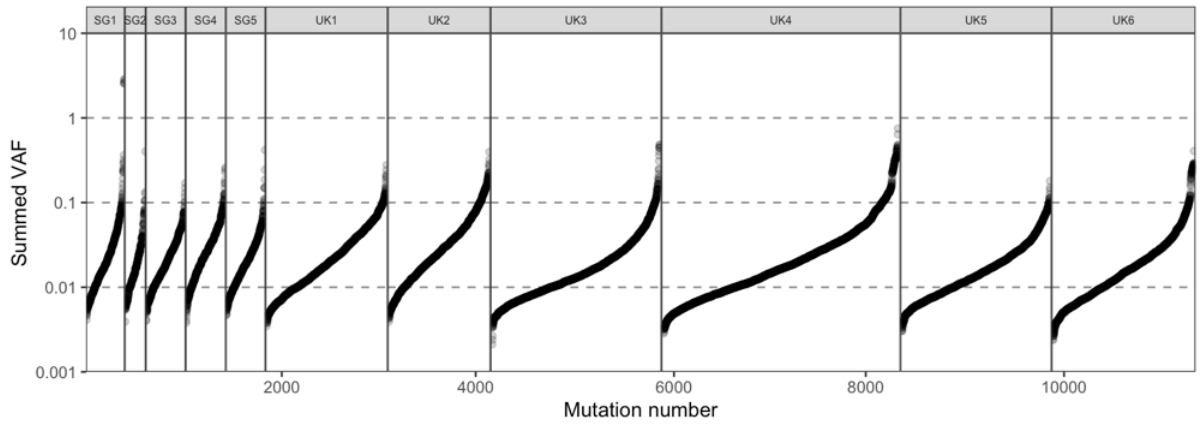


Figure 2.7: Distribution per donor in the summed variant allele fraction (VAF) for each mutation (after merging mutations which span multiple samples, n = 11,356). Median summed VAF = 0.016.

I detected one very large clone as an outlier, driven by a *TP53* P278S mutation in donor SG1. This clone had a summed VAF of 2.86 and occupies 16 adjacent samples, however, the true size of this clone was likely to be larger since it is found on the edge of the piece of tissue sampled (**Fig. 2.8**).

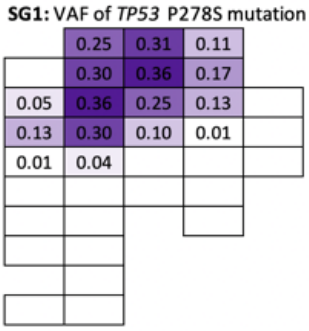


Figure 2.8: A large clone spanning sixteen samples in donor SG1. The VAF for each sample of a single *TP53* P278S mutation is shown. Each rectangle represents a 2 x 1 mm sample.

Surprisingly, I found both the mean and median summed VAF to be larger in Singaporean donors (mean = 0.037, median = 0.019) than in UK donors (mean = 0.029, median = 0.016, Welch's t-test: $p = 4.27 \times 10^{-14}$, **Fig. 2.9**), even after removal of the large outlier clone in SG1 (Welch's t-test: $p = 5.70 \times 10^{-13}$).

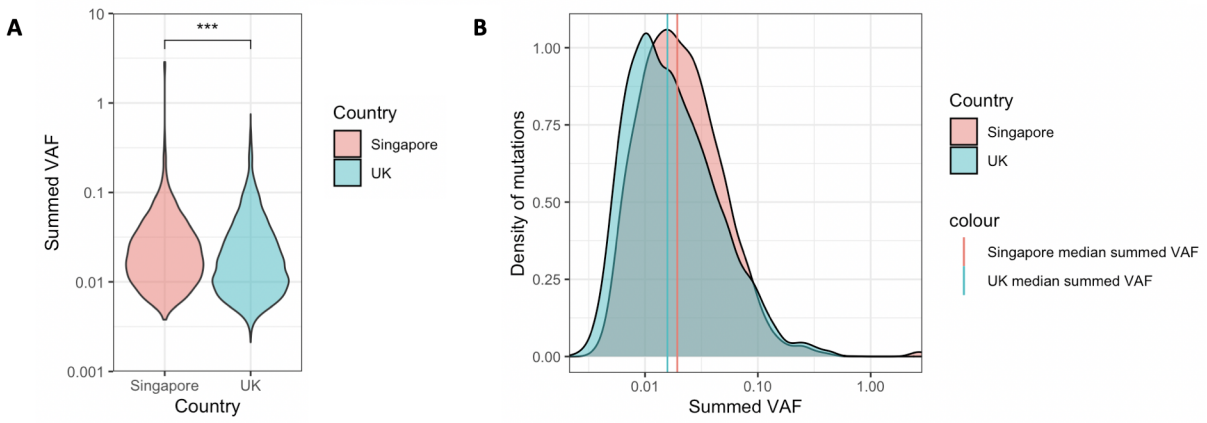


Figure 2.9: A Violin plot comparing the distributions of summed VAF per mutation in patients of each country. Mutations from UK donors had a lower mean summed VAF (Welch's t-test: $p = 4.27 \times 10^{-14}$). **B** Density plot of summed VAF for each mutation by country.

An estimate of the percentage of cells in each donor harbouring at least one non-synonymous mutation can be used as a proxy for the level of competition present in that tissue. Assuming non-synonymous mutations occur as exclusively in cells of a sample as possible, skin from UK donors had a mean 95.7% of cells carrying a mutation in at least one of the 74 targeted genes, compared to a mean of 49.7% in Singaporean donors (**Fig. 2.10**). This is an upper bound estimate, however, it is possible this value is even higher since it does not account for clones too small to be detected or driven by genes not targeted here. If the level of mutation in skin from UK donors is nearing saturation, this could indicate increased competition as an explanation for the smaller mean clone size observed (**Fig. 2.9**).

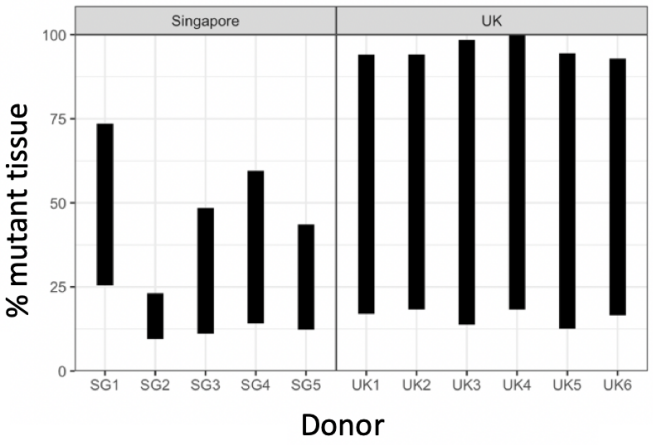


Figure 2.10: Plot per donor of the percentage of cells sampled containing a non-synonymous mutation. Lower bound assumes all mutations in a sample occur within the same subset of cells. Upper bound assumes mutations are as exclusive as possible within the sample. Samples with known CNA were excluded from this estimation.

Skin from UK donors showed more damage by ultraviolet light

Across all donors, 10,311 single-base substitutions (SBS) were detected, of which, 65.6% were C>T changes. **Figure 2.11** shows the trinucleotide context of each SBS detected in donors of each country.

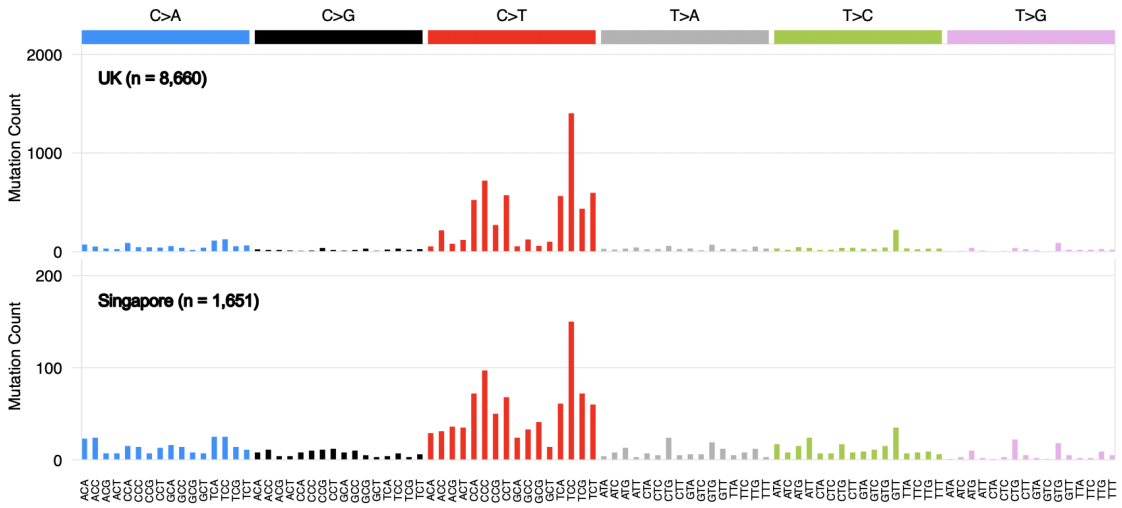


Figure 2.11: The trinucleotide context for each single base substitution in skin from UK and Singapore donors.

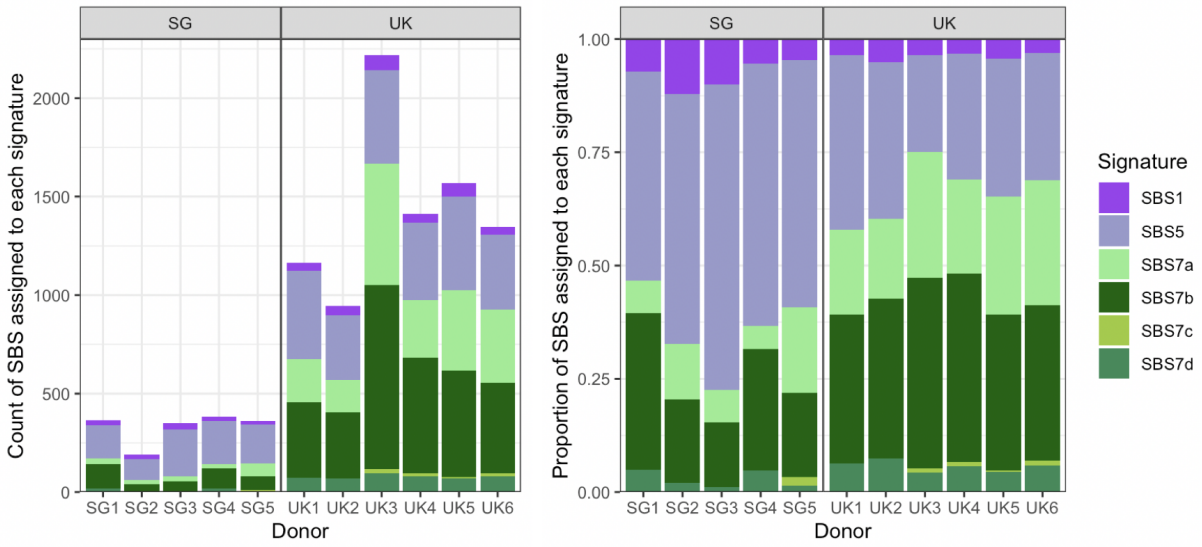


Figure 2.12: Count and proportion, respectively, of single base substitutions assigned to each signature per donor (purple = ageing signatures; green = UV signatures).

Six SBS reference signatures (L. B. Alexandrov et al. 2020b) were identified as contributing to these mutational spectra: SBS1, SBS5 and SBS7a-d (**Fig. 2.12**). Both SBS1 and SBS5 are associated with tissue ageing and are ubiquitous amongst normal tissues and cancers in humans. SBS1 is caused by the endogenous deamination of 5-methyl-cytosine to thymine and rates of SBS1 acquisition in different cell types correlate with estimated rates of stem

cell division (L. B. Alexandrov et al. 2020b). The majority (64.2%) of SBS detected in Singaporean donors were attributed to ageing signatures SBS1 and SBS5. In contrast, the majority (66.1%) of SBS in UK donors were attributed to UV signatures SBS7a-d. It is hypothesised that SBS7a and SBS7b are each the consequence of the two major known UV photoproducts: cyclobutane pyrimidine dimers and 6-4 photoproducts, whilst SBS7c and SBS7d may be the consequence of translesion DNA synthesis by error-prone polymerases inserting T or G respectively, rather than A, opposite UV induced photodimers (L. B. Alexandrov et al. 2020b). The difference in contributions of SBS signatures to the combined spectra of donors from each country was statistically significant (**Fig. 2.13A**, SBS7c excluded due to low counts, Pearson's chi-square: χ^2 = 644, df = 4, p < $2x10^{-16}$), with the over-representation of SBS5 in Singaporean donors having the greatest contribution to the test statistic and all UV signatures over-represented in UK skin. In terms of absolute number of mutations detected, I observed an increase of mutations as a consequence of SBS1 and SBS5 in UK skin, in addition to SBS7, compared to Singapore skin. In fact, the burdens of SBS1 and SBS5 mutations correlated with SBS7 burden (Pearson's R = 0.90 and 0.91, respectively).

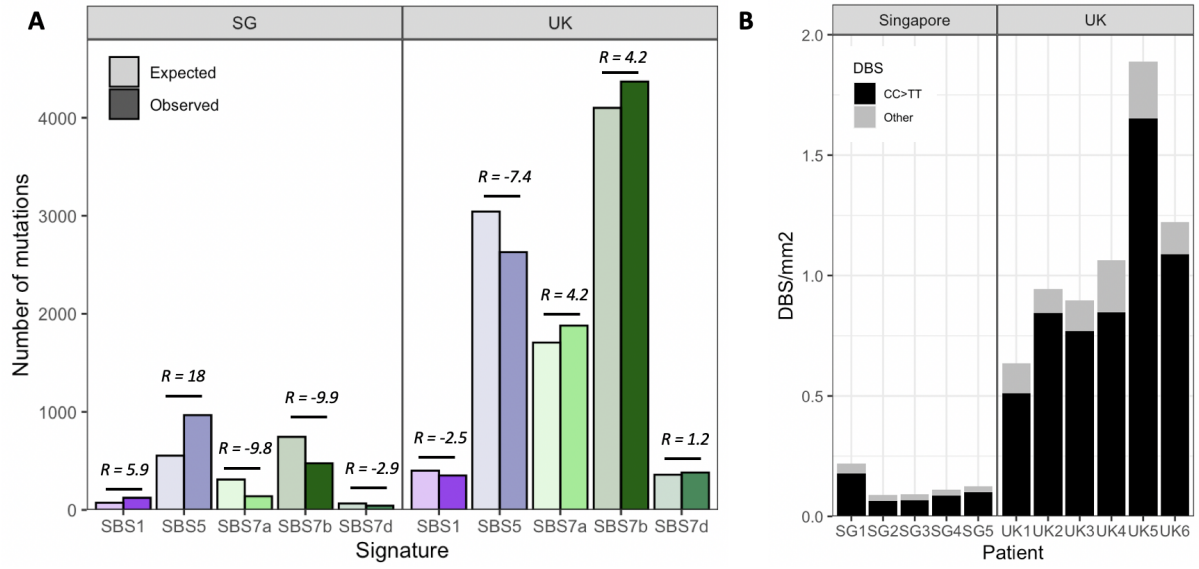


Figure 2.13: A Expected and observed counts of mutations assigned to each SBS signature, R = Pearson's residual (SBS7c excluded due to low counts, Pearson's chi-square: χ^2 = 644, df = 4, p < $2x10^{-16}$). **B** Counts of double-base substitutions (DBS) per mm^2 of skin of donors from each country (UK mean = 1.08 DBS/mm^2, SG mean = 0.126 DBS/mm^2, Welch t-test: p = 2.3 $x10^{-3}$).

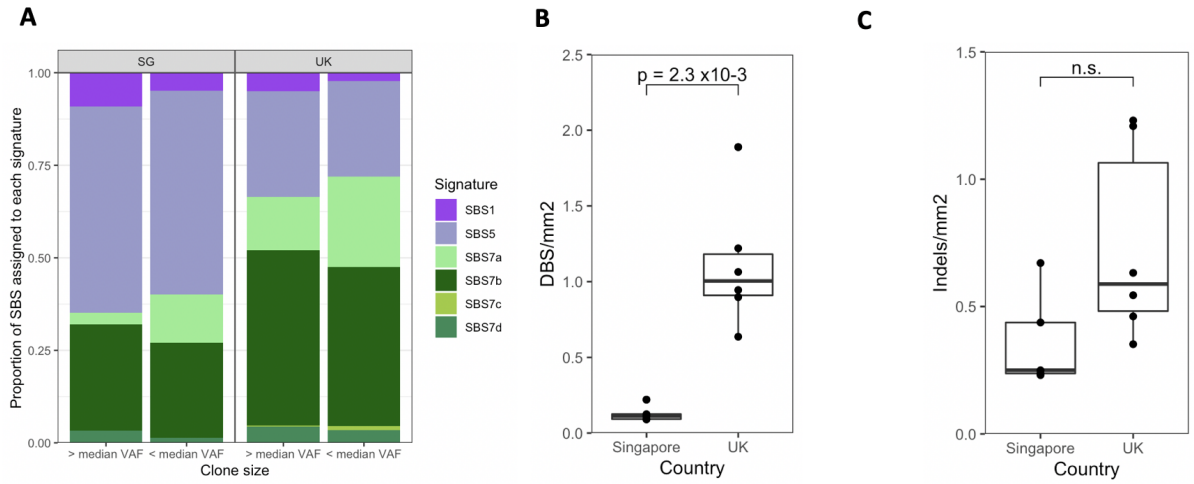


Figure 2.14: A Proportion of SBS assigned to each signature, split by mutations above and below median VAF for each country (Pearson's chi-square UK: $p < 2.2x10^{-16}$; SG: $p = 1.3x10^{-14}$). **B** Difference in double-base substitutions (DBS) per mm^2 in each donor by country (UK mean = 1.08 DBS/mm^2, SG mean = 0.126 DBS/mm^2, Welch t-test: $p = 2.3\ x10^{-3}$). **C** Count of insertions and deletions per mm^2 in each donor by country (UK mean = 0.72 indels/mm^2, SG mean = 0.37 indels/mm^2, Welch t-test: p = 0.07).

C>T mutations caused by UV damage are more frequently observed on the untranscribed strand of genes, due to the repair of lesions on the transcribed strand by transcription-coupled nucleotide excision repair (Shuck, Short, and Turchi 2008). However, C>T mutations caused by ageing (SBS1) do not exhibit strong transcriptional strand bias (L. B. Alexandrov et al. 2020b). I observed a stronger transcriptional strand bias of C>T mutations in skin from donors of the UK compared to Singapore, consistent with increased sun damage (transcribed/untranscribed = 0.829 and 0.896, respectively). Of note, in both the UK and Singapore, I found a significant difference in SBS signature contribution with clone size (**Fig. 2.14A**, Pearson's chi-square: UK $p < 2.2x10^{-16}$ and SG $p = 1.3x10^{-14}$). In both countries, SBS7a was over-represented in mutations below the median VAF and SBS1 over-represented in mutations above the median VAF. One possible explanation could be that larger clones have undergone more cell division, leading to an increase of SBS1 mutations.

Across all donors, I detected 561 double-base substitutions (DBS), 388 deletions and 94 insertion events. Over eight times as many DBS were called in skin of UK donors compared to Singaporean donors (**Fig. 2.14B**, UK mean = 1.08 DBS/mm^2, SG mean = 0.126 DBS/mm^2, Welch t-test: $p = 2.3\ x10^{-3}$). The majority of DBS (85.0%) were CC>TT substitutions and show a transcriptional strand bias consistent with UV damage (transcribed/untranscribed = 52.7), a bias not observed in other DBS types called. The

burden of insertions and deletions was comparatively low and variable across donors and I did not find a significant difference between countries (**Fig. 2.14C**, UK mean = 0.72 indels/mm^2, SG mean = 0.37 indels/mm^2, Welch's t-test: p = 0.07).

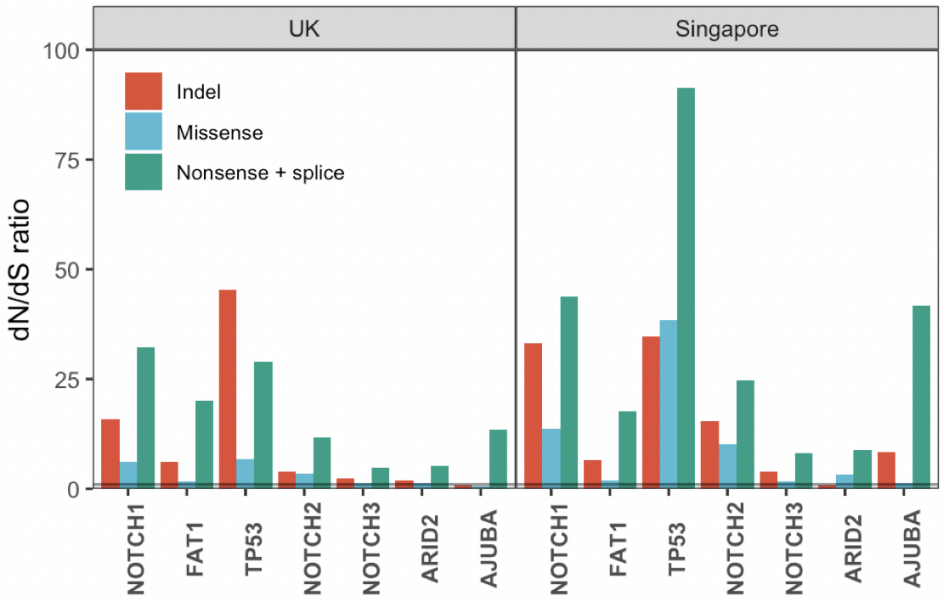


Figure 2.15: Ratio of observed/expected non-synonymous mutations for positively selected genes by country. No synonymous mutations were detected in Singapore skin for TP53 and AJUBA, leading to high dN/dS ratios. Line drawn at y = 1.

Clonal selection and competition differed by country

I found seven genes (*NOTCH1*, *FAT1*, *TP53*, *NOTCH2*, *NOTCH3*, *ARID2* and *AJUBA*) with a significant disproportionately high number of non-synonymous mutations relative to synonymous mutations (the *d*N/*d*S ratio, q < 0.01), suggesting protein-altering mutations in these genes play a role in driving clonal expansion (**Fig. 2.15**). This is consistent with previous studies of the somatic clonal landscape of sun-exposed skin (Martincorena et al. 2015; Fowler et al. 2021). A comparison of the *d*N/*d*S ratio per gene by country finds *TP53* non-synonymous mutations over-represented and *NOTCH1* and *NOTCH2* non-synonymous mutations under-represented in UK donors (with respect to synonymous mutation rates) compared to Singaporean donors (**Fig. 2.16 & Fig. 2.17**). As may be expected with a lower mutation burden, I estimated fewer cells to be carrying a non-synonymous mutation in positively selected genes in Singaporean skin than in the UK (**Fig. 2.18**). However, of all positively selected genes, *NOTCH1* and *NOTCH2* had the smallest difference between the two countries, with the estimated proportion of cells mutant for these two genes in the UK and Singapore having an overlapping range. In fact, I estimated non-synonymous mutations in *NOTCH2* to be present in 6-8% of cells in Singaporean skin, higher than the 4-6%

estimate for UK skin. After removing the donor with a large outlier clone, SG1, the Singaporean estimate fell to 3-5%. In contrast to *NOTCH2*, I estimated that 7-13% of cells in UK donor skin carry a protein-altering mutation in *TP53*, compared to just 4-5% of cells in Singaporean skin. After removing SG1, this fell to just 1-2% of cells in Singaporean skin.

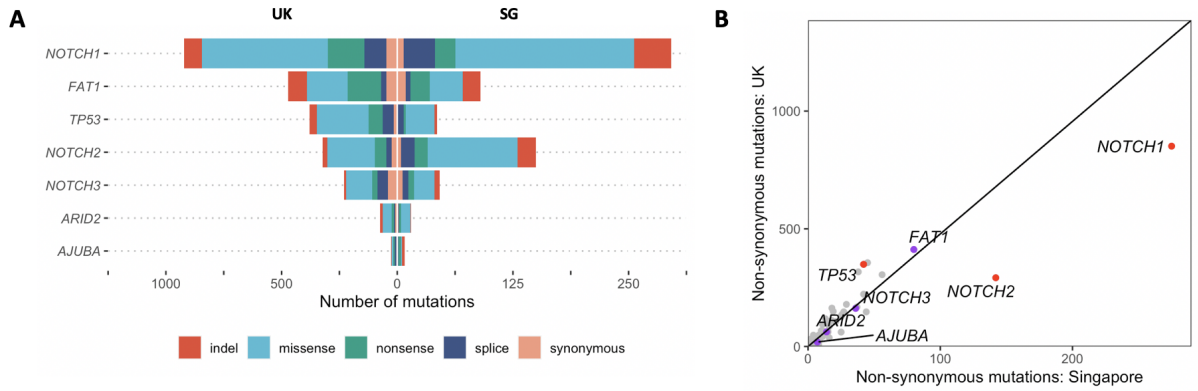


Figure 2.16: A The number of mutations of each consequence for positively selected genes (q < 0.01) by country. *ARID2* was not significantly positively selected in Singaporean samples. **B** Plot of non-synonymous mutations per gene in Singapore vs. UK samples. Gradient of line = total number of non-synonymous mutations in UK/SG = 6846/1432. Positively selected genes (purple) are labelled. Red indicates positively selected genes with a significant (p < 0.001) difference in dN/dS ratio with country, after accounting for global differences.

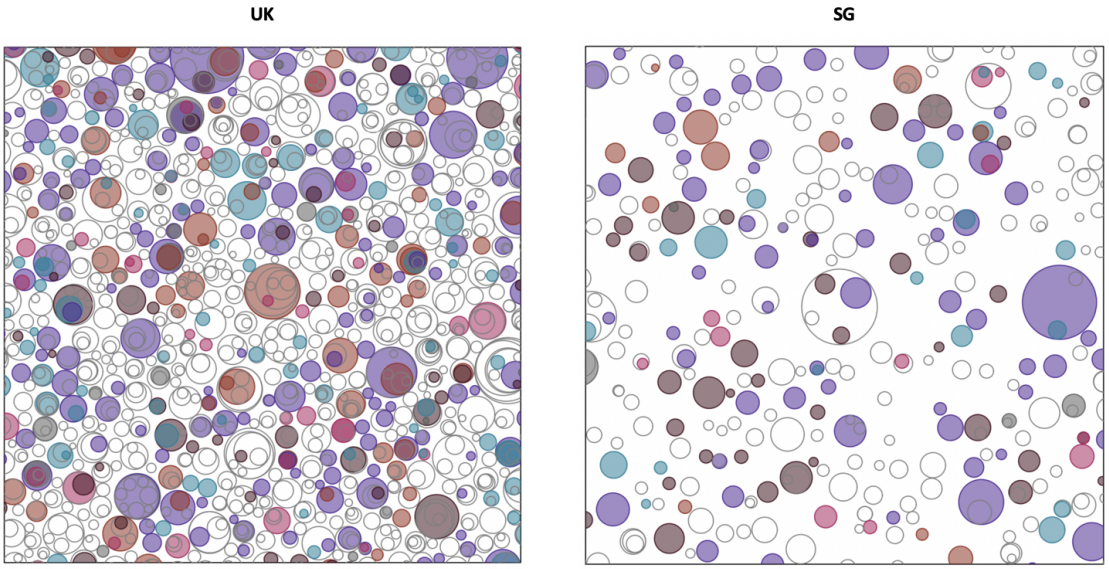


Figure 2.17: A representation of protein-altering mutations in 1 cm^2 of skin from donors of the UK and Singapore. Samples from the two countries were randomly selected and mutations displayed as circles, randomly distributed in the space. Sequencing data, including copy number, was used to infer the size and number of clones and, where possible, the nesting of sub-clones. Otherwise, sub-clones are nested randomly. *NOTCH1* (purple), *NOTCH2* (brown), *TP53* (orange), *NOTCH3* (pink), *FAT1* (green).

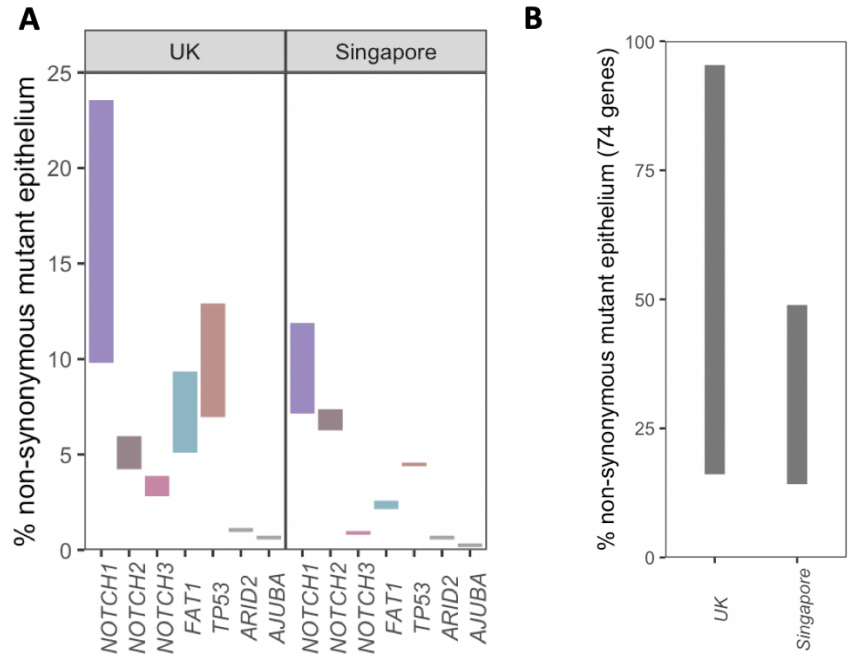


Figure 2.18: A Estimated range of the percentage of cells harbouring at least one non-synonymous mutation per positively selected gene, by country (samples with known CNA were removed). **B** Estimated range of the percentage of cells harbouring at least one non-synonymous mutation across all 74 genes, by country (samples with known CNA were removed).

Codons common in cancer were more prevalent in UK skin

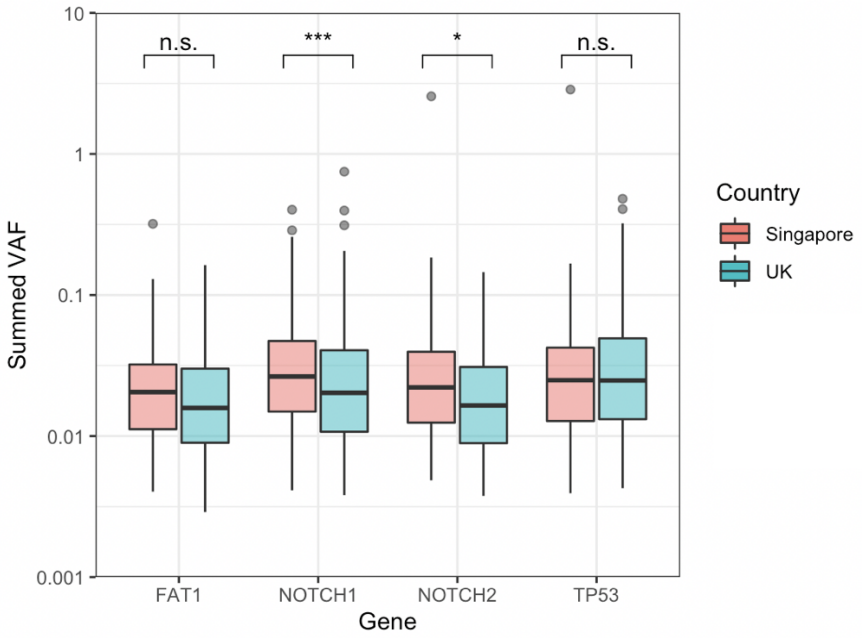


Figure 2.19: The total summed variant allele fraction (VAF) of protein-altering mutations in each of the top four positively selected genes, by country (samples with known CNA were removed). T-test p = $2x10^{-4}$ (*NOTCH1*) and $3x10^{-3}$ (*NOTCH2*).

I found the mean summed VAF of *NOTCH1* and *NOTCH2* mutant clones to be larger in Singaporean skin than in the UK (**Fig. 2.19**, t-test p = $2x10^{-4}$ and $3x10^{-3}$ respectively). I hypothesised that this was because UK skin is nearing saturation (**Fig. 2.10 & Fig. 2.18B**), as described earlier, where it is not inconceivable that every cell harbours at least one protein-altering mutation. It is likely that this increased competition in UK skin is restricting the growth of weak competitors, such as clones mutant for *NOTCH1* or *NOTCH2*. However, in Singaporean skin, I estimated roughly half of the cells to still be wild-type, reducing the level of competition and enabling *NOTCH1* and *NOTCH2* mutant clones to grow larger. In the same way, if there is increased competition in UK skin, we may expect that *TP53* mutant clones are limited in their growth. However, I did not find a difference in the mean summed VAF for *TP53* mutant clones by country (**Fig. 2.19**). This suggests that *TP53* mutants are stronger competitors than other non-wild-type clones present in the tissue (such as *NOTCH1* and *NOTCH2* mutants).

It could also be that UK *TP53* mutants are stronger competitors than the type of *TP53* mutations observed in Singaporean skin and, indeed, there was evidence to suggest this. The two most frequent codon changes observed in UK skin were *TP53* R248W and R282W (at 3.2 mutations/cm^2 and 2.5 mutations/cm^2 respectively). Both these specific codon changes target the p53 DNA-binding domain and there is evidence to suggest they have gain-of-function properties that can lead to chromosomal instability (H. Song, Hollstein, and Xu 2007); (Murai et al. 2018); (Zhang et al. 2016). R248W and R282W are the second and fourth most frequent *TP53* codon changes, respectively, observed in human cancers, with R248W the most frequent codon change in keratinocyte cancers (Giglia-Mari and Sarasin 2003; Zhang et al. 2016). Adjusting for the difference in the estimated genome-wide mutation burden between countries, I would expect to observe 0.8 mutations/cm^2 and 0.6 mutations/cm^2 at *TP53* R248 and R282, respectively, in Singaporean skin and therefore, across the sample area, approximately 3 R248W and 2-3 R282W mutations. However, I did not observe any mutations at either of these codons in the Singaporean skin sampled. Other oncogenic hotspot mutations I observed in UK skin include: *FGFR3* K652M (n = 2), G382R (n = 2), R248C (n = 1), S249C (n = 1), G372C (n = 1) and Y375C (n = 1); *KRAS* G12D (n = 1); *NRAS* Q61L (n = 1) and *HRAS* E143K (n = 2) and G12D (n = 1). In Singaporean skin, I observed a single occurrence of the oncogenic mutation *KRAS* G12V.

Genetic background of donors

In order to gain more insight into differences that may exist in the genetic background of the donors in this chapter, I used off-target reads to genotype each individual for 41 single nucleotide polymorphisms (SNPs) associated with increased risk for cSCC (Green and Olsen 2017) and BCC (Verkouteren et al. 2017) (**Fig. 2.20**). At 13 risk loci I found evidence of polymorphism in at least one donor of the UK but none in Singaporean donors. The majority of these SNPs are associated with genes linked to pigmentation, such as *SLC45A2*, *IRF4*, *BNC2* and *HERC2*, but also SNPs in *FOXP1* (involved in immune response), the proto-oncogene *SRC* and the oncogene *SEC16A*. All risk loci that showed evidence of polymorphism in the Singaporean donors were also present in the UK donors. In addition, four of the five Singaporean donors and zero UK donors had introgression of the *HYAL2* region at chromosome 3p21 described by (Ding et al. 2014).

Discussion

This chapter describes the somatic mutational landscape of sun-exposed eyelid skin from donors of Singapore and compares this to the landscape of published skin samples from the eyelid and brow of donors of the UK (Fowler et al. 2021). The UK has a 17-fold increased incidence of keratinocyte cancers relative to Singapore and I identified multiple features of the mutational landscape to reflect this difference in cancer incidence. Not only did I estimate a 4-fold increase in genome-wide mutational burden in UK skin, but I found a disproportionate increase in *TP53* protein-altering mutations, particularly in codons common to cancers (such as R248W and R282W). This is interesting as it supports work from previous studies which have suggested that UV radiation acts to promote carcinogenesis twice, first as a mutagen and second in promoting the expansion of *TP53* mutant clones (Brash et al. 1996), particularly those which provide a strong cell fate imbalance, such as *TP53* R248W (Murai et al. 2018). I identified further features common to cancers which are more prevalent in UK donors compared to those of Singapore, such as a higher number of oncogenic hotspot mutations and a 10-fold increase in chromosomal CNA. I found protein-altering mutations in both *NOTCH1* and *NOTCH2* to be overrepresented in Singaporean skin compared to the UK. This substantiates previous work which identified a disproportionately high number of *NOTCH2* protein-altering mutations in skin of a donor with South Asian ancestry compared to three West European donors (Martincorena et al. 2015). However, by now taking samples as a gridded array across a larger area of skin from a greater number of donors, I was able to more accurately estimate clone size and burden per donor. I also showed that both *NOTCH1* and *NOTCH2* mutant clones colonised a

comparable proportion of the tissue across the two countries, despite the mutation burden of Singaporean skin being around a quarter of that of the UK. Mean clone size, including that of *NOTCH1* and *NOTCH2* mutants specifically, was larger in Singaporean skin.

SNP ID	Cancer	Gene	Function	SG1	SG2	SG3	SG4	SG5	UK1	UK2	UK3	UK4	UK5	UK6
rs16891982[G]	cSCC	*SLC45A2*	Pigmentation	1	2	2	2	3	1			1	2	1
rs12916300[T]	cSCC	*HERC2*	Pigmentation	1	2	3	3	2	1		1	3	2	4
rs74664507	cSCC	*BNC2*		1	1	3	3	1	3	3	2	1	1	2
rs62246017[A]	cSCC	*FOXP1*	Immune response			1	2	2	1		1	2	1	6
rs12203592[T]	cSCC	*IRF4*	Pigmentation		1	1	2	1				2	2	3
rs754626[G]	cSCC	*SRC*	Proto-oncogene	8	3	8	5	5	6	6	2	7	3	5
rs57994353[C]	cSCC	*SEC16A*	Oncogene	3	2	8	3		14	>20	12	18	7	20
rs1805008[T]	cSCC	*MC1R*	Pigmentation	3	1	3	1	3	8	1	1	3	3	4
rs12210050[T]	Both	*EXOC2*		1	4	1	2	2	7	1	4		2	2
rs6059655[A]	cSCC	*RALY*	Pigmentation		2	2	1	3	1		2		2	
rs11170164[T]	BCC	*KRT5*		2	4	5	2	6	5	4	5	5	3	3
rs1805007[T]	Both	*MC1R*	Pigmentation	3	1	3	1	4	7	3	2	3	4	6
rs214782[G]	BCC	*TGM3*		2	5	7	7	3	7	2	4	9	1	2
rs192481803	cSCC	*AHR*	Anti-apoptopic pathways		1	3	2	1	3		2	2		1
rs6791479	cSCC	*TP63*			1		2		1	1				3
rs17247181[T]	cSCC	*ERBB2IP*	Ras signalling			2	1	1	2	2	6	1	1	1
rs117132860[A]	cSCC	*AHR*	Anti-apoptopic pathways	1	2	1		2	5		1	1	2	2
rs1126809[A]	cSCC	*TYR*	Pigmentation	3	1	1	1	1	2	1			1	3
rs74899442[C]	cSCC	*CADM1*	Cell-mediated immunity		1			2			1			1
rs1805005[T]	cSCC	*MC1R*	Pigmentation	1		3	7		2	7	4	3	1	6
rs1805006[T]	cSCC	*MC1R*	Pigmentation	2	2	8	3		9	11	3	8	6	6
rs11547464[A]	cSCC	*MC1R*	Pigmentation	3	1	2	1	2	5	3	3	4	3	4
rs1110400[C]	cSCC	*MC1R*	Pigmentation	2	1	2	1	4	7	2	1	2	3	5
rs1805009[C]	cSCC	*MC1R*	Pigmentation	4	3	1	4	4	4	4	7	5	3	2
rs78378222[C]	BCC	*TP53*		>20	>20	>20	>20	>20	>20	>20	>20	>20	>20	>20
rs7335046[G]	Both	*EXOC2*		2	2	4	5	1	>20	11	17	>20	17	20
rs13014235[C]	BCC	*ALS2CR12*		1	1	5	1	2	5				3	1
rs4455710[T]	cSCC	*HLA_DQA1*	Immune response		2	3		2	2			2	1	
rs59586681[T]	BCC			1	1	1	3	3	5	3	1	1	2	1
rs4268748[C]	cSCC	*DEF8*	Pigmentation	8	7	14	7	8	11	15	10	14	13	20
rs2228479[A]	cSCC	*MC1R*	Pigmentation	2	4	5	3		11	11	1	6	4	5
rs7538876[A]	BCC	*PADI6*				2	4	3		2	3	3		
rs9689649[C]	cSCC	*PARK2*		1	3	6					1	3	4	1
rs2151280[G]	BCC	*CDKN2A*			3	2	1	1	2	2	1	3	2	1
rs157935[T]	BCC				2	4	2	5	8	3	1		3	3
rs73635312[G]	BCC				2	1	1	1	3			2		1
rs801114[G]	BCC	*RHOU*		5	2	2	2	2			3	1	2	1
rs57244888[T]	BCC			1	1	3	1			1		2		3
rs7006527[A]	BCC	*RGS22*			3		1	1		4	2	1	1	2
rs28727938[C]	BCC	*MRPL9P1*		3	3	5	3	4	9	1	1	1	3	2
rs401681[C]	BCC	*CLPTM1L*		1	3	2	1	6	2	3	2	1	2	4

Figure 2.20: Genotyping of donors using off-target reads at loci of known risk SNPs for cSCC and BCC. Red = evidence of homozygosity for risk SNP, orange = evidence of heterozygosity, green = no evidence of risk SNP, grey = no data available, number = read count at that locus.

I hypothesise that the growth of *NOTCH1* and *NOTCH2* mutant clones is more restricted in UK skin due to the increased mutation burden of the tissue, leading to increased competition with fitter, non-wildtype cells, such as *TP53* mutants. I estimated that the skin of the UK donors is nearing saturation, with a mean upper bound estimate of 95.7% of cells carrying a non-synonymous mutation, approximately double that estimated for Singaporean skin. I found no difference in *TP53* mutant clone size between countries, despite the increased

burden and competition present in UK skin. This could be further evidence that UV light promotes the expansion of *TP53* mutants, giving them an advantage over other non-wildtype clones, such as *NOTCH1* and *NOTCH2* mutants, in such an environment.

The majority of the increased burden of mutations observed in UK skin can be explained by damage by UV light (SBS7 and CC>TT double-base substitutions). However, I did observe an increase in the number of mutations assigned to the non-UV signatures SBS1 and SBS5 in UK skin. It could be that these mutations are more likely to be above the clone size limit of detection in UK skin due to the increased number of cells mutant for a driver of clonal expansion. It is also possible that cells within UK skin have undergone more divisions due to increased exposure to UV radiation promoting cell proliferation.

In summary, I observed multiple differences in the mutational landscape of skin across the two countries and this helps to explain the difference in skin cancer risk. The main limitation of this study, however, is that this difference in skin cancer risk is multifactorial and it is not possible to ascertain the impact each factor individually has on the landscape. I have shown that Singaporean skin is exposed to less damage by UV light than UK skin, but it is likely that multiple processes are acting to prevent this damage. Firstly, the Singaporean donors have SNPs associated with the production of melanin (which acts to absorb UV light) that are absent in the UK donors. Secondly, 80% of the Singaporean donors and none of the UK donors have introgression at a region of chromosome 3p which includes *HYAL2*, a gene thought to be involved in hyaluronan metabolism and cellular response to UVB radiation (Ding et al. 2014). It has been suggested that introgression of this region is latitude-dependent in Asian populations and *HYAL2* expression is dependent on UV exposure in HaCaT cells (immortalised keratinocytes) (Ding et al. 2014). This, in addition to the presence of additional keratinocyte cancer risk SNPs in the UK donors, shows that there are likely to be cellular mechanisms unrelated to pigmentation which have protected Singaporean skin here from UV damage. Thirdly, it is possible that behavioural differences around sun exposure have meant that Singaporean skin is less exposed to UV light in the first place, despite a greater year-round UV index than that of the UK.

Finally, it is possible that differences in the genetic background between individuals of the two countries favour the growth of particular mutants, whilst perhaps decreasing the fitness of others. One way to determine this, without requiring the sequencing of a prohibitively large number of donors and samples, would be to analyse the cancer genomes of keratinocyte cancers in Singapore to determine if the same drivers, *TP53* codons and mutational

signatures are present as in high risk countries, with reference to the known landscape of normal skin.

Polygenic risk scores created from genome-wide association studies are currently able to explain a 2-fold increase in risk for cSCC (Roberts, Asgari, and Toland 2019). Future work should focus on expanding these studies to populations where skin cancer rates are low, in order to identify SNPs which may confer protection. Such work could use these SNPs to stratify individuals within a country in order to assess the impact skin pigmentation and other consequences of genetic background have on the mutational landscape of donors, whilst minimising the effects of differences in UV exposure and environment. The identification of cellular mechanisms at work in some populations to minimise UV damage could enable clinical interventions to help prevent skin cancers in at-risk individuals.

Chapter 3: The Effect of Body Site on the Mutational Landscape of Normal Skin

Introduction

The incidence of keratinocyte cancers is increasing worldwide, likely due to improved screening, a growing ageing population and a depleting ozone layer (Kim, Del Rosso, and Bellew 2009). As discussed in the previous chapter, keratinocyte cancer incidence is highest in populations with low levels of skin pigmentation (Que, Zwald, and Schmults 2018). In these populations, skin cancers are most frequently found on sun-exposed skin sites (**Fig. 3.1**), such as the face and forearms (Que, Zwald, and Schmults 2018). Furthermore, global cSCC incidence is three times higher in men than in women (Que, Zwald, and Schmults 2018), which may at least be partially explained by differences in preferences for outdoor work and the use of cosmetics, consequently resulting in differing levels of UV exposure between the two sexes (Kim, Del Rosso, and Bellew 2009).

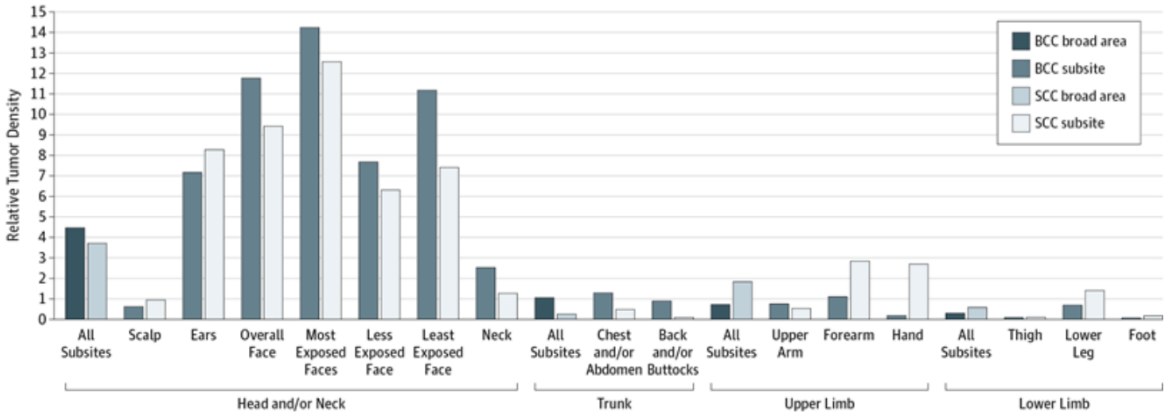


Figure 3.1 (from Subramaniam *et al.*, 2017): Relative tumour density (ratio of tumours to skin surface area) by body site for cSCC and BCC lesions (Subramaniam et al. 2017). Relative tumour density of the whole body = 1.

In Chapter 2, I characterised the mutational landscape of sun-exposed skin from the eyelid or brow of six UK donors. In order to expand our knowledge of how the mutational landscape of skin varies by body site, in this chapter I combine samples with skin taken from a range of body sites from a further 29 donors of the UK, published in Fowler *et al.* (2021) (**Table 3.1**). Abdominal skin was obtained through warm autopsy of organ transplant donors. Normal skin from other sites was obtained through wide local excision after diagnosis of melanoma or, in one case, from a donor undergoing cheek surgery. The same methodology of sample

preparation and DNA sequencing of 2-mm^2 contiguous grids was applied across both Chapters 2 and 3, allowing a direct comparison of the mutational landscape of skin by body site (**Fig. 2.2**).

Donor	Body Site	Age	Sex	Smoker	Tissue Origin	Other Cancer	Years working outdoors (UK)	Years living abroad	No. 2 mm^2 grid samples	No. punch samples
PD30272	Abdomen	36	F	No	Warm autopsy	None	-	-	26	-
PD30273	Abdomen	68	M	Ex	Warm autopsy	None	-	2 (Spain)	24	-
PD30274	Abdomen	54	M	No	Warm autopsy	None	-	-	24	-
PD36126	Forearm	68	F	Ex	Melanoma WLE	BCC (foot)	0	0	39	32
PD37182	Abdomen	67	F	Ex	Warm autopsy	None	-	-	45	-
PD37184	Abdomen	63	M	Ex	Warm autopsy	None	-	-	27	-
PD37185	Abdomen	26	F	Yes	Warm autopsy	None	-	-	20	-
PD37275	Abdomen	75	M	Ex	Warm autopsy	None	-	-	23	-
PD37576	Forearm	69	M	No	Melanoma WLE	None	0	0	40	-
PD37577	Forearm	41	F	No	Melanoma WLE	None	0	0	34	32
PD37578	Trunk	57	F	No	Melanoma WLE	Lynch syndrome	20	0	23	-
PD37579	Forearm	56	M	No	Melanoma WLE	None	23	0	39	-
PD37614	Head	74	M	No	Melanoma WLE	None	1	0	36	-
PD37615	Head	70	M	No	Melanoma WLE	None	0	4 (Singapore)	40	-
PD37617	Head	47	M	Ex	Melanoma WLE	None	12	0	36	-
PD37619	Head	62	M	Yes	Blepharoplasty	None	0	0	44	-
PD38215	Trunk	71	M	Ex	Melanoma WLE	None	40	0	39	26
PD38216	Trunk	30	M	Yes	Melanoma WLE	None	15	0	39	-
PD38217	Trunk	76	M	No	Melanoma WLE	None	50	0	39	61
PD38218	Trunk	71	F	No	Melanoma WLE	Lung + liver metastases	0	15 (Spain)	97	-
PD38219	Forearm	58	M	No	Melanoma WLE	Leukaemia + cSCCs	39	0	23	-
PD38220	Trunk	65	M	No	Melanoma WLE	None	0	0	37	-
PD38330	Leg	67	F	Ex	Melanoma WLE	None	0	0	39	-
PD38331	Leg	59	F	No	Melanoma WLE	None	0	0	39	27
PD38332	Leg	70	F	Ex	Melanoma WLE	BCCs	0	0	39	-
PD38333	Leg	44	F	No	Melanoma WLE	None	0	0	39	-
PD38334	Leg	71	F	Ex	Melanoma WLE	None	0	0	38	31
PD38335	Leg	32	F	No	Melanoma WLE	None	0	0	32	-
PD38336	Leg	48	M	Ex	Melanoma WLE	None	0	0	31	-
PD43943	Head	72	M	Ex	Cheek surgery	Prostate	8	0	7	-
PD43991	Head	67	F	Ex	Brow surgery	None	0	5 (Spain)	39	-
PD43992	Head	79	M	-	Brow surgery	Leukaemia + cSCCs	0	22 (Australia)	39	-
PD43994	Head	73	M	No	Brow surgery	Oral SCC + pharyngeal cancer	30	0	36	-
PD43995	Head	77	F	-	Blepharoplasty	None	0	0	45	-
PD43996	Head	51	F	Yes	Brow surgery	None	3	0	34	-

Table 3.1: Demographics of 35 UK donors across all samples, including the six donors analysed in Chapter 2 (UK1-6 = PD37619, PD43995, PD43991, PD43992, PD43994, PD43995 and PD43996, respectively). WLE = wide local excision.

In addition to 1,251 2-mm^2 grid samples, 209 smaller samples (with an area a 40th of that of a grid sample) were collected using a circular punch across six donors. Each punch sample is estimated to contain approximately 2,400 nucleated cells, including 750 basal cells (Fowler et al. 2021), which allows the mapping of clones at a higher spatial resolution, possibly giving insight into a more recent mutational landscape. Furthermore, 46 of these punch samples were whole-genome sequenced to obtain a more accurate measure of genome-wide mutation burden, mutational signatures, copy number aberration and telomere length. Results from this chapter are published in Fowler *et al.* (2021).

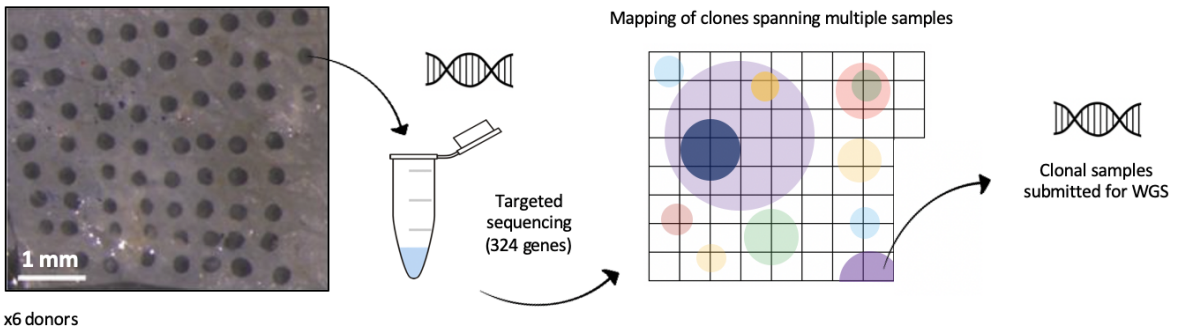


Figure 3.2: Method of punch sampling in skin epidermis (photo credit: J Fowler). 209 punch samples were submitted for targeted sequencing (~45X). Samples dominated by a mutant clone were then submitted for whole-genome sequencing (~33X).

Materials & Methods

Targeted 2-mm^2 grid samples

Skin tissue was collected from patients undergoing wide local excision after initial melanoma excision or from deceased organ donors whose organs were being retrieved for transplantation. In one case, cheek skin was obtained through a donor undergoing surgery for zygomaticus plication. Written informed consent was obtained in all cases under ethically approved protocols (Research Ethics Committee references 15/EE/0152 NRES Committee East of England—Cambridge South and 15/EE/0218 NRES Committee East of England—Cambridge East). All sample collection and preparation for DNA sequencing was carried out by J Fowler as described in Chapter 2. In addition, fat and dermis was collected as a germline sample from each donor. All samples were sequenced by Wellcome Sanger Institute Sequencing Pipelines using the 74-gene targeted bait as in Chapter 2. ShearwaterML was used to call somatic mutations across a panel of germline samples (providing a mean coverage depth of >24,000X) as part of the Jones group in-house pipeline as described in Fowler et al. (2021). I used this set of somatic mutations to carry out the

downstream analysis shown in this thesis and used SNP phasing to call allele copy number, as described in Chapter 2.

Targeted punch samples

For six donors, 232 additional samples were taken with a circular punch (0.25 mm diameter, Stoelting Europe), a 40th of the area of a 2-mm^2 grid, by J Fowler. Within each donor, skin was sampled as a rectangular array of punches, with approximately a punch diameter of unsampled space between each (**Fig. 3.2**). DNA was extracted by J Fowler using the Arcturus Picopure Kit (Applied Biosystems) and 23 punch samples failed due to low DNA input. All remaining samples were sequenced for 324 genes (to include a broader range of genes frequently mutated in cancers) by the Wellcome Sanger Sequencing Low-Input Pipeline with an average on-target coverage of 45X. Variants were called using the ShearwaterML algorithm (Gerstung, Papaemmanuil, and Campbell 2014) as part of the Jones group in-house pipeline against a germline panel of low-DNA input fat and dermis. This low-DNA input germline panel provided a mean sequencing coverage of at least 12,000X. Due to the relatively low coverage of these samples, I also called variants using the Cancer Variants through Expectation Maximisation (CaVEMan) algorithm (Jones et al. 2016) as a comparative method. For each skin sample, the corresponding donor's germline sample was provided as a matched reference and, to maximise sensitivity, CaVEMan copy-number inputs were set to a major copy number of 10 and a minor copy number of 2. Only variants which passed all of CaVEMan's post-processing filters were kept. I found that ShearwaterML called more mutations above a read depth of approximately 40X. This is to be expected, due to the increased sensitivity afforded by using a panel of germline samples. However, after filtering, CaVEMan called 3,465 more passed substitutions than the 6,224 passed by ShearwaterML. I found that 2,883 of these variants were called at loci where only reads of a single direction were present. In addition, CaVEMan only requires a depth of two reads for a mutation to be called. To conservatively avoid false positive calls, I only used mutations called by ShearwaterML with a variant allele fraction (VAF) of at least 0.2 to spatially map the clones present in the tissue of each donor.

Whole-genome sequencing (WGS)

I selected 46 of the 209 successful punch samples for WGS, to more accurately estimate genome-wide mutation burden, signatures and copy number aberration, if at least half of the sample was dominated by a single clone. WGS may also reveal additional genes not on the bait panel that may be under selection at this higher spatial resolution. DNA from these samples, along with the donor's matched germline sample, was whole-genome sequenced to a mean coverage of 33X (range = 23-55X). In addition, DNA from eight clonal 2 mm^2 grid

samples and matched germline samples from three donors was whole-genome sequenced to a mean coverage of 49X (range = 41-61X). Sequences were aligned to the human reference genome (NCBI build37) using BWA-MEM by Wellcome Sanger Sequencing Pipelines.

Whole-genome substitution calling

I called somatic single-base substitutions using the CaVEMan algorithm (Jones et al. 2016), as described above for the targeted sequencing. I merged two SNVs called at adjacent positions within the same sample to form a doublet-base substitution if 90% of the mapped DNA reads contained both SNVs. I removed mapping artefacts associated with BWA-MEM: the median alignment score of reads supporting a variant had to be at least 140 and the number of clipped reads equal to zero. In addition, the proportion of mutant reads present in the matched sample also had to be zero.

Whole-genome small indel calling

I called small (<200 bp) insertions and deletions against each donor's matched germline sample using cgpPindel (Raine et al. 2015). Only indels which passed all cgpPindel filters were kept. For the WGS punch samples only, I then filtered to remove a large excess of single base pair insertions at homopolymers of length five or more, an artefact likely caused by PCR amplification of low-input DNA concentrations during WGS. I classed indels as clonal if VAF >= 0.3.

Whole-genome copy-number estimates

I called somatic copy number aberrations in each WGS sample against each donor's matched germline sample using ascatNgs (Raine et al. 2016). This method compares the read depth and B-allele fraction at each SNP position in the sample and the germline in order to estimate the sample ploidy and purity, using a goodness-of-fit model.

Telomere length

I estimated telomere length using the Telomerecat software package (Farmery et al. 2018). The telomere length given is a median of all chromosomes for all cells in that sample.

Whole-genome phylogeny construction

Due to the limited number of clonal grid samples, I constructed phylogenetic trees for punch genomes only. Prior to phylogeny construction, I genotyped each substitution. For each sample in a donor, a pile-up of all quality reads was constructed and the number of mutant and wild-type reads counted for every locus that had a mutation called in any sample of that

donor. Only reads with a mapping quality of at least 30 and bases with a base quality of at least 25 were included. I gave mutations with a VAF less than 0.2 a genotype of 0 and those with a VAF greater than or equal to 0.3 a genotype of 1. For mutations with a VAF between 0.2 and 0.3, I set the genotype to NA (not applicable) for the purposes of phylogeny construction. I drew maximum parsimony trees for each donor after 1000 iterations with MPBoot (Hoang et al. 2018). Due to the large number of CC>TT substitutions present in sun-exposed skin, I drew phylogenies for each donor with single-base and double-base substitutions combined and with single-base substitutions only. The addition of double-base substitutions however, did not alter the structure of the tree, only the branch lengths. I then adjusted the branch lengths of the final phylogenies so that each double-base change counted as a single substitution. I then re-assigned substitutions to each branch of a phylogeny. I annotated each branch with exonic non-synonymous (missense, nonsense, essential splice, insertion or deletion) mutations in genes previously identified as being under positive selection by *dNdScv* in the targeted 2-mm^2 grid dataset. To reduce the number of false negatives, I only annotated phylogenies with a driver mutation in these genes if it was present in either the filtered whole-genome CaVEMan and Pindel calls, the targeted ShearwaterML calls or the targeted CaVEMan and Pindel calls with a VAF of at least 0.3. I then manually assigned copy number aberrations to each branch, where present.

Mutational signature assignment to branches

I called mutational signatures with each branch of a phylogeny treated as an independent sample. I first removed double-base substitutions before assigning SBS reference signatures with SigProfiler (L. B. Alexandrov et al. 2020b).

Results

Variation in clonal burden across 2-mm^2 grid samples

Normal skin from chronically and intermittently sun-exposed areas of the body was sampled from 35 UK donors across a range of ages between 26 and 79 years (**Fig. 2.2 and Table 3.1**). Skin samples were grouped into five body sites (head, forearm, leg, abdomen and trunk), based on evidence of differing keratinocyte tumour density at these sites (**Fig. 3.1,** (Subramaniam et al. 2017)). A total of 53,891 mutations (comprising 47,977 SBS, 3,824 DBS and 2,090 indels) were detected across the 25 cm^2 of epidermis sampled, with high variation noted in the number of mutations between samples of a donor (**Fig. 3.3**). In the case of tissue sampled across an ear lobe of donor PD37615, an area of high clonal density correlated with an area of darker pigmentation in the tissue, consistent with these samples having received a higher level of UV radiation (**Fig. 3.4A-B**).

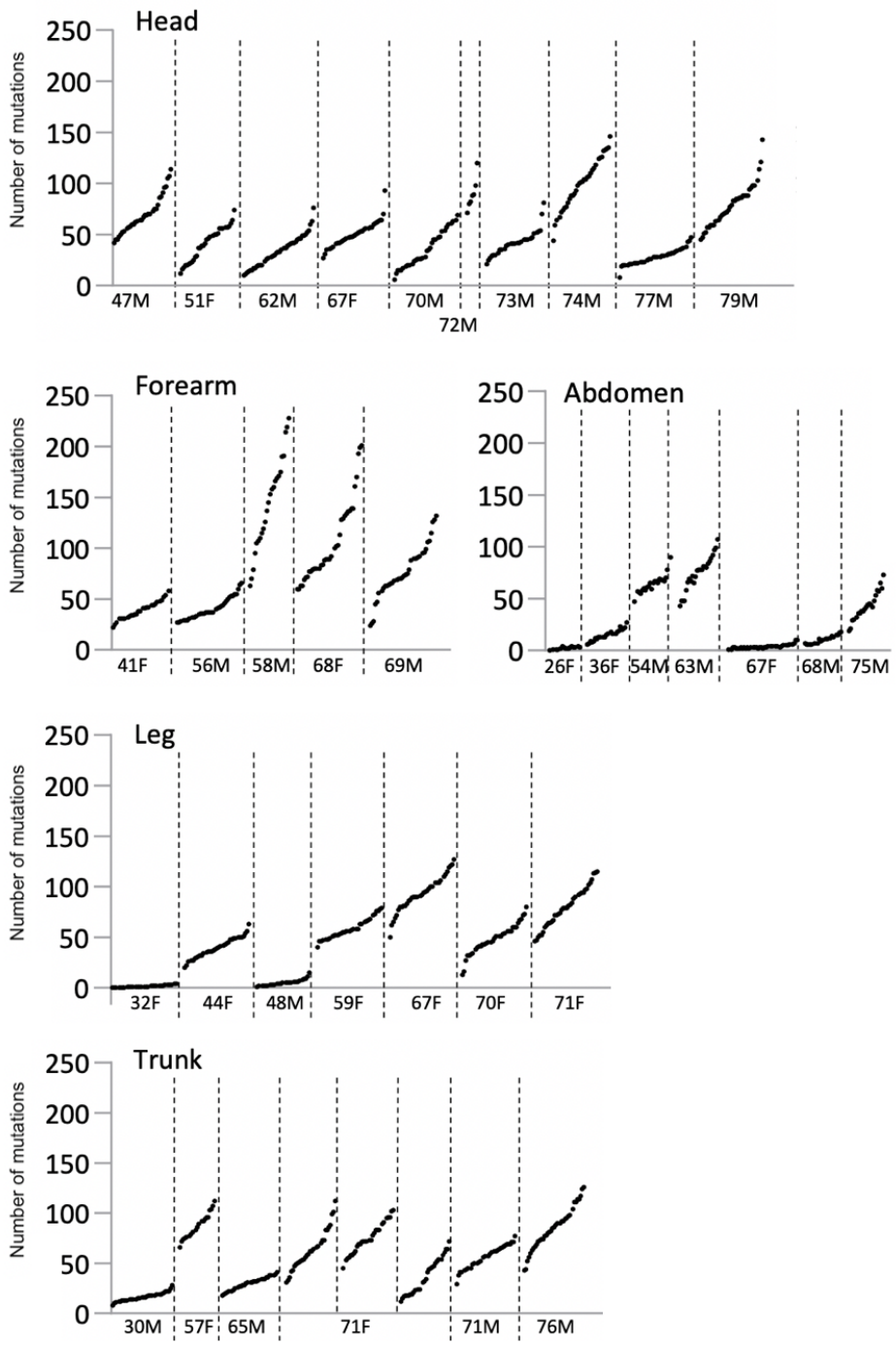


Figure 3.3: Distribution of the number of mutations detected per 2 mm^2 sample of epidermis (n = 1,251) across 74 genes per donor, adapted from (Fowler et al. 2021).

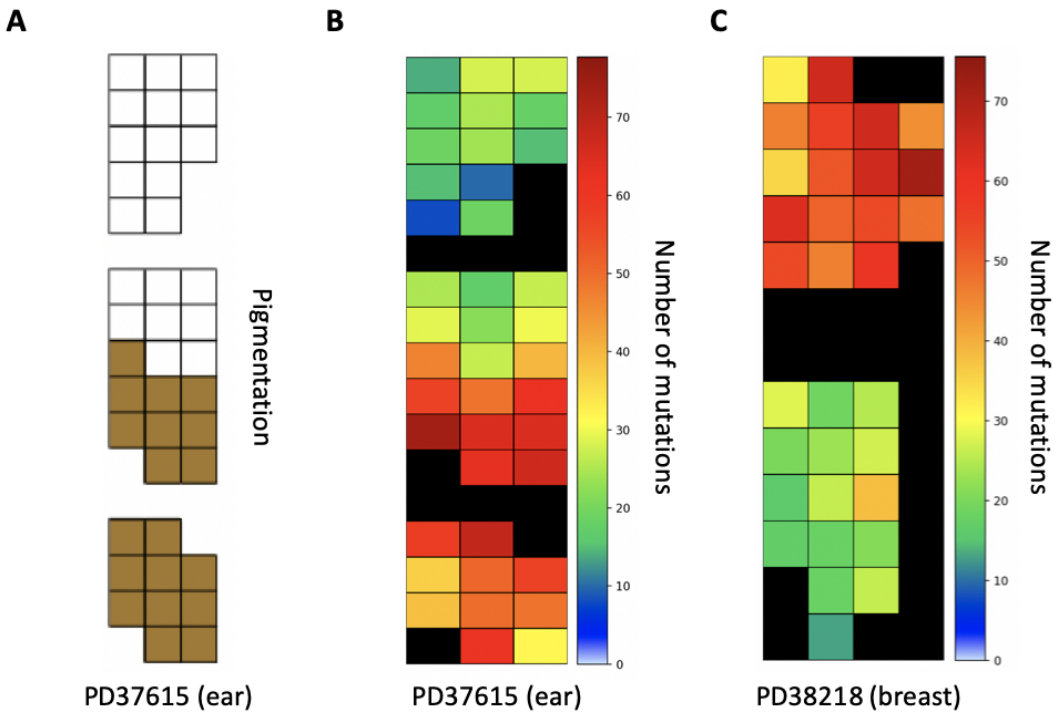


Figure 3.4: A Mapping of an area of higher pigmentation in the 40 2-mm^2 samples of tissue from donor PD37615 (70M - ear), drawn by J Fowler. **B** The number of mutations called per sample of PD37615. Each rectangle represents a 2-mm^2 sample, where black indicates tissue not sampled. An area of high clonal density correlated with the area of higher pigmentation in **A**. **C** The number of mutations called in 32 2-mm^2 samples of tissue from donor PD38218 (71F - breast). Here, clonal density correlates with two distinct areas of tissue that were sampled.

In donor PD38218, the clonal density of samples was found to cluster by two distinct sampling areas taken across breast epidermis (**Fig. 3.4C**). The difference in clonal density between the two areas in this donor may also stem from differences in chronic UV exposure.

Mutational signatures in 2-mm^2 grid samples

The majority of mutations observed were consistent with damage by UV light, with C>T (or G>A) substitutions contributing to 77% of all SBSs (**Fig. 3.5A**) and CC>TT (or GG>AA) substitutions contributing to 88% of all DBSs. Both C>T and CC>TT substitutions were more frequently found on the untranscribed, rather than the transcribed, strands of genes, compatible with the activity of transcription-coupled nucleotide excision repair (**Fig. 3.6**).

All SBS from each donor were combined and decomposed by non-negative matrix factorisation into their contributions to each reference signature (L. Alexandrov et al. 2018). Four donors with fewer than 200 mutations each (PD37182, PD37185, PD38335 and PD38336) were excluded from mutational signature analysis (**Fig. 3.5B**).

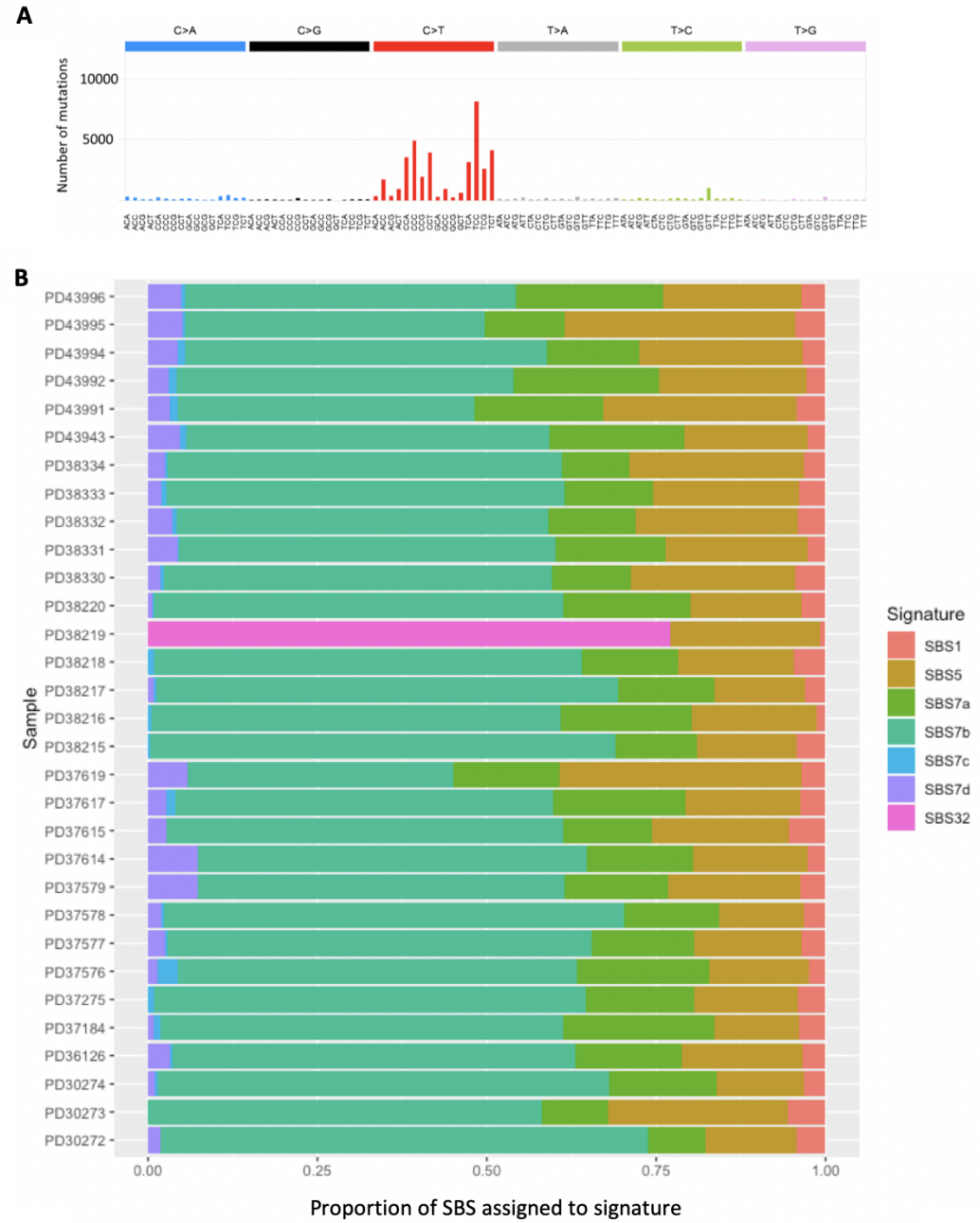


Figure 3.5: A Trinucleotide context of single-base substitutions (SBS) across all 2-mm^2 grid donors. **B** Proportion of SBS per donor attributed to each reference signature. Four donors were excluded due to too few mutations.

One donor, PD38219 (forearm), had 77% of their mutations attributed to reference signature SBS32. SBS32 was first identified in patients treated with azathioprine to induce immunosuppression (Inman et al. 2018). This signature predominantly consists of C>T mutations (75%) with a very strong transcriptional strand bias (Inman et al. 2018), opposite to the strand bias observed with UV damage (**Fig. 3.7**). This donor (58M) was a kidney transplant recipient and had been taking medication for immunosuppression since age 24.

From the age of 54 he developed multiple cSCCs and was diagnosed with acute lymphoblastic leukaemia at age 57, where he received non-cisplatin chemotherapy. Samples from this donor had the highest clonal density of all the 2-mm^2 grids sampled (**Fig. 3.3**). However, other than SBS32 and the ageing signatures SBS1 and SBS5, no other signatures were attributed to this donor. No SBS were attributed to damage by UV light, however, it is likely that SBS32 itself is contaminated slightly by the SBS7 (UV) signatures as SBS32 was first characterised in cSCCs (Inman et al. 2018). The majority (58%) of DBS in this donor are CC>TT substitutions, consistent with damage by UV light, however, they do not exhibit the transcriptional strand bias commonly observed with UV damage, suggesting a proportion of these DBS were caused by azathioprine treatment (**Fig. 3.7**).

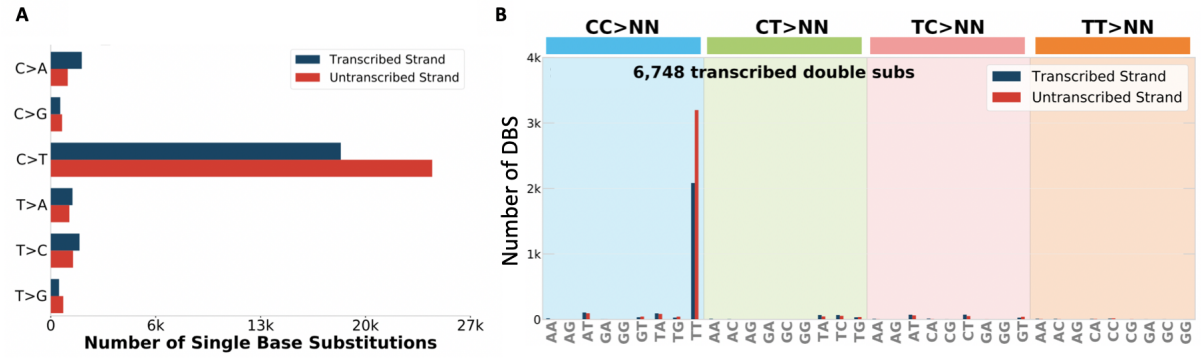


Figure 3.6: Transcriptional strand bias for single-base substitutions (**A**) and double-base substitutions (**B**) in mutations called from all donor 2-mm^2 grid samples.

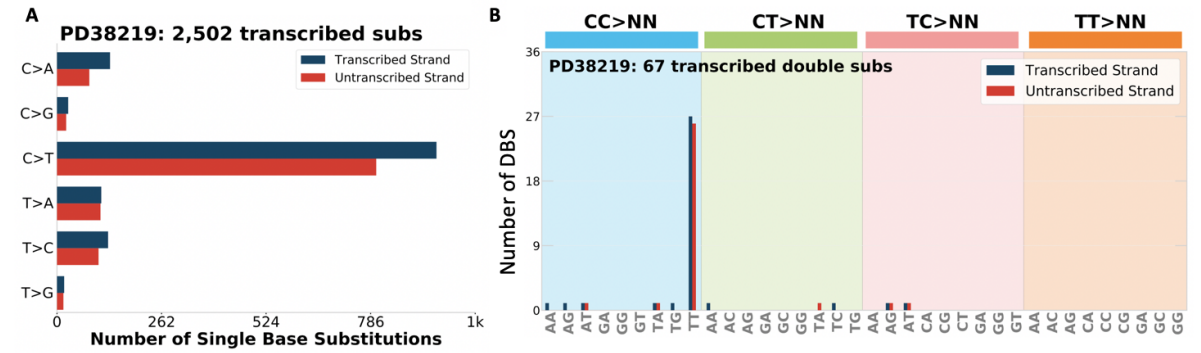


Figure 3.7: Transcriptional strand bias for single-base substitutions (**A**) and double-base substitutions (**B**) in mutations called from a donor, PD38219 (58M - forearm), who was taking medication for immunosuppression for 34 years.

Mutational signatures vary by body site

The SBS across the remaining donors were attributed to SBS7a-d, SBS1 and SBS5 (**Fig. 3.5B**). Signatures SBS7a-d are found in cancers of the skin and are likely to be due to

exposure to UV light (L. Alexandrov et al. 2018). It is hypothesised that SBS7a and SBS7b are each the consequence of the two major known UV photoproducts: cyclobutane pyrimidine dimers and 6-4 photoproducts (L. B. Alexandrov et al. 2020b). SBS7c is possibly the consequence of translesion DNA synthesis by error-prone polymerases inserting T, rather than A, opposite UV induced photodimers and SBS7d possibly the consequence of inserting G rather than A (L. B. Alexandrov et al. 2020b).

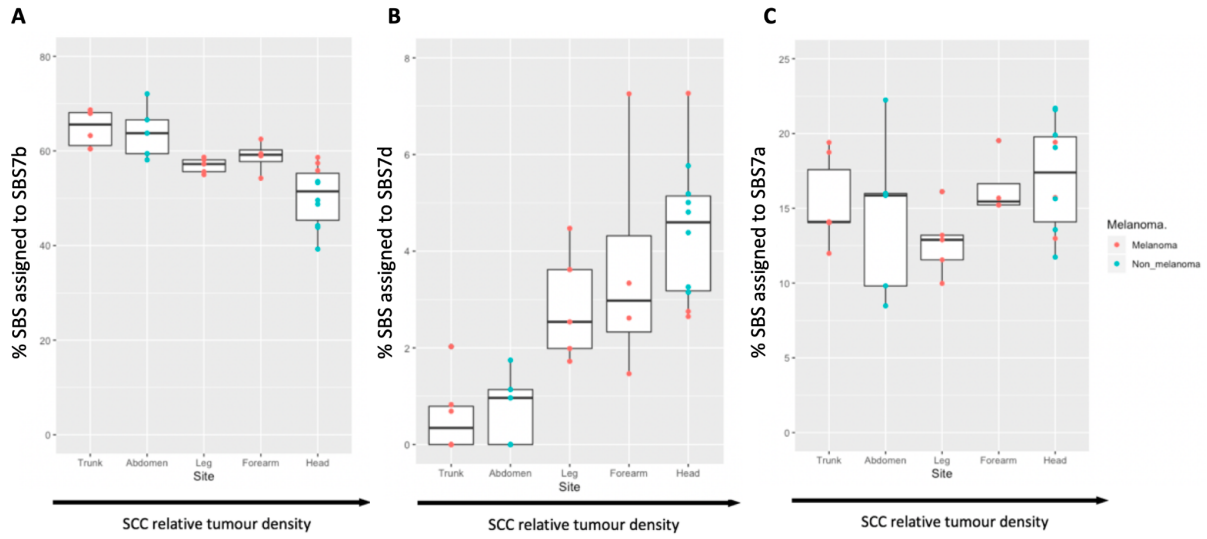


Fig 3.8: Percentage of SBSs assigned to each reference signature SBS7b (**A**), SBS7d (**B**) and SBS7a (**C**) per donor, by body site. Body site is ordered by relative cSCC density (trunk = 0.26, abdomen = 0.49, leg = 0.58, forearm = 2.85, head = 3.71). Colour indicates donors whose samples were obtained through wide-local excision of a melanoma (red), or those from donors with no history of melanoma (blue).

A one-way ANOVA shows the percentages of SBS7b and SBS7d to differ significantly with body site ($p = 4.29 \times 10^{-5}$ and $p = 5.92 \times 10^{-5}$ respectively). Using the classification of cSCC tumour density described by Subramaniam *et al.* (2017), I found a higher proportion of SBS7d mutations and a lower proportion of SBS7b mutations in sites with a higher cSCC tumour density (**Fig. 3.8A-B**; relative cSCC densities: trunk = 0.26, abdomen = 0.49, leg = 0.58, forearm = 2.85, head = 3.71). Tukey tests found percentage SBS7b to be significantly lower in head than in trunk ($q = 7.8 \times 10^{-5}$) and abdomen ($q = 3.8 \times 10^{-4}$), while percentage SBS7d was found to be significantly higher in forearm than in trunk ($q = 0.018$) and abdomen ($q = 0.038$) as well as significantly higher in head than in trunk (**Fig. 3.9**, $q = 1.7 \times 10^{-4}$) and abdomen ($q = 6.5 \times 10^{-4}$). The proportion of SBS7a mutations did not differ by site (**Fig. 3.8C**; one-way ANOVA, $p = 0.25$). In total, only 201 mutations were attributed to SBS7c, too few to assess differences by site.

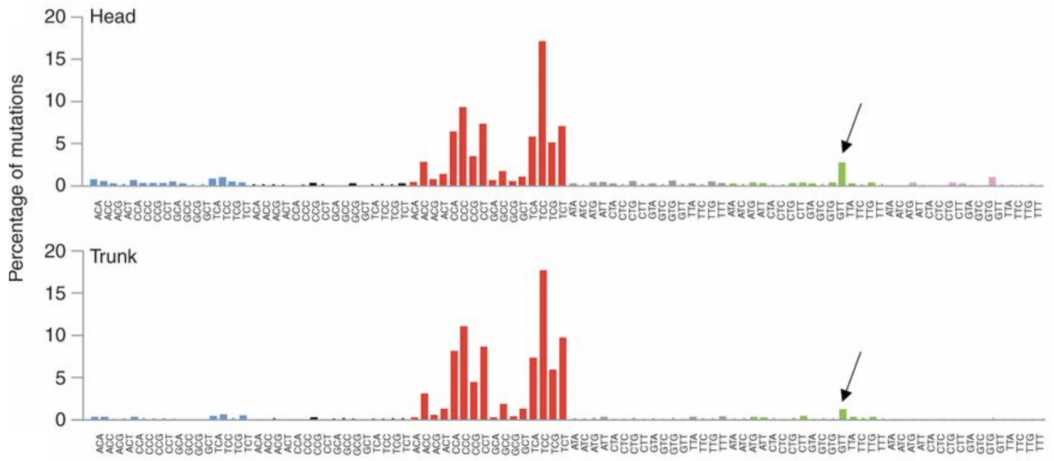


Figure 3.9: Combined mutational spectra of head and trunk samples, illustrating the significant reduction in the contribution of SBS7d (arrow) in trunk samples, adapted from (Fowler et al. 2021).

Factors influencing mutation burden

Independent synonymous mutations were used to estimate the genome-wide mutation burden per donor (**Fig. 3.10**). The highest estimated burden, 27.6 substitutions/Mb, was observed in the patient treated with azathioprine (PD38219 - forearm). This patient was a transplant recipient on immunosuppressant treatment for over 20 years, with a history of multiple cSCCs. Analysis of mutational signatures attributed 77% of mutations in PD38219 to the azathioprine signature (SBS32), whilst most of the mutations observed across all other donors were attributed to UV damage (SBS7a-d, **Fig. 3.5**). In these donors, it is therefore expected that genome-wide mutation burden is a reflection of the cumulative lifetime exposure of the tissue to UV radiation. A person's cumulative UV exposure, a principal risk factor for keratinocyte cancers in fair-skinned populations, is notoriously difficult to estimate. However, cumulative UV exposure may be expected to correlate with other measurable factors such as a donor's age, time working outdoors, time living in a country with a higher UV index and previous history of skin cancer (**Table 3.1**). Furthermore, keratinocyte cancer risk has been shown to differ by body site (Subramaniam et al. 2017).

I found there was large variation in estimated burden between donors of the same site relative to the variation across body sites, particularly for forearm donors, where there was a greater than 9-fold difference in estimated burden (**Fig. 3.10**). There was no significant difference in burden between samples obtained through wide local excision for melanoma (where the donor may be expected to be at higher risk of keratinocyte skin cancer) with samples obtained from donors with no previous history of melanoma, whether comparing all donors across the two groups or head samples alone.

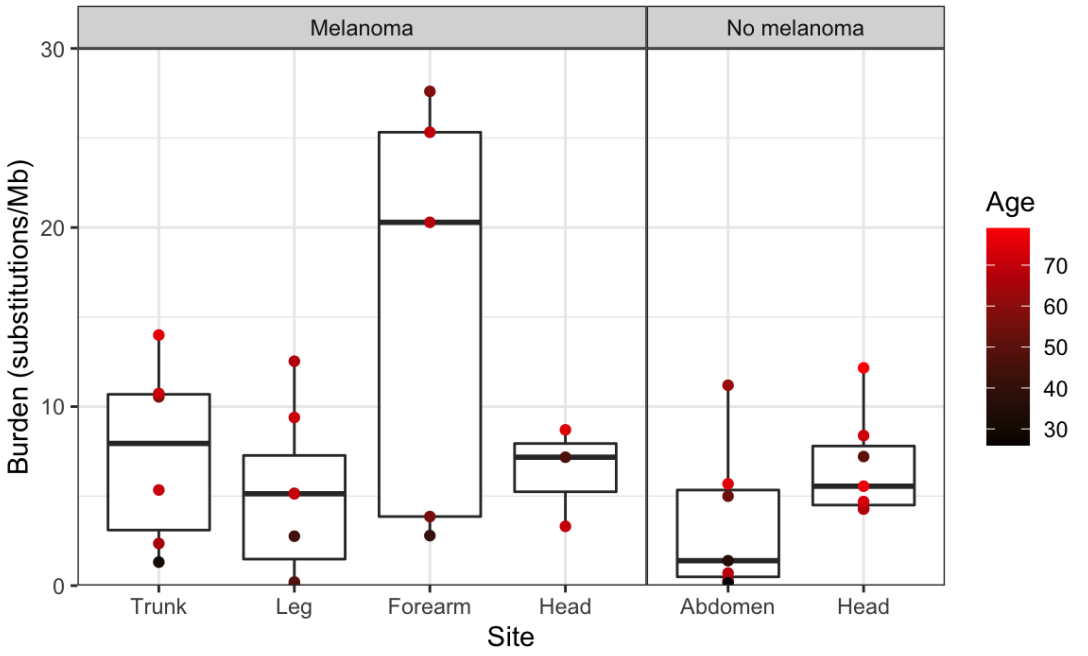


Figure 3.10: Boxplot of mutation burden (substitutions/Mb) per donor estimated using synonymous mutations only. Burdens are grouped by melanoma status of the donor and coloured by donor age (years). Body sites within groups are ordered by cSCC risk (or the density of observed cSCCs) as described by Subramaniam *et al.* (2017).

Due to the nature of sample collection, body site is confounded by donor melanoma status, for example, abdominal skin samples were all collected from organ transplant donors with no previous cancer history. Therefore, in order to investigate possible relationships between clinical factors associated with cumulative UV exposure (**Table 3.1**) on estimated burden, I grouped the data into five groups, based on body site and melanoma status of the donor: Trunk-melanoma (n = 6), Leg-melanoma (n = 7), Forearm-melanoma (n = 4), Abdomen-no_melanoma (n = 7) and Head-no_melanoma (n = 7). Due to low numbers, I removed three Head-melanoma samples (PD37614, PD37615 and PD37617), in addition to the azathioprine treated donor (PD38219), from this analysis.

Donor age is positively correlated with genome-wide burden (Pearson's r = 0.48, **Fig. 3.11**), however, the variance in burden also increases with age. For example, two forearm samples have a burden >20 substitutions/Mb and two abdominal samples have a burden <1 substitutions/Mb, despite these donors all being in their late 60s. An analysis of covariance (ANCOVA) identified a significant interaction between the effect of age on burden with body site ($p = 0.032$), specifically, that the effect of age on burden is accelerated in the forearm samples (**Fig. 3.11**). It should be noted that this forearm group is of small sample size (n = 4). A Johnson-Neyman analysis concluded that the effect of age is only significant for the

Forearm-melanoma and Leg-melanoma groups ($p < 0.01$ and $p = 0.03$, respectively). This is interesting as it suggests that in normal skin taken from melanoma patients, the effect of age on mutation burden is greatest at body sites with a higher cSCC risk (where forearm > leg > trunk, based on reported tumour densities). This is most likely due to more frequent sun exposure at these sites, consistent with age being a risk factor for keratinocyte cancers due to increased cumulative UV exposure. Age appears to have the least effect on burden at body sites taken from donors without a previous melanoma, however, greater sample sizes are necessary to confirm the findings discussed here.

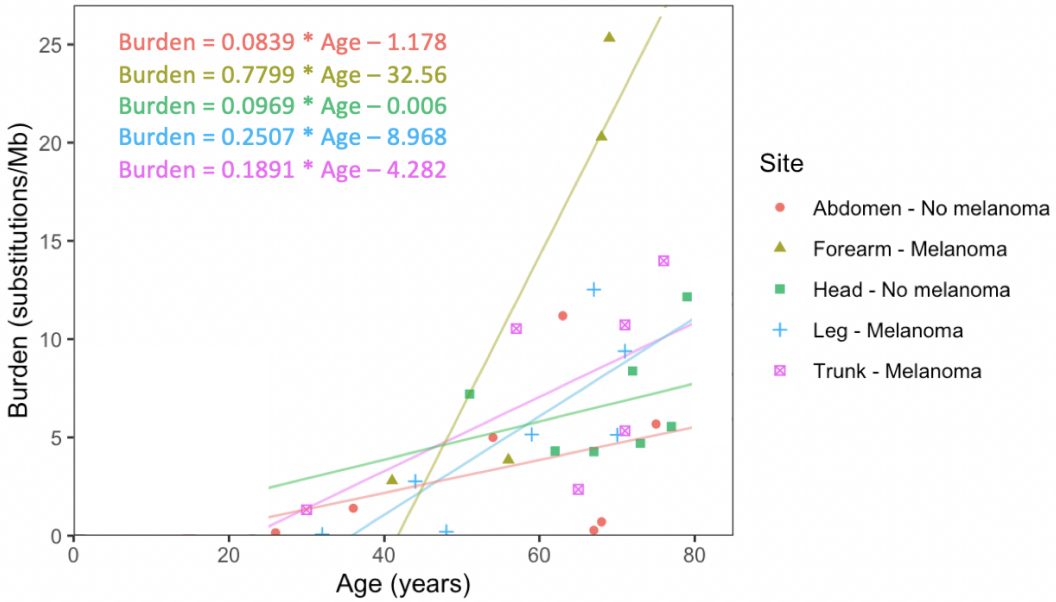


Figure 3.11: Estimated genome-wide burden using synonymous sites by donor age (All sites: Pearson's r = 0.48), coloured by body site and melanoma status of donor. Lines show linear regression of each group.

Unfortunately, due to the sparsity of data available on donor years working outside or living abroad (**Table 3.1**), the influence of such risk factors on mutation burden could not be determined.

Clone size and selection in 2-mm^2 grid samples

The mutational landscape of skin from the majority of donors comprised a large number of many small clones (median VAF = 0.015) but no significant difference by body site was observed (**Fig. 3.11**). 11 genes were found to be under positive selection, as previously reported: *NOTCH1*, *TP53*, *NOTCH2*, *FAT1*, *NOTCH3* and *RBM10* (Martincorena *et al.*, 2015) and novel drivers: *ARID2*, *AJUBA*, *KMT2D*, *TP63* and *RB1*.

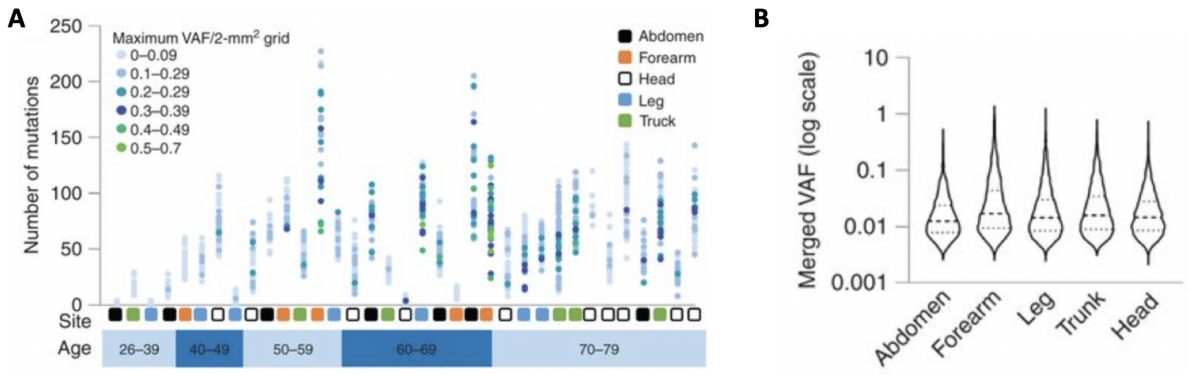


Figure 3.11: Taken from (Fowler et al. 2021). **A** Number of mutations called per 2-mm^2 grid sample by donor, ordered by donor age in years. Points are coloured by maximum VAF of mutations in that sample. **B** Violin plot of the log summed VAF distribution of all independent clones by body site. The median at each site is drawn as a thick dashed line.

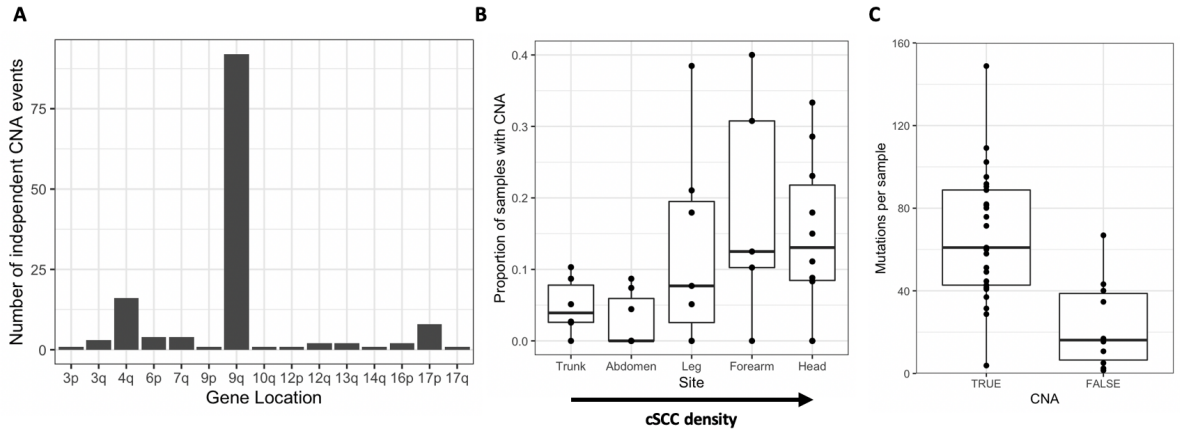


Figure 3.12: Copy number aberration as called by phasing of heterozygous SNPs from targeted sequencing in 2-mm^2 grid samples. **A** Number of independent CNA events across all samples by chromosome arm. **B** Proportion of samples of a donor with CNA detected, by body site, ordered by cSCC tumour density (Subramaniam *et al.*, 2017). **C** Mean number of mutations per sample for donors with (TRUE) and without (FALSE) CNA (Student's t test: p = 3.8x10^{-4}).

Copy number aberration in 2-mm^2 grid samples

I estimated allele frequencies using heterozygous SNP sites across the 74 targeted genes to call CNA across all 2-mm^2 grid samples. In total, 139 independent CNA events were detected, after merging those which spanned multiple samples (**Fig. 3.12A**). By far the most frequent event was loss of heterozygosity (LOH) at the *NOTCH1* locus on 9q (81 events, 58%). 11 of these LOH events on 9q extended to cover the *PTCH1* locus. *PTCH1*, as a gene in the Sonic Hedgehog pathway, is recurrently mutated in BCC (Bonilla et al. 2016). Other recurrent CNA included LOH at the *FAT1* locus on 4q (11 events, 8%) and LOH at the *TP53* locus on 17p (8 events, 6%). CNA was detected more frequently in samples of the leg,

forearm and head (body sites with a higher cSCC and BCC density) than the abdomen or trunk (**Fig. 3.12B**). In fact, there was a significant difference in the mean number of mutations per sample of a donor and the detection of CNA in that donor (**Fig. 3.12C**, Student's t test: $p = 3.8x10^{-4}$), suggesting clonal density and/or exposure to UV light increases the frequency of CNA persisting in the tissue.

Whole-genome sequencing of 2-mm^2 grid samples

Eight (0.4%) of the 2-mm^2 grid samples, across three donors, were found to be clonal and whole-genome sequenced. These large clones were found to harbour CNA and mutations at genes recurrently lost in keratinocyte cancers (**Fig. 3.13**).

Targeted sequencing of punch samples

Across six donors, 232 additional samples were taken with a circular punch, a 40th of the area of a 2 mm^2 grid. The SBS, DBS and indels called were of a similar mutational spectrum as that observed in 2-mm^2 samples, indicative of damage by UV light (**Fig. 3.14A** and **C**) and there was large variation in the number of mutations called across punches of the same donor (**Fig. 3.14B** and **Fig. 3.15A**). Analysis of selection by *d*N/*d*Scv identified *NOTCH1*, *FAT1*, *PPM1D*, *TP53*, *ASXL1* and *NOTCH2* as being under positive selection in this expanded gene set (**Fig. 3.14D** and **E**). Interestingly, *PPM1D* is a regulator of *TP53* under positive selection in normal oesophagus and mutations of the epigenetic regulator *ASLX1* are selected in squamous head and neck cancer. I used the targeted sequencing data of these punch samples to spatially map clones across the tissue (**Fig. 3.15B**).

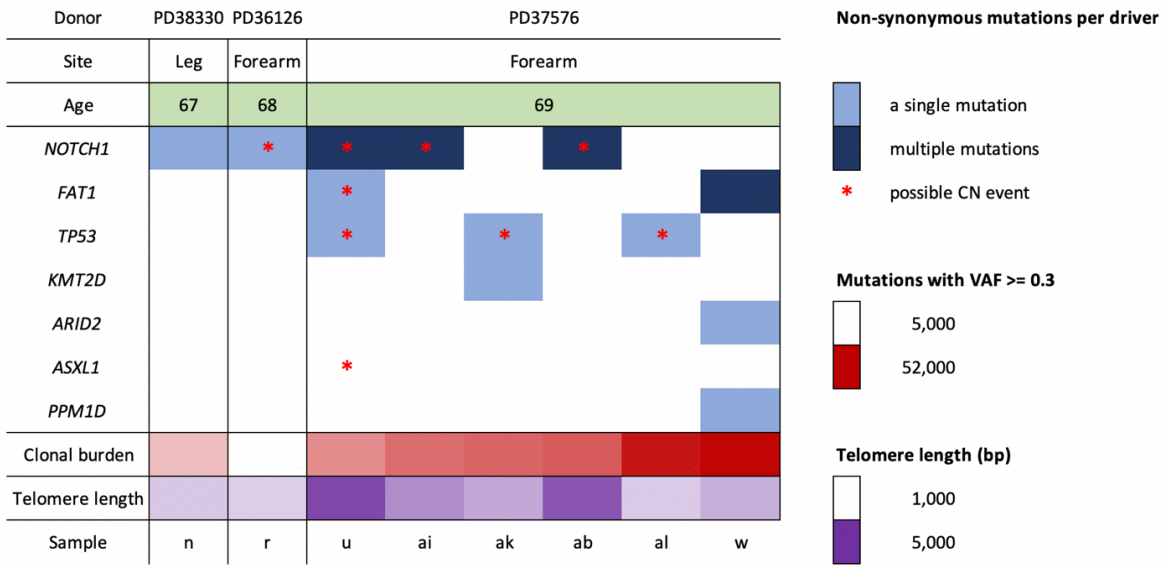


Figure 3.13: Summary of drivers, burden and telomere length across eight clonal 2-mm^2 WGS samples across three donors.

Whole-genome sequencing of punch samples

In order to better understand the genome-wide mutation burden, signatures, copy number aberrations and genes under selection that may not be present in the targeted bait sets, a sample of 46 of the 209 successful punch samples were submitted for whole-genome sequencing. Consistent with results from the targeted data, I identified *NOTCH1*, *FAT1*, *TP53* and *ASXL1* as being significantly ($q < 0.01$) positively selected drivers of clonal expansion through genome-wide analysis of non-synonymous to synonymous mutations per gene with *dNdScv* across all samples.

I constructed phylogenetic trees using shared clonal single-base and double-base substitutions across all whole-genome samples for each donor (see Methods, **Fig. 3.15C** and **Fig. 3.16**). Non-synonymous mutations in genes that I identified as positive drivers of clonal expansion (**Fig. 3.14D**) were added to branches of the tree in addition to copy number aberrations. Two samples with clonal non-synonymous mutations in *NOTCH1* displayed loss of heterozygosity at 9q, the *NOTCH1* locus. In addition, a clone spanning samples PD38217cj and bz had a clonal deletion of one whole allele of chromosome 12. This clone harboured a missense mutation in *KMT2D*, a positively selected driver of clonal expansion located on chromosome 12. In contrast, PD36126dy had an amplification for one whole allele of chromosome 12. Finally, one sample, PD38217dk, has a deletion of one allele at 17p, the *TP53* locus.

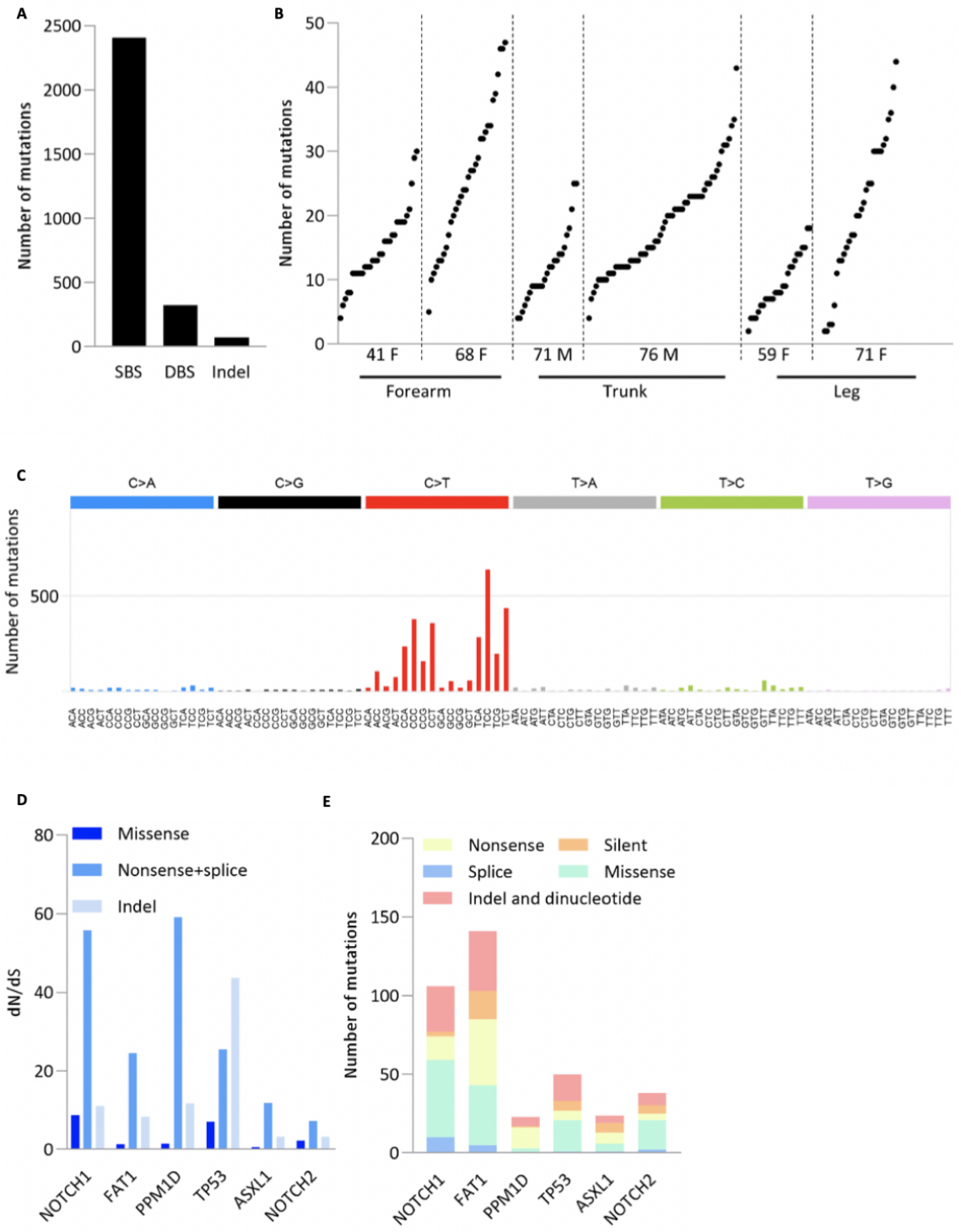


Figure 3.14: Adapted from (Fowler et al. 2021). Mutations called through sequencing of 324 genes across 209 punch samples. **A** The number of single-base substitutions, double-base substitutions and indels called across all samples. **B** The number of mutations called per sample per donor. **C** The spectrum of trinucleotide contexts for all SBS, consistent with damage by UV light. **D** *d*N/*d*S ratio for genes found to be under positive selection across all donors combined. **E** The mutation consequence for all mutations across the six positively selected genes.

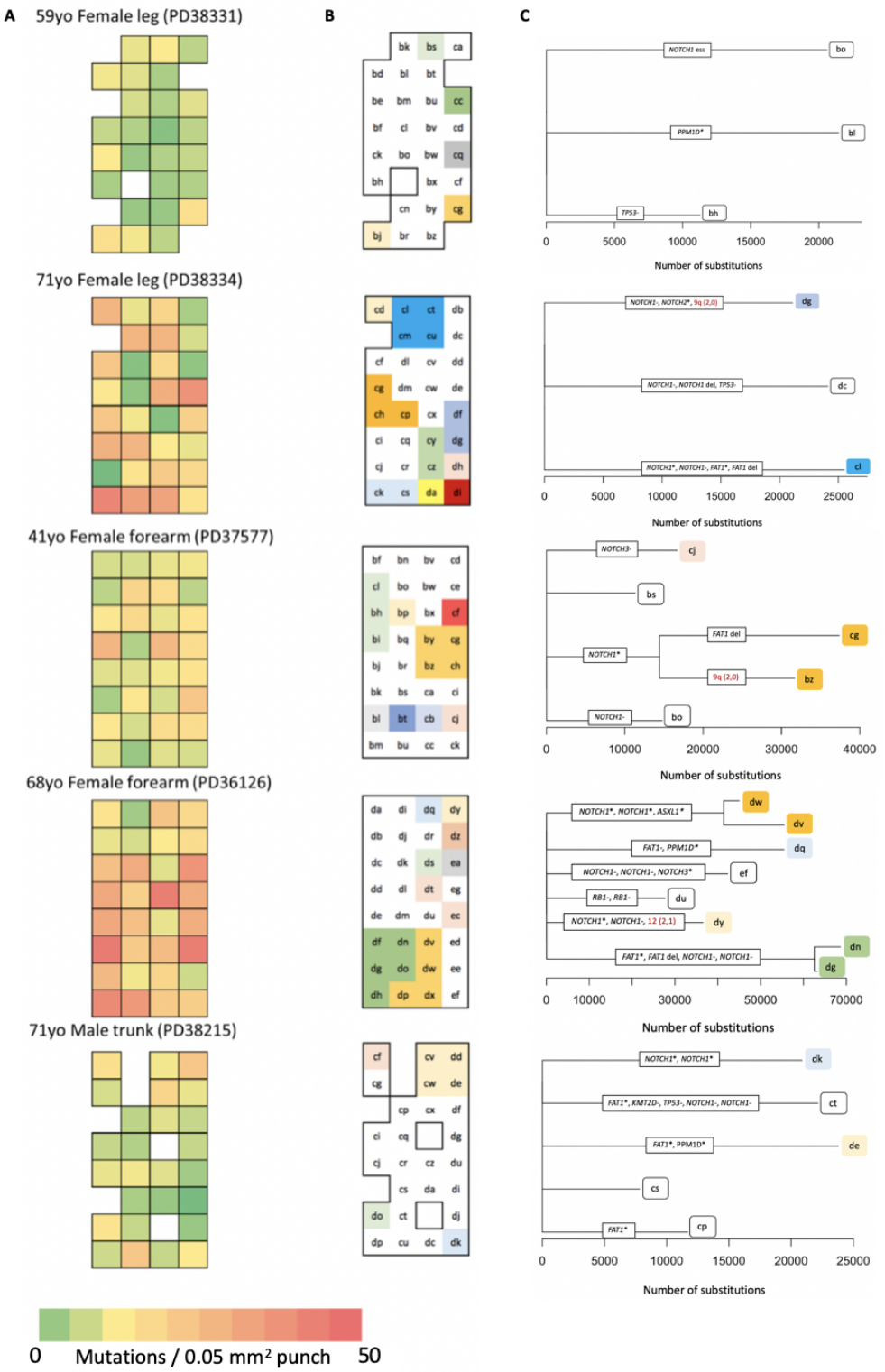


Figure 3.15: Clonal mapping in 0.05 mm^2 punch samples, across five donors. **A** Adapted from (Fowler et al. 2021). Number of mutations called across 324 genes per sample. **B** Mapping of clones (VAF >= 0.2) across the tissue identified through targeted sequencing. Letters indicate individual punch samples, with each colour denoting a separate clone. White is used for polyclonal samples. Failed samples have been removed from the map. **C** Phylogenetic trees drawn using WGS data, where branch length is equal to the number of shared substitutions (SBS + DBS) between samples (tips). Known drivers (* = nonsense, - = missense, ess = essential splice) and CNA (red) are labelled.

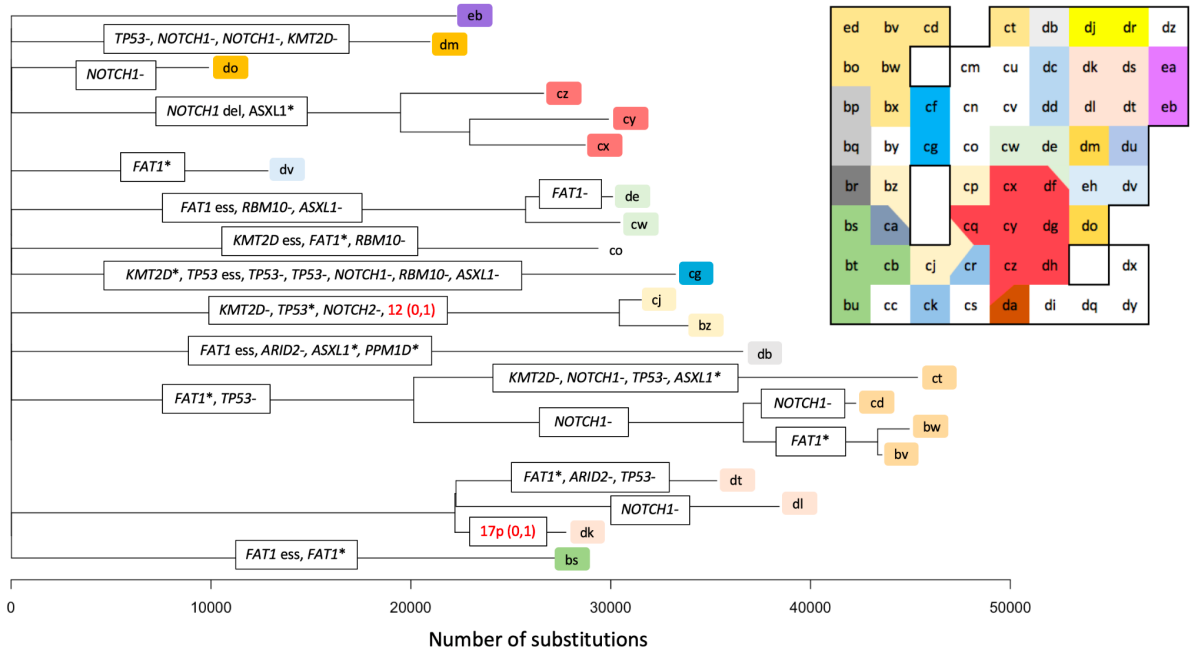


Figure 3.16: Clonal mapping (from targeted sequencing of 324 genes) and phylogenetic tree (from WGS) of 0.05-mm^2 punch samples across donor PD38217 (76M, shoulder), as in **Fig. 3.15**. Letters indicate individual punch samples, with each colour denoting a separate clone. White is used for polyclonal samples. Failed samples have been removed from the map. Phylogeny branch length is equal to the number of shared substitutions (SBS + DBS) between samples (tips). Known drivers (* = nonsense, - = missense, ess = essential splice) and CNA (red) are labelled.

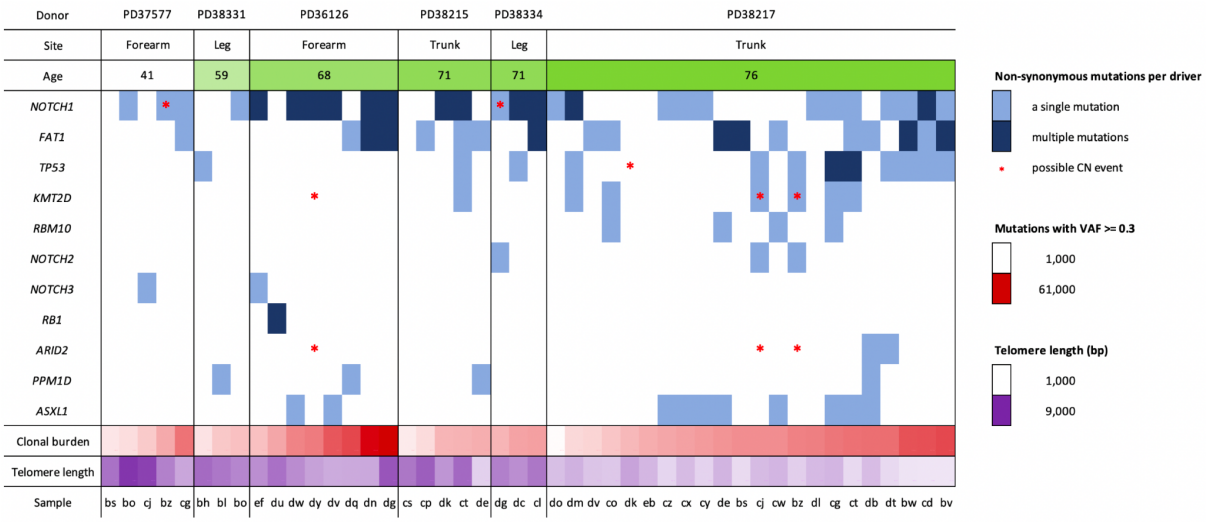
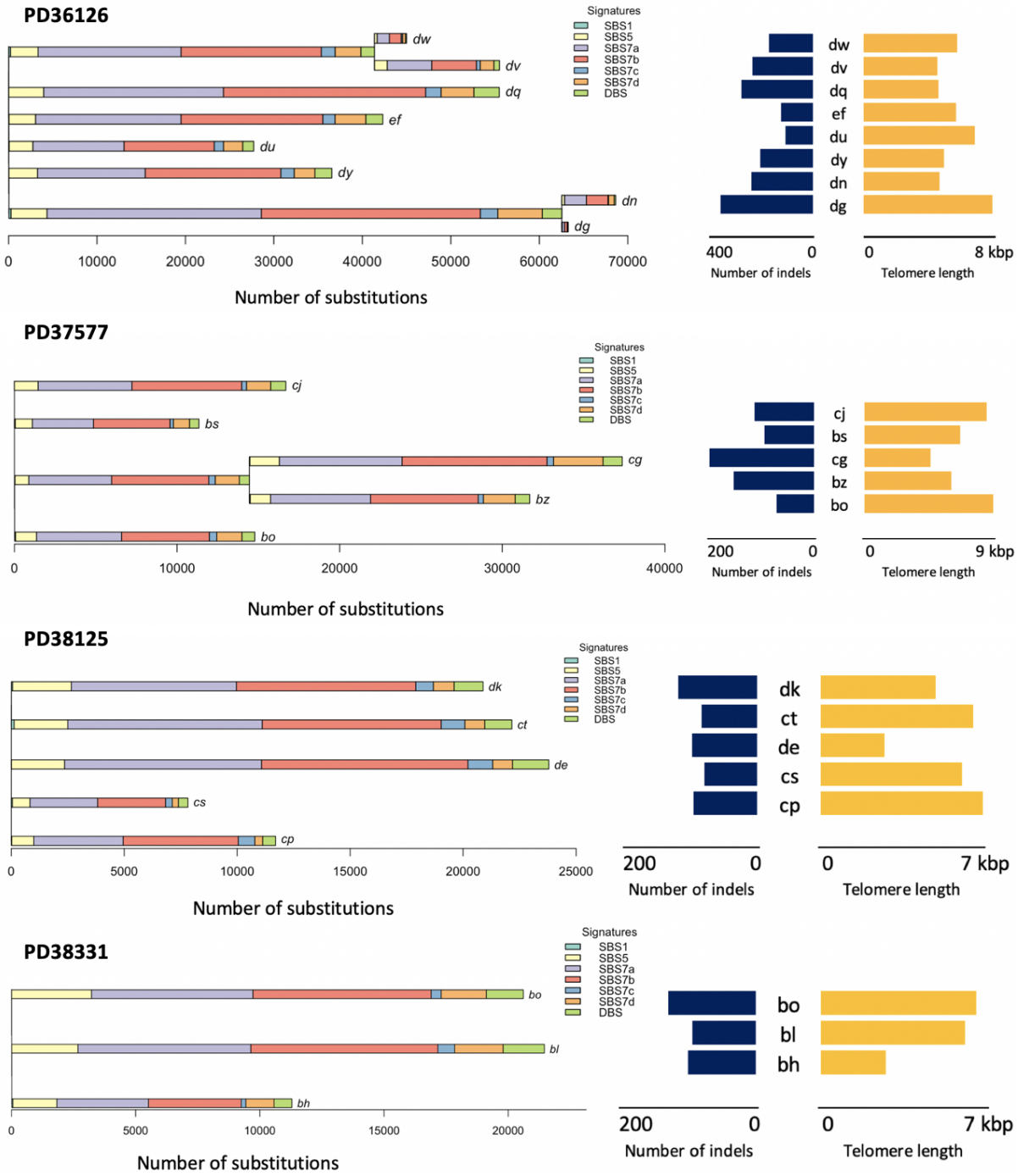


Figure 3.17: Summary of drivers, burden and telomere length across 46 clonal 0.05-mm^2 WGS samples across six donors.

Clonal driver mutations and copy number aberrations for all donors are summarised in Figure 3.17, along with clonal mutation burden and telomere length per sample.

I next wanted to assess if different mutational processes were operative in clones of the same donor. Each branch of the phylogenetic tree was treated as a 'sample' and mutational signatures were extracted using non-negative matrix factorisation. I plotted the proportion of clonal single-base substitutions assigned to each reference signature along each branch, in addition to the number of clonal double-base substitutions (**Fig. 3.18**).



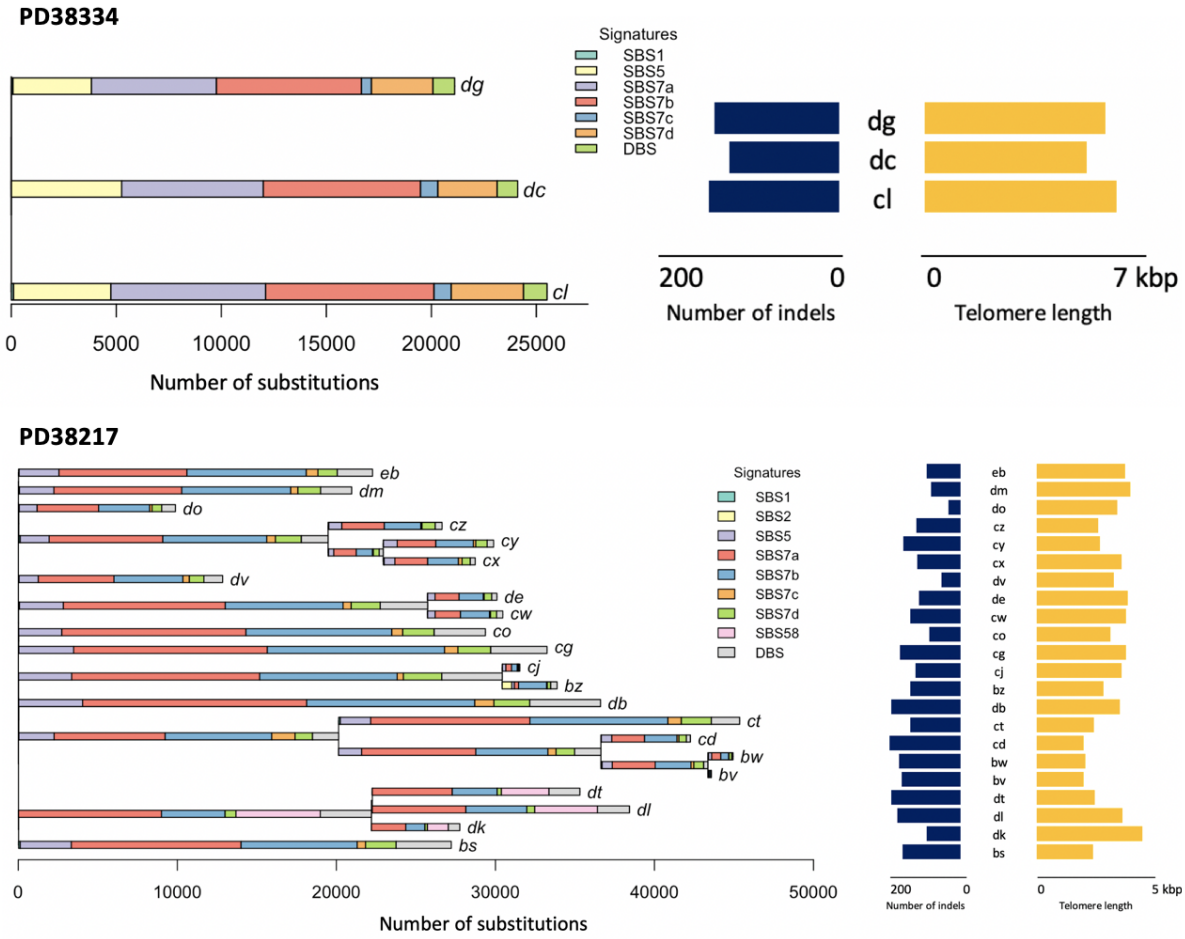


Figure 3.18: Phylogenetic trees drawn from 46 WGS 0.05-mm^2 punch samples across six donors. Contributions of reference SBS mutational signatures are plotted along branch lengths. Bar charts display the number of clonal (VAF >= 0.3) indels and the average telomere length for all chromosomes of all cells in a sample.

Across the majority of clones, SBSs were decomposed into six reference signatures: the ageing signatures SBS1 and SBS5 and the UV signatures SBS7a-d. The proportions of these signatures were consistent across early and late branches of a donor, suggesting no detectable difference in the effect of UV radiation in the skin over time. A single clone spanning samples PD38217dk, dl and dt had a mutational spectrum different to all other clones sampled (**Fig. 3.18** and **Fig. 3.19**). I estimated the spectrum of this clone to have a composition of 39% SBS7a, 21% SBS7b, 3% SBS7d, 24% SBS58 and 14% DBS. No mutations in this clone were assigned to either SBS5 or SBS7c, both ubiquitous signatures in all other punch samples. Instead, mutations were attributed to SBS58, an unknown signature, previously identified in 87 tumour samples and suggested as being a possible sequencing artefact (L. B. Alexandrov et al. 2020b). Here, this signature is unlikely to be a sequencing artefact as the three samples it covers are spatially adjacent (**Fig. 3.16**) and share thousands of substitutions attributed to UV damage. In addition, since each branch of

the phylogeny is treated as an independent sample for the signature analysis, I am confident this spectrum is a biological feature of this clone. I did not find any non-synonymous mutations in known DNA repair genes (Wood et al. 2001) private to this clone. Another sample, PD38217bz, had 17% of substitutions assigned to SBS2. SBS58, SBS2, SBS7a and SBS7b are all characterised by large peaks at T[C>T]N. In all samples at these peaks, I found a strong transcriptional strand bias, consistent with transcription-coupled repair of UV damage. I believe that small but real biological differences in UV damage or repair in these samples has led to a misassignment of mutations to SBS58 and SBS2.

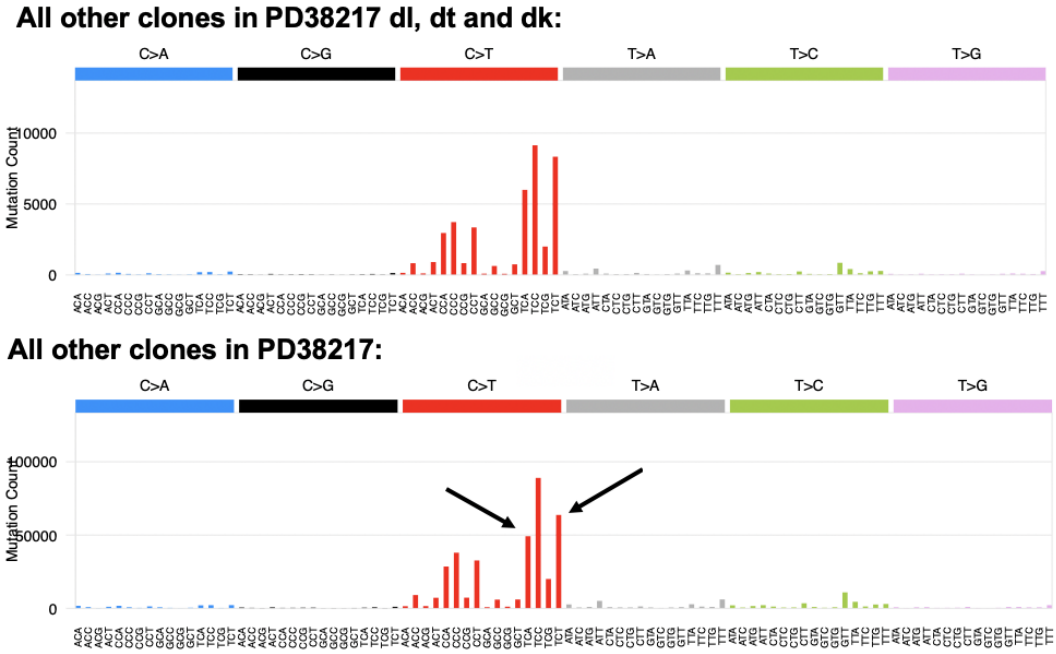


Figure 3.19: Adapted from (Fowler et al. 2021). Differences in spectrum of SBS trinucleotide context in samples dl, dt and dk of a clone in PD38217 compared to that of all other samples in this donor combined (indicated by arrows). (χ^2 = 397, df = 4, p = 0). Clone dt, dl, dk = 49,144 mutations; all other clones = 458,479 mutations.

All other branches showed very little variation in signature contribution with, on average, 36% of substitutions attributed to SBS7a, 35% to SBS7b, 11% to SBS5, 9% to SBS7d, 3% to SBS7c and less than 1% to SBS1. Overall, 6.4% of substitutions were doublets with 116,534 (90.0%) of these being a CC>TT substitution. This is consistent with DNA damage by UV light and has been classified previously in skin melanoma samples as signature DBS1 (L. Alexandrov et al. 2018). All samples, except for PD38331bh, showed CC>TT dinucleotides to be more frequently found on the untranscribed strand of genes, consistent with the activity of transcription-coupled repair of this UV damage. PD38331bh had a non-synonymous substitutions in *PRKDC* (a DNA repair gene involved in non-homologous end-joining) and

POLG2 (which interacts with mitochondrial DNA polymerase *POLG*)(Wood et al. 2001) but in no other DNA repair genes.

Analysis of the 10,769 indels detected reveals a proportion of indels attributed to ID8, suggested to be a consequence of repair of double-strand breaks by non-homologous end-joining, likely induced by UV radiation (L. Alexandrov et al. 2018). Surprisingly, I do not observe any evidence of ID13 in any samples, an indel signature common in melanoma, which is characterised by a deletion of a thymine at homopolymers of length two and is thought to be a consequence of UV damage due to its association with CC>TT mutations (L. Alexandrov et al. 2018).

Discussion

Overall, the trinucleotide spectra of mutations revealed the very high burden of mutant clones present in UK skin to be overwhelmingly caused by damage from UV radiation. Whole-genome sequencing showed that cells of physiologically normal epidermis are able to withstand a burden of over 60,000 substitutions, in addition to multiple copy number aberrations at loci of positively selected genes commonly mutated in cancers, whilst persisting long enough in the tissue to grow to a size large enough to be detected. Such copy number aberration was observed more frequently at body sites with a higher clonal density and those where cSCCs are more common (Subramaniam et al. 2017).

Large-scale clonal mapping of the epidermis across donors revealed a high clonal density, with largely unexplained variation between samples of the same donor. In one example, clonal density correlated with levels of pigmentation observed in the skin, providing an interesting opportunity to observe the effect on tissue local environment within the same donor and genetic background. In this case, the tissue was sampled across the ear lobe, from back to front. Levels of the light-absorbing pigment melanin in the skin increase in response to UV radiation (Coelho et al. 2009), suggesting these samples of increased pigmentation have received a higher exposure to UV radiation, resulting in this increased clonal density. In other cases, differences in the number of mutations detected per sample may partly be explained by the clonality of that sample and through the stochastic growth of mutant clones within the tissue (Murai et al. 2018); (Fowler et al. 2021). Estimates of genome-wide burden from targeted data using synonymous mutations also identified greater variability within donors of a body site than between them. However, it is likely that these estimates are unreliable due to the very high clone density observed. It is likely there are

many clones residing in the tissue that are of too small a size to be detected, despite the depth of sequencing used here.

One donor, treated with azathioprine, was found to have a mutational signature and transcriptional strand bias distinct from other donors and the highest clonal density, with some 2-mm^2 samples harbouring over 200 independent clones. However, the clones observed in this donor were not as large as those of some other donors suggesting the tissue is becoming saturated with mutant clones, similar to that observed in Chapter 2. High levels of saturation in the tissue may explain why there was no strong correlation observed between clone size and mutation burden or body site.

The large area of skin sampled in this chapter enabled the identification of a difference in UV signature by body site. The contribution of SBS7d was higher at body sites with a higher cSCC density, being highest in samples of the head. This may be a reflection of a difference in repair of UV damage at more continuously UV light exposed skin on the face and forearms, in comparison to more intermittent exposure at sites such as the trunk and abdomen. Finally, the analysis of phylogenetic trees did not detect a difference in UV signature over time. However, in some cases, there was variation in mutational signatures and both mutation burden and telomere length varied as much as two-fold between clones growing less than a millimetre apart within the tissue of the same donor.

Chapter 4: The Effect of Cancer Risk Factors on the Mutational Landscape of Aged Oesophagus

Introduction

To date, there have been two major studies on the mutational landscape of normal human esophageal epithelium. The first, in 2018, used deep targeted sequencing of 857 2-mm^2 epithelial samples across nine UK donors to estimate mutation burdens ranging from several hundred to several thousand mutations per cell, with both this burden and the size of mutant clones increasing with the age of the donor (Martincorena, Fowler, et al. 2018). 14 genes were found to be under positive selection in the tissue (*NOTCH1, TP53, NOTCH2, FAT1, NOTCH3, ARID1A, KMT2D, CUL3, AJUBA, PIK3CA, ARID2, TP63, NFE2L2* and *CCND1*), with more than half the tissue being colonised by mutant clones, predominantly in *NOTCH1*, by middle-age (**Fig. 4.1**).

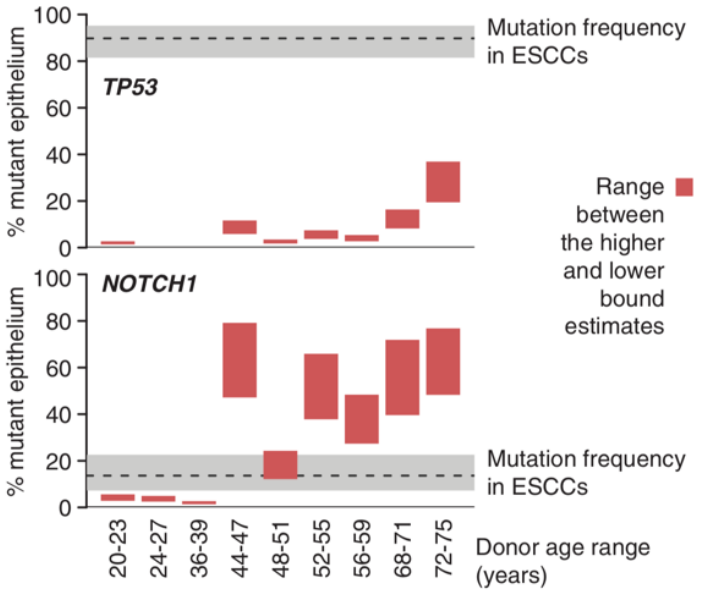


Figure 4.1: Figure 2G from Martincorena, Fowler, *et al.* (2018) which highlights the dominance of *NOTCH1* mutations in oesophageal epithelium by middle-age. In contrast, the percentage of tissue mutant for *TP53* remains at a much lower frequency than that observed in ESCCs. However, in the oldest donor, this proportion of *TP53* mutant tissue was estimated to be as much as 37%.

In the oldest donor of this study, a 73 year old non-smoker, up to 37% of the epithelium sampled was mutant for *TP53*. In ESCCs, *TP53* mutations are almost ubiquitous, suggesting they are an early step required for oncogenesis. In contrast, *NOTCH1* mutations appear to

be more prevalent in normal epithelium than in ESCCs, raising questions about the role of *NOTCH1* in the development of cancers. Due to the small number of older donors sampled in this study, it is unclear whether colonisation of the tissue by mutant *TP53* clones is a normal process of ageing (helping to explain the increased incidence of ESCCs with age), or if this one donor is an anomaly. Furthermore, large differences in mutant density, clone size and driver frequency were detected across donors, however, due to the small sample size it was not possible to correlate these with known clinical factors.

The second major study of histologically normal oesophageal epithelium was that of Yokoyama *et al.* in 2019. In addition to samples taken from ESCC patients, Yokoyama *et al.* undertook whole-exome sequencing of 106 samples of varying sizes (ranging from 0.2 mm^2 to 4 mm^2) taken by endoscopic biopsy from 40 healthy Asian donors (of which, eight were heavy smokers and/or drinkers). This study was the first to detect mutations in normal oesophageal epithelium in the trinucleotide context associated with APOBEC activity - a mutational signature seen at a high frequency in ESCCs. These APOBEC mutations were more frequently observed in samples from heavy smokers and drinkers (Yokoyama et al. 2019). Mutations in the context of SBS16 (associated with alcohol consumption) were also more frequently observed in heavy drinkers and smokers. Finally, Yokoyama *et al.* argue that the expansion of mutant clones, particularly for *NOTCH1*, is substantially accelerated by donor alcohol consumption and smoking. In addition, this study identified mutant *PPM1D* as a driver of clonal expansion. However, despite mutations in *PPM1D* being more frequent with heavy drinking and smoking, they are over-represented in normal epithelium compared to ESCCs (Yokoyama et al. 2019). In fact, the observation that both mutant *PPM1D* and *NOTCH1* clone sizes are increased with heavy drinking and smoking, yet these mutations are under-represented in normal tissue compared to cancers, adds confusion to the mechanism by which heavy smoking and alcohol act as risk factors for ESCC. However, the small number and inconsistent area of samples taken per donor and the relatively low depth of sequencing in this study potentially precludes accurate estimation of clone sizes.

In this chapter, I expand on the work of Martincorena, Fowler *et al.*, (2018), using their method of deep targeted sequencing of a contiguous array of 2-mm^2 grids to analyse the mutational landscape of 19 additional donors (**Fig. 4.2**). I aim to characterise the landscape of aged oesophageal epithelium and determine if the proportion of tissue mutant for *TP53* observed in one 73 year old donor is typical of normal ageing of the oesophagus. Furthermore, I aim to measure the level of intra- and inter- donor variation in the mutational landscape and link this variation, where possible, to cancer risk factors.

I use Nanorate sequencing (Nano-seq), a method which allows the calling of mutations from a single molecule, in order to obtain an accurate mutation burden per donor without the potential bias of targeted data and with a lower error-rate than whole-genome sequencing (Abascal et al. 2021). Finally, as in Chapter 3, I take an array of punch biopsies (a 40th the area of a 2-mm^2 grid sample) from seven donors, in order to map aged tissue at a higher spatial resolution. 173 of these punch samples found to be clonal through targeted sequencing are then whole-genome sequenced, to allow the drawing of phylogenetic trees and the analysis of mutational signatures over time in selected clones.

Materials & Methods

Sample Collection (2-mm^2 grids) - completed by J Fowler

Sample collection was carried out as described in (Martincorena, Fowler, et al. 2018). Mid-oesophagus was excised within 60 minutes of circulatory arrest from 20 deceased organ donors over 60 years of age at Addenbrooke's Hospital (Cambridge, UK). Informed consent was obtained in all cases from donor families. Organ tissue was preserved in University of Wisconsin organ preservation solution until processing. Each oesophagus was cut open longitudinally and the muscle and submucosa removed. The remaining tissue was cut into approximately 0.25 cm^2 pieces and each piece was incubated in 20 mmol/L EDTA for two hours at 37°C. Epithelium was then peeled from the remaining submucosa using fine forceps under a dissecting microscope and fixed for 30 minutes with 4% paraformaldehyde (PFA; FD Neurotechnologies) before being washed three times in 1x PBS.

The fixed epithelium was cut into a contiguous array of approximately 80 samples per donor, each measuring 2 x 1 mm. DNA was extracted from each sample using the QIAamp micro DNA extraction kit (Qiagen) by digesting overnight and following the manufacturer's instructions. DNA was eluted using pre-warmed AE buffer where the first eluent was passed through the column twice more. Germline samples were collected from flash frozen oesophageal muscle and DNA was extracted as for epithelial samples.

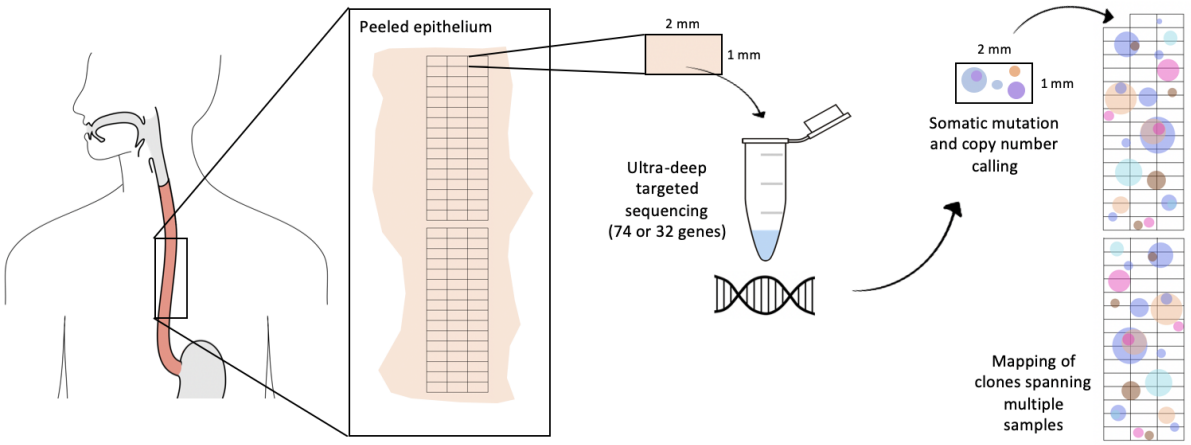


Figure 4.2: Sampling method to allow mapping of clones which span multiple samples of aged oesophagus.

DNA Sequencing (2 mm^2 grids) - completed by Wellcome Sanger Sequencing Pipelines

For eight donors, deep (~700x), targeted sequencing was performed across all samples using the 74-gene bait panel used in Chapter 2 and 3 of this thesis and described by Martincorena *et al.* (2015). The remaining 12 donors were sequenced using a reduced bait set of 32 genes commonly mutated in squamous cell carcinoma and normal squamous epithelium (ADAM10, AJUBA, ARID1A, ARID2, CCND1, CUL3, DICER1, ERBB4, FAT1, FAT4, FBXW7, HRAS, KMT2C, KMT2D, KRAS, NFE2L2, NF1, NOTCH1-3, NRAS, NSD1, PIK3CA, PREX2, PTCH1, PTEN, RB1, SMAD4, SMO, TP53, TP63 & ZNF750). Samples were multiplexed and sequenced on HiSeq 2000 machines (Illumina) with version 4 chemistry to generate 75 bp paired-end reads. BAM files were mapped to the GRCh37d5 reference using BWA-MEM (Li and Durbin 2009). Duplicate reads were marked using Biobambam2 (https://gitlab.com/german.tischler/biobambam2). I found all samples from one donor sequenced with the 32-gene bait to be affected by a C>A sequencing artefact and I therefore removed them. In addition, BAM files from samples of the nine donors published in Martincorena, Fowler *et al.* (2018) were included in downstream analysis (**Table 4.1**).

Mutation Calling

I called mutations in the oesophageal epithelial samples sequenced with the 74-gene bait panel using ShearwaterML (Gerstung, Papaemmanuil, and Campbell 2014) against an unmatched normal panel of muscle from 47 donors, sequenced with the same targeted panel. Epithelial samples sequenced with the 32-gene bait panel had mutations called using ShearwaterML against an unmatched normal panel made up of bulk muscle samples from the remaining 32-gene sequenced donors. I then used the matched muscle sample from

each donor to remove germline mutations in both cohorts. After variant calling, filters were applied and mutations annotated as described in Chapter 2.

Donor	Published	Bait Panel	Age	Sex	Alcohol (units/day)	Smoker	No. 2 mm^2 grid samples	No. punch samples
PD30272	Yes	74 gene	36	Female	1-2	No	96	-
PD30273	Yes	74 gene	68	Male	<1	Ex	95	-
PD30274	Yes	74 gene	54	Male	<1	No	98	-
PD30986	Yes	74 gene	45	Male	<1	Yes	94	-
PD30987	Yes	74 gene	50	Male	1-2	Yes	95	-
PD30988	Yes	74 gene	57	Female	0	Yes	95	-
PD31182	Yes	74 gene	75	Male	<1	No	94	-
PD36712	Yes	74 gene	26	Female	<1	Yes	95	-
PD36806	Yes	74 gene	22	Male	2-3	Yes	95	-
PD37275	No	74 gene	75	Male	1-2	Ex	79	-
PD39463	No	74 gene	73	Male	0	No	79	-
PD40289	No	74 gene	77	Male	1-2	Yes	79	-
PD40290	No	74 gene	77	Female	3-6	Ex	79	-
PD40291	No	74 gene	77	Female	<1	No	79	-
PD40292	No	74 gene	69	Female	<1	No	79	-
PD41049	No	-	66	Male	<1	No	NanoSeq only	60
PD41063	No	74 gene	66	Male	3-6	Ex	79	-
PD41064	No	74 gene	73	Female	1-2	Ex	69	-
PD42785	No	32 gene	71	Female	1-2	No	79	62
PD42786	No	32 gene	61	Male	<1	No	55	-
PD43382	No	-	45	Female	0	Yes	NanoSeq only	-
PD43383	No	32 gene	70	Male	<1	Ex	79	89
PD43384	No	32 gene	75	Female	1-2	No	79	61
PD44715	No	32 gene	70	Male	3-6	No	79	62
PD44716	No	32 gene	63	Male	>9	Yes	79	63
PD44717	No	32 gene	61	Female	0	No	79	58
PD47560	No	32 gene	66	Male	3-6	Yes	79	-
PD48811	No	32 gene	64	Male	1-2	Yes	77	-
PD48812	No	32 gene	61	Female	<1	Ex	78	-
PD48813	No	32 gene	78	Female	<1	Ex	79	-
PD48814	No	-	37	Male	3-6	Ex	NanoSeq only	-

Table 4.1: Demographics of nine donors published in Martincorena, Fowler *et al.* (2018) and 22 newly sequenced donors.

Analysis of targeted 2 mm^2 grid samples

Mutations were merged across adjacent grid samples and downstream analysis of clone size, percentage mutant epithelium and selection were conducted as described in Chapter 2.

Nano-seq of targeted 2 mm^2 grid samples

Sequencing a cohort of 2 mm^2 grid samples with a reduced bait panel precludes an accurate comparison of mutation burden across all donors. Furthermore, the number of mutations called across 32 genes is, in some cases, too low to allow reliable decomposition of

substitutions into mutational signatures. In order to overcome both of these limitations, one sample from each donor was sequenced using Nano-seq (Abascal et al. 2021), a method of single DNA molecule sequencing with an error rate less than five per billion base pairs. In addition to the very low error rate achieved, another benefit of calling mutations from a single molecule of DNA is that the mutations called do not depend on the clonality of the sample. After middle-age, it has been shown that a cell in normal human oesophagus is more likely to be mutant for *NOTCH1* than not (Martincorena, Fowler, et al. 2018). In light of this, for each donor, I chose DNA from a 2 mm^2 grid sample known to harbour a *NOTCH1* mutant clone with a VAF of at least 0.3 to be submitted for Nano-seq. Nano-seq was performed using 0.3 fmol DNA in order to achieve a 0.5-0.7x genome wide coverage. 14 PCR cycles were performed and germline mutations were filtered using standard whole genome sequencing from a matched germline sample. Mutational signature decomposition and *de novo* analysis of Nano-Seq data was performed using MutationalPatterns (Blokzijl et al. 2018).

Targeted punch samples

Across seven donors, I took 455 samples from the peeled oesophageal epithelia with a circular biopsy punch (0.25 mm diameter, Stoelting Europe), each a 40^{th} of the area of a 2 mm^2 grid. Within each donor, I sampled the tissue as a rectangular array of punches, with approximately a punch diameter of unsampled space between each (**Fig. 3.2**). In addition, I took a punch sample of muscle from each donor as a germline sample. For both epithelium and muscle samples, I extracted DNA using the Arcturus picopure kit following manufacturer's instructions (ThermoFisher). DNA sequencing and all downstream analysis was performed as described for targeted punch samples in Chapter 3.

Whole-genome sequencing (WGS)

I selected 173 clonal punch samples for WGS, to more accurately estimate genome-wide mutation burden, signatures and copy number aberration. Samples were defined as clonal if at least half of the sample was dominated by a single clone. WGS may also reveal additional genes not on the bait panel that may be under selection at this higher spatial resolution. DNA from these samples, along with the donor's matched germline sample, was whole-genome sequenced to ~20X coverage on NovaSeq machines by Wellcome Sanger Sequencing Pipelines. All downstream analysis of whole-genome sequencing was performed as described in Chapter 3.

Results

Nano-seq: substitution burden increases with age and smoking status

Estimates of mutation burden across all 31 donors showed a positive linear relationship with age through the ages of 20-80 years old (**Fig. 4.3A**). Linear regression showed mutation burden (Mutations/Mb) was equal to 0.005 per year of age + 0.122. Estimates of burden obtained through Nano-seq were lower than those estimated through the targeted sequencing of nine donors (**Fig. 4.3B**) reported in Martincorena, Fowler *et al.* (2018). This may be because Nano-seq has a lower error rate than the method used to call mutations in targeted sequencing. However, the burdens reported in Martincorena, Fowler *et al.* (2018) are likely to be an overestimate, despite using synonymous mutations only, since the bait regions targeted cover a high proportion of genes which drive clonal expansions, are likely to be transcriptionally active, in areas of open chromatin and consequently more susceptible to sustained DNA damage.

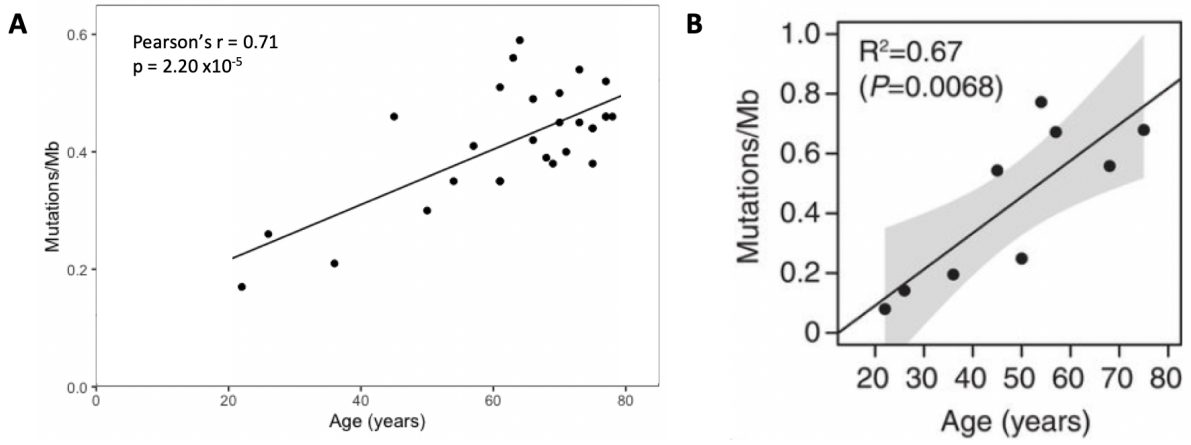


Figure 4.3: A Mutation burden for all 28 donors by age, obtained through Nano-seq. **B** Mutation burden per donor by age for the nine donors as published in (Martincorena, Fowler, et al., 2018, Figure 1C), where mutation burden was estimated using synonymous mutations in targeted sequencing data.

A linear regression of mutation burden by age for all donors found that smokers have a higher mutation burden than ex-smokers, with samples from non-smokers having the lowest burden (**Fig. 4.4A**). After excluding donors under the age of 60 years, samples from both current smokers and ex-smokers were found to have a mutation burden significantly higher than samples from non-smokers (**Fig. 4.4B**, ANOVA $p = 1.2 \times 10^{-3}$; smokers vs. non-smokers: $p = 1.6 \times 10^{-3}$; ex-smokers vs. non-smokers: $p = 0.022$).

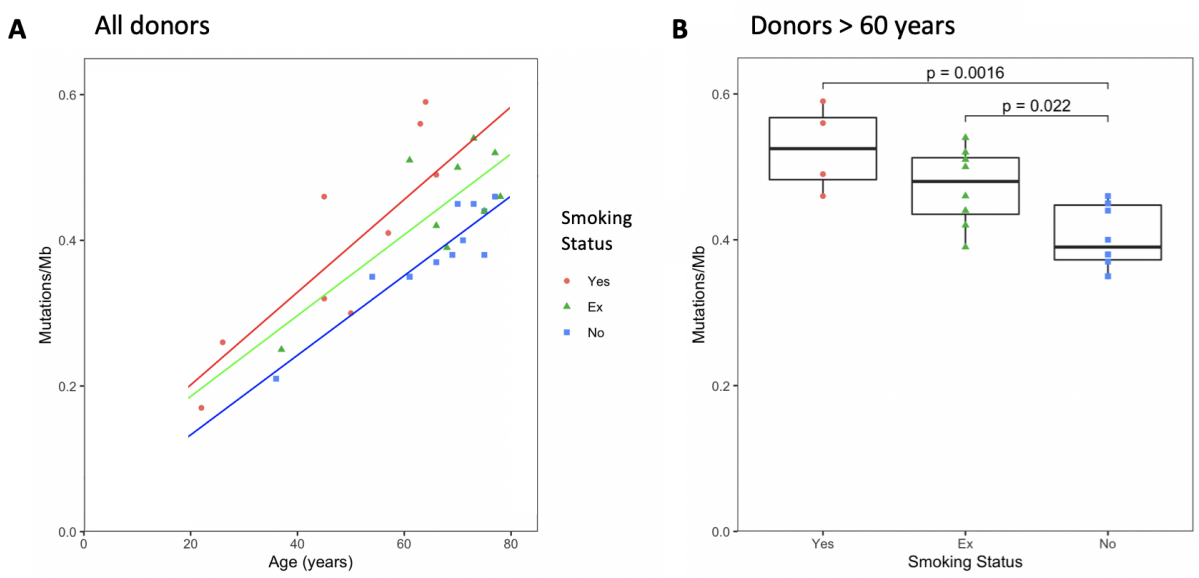


Figure 4.4: A Linear regression of mutation burden with age for smokers (red), ex-smokers (green) and non-smokers (blue). **B** Boxplot of mutation burden by smoking status for all donors over 60 years old. ANOVA, p = 1.2 x10^{-3}. Tukey test, smokers vs. non-smokers: p = 1.6 x10^{-3}; ex-smokers vs. non-smokers: p = 0.022.

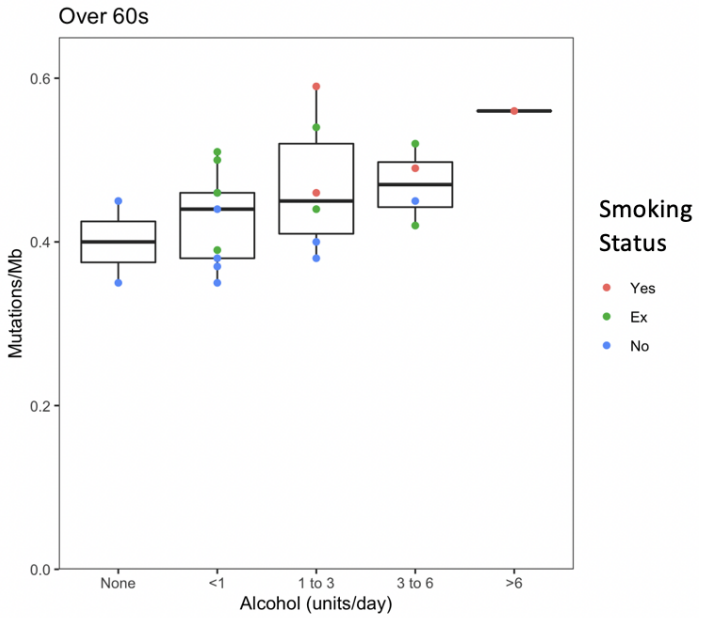


Figure 4.5: Boxplot of mutation burden by reported alcohol consumption for all donors over 60 years, coloured by smoking status.

In donors over the age of 60, there appeared to be a modest increase in mutation burden with reported alcohol consumption (**Fig. 4.5**). However, the effect of reported alcohol consumption was non-significant and not as large as that observed with smoking status. However, reported alcohol consumption may be an unreliable measure of a person's

exposure to alcohol throughout their lifetime, particularly in these samples, where questionnaires were completed by donor relatives soon after the patient's death. In order to investigate a possible interaction between smoking and alcohol (and other factors), I fit a linear model across all donors to describe mutation burden with age, smoking status, reported alcohol consumption and sex as predictor variables, including the interactions between them as terms. I then used the Akaike Information Criterion method to remove terms through backwards stepwise elimination that did not offer useful predictive power. The final linear model to describe burden (in mutations/Mb) was as follows:

$$Burden = -0.00390 + 0.00588 * Age + \begin{cases} 0.0572 \text{ if } ex\ smoker \\ 0.103 \text{ if } current\ smoker \end{cases}$$

Both donor age and smoking were found to be significant positive predictors of burden ($p = 1.05 \times 10^{-7}$ and $p = 1.97 \times 10^{-3}$ respectively), with no significant interaction between them (**Fig. 4.4A**). The contribution of alcohol consumption, donor sex and their interactions with other terms, were found to be non-significant.

Previous study of mutational signatures in over 500 ESCC genomes reported an increased indel signature (ID3) with tobacco smoking. ID3 was identified in 10 out of 261 smokers and 2 out of 291 non-smokers (Moody et al. 2021).

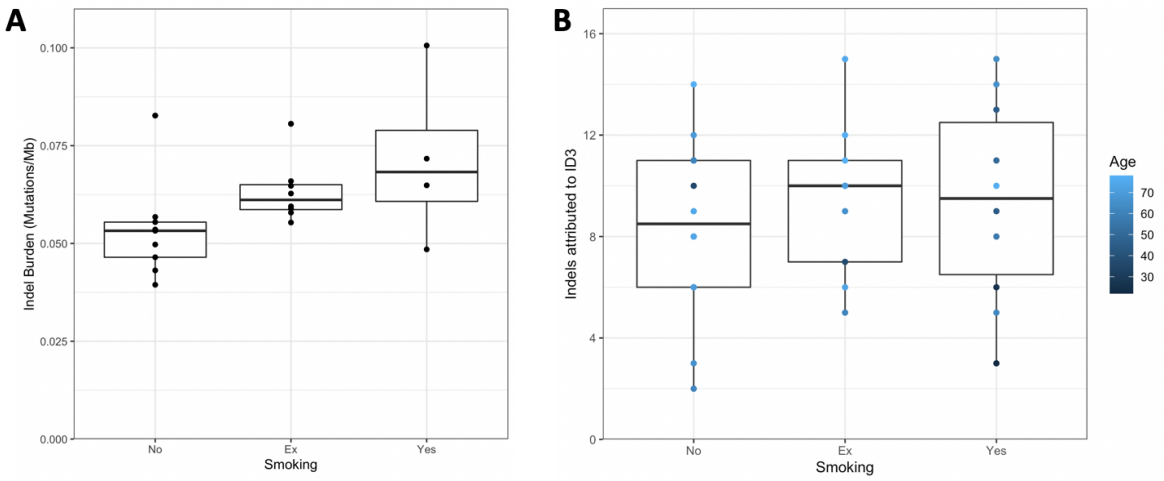


Figure 4.6: A Boxplot of indel burden estimated from Nano-seq data for all donors over 60 years, by smoking status. **B** Boxplot of indels from Nano-seq data for all donors attributed to ID3, by smoking status, coloured by donor age (years).

Here, I find the median indel burden is higher in smokers and ex-smokers than in non-smokers (**Fig. 4.6A**). However, two outliers (non-smoker PD40291 and ex-smoker PD48812) mean the effect of smoking status on indel burden is non-significant (ANOVA: p = 0.082). 279 of the indels called by Nano-seq were attributed to the ID3 signature, however, no significant difference was identified with either donor age or smoking status (**Fig. 4.6B**).

Mutations called across targeted 2 mm^2 grid samples

A total of 19,979 mutations were detected across 1,463 samples of epithelia from the unpublished donors (**Fig. 4.7**). There was large variation in the number of mutations called per sample, ranging from 0.5 to 21 mutations/mm^2 (mean = 13.6 mutations/mm^2). After merging mutations which spanned adjacent samples of the same donor, the number of independent mutations was 15,912 (**Fig. 4.8**). There was no significant difference in mean mutations called per mm^2 of tissue between the two bait sets (Welch's t-test, p = 0.23). This suggests that the majority of clones of a large enough size to be detected are carrying mutations within this smaller set of 32 genes.

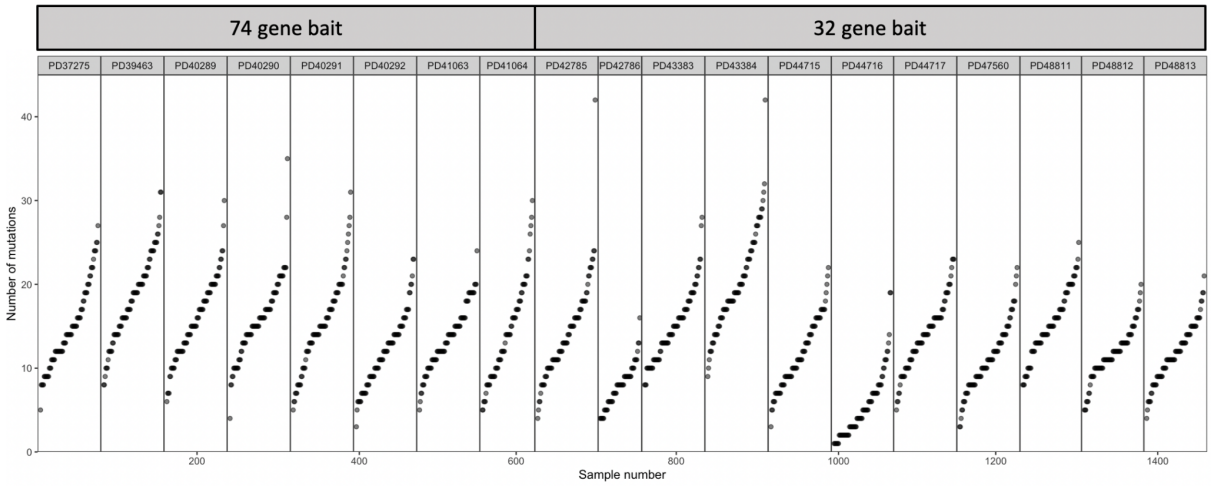


Figure 4.7: Distribution of the number of mutations detected across 19 newly sequenced oesophagus donors per 2 mm^2 sample of epithelium (n = 1,463) for 74 or 32 targeted genes.

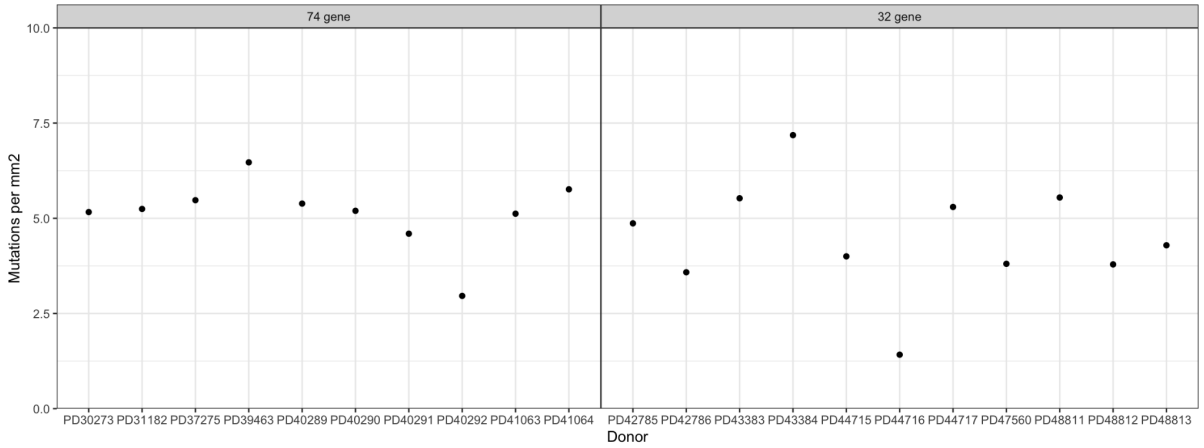


Figure 4.8: Number of independent mutations (n = 15,912) detected per mm^2 of oesophageal epithelium per donor across either 74 or 32 targeted genes, including the two donors (PD30273 & PD31182) aged over 60 years published in Martincorena, Fowler et al. (2018).

Mutational signatures

Mutational signature analysis of the Nano-seq SBS mutational spectra resulted in decomposition into 12 reference signatures (**Fig. 4.9**). The largest contribution to donor spectra was due to signatures associated with ageing (SBS1, SBS5 and SBS40). In 16 donors, these were the only signatures assigned. SBS16, associated with alcohol consumption, was identified in three donors. SBS16 contribution was greatest in PD44716, the one donor reported as a heavy drinker (>9 units of alcohol/day) and a smoker. Interestingly, SBS92 was identified with a high relative contribution in four donors. SBS92 is associated with tobacco smoking in normal and cancerous bladder, however, only one of these four donors has a history of smoking. SBS92 has a strong T>C peak at ATA and it may be that this signature is being misassigned from SBS16. Across all 31 donors, 93 mutations were called as double-base substitutions (DBS). DBS2, a reference mutational signature predominantly characterised as CC>AA substitution, has previously been correlated with smoking in ESCC genomes (Moody et al. 2021). Here, the small number of DBS precludes deconstruction into mutational signatures, however, only five DBS were called in the CC>AA context.

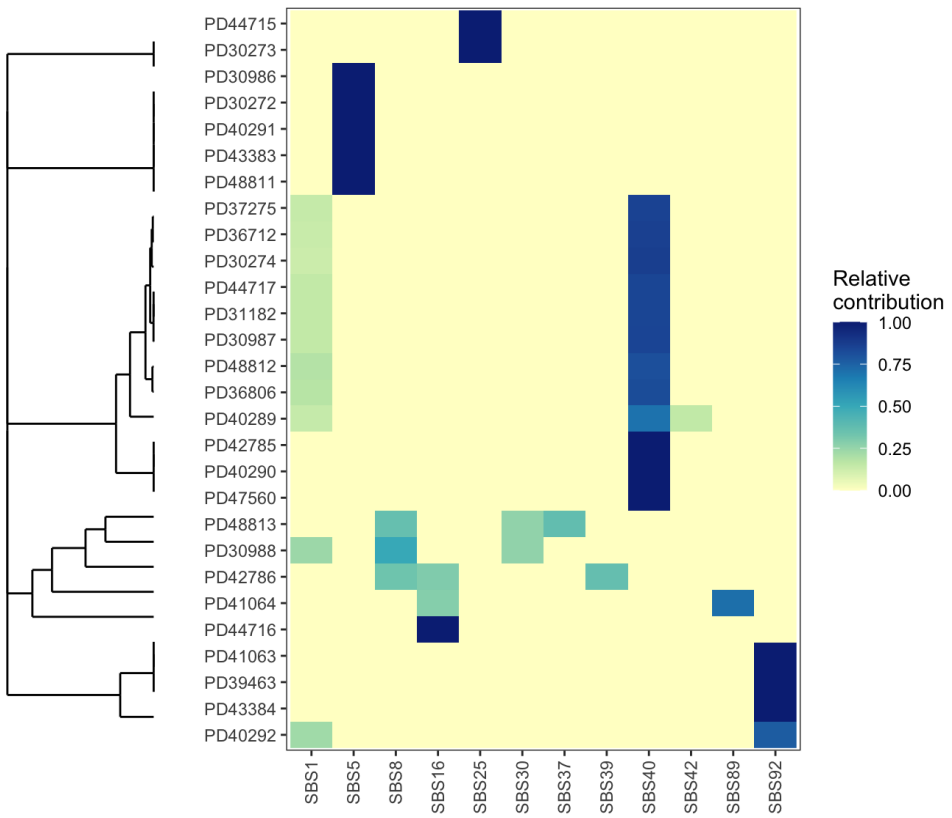


Figure 4.9: Decomposition of Nano-seq mutation spectra into PCAWG reference signatures (maximum delta = 0.1).

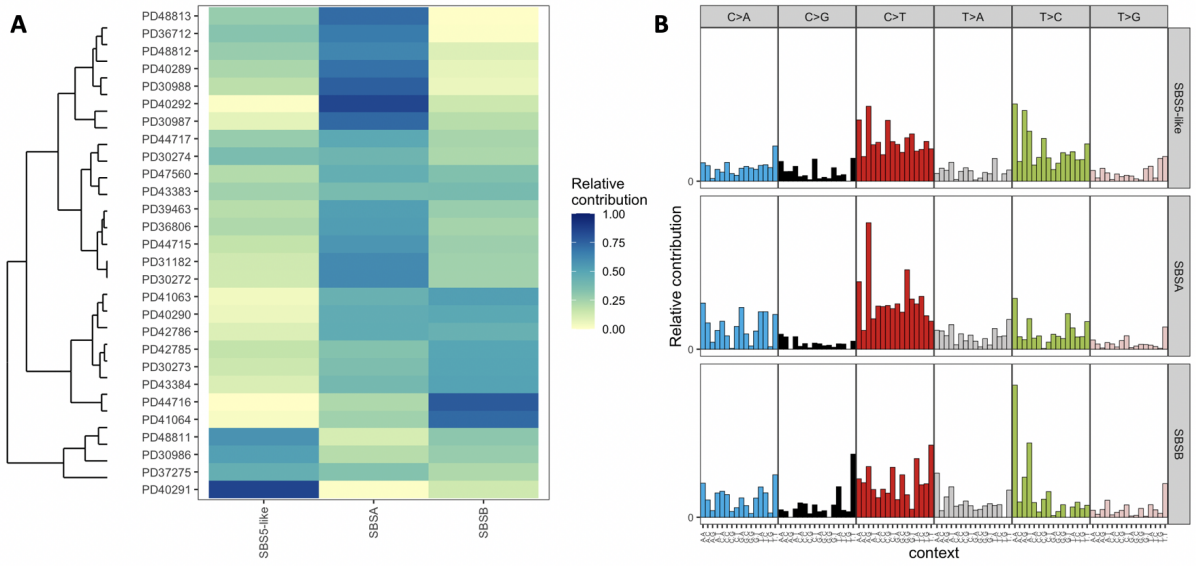


Figure 4.10: *De novo* analysis of mutational signatures in Nano-seq data.

I also ran a *de novo* mutational signature analysis on the Nano-seq data, using non-negative matrix factorisation, in order to identify the action of a novel signature that may be absent in the reference. Three *de novo* signatures were extracted, with one having more than 0.9

cosine similarity to SBS5 and therefore is labelled 'SBS5-like' (**Fig. 4.10**). SBSA is a relatively flat signature, with distinctive C>T peaks at ACA and AGA. SBSB has strong T>C peaks similar to that of SBS16 (alcohol consumption) but peaks with similarity to SBS2 and SBS13, the two reference signatures associated with APOBEC activity (**Fig. 4.11**). SBSB has the highest contribution in PD44716, the heavy drinker and smoker.

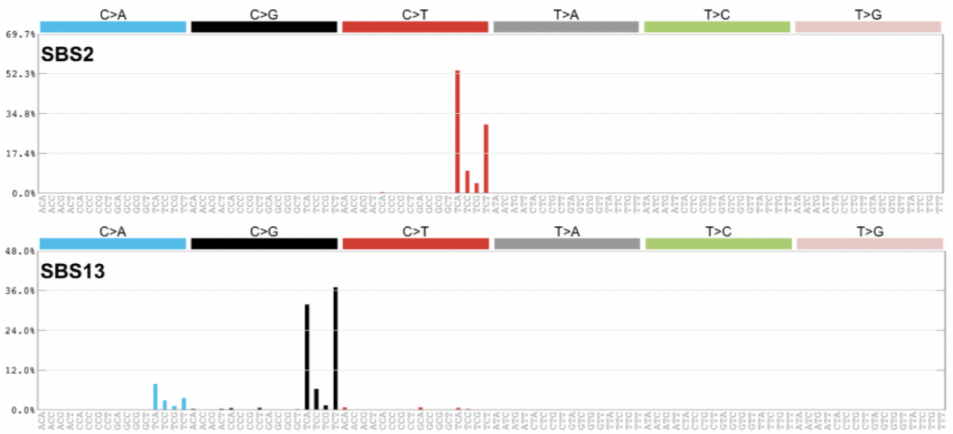


Figure 4.11: The distinctive trinucleotide contexts associated with APOBEC mutagenesis (SBS2 and SBS13), taken from (L. Alexandrov et al. 2018).

Mutations called across targeted 2-mm^2 grids had similar mutation profiles to those called by Nano-sequencing (**Fig. 4.12**). However, since SBS16 mutations are most commonly found in transcribed regions, these T>C peaks are more pronounced in the targeted data. In addition, C>T peaks associated with SBS1 are more clearly observed in the targeted data. Previous study of mutational signatures in ESCC genomes reported a *de novo* signature, SBS288I (described as a composite of the reference signatures SBS4 and SBS5), as correlating with smoking (Moody et al. 2021). SBS288I was identified in 20 out of 261 smokers and 2 out of 291 non-smokers. Here, however, mutational signature decomposition does not identify SBS288I across any of the 28 targeted donors. This is perhaps unsurprising. Firstly, if SBS288I is associated with smoking, it appears to be rare. Secondly, SBS288I has been characterised as contributing mutations predominantly to untranscribed regions, so the number of mutations occurring across the targeted bait used here is expected to be small.

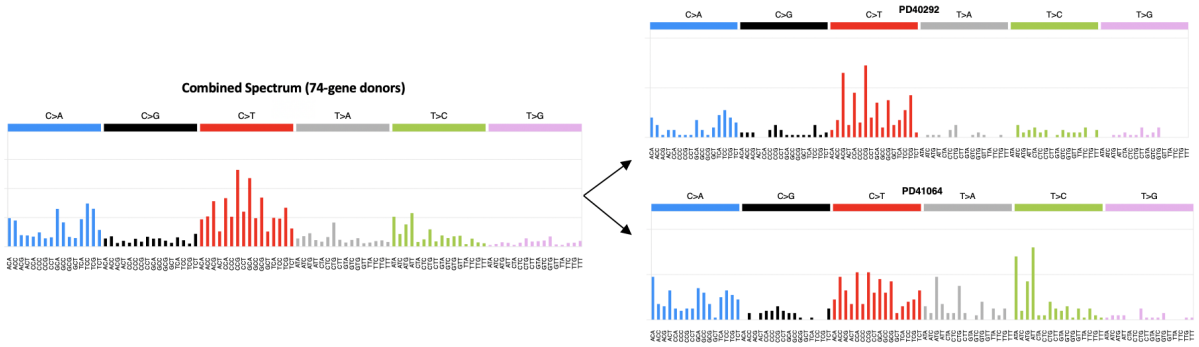


Figure 4.12: Combined spectrum of trinucleotide contexts for all mutations called across the 74-gene bait donors. Two donor spectra are shown to highlight high contributions of SBS16 (T>C peaks) associated with alcohol consumption in donor PD41064 (occasional reported drinker, ex-smoker) and high contributions of SBS1 (C>T peaks), a clock-like process, in PD40292 (non-drinker, non-smoker).

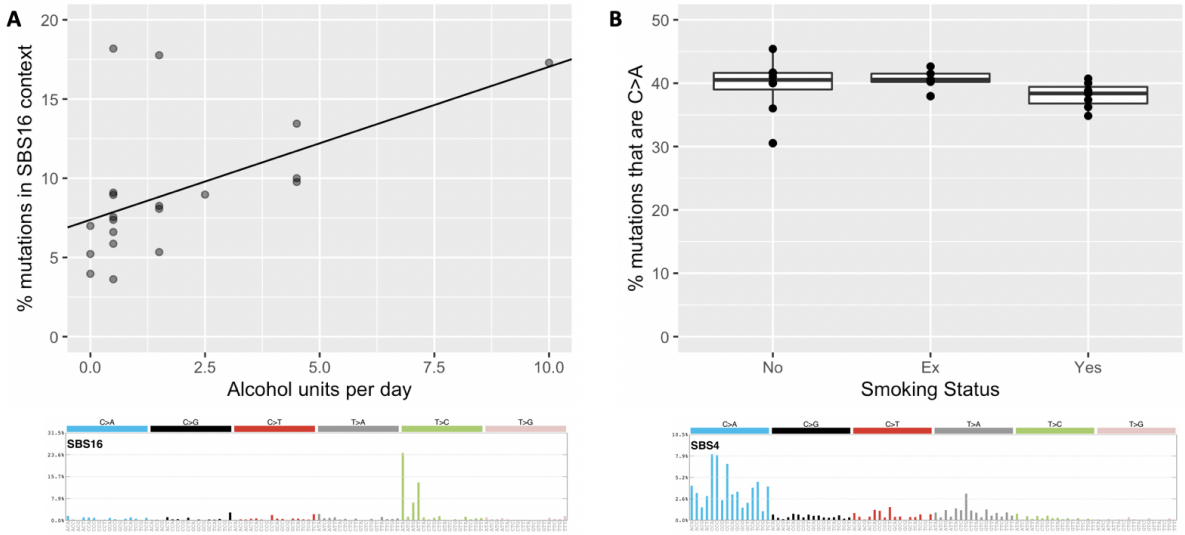


Figure 4.13: Mutational signatures in 74-gene targeted donors by reported cancer risk factor exposure. **A** Correlation between percentage of mutations attributed to SBS16 and reported daily alcohol consumption. **B** SBS4 is a C>A signature associated with tobacco smoking in lung. There is no effect of donor smoking status on the percentage of C>A mutations per donor.

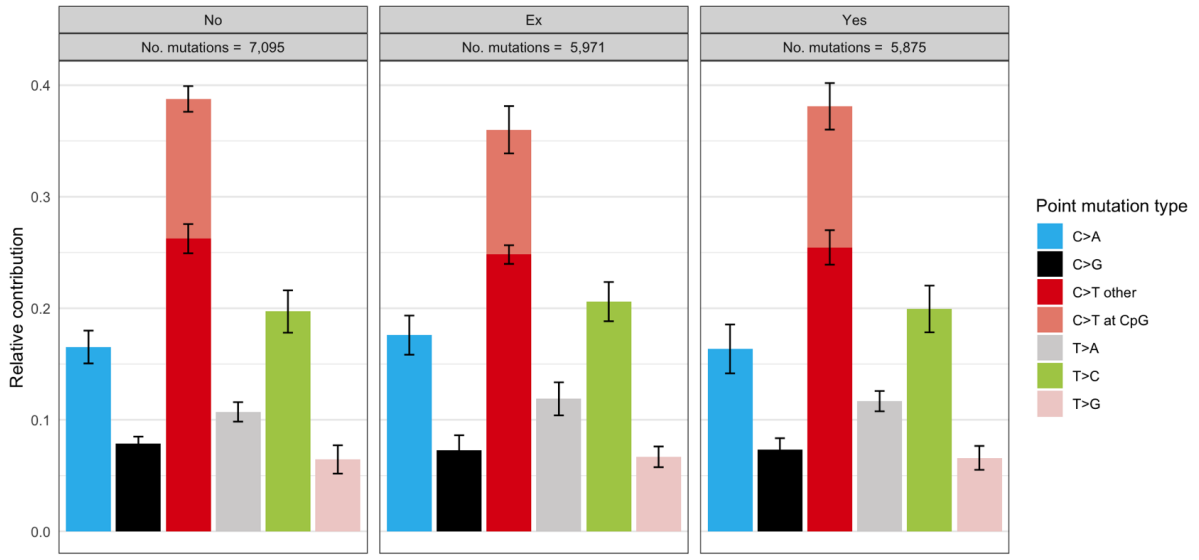


Figure 4.14: There is no evidence to suggest that smoking status affects the relative contribution of each of the six base substitution types. Error bars represent 95% confidence intervals.

In summary, there is no evidence to suggest that smoking leaves a mutational signature in oesophageal epithelium in mutations called either by targeted sequencing (**Fig. 4.13**) or Nano-seq (**Fig. 4.14**). In contrast to alcohol consumption, smoking status appears to have no effect on the relative contribution of substitution type.

Variation in clone size by donor

After merging mutations spanning multiple adjacent samples, I summed the VAF across each sample in order to estimate the clone size of each mutation. The number of clones detected per mm^2 of epithelium in each donor is shown in Figure 4.8. The sensitivity of mutation calling is base-dependent, however in some cases, mutations were detected down to a variant allele fraction of 0.002 (**Fig. 4.15**).

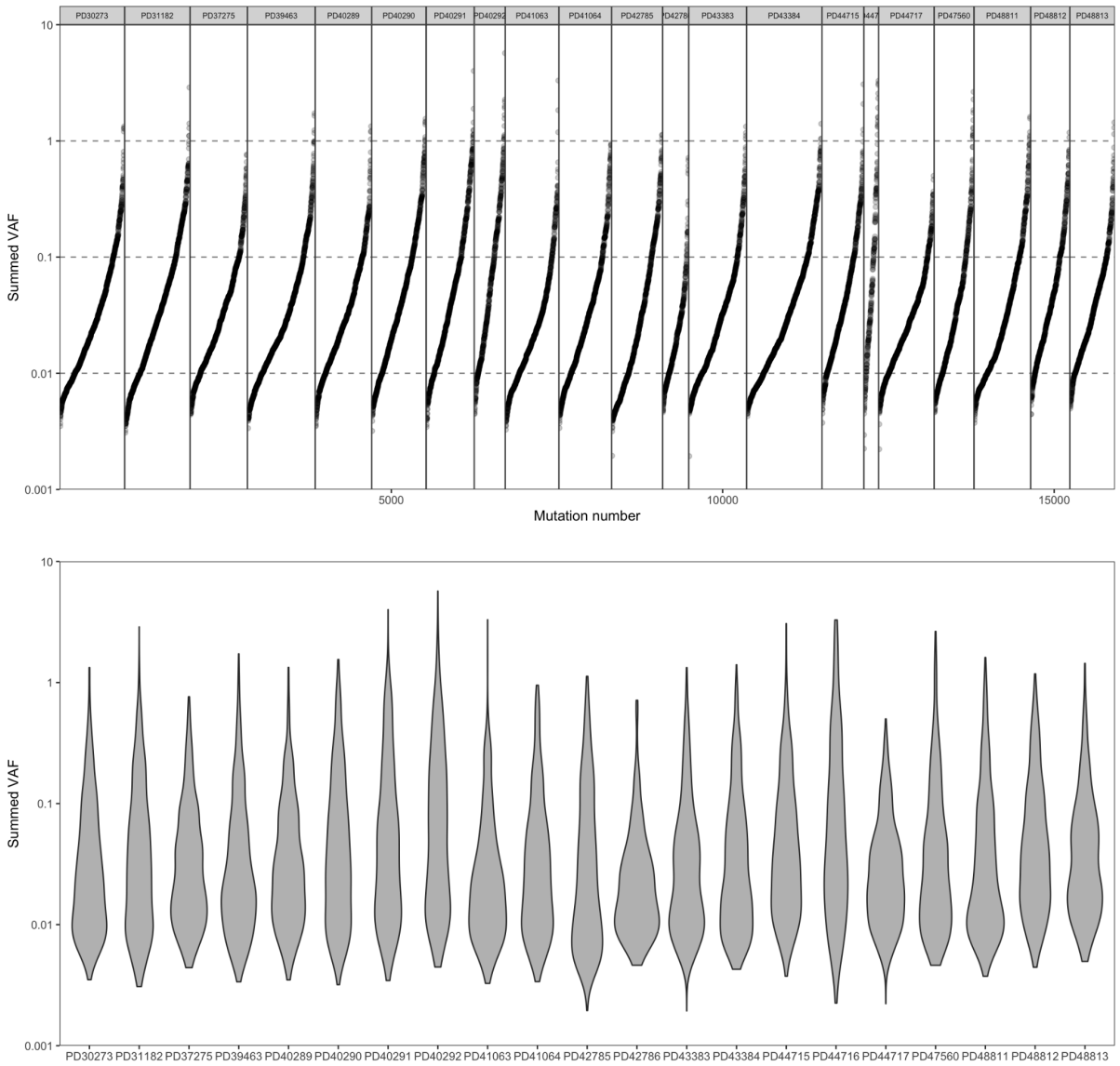


Figure 4.15: Distribution per aged (over 60 years) donor in the summed variant allele fraction (VAF) for each mutation (after merging mutations which span multiple samples), shown both as a plot of the summed VAF for every clone and as a violin plot. Median summed VAF = 0.026.

Figure 4.15 highlights the variation in the distribution of clone sizes by donor. For example, 5.30 mutations/mm^2 were detected across samples of PD44717, with the majority of these mutant clones being small (summed VAF < 0.05) and none growing larger than a summed VAF of 0.5. In contrast, only 1.42 mutations/mm^2 were detected across samples of PD44716, the lowest burden of all donors, but the distribution of clone sizes is much greater, with tens of clones growing larger than a summed VAF of 0.5 and consequently PD44716 had a mean summed VAF over four times that of PD44717 (PD44716 = 0.22, PD44717 = 0.045). The largest clone, driven by a S385F mutation in *NOTCH1*, was found to span 14 samples of PD40292, with a summed VAF of 5.70 (**Fig. 4.16**). As the height of each grid sampled was 1 mm, this clone therefore had a diameter of at least 9 mm, however, the true size is likely to

be larger since it was found on the edge of the piece of tissue sampled. Kruskal-Wallis (and Dunn) tests found no significant difference in clonal density (mutations per mm^2) or clone size (mean summed VAF) by smoking status or alcohol consumption.

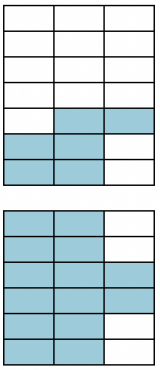


Figure 4.16: A S385F mutation in *NOTCH1* spanning fourteen samples across two pieces of tissue sampled from donor PD40292. Each rectangle represents a 2 x 1 mm sample.

Clonal selection and competition with age

Across all 28 donors (9 published and 19 newly sequenced), I found 17 genes (*TP53*, *NOTCH1-3*, *FAT1*, *AJUBA*, *CCND1*, *KMT2D*, *CUL3*, *ARID1A*, *PIK3CA*, *NFE2L2*, *EPHA2*, *ZNF750*, *CREBBP*, *TP63* and *ARID2*) with a significant *d*N/*d*S ratio ($q < 0.01$), suggesting these genes are under positive selection and play a role in driving clonal expansion in normal oesophagus (**Fig. 4.17**). This work is the first to identify *EPHA2* and *CREBBP* as being under positive selection in oesophageal epithelium. A comparison of the *d*N/*d*S ratio per gene, for genes present on both bait panels, found *TP53* and *FAT1* non-synonymous mutations significantly over-represented and *NOTCH3* non-synonymous mutations significantly under-represented in donors over 60 years old (with respect to synonymous mutation rates) compared to donors under 60 years of age ($p < 0.001$, **Fig. 4.18**).

Interestingly, missense mutations in both *NOTCH1* and *NOTCH2* were significantly under-represented in donors over 60 years ($p < 0.01$), however, there was no significant difference in truncating mutations in these genes with age. This suggests that clones mutant for *NOTCH3* and those harbouring a missense mutation in *NOTCH1/2* are relatively weak drivers of clonal expansion. As the tissue ages and the burden of mutant clones increases, stronger drivers of clonal expansion, such as clones mutant for *FAT1* and *TP53,* are able to outcompete weaker clones present in younger tissue.

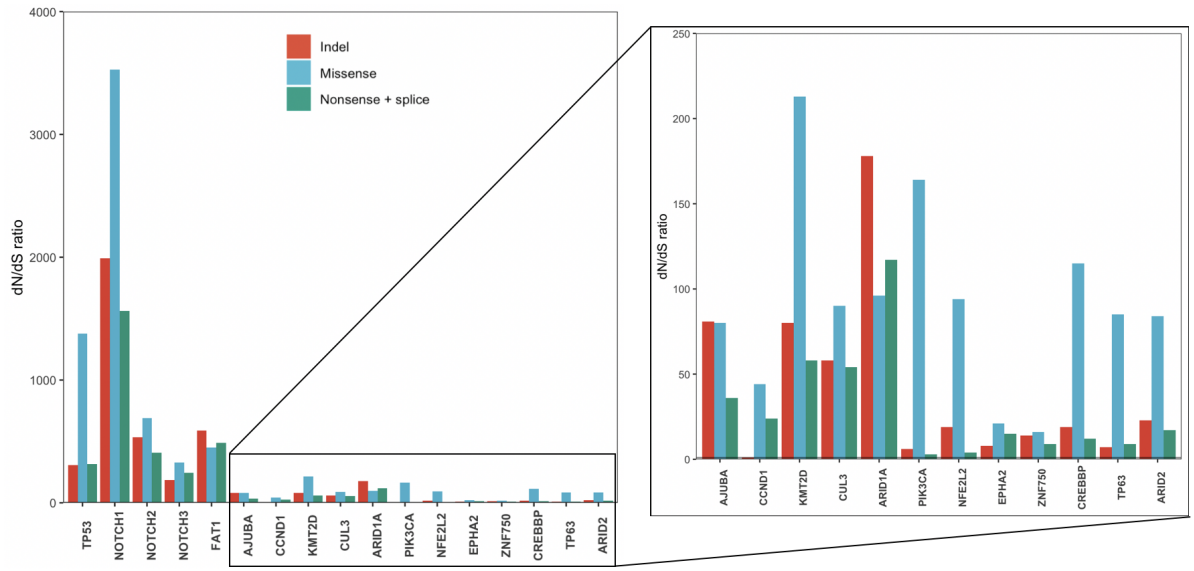


Figure 4.17: Ratio of observed/expected non-synonymous mutations for positively selected genes across all 28 donors. Line drawn at y = 1.

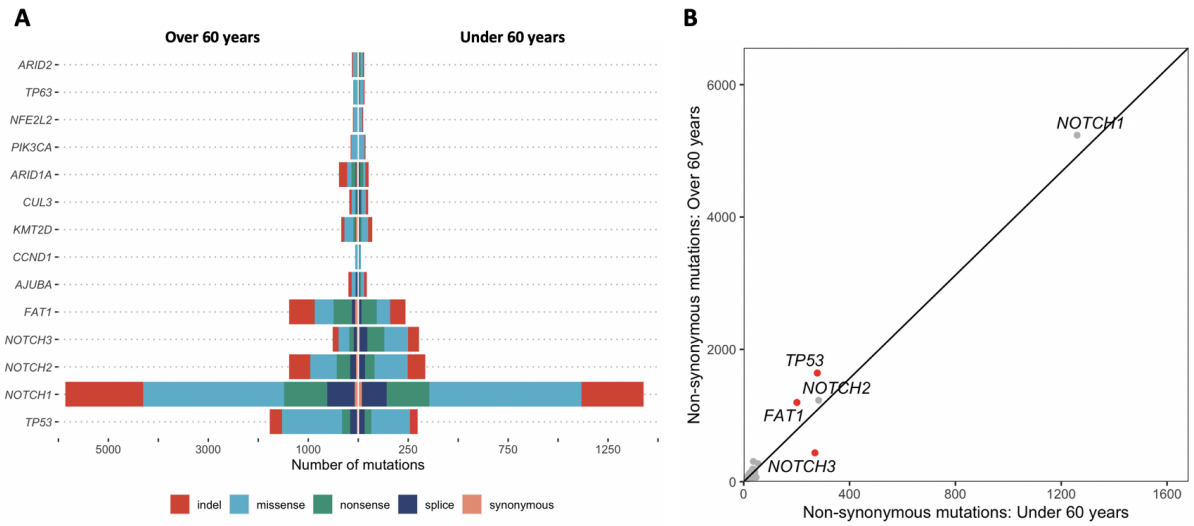


Figure 4.18: A The number of mutations of each consequence for positively selected genes (q < 0.01) by age. **B** Plot of non-synonymous mutations per gene by age. Gradient of line = total number of non-synonymous mutations in donors >60 years/<60 years = 13,112/3,361. Red indicates positively selected genes with a significant (p < 0.001) difference in dN/dS ratio by donor age, after accounting for global differences.

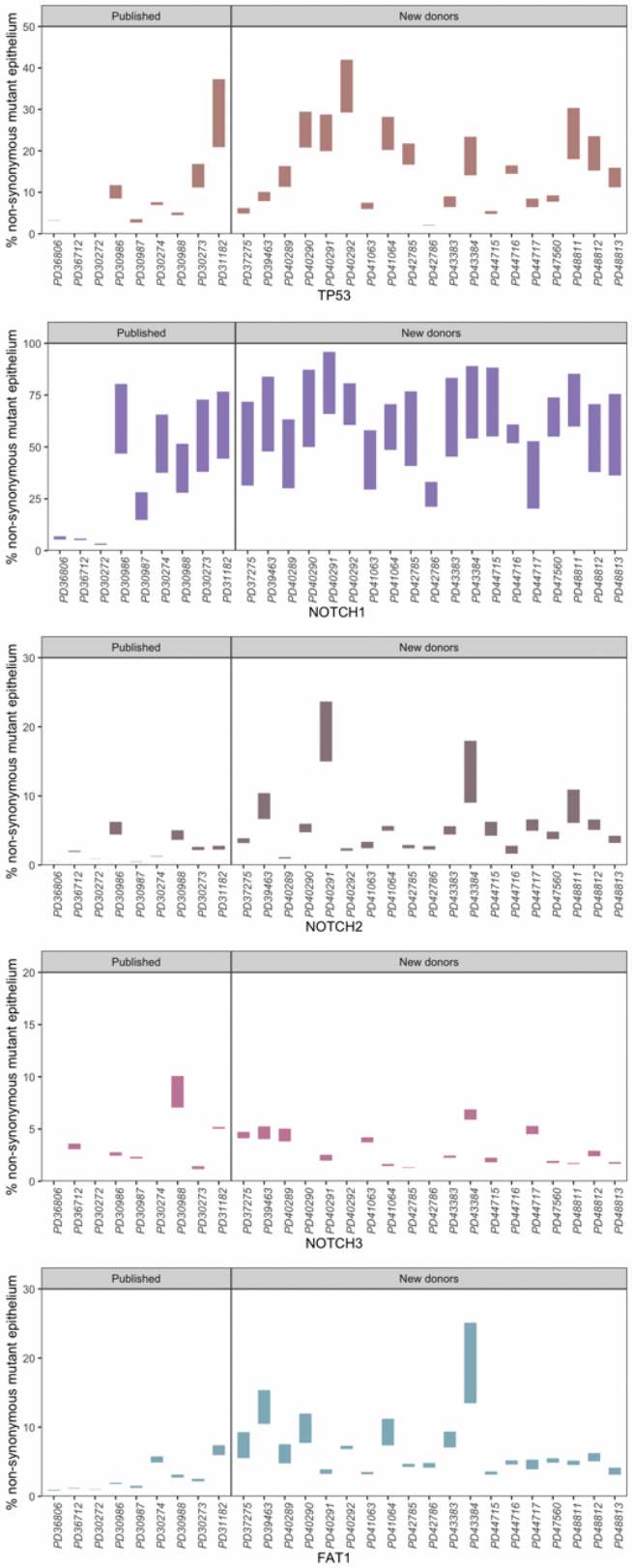


Figure 4.19: Estimated range of the percentage of cells sampled from each donor harbouring at least one non-synonymous mutation for the five genes with greatest mutant coverage of the tissue. Lower bound assumes all mutations in a sample occur within the same subset of cells. Upper bound assumes mutations do not occur within the same cells of a sample, where possible.

I estimated the percentage of tissue in each donor carrying a nonsynonymous mutation for each of the five genes that have the greatest mutant coverage across the samples (**Fig. 4.19**). The percentage of tissue covered by at least one protein-altering mutation in each of these genes was highly variable across donors, however, other than donor age, this variation did not correlate with reported smoking status or alcohol consumption across any of these five genes.

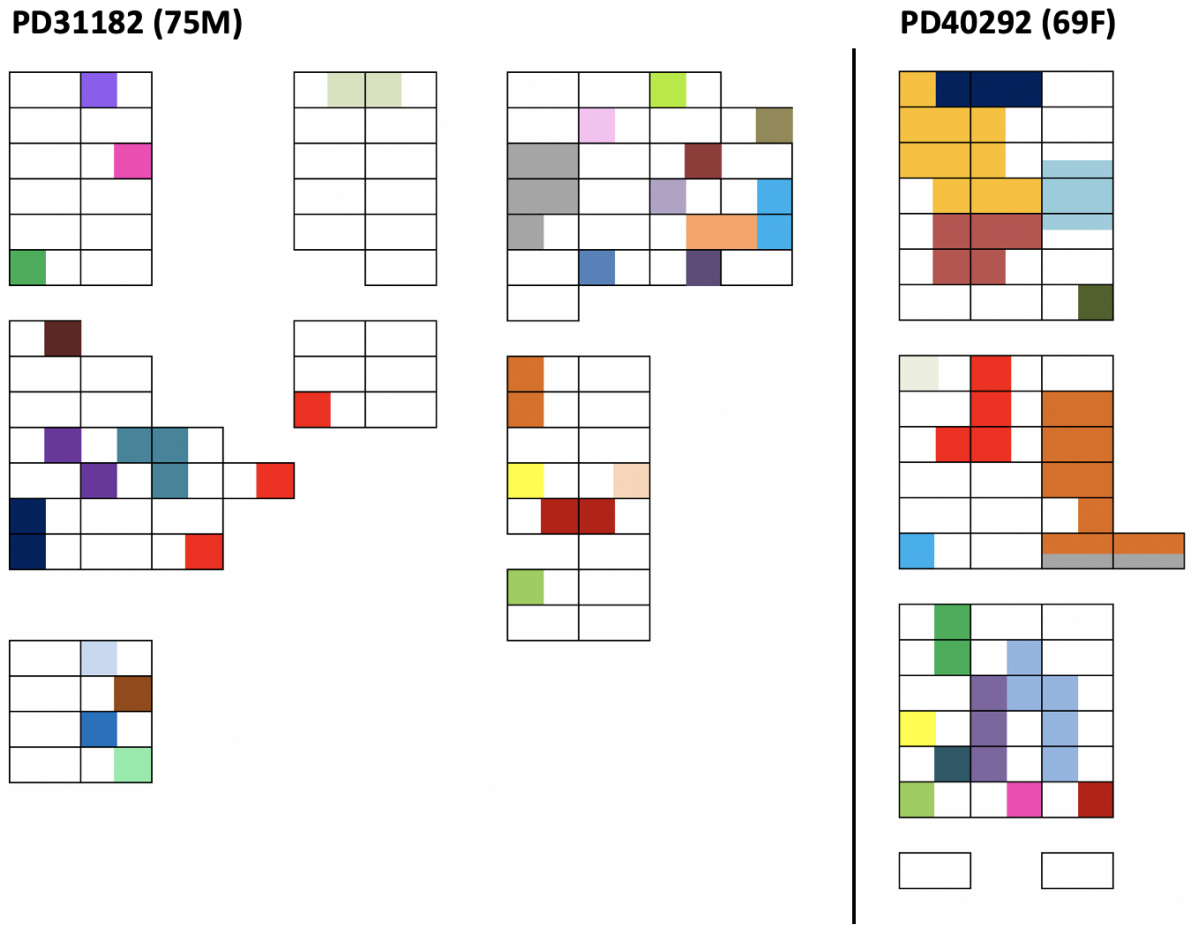


Figure 4.20: A spatial representation of the size and frequency of *TP53* mutant clones mapped across samples of two donors with the greatest proportions of tissue mutant for this gene. Both donors have a similar proportion of their tissue mutant for *TP53* (~30-35%) yet show differing clone size and frequency. Each colour represents an independent *TP53* clone. Each rectangle represents a 1 x 2 mm grid of epithelium.

I observed variation in the size and frequency of mutant clones, even across donors with a comparable proportion of their tissue mutant for that gene (**Fig. 4.20**). For example, donors PD40292 and PD31182 had a similar proportion of tissue mutant for *TP53*, approximately 30-40% and 20-37% respectively. Despite this, PD31182 had a *TP53* mutation burden double that of PD40292 (0.94 and 0.46 mutations/mm^2 respectively) and the mean *TP53*

summed VAF in PD31182 was over half that observed in PD40292 (0.11 and 0.25 respectively).

Targeted punches

Across all targeted punches, a total of 1,805 mutations were called, ranging from 0 to 12 mutations per sample. Of the 455 punches sequenced, 295 (65%) were found to be clonal (VAF > 0.25). Across the seven donors, 27 samples (across 14 independent clones) were clonal for mutant *TP53.* No samples were found to be clonal for both mutant *TP53* and mutant *NOTCH1*. As with the 2 mm^2 grid data, there was large variation in clone size by donor (**Fig. 4.21**).

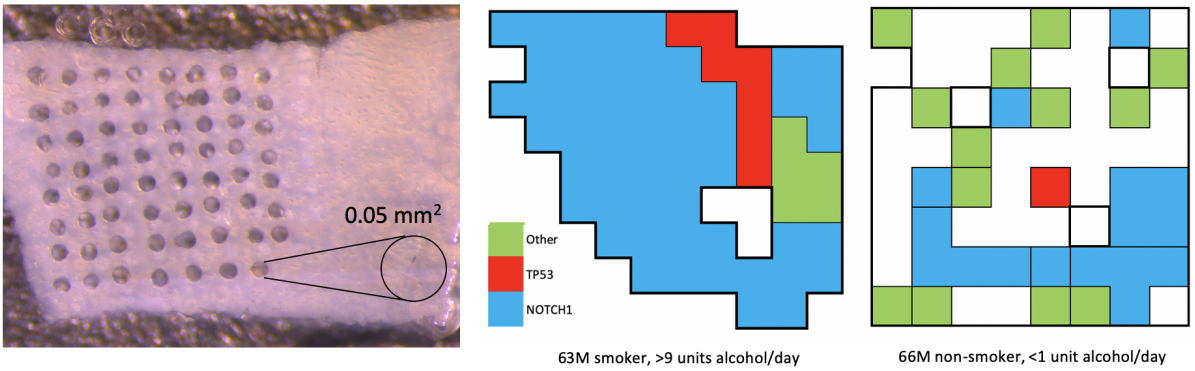


Figure 4.21: Image of peeled oesophageal epithelium showing punches taken in a gridded array. Two examples (PD44716 and PD41049 respectively) of spatial clone mapping to illustrate the variation in clone size by donor.

Whole-genome sequencing

As described in the methods of this chapter, clonal targeted punch samples were selected from whole-genome sequencing. Both between donors and within samples of the same donor, there was large variation in the number of mutations called per genome (**Fig. 4.22**). I performed mutational signature analysis on each sample individually by deconstructing the trinucleotide contexts of single-base substitutions in each sample into the reference signatures (**Fig. 4.23**). As in the targeted data, the majority of substitutions were attributed to ageing signatures SBS1, SBS5 and SBS40. Both SBS5 and SBS40 are flat signatures and so difficult to assign, but there appeared to be a preference for SBS40 instead of SBS5 in samples of some donors over others, suggesting a biological difference between the two. SBS42, a predominantly C>T signature linked with exposure to the haloalkane 1,2,3-trichloropropane, a common pollutant of drinking water (Riva et al. 2020), was assigned to samples in PD41049 in particular. A consistent contribution of SBS16 was

assigned to every genome sample of PD44716, consistent with this donor as a heavy alcohol drinker.

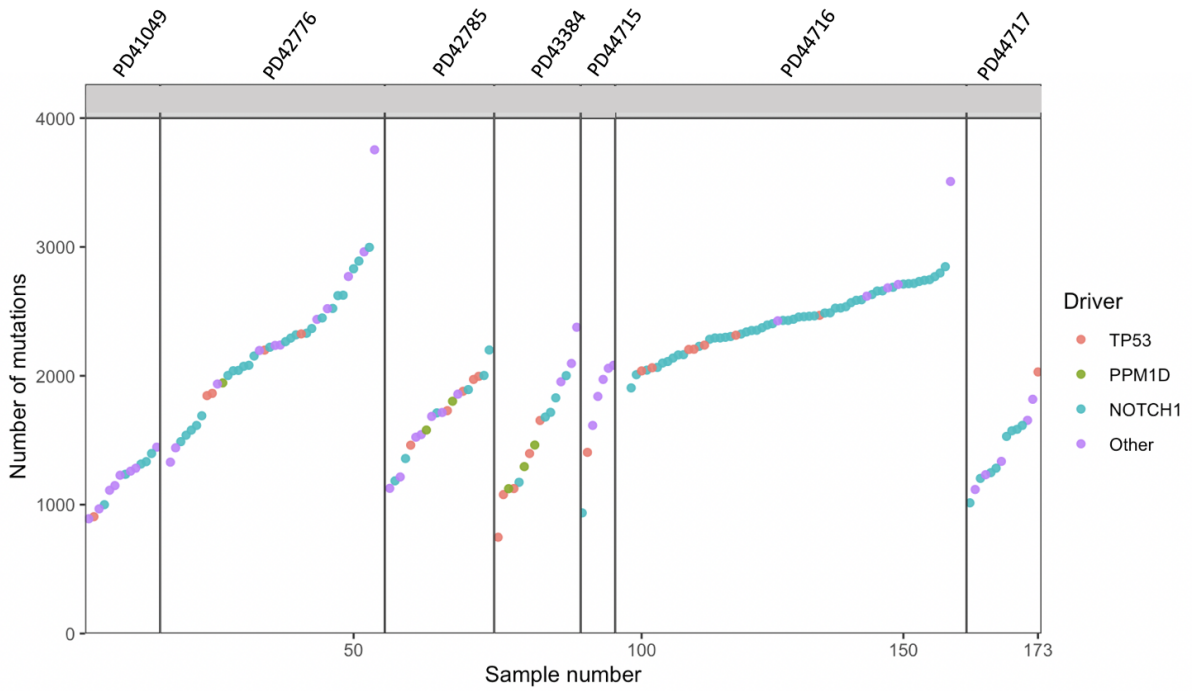


Figure 4.22: The number of substitutions called per whole genome sample per donor, coloured by the clonal driver identified through targeted sequencing.

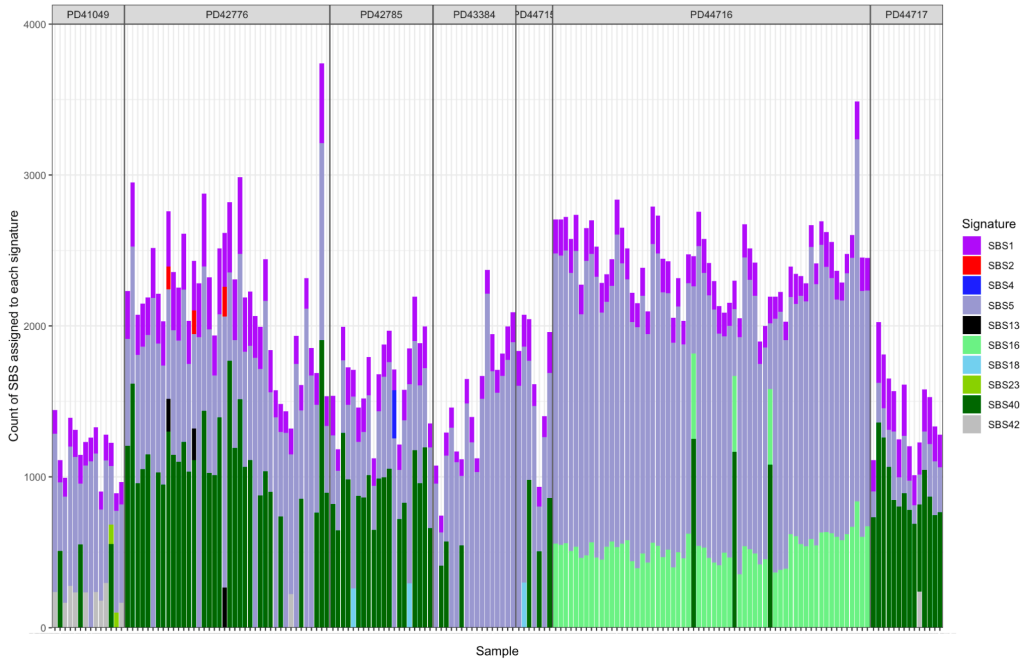


Figure 4.23: Count of single-base substitutions assigned to each reference signature for each whole genome punch per donor. Most signatures are consistently called in independent samples of the same donor, suggesting a biological aetiology.

Finally, SBS2 and SBS13, associated with APOBEC mutagenesis (**Fig. 4.11**) were identified in three samples of PD42776 (70M, ex-smoker, light drinker). These three samples (as, bb and bj) were adjacent to each other in the tissue, as part of the same *NOTCH1* mutant clone which had gained a mutation in *LRP1B*, one of the ten most frequently mutated genes in cancer (**Fig. 4.24**). Samples as, bb and bj have a distinct signature from samples adjacent in the tissue (for example, bc), suggesting that APOBEC mutagenesis is contained within cells of a clone, as opposed to being a local event in the tissue.

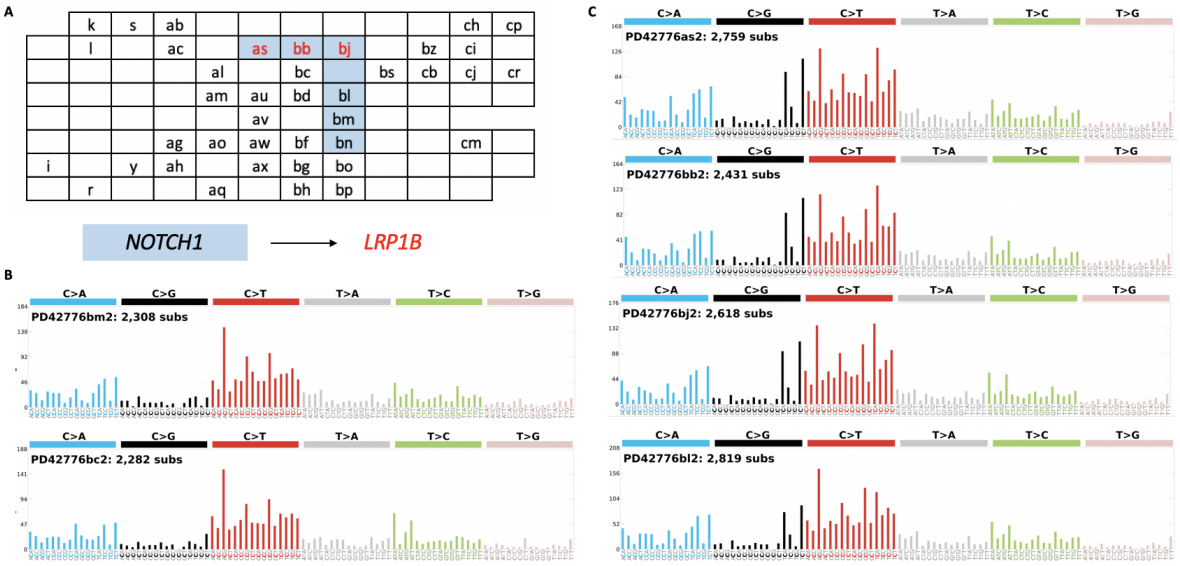


Figure 4.24: A Spatial array of punch samples in PD42776. Each rectangle represents a single punch biopsy and those with letters are whole-genome samples. Blue rectangles show the expansion of a *NOTCH1* mutant clone in the tissue and red text indicates samples which share the same mutation in *LRP1B.* **B** Example of mutational spectra from two samples (bm and bc) which lack an APOBEC signature, despite being spatially adjacent to samples that do. **C** Spectra of samples with a distinctive APOBEC signature (as, bb, bj), as called by SigProfiler and a sample (bl) where a low contribution of APOBEC can be seen, despite not being called by SigProfiler.

Phylogenies

For each donor, I drew a phylogenetic tree, with branch lengths equalling the number substitutions shared between samples of that donor (**Fig. 4.25**). Clades of samples determined by the tree are spatially adjacent when mapped to the tissue. Four distinct clones are identified in PD44716 (63M, smoker, heavy drinker). The initial branch of each distinct clone is made up of over 1,000 substitutions, before later branching occurs, suggesting these clones have persisted in the tissue for a long time, before a more sudden clonal expansion, perhaps due to acquisition of the driver mutations (labelled in **Fig. 4.25**) and/or a change in the environment of the tissue with age.

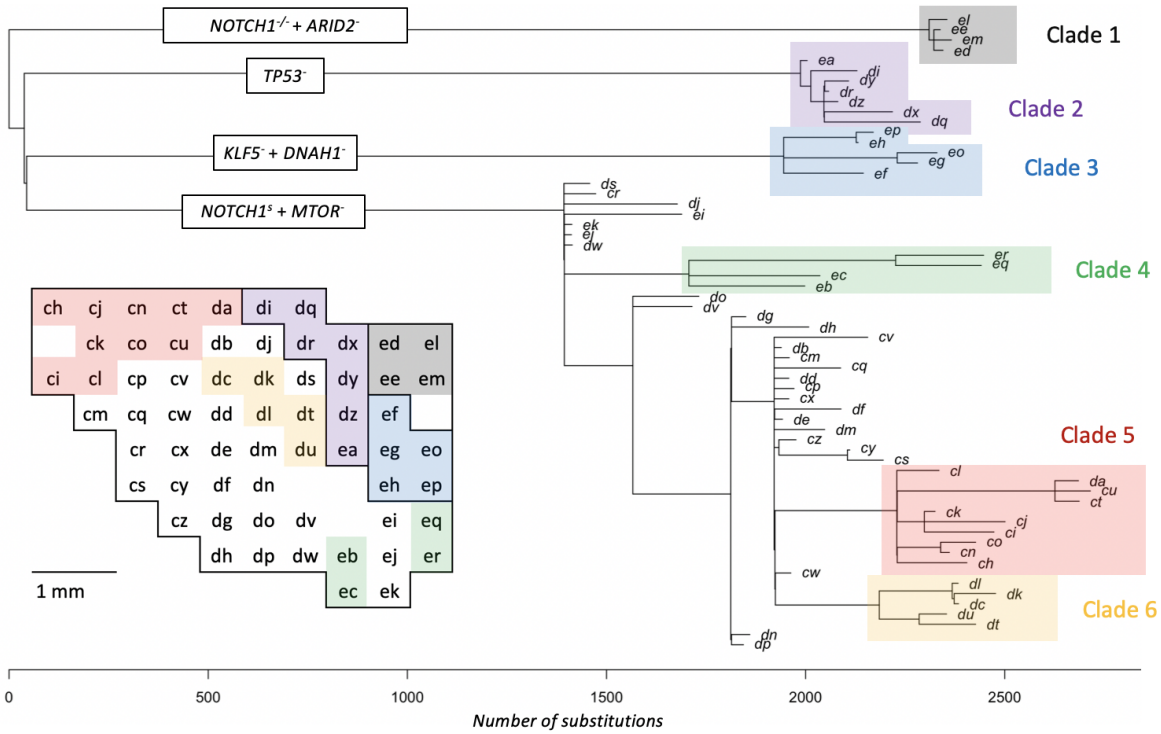


Figure 4.25: A phylogenetic tree depicting the relatedness between 62 whole-genome punch samples from donor PD44716 (63M, smoker, heavy drinker). Branch lengths equal the number of substitutions shared between samples. The position of each sample in the tissue is mapped, where each two-letter code (e.g. 'ch') represents a single circular punch measuring 0.05 mm^2 in area and colours highlight individual clades as determined by the tree. Solid borders within the map outline the four distinct clones in this donor and possible drivers of clonal expansion in each clone are labelled (missense = - ; nonsense = * ; essential splice = s). No copy number events were detected in this donor.

Decomposition of the substitutions which make up the four main branches and six clades of this phylogenetic tree into reference mutational signatures showed that the proportion of SBS16 (alcohol consumption) did not change over time (**Fig. 4.26**). The majority of all other substitutions were assigned to the clock-like signatures SBS1 and SBS5. However, alternative signatures begin appearing in substitutions assigned to later branches. Substitutions assigned Clade 6, part of the *NOTCH1* and *MTOR* mutant clone, are assigned to SBS40 (associated with age) instead of to SBS5. Clade 2, driven by a *TP53* mutation, has substitutions assigned to SBS8, a C>A and T>A signature with unknown aetiology. Phylogenetic trees for six additional donors and the contribution of reference mutational signatures assigned to each branch are shown in Figure 4.27.

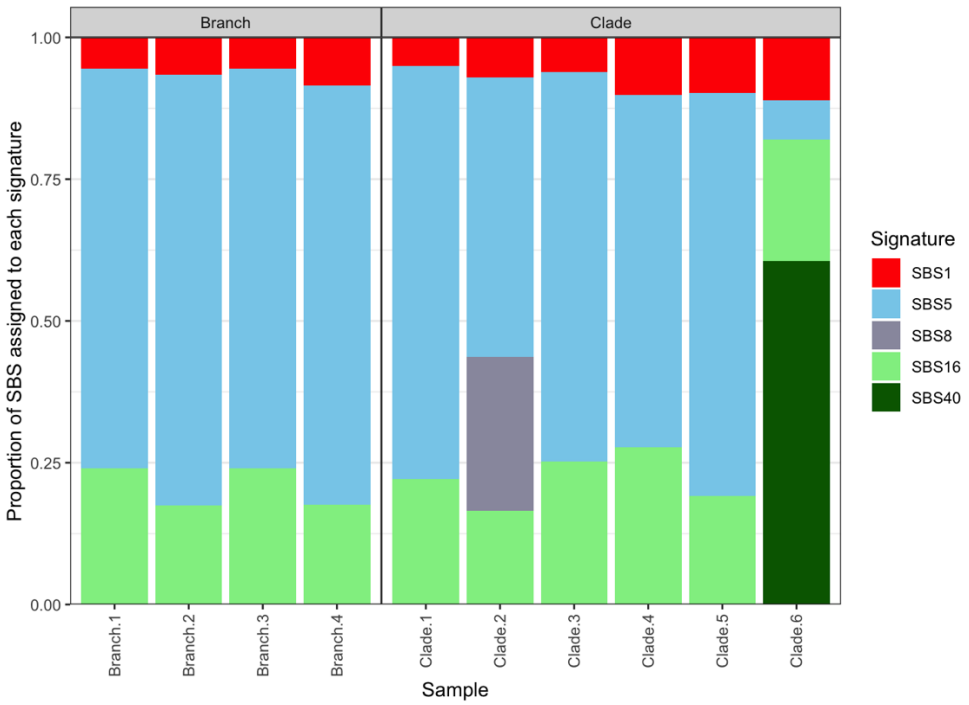


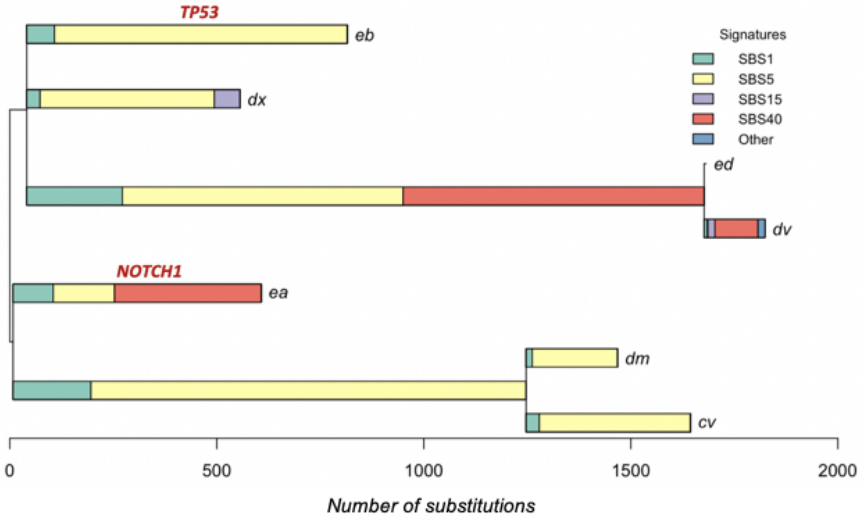
Figure 4.26: Contribution of each reference signature to the substitutions which make up the four main branches and six clades of the phylogenetic tree for donor PD44716 (**Fig. 4.25**). Branch 1 = *NOTCH1*$^{-/-}$ and *ARID2*$^{-}$; Branch 2 = *TP53*$^{-}$; Branch 3 = *KLF5*$^{-}$ and *DNAH1*$^{-}$; Branch 4 = *NOTCH1*^{s} and *MTOR*$^{-}$.

Discussion

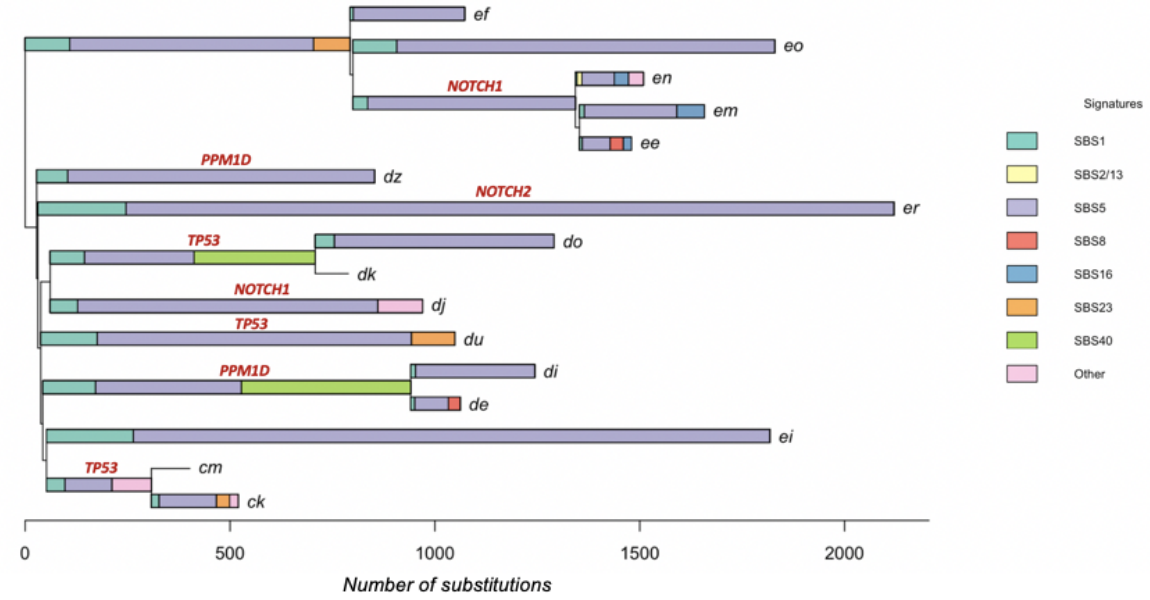
This chapter describes the somatic mutational landscape of 29.5 cm^2 of normal oesophageal epithelium sampled across 20 donors over the age of 60 and compares this landscape to that of published samples taken from a range of donor ages by Martincorena, Fowler *et. al* (2018). Mutation burden, clone size and the percentage of tissue mutant for drivers of clonal expansion were highly variable across the aged donors.

The sampling of tissue within this chapter comprises the largest dataset of somatic mutations in normal oesophageal epithelium to date, building on previous work to enable the detection of two additional genes likely to be under positive selection in the tissue, *EPHA2* and *CREBBP*. *EPHA2* encodes the tyrosine kinase EphA2, involved in a complex network of cell signalling including control over cell proliferation and adhesion (Xiao et al. 2020). *CREBBP* encodes the CREB-binding protein which has roles in modifying transcription factors and promoting cell proliferation in squamous cell carcinomas of the head and neck (Kartha et al. 2018).

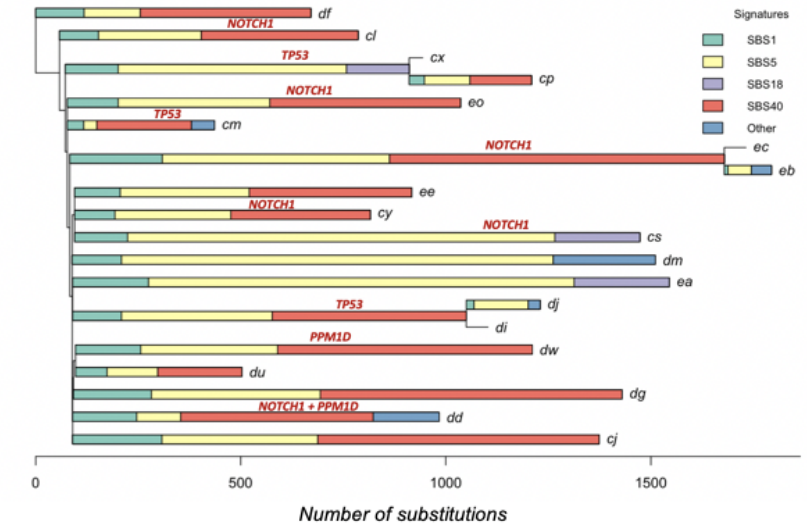
PD44715: 75M, non-smoker, moderate drinker



PD43384: 75F, non-smoker, occasional drinker



PD42785: 61M, non-smoker, occasional drinker



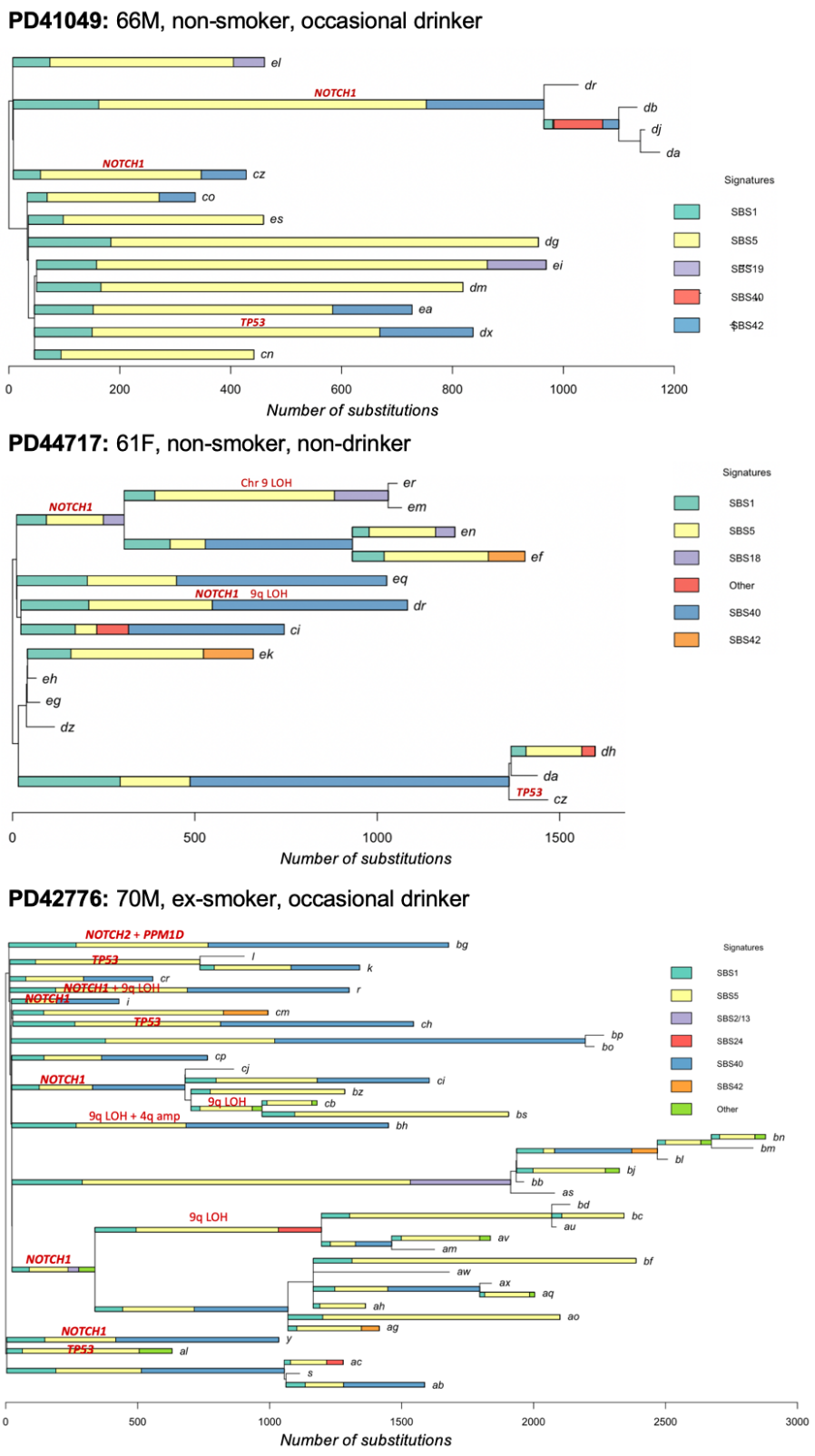


Figure 4.27: Phylogenetic trees depicting the relatedness between whole-genome punch samples of six additional donors. Branch lengths equal the number of substitutions shared between samples, where each tip of the tree is a single circular punch measuring 0.05 mm^2 in area and colours indicate the contribution of substitutions assigned to reference signatures. Known copy number changes and drivers of clonal expansion in each clone are labelled.

A comparison of genes under selection in donors under and over the age of 60 revealed differences in competition within the mutational landscape with donor age. Non-synonymous mutations in *TP53* and *FAT1* were over-represented in donors over 60 years. Mutations in *TP53* are found in over 90% of ESCCs, suggesting mutant *TP53* is a key step required for carcinogenesis. Non-synonymous mutations in *NOTCH3* and missense mutations in both *NOTCH1* and *NOTCH2* were significantly under-represented in donors over 60 years. This suggests clones mutant for *NOTCH3* and those harbouring a missense mutation in *NOTCH1* and *NOTCH2* are relatively weak drivers of clonal expansion. In contrast, as the tissue ages and the burden of mutant clones increases, stronger drivers of clonal expansion, such as clones mutant for *FAT1* and *TP53,* are able to outcompete weaker clones, at least partially explaining the increased risk of carcinogenesis in this age group.

I used the recently developed method of Nanorate sequencing (Abascal et al. 2021) in order to gain the most accurate measure of mutational burden to date. Estimated substitution burden showed a linear increase with age, consistent with previous studies (Martincorena, Fowler, et al. 2018; Yokoyama et al. 2019). Aside from one donor (PD44716) who was a heavy alcohol drinker (>9 units/day), there was little evidence to suggest that alcohol contributes to genome-wide burden in either Nano-seq or WGS data of moderate to light drinkers. However, mutational signature analysis of targeted data identified contribution to substitutions by SBS16 (a signature characterised in liver cancers and associated with alcohol consumption) in moderate and light alcohol drinkers, consistent with the bias of this signature to the transcribed strand only of transcriptionally active regions (L. Alexandrov et al. 2018). Furthermore, this contribution of SBS16 showed an increase with reported alcohol consumption.

Mapping of targeted data in both 2 mm^2 grids and 0.5 mm^2 punches showed the donor with high alcohol consumption to have a mutational landscape dominated by very large clones in comparison to other donors. Through subsequent WGS, it seems these clones have persisted in the tissue for a long time, before undergoing a more sudden clonal expansion, perhaps due to the acquisition of a driver mutation and/or due to changes in the tissue environment with age. However, one main limitation of this study is that this donor serves as an anomaly, and otherwise, there is no evidence to suggest alcohol changes selection and competition in the mutational landscape of the tissue. Future studies could expand this dataset to include more samples from donors that are heavy alcohol drinkers (>6 units/day).

Smoking was found to significantly increase both genome-wide substitution and indel burden, with current smokers and then ex-smokers having an increased burden compared to

non-smokers. However, there is no evidence that smoking leads to a difference in mutational base contexts and signatures. This suggests smoking doesn't increase cancer risk in the oesophagus through mutagenesis, but through an indirect process, such as changing cell behaviour or through epigenetic mechanisms. It is also possible that smoking status is confounded by other lifestyle factors or behaviours which may contribute to a higher global mutation burden in the tissue. The finding that alcohol, but not smoking, acts a mutagen in oesophageal epithelium is consistent with both the study of mutational signatures in normal oesophagus and ESCCs by Yokoyama *et al.* (2019) and work by the Mutographs team, which sequenced ESCCs from eight countries of varying incidence (Moody et al. 2021). Both studies found no difference in substitution signature with smoking but an association of SBS16 with alcohol consumption in Japanese donors.

Chapter 5: Discussion

Keratinocyte skin cancers and oesophageal squamous cell carcinoma both develop from mutant clones of squamous epithelium, yet are remodelled during ageing by different mutagenic processes and environmental exposures. In this thesis, I characterise the mutational landscape of normal skin and oesophagus and find differences associated with respective cancer risk factors and incidence.

I compare the mutational landscape of sun-exposed normal skin in the UK to that of Singapore, which has a 17-fold lower age-adjusted keratinocyte cancer incidence. I find a 4-fold higher mutation burden and 10-fold increase in copy number aberrant clones in the UK compared to Singapore. The majority of mutations in UK skin are caused by UV damage however, in Singapore, endogenous processes predominate. I estimate UK skin to be nearing saturation with mutant clones and observe a larger median clone size in Singaporean skin consistent with this. Mutations in *TP53* are more strongly selected in the UK, whilst those in *NOTCH1* and *NOTCH2* are preferentially selected in Singapore, which further evidences increased competition in UK skin. I conclude that differences in keratinocyte cancer incidence between countries are reflected in the mutational landscape of normal skin. Future study could compare DNA sequencing of keratinocyte cancers from lower-incidence countries, such as Singapore, to determine if the differences observed in the mutational landscape of normal epidermis in this thesis are reflected in the frequency of genomic aberrations in the tumour. In normal epidermis, I observe mutations in specific *TP53* codons that are common to cancers in the UK epidermis, but not in Singapore. Observing if these specific *TP53* codons are prevalent in keratinocyte cancers of low-incidence countries would give insight into if they drive carcinogenesis or their prevalence in cancer is merely a reflection of the frequency in normal sun-exposed epithelium. In general, our understanding of how genetic variation between individuals confers protection against cancer is poor. Polymorphisms specific to the *HYAL2* region in Singaporeans (Ding et al. 2014) and DNA repair in African populations (Crawford et al. 2017) are two examples of suggested biological mechanisms, unrelated to pigmentation, that confer protection against keratinocyte cancer. Better study of low-risk populations could aid in the development of protective therapies for high-risk groups.

A comparison of UK skin by body site reveals differences in UV signature with the incidence of keratinocyte cancer at that site, perhaps reflecting a difference in DNA repair processes at more frequently UV-exposed sites. Targeted sequencing of forearm epidermis reveals

burden estimates of 10s of mutations/Mb, in the range observed in some cSCCs. High resolution mapping of punch whole genomes shows that cells of physiologically normal epidermis are able to withstand a burden of over 60,000 substitutions each, again comparable to burden estimates for keratinocyte cancers, in addition to multiple copy number aberrations at loci of positively selected genes. Previous study of *TP53* mutant keratinocyte clones suggests such clones switch cell fate to increased proliferation in response to UV radiation, however, over time, cell fate reverts back to equal daughter production (Jonason et al. 1996). This reversion to homeostasis could explain how such heavily mutated epidermis is able to retain physiological function. This does mean, however, that if normal epidermis is able to sustain a high burden of mutation and copy number aberration, and mutational signatures are comparable to those observed in cancers, it is not clear what further steps are required for carcinogenesis.

The large area of aged oesophageal epithelium sampled in this thesis has allowed the detection of two additional genes, *EPHA2* and *CREBBP,* under positive selection in normal tissue. Both have roles in cell proliferation and are therefore plausible drivers of clonal expansion. Furthermore, I find that mutations in both *TP53* and *FAT1* are more strongly selected, whilst those in *NOTCH3* more weakly selected, in oesophagus over the age of 60 years, suggesting changes to selection and levels of competition within the tissue with age.

The use of Nano-seq in this thesis has provided the most accurate estimates of mutation burden in normal oesophageal epithelium to date. Mutation burden increases with smoking, but I find no evidence of a smoking-associated mutational signature, suggesting that, if tobacco smoke increases oesophageal cancer risk, it does so without direct action as a mutagen. I do not detect a difference in clone size or selection with smoking, however, it could be that subtle changes in clonal dynamics are revealed through larger sample sizes of smokers and non-smokers. Future studies should look to characterise the epigenetic landscape of normal oesophagus, as it is possible for example, that smoking acts to increase ESCC risk through altering DNA methylation. It is also possible that smoking status is confounded by other lifestyle factors or behaviours which contribute to a higher global mutation burden in the tissue.

I observe a moderate increase in mutations of the alcohol-associated signature, SBS16, with reported alcohol consumption. This finding supports previous analyses of oesophageal epithelium and ESCCs from both European (Martincorena, Fowler, et al. 2018) and Asian (Chang et al. 2017; Yokoyama et al. 2019; Moody et al. 2021) donors. However, it is unlikely this small increase in burden explains the relatively large increased ESCC risk with alcohol

consumption. Unfortunately, only one donor in this study was a heavy drinker, but this donor has a high proportion of very large clones, unlike any other individual. It would be extremely interesting to continue this sampling of aged epithelium in the oesophagus of heavy drinkers to determine if this variation in mutational landscape in one donor is an anomaly or an effect of alcohol on clonal expansion. As in previous work on normal oesophageal epithelium (Yokoyama et al. 2019; Martincorena, Fowler, et al. 2018), I observe a high proportion of C>A mutations. The trinucleotide context of these mutations has similarities with SBS18, which is possibly a consequence of damage by reactive oxygen species (L. Alexandrov et al. 2018), although there is as of yet no direct evidence of this.

High resolution mapping of a heavy smoker and drinker found a substitution burden of some cells to be ~2,500 mutations/genome. This is within the lower range of that observed in some ESCCs, which is surprising, considering no APOBEC or DNA repair related mutational signatures were present in this donor. Across all donors, three samples of the same clone show evidence of APOBEC mutagenesis. This is consistent with the hypothesis of APOBEC activity being an early event on the path of ESCC carcinogenesis (Moody et al. 2021), however, there was no evidence of *TP53* mutation in this clone. Analysis of ESCC genomes from countries of varying incidence did not reveal a difference in mutational signature to explain this increased risk (Moody et al. 2021). This thesis characterises the aged oesophageal epithelium from a relatively low ESCC-risk country, the UK. Future studies can now expand this sampling to normal, aged esophageal epithelium from countries of high ESCC-risk. If environmental exposures increase ESCC risk through mutagenesis, any difference in signature is more likely to be detected in normal tissue, before the onset of APOBEC activity or aberrant DNA repair. Furthermore, the sampling of somatic mutations across normal epithelium can give insight into genes under selection, clone size and insight into environmental risk factors that act by altering clonal dynamics.

References

(1) https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer

Abascal, Federico, Luke M. R. Harvey, Emily Mitchell, Andrew R. J. Lawson, Stefanie V. Lensing, Peter Ellis, Andrew J. C. Russell, et al. 2021. “Somatic Mutation Landscapes at Single-Molecule Resolution.” *Nature* 593 (7859): 405–10.

Abnet, Christian C., Melina Arnold, and Wen-Qiang Wei. 2018. “Epidemiology of Esophageal Squamous Cell Carcinoma.” *Gastroenterology* 154 (2): 360–73.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020a. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Alexandrov, Ludmil B., and Michael R. Stratton. 2014. “Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes.” *Current Opinion in Genetics & Development* 24 (February): 52–60.

Alexandrov, Ludmil, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin W. T. Ng, Arnoud Boot, Kyle R. Covington, et al. 2018. “The Repertoire of Mutational Signatures in Human Cancer.” https://doi.org/10.1101/322859.

Armstrong, Bruce K., and Anne Kricker. 2001. “The Epidemiology of UV Induced Skin Cancer.” *Journal of Photochemistry and Photobiology B: Biology*. https://doi.org/10.1016/s1011-1344(01)00198-1.

Arnold, Melina, Isabelle Soerjomataram, Jacques Ferlay, and David Forman. 2015. “Global Incidence of Oesophageal Cancer by Histological Subtype in 2012.” *Gut* 64 (3): 381–87.

Asombang, Akwi W., Nathaniel Chishinga, Alick Nkhoma, Jackson Chipaila, Bright Nsokolo, Martha Manda-Mapalo, Joao Filipe G. Montiero, Lewis Banda, and Kulwinder S. Dua. 2019. “Systematic Review and Meta-Analysis of Esophageal Cancer in Africa: Epidemiology, Risk Factors, Management and Outcomes.” *World Journal of Gastroenterology: WJG* 25 (31): 4512–33.

Barbera, Mariagnese, Massimiliano di Pietro, Elaine Walker, Charlotte Brierley, Shona MacRae, Benjamin D. Simons, Phil H. Jones, John Stingl, and Rebecca C. Fitzgerald. 2015. “The Human Squamous Oesophagus Has Widespread Capacity for Clonal Expansion from Cells at Diverse Stages of Differentiation.” *Gut* 64 (1): 11–19.

Batai, Ken, Zuxi Cui, Amit Arora, Ebony Shah-Williams, Wenndy Hernandez, Maria Ruden, Courtney M. P. Hollowell, et al. 2021. “Genetic Loci Associated with Skin Pigmentation in African Americans and Their Effects on Vitamin D Deficiency.” *PLoS Genetics* 17 (2): e1009319.

Blokzijl, Francis, Roel Janssen, Ruben van Boxtel, and Edwin Cuppen. 2018. “MutationalPatterns: Comprehensive Genome-Wide Analysis of Mutational Processes.” *Genome Medicine* 10 (1): 33.

Bonilla, Ximena, Laurent Parmentier, Bryan King, Fedor Bezrukov, Gürkan Kaya, Vincent Zoete, Vladimir B. Seplyarskiy, et al. 2016. “Genomic Analysis Identifies New Drivers and Progression Pathways in Skin Basal Cell Carcinoma.” *Nature Genetics* 48 (4): 398–406.

Brash, D. E., A. Ziegler, A. S. Jonason, J. A. Simon, S. Kunala, and D. J. Leffell. 1996. “Sunlight and Sunburn in Human Skin Cancer: p53, Apoptosis, and Tumor Promotion.” *The Journal of Investigative Dermatology. Symposium Proceedings / the Society for Investigative Dermatology, Inc. [and] European Society for Dermatological Research* 1 (2): 136–42.

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2018. “Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” *CA: A Cancer Journal for Clinicians* 68 (6): 394–424.

Brooks, Philip J., Mary-Anne Enoch, David Goldman, Ting-Kai Li, and Akira Yokoyama. 2009. “The Alcohol Flushing Response: An Unrecognized Risk Factor for Esophageal Cancer from Alcohol Consumption.” *PLoS Medicine* 6 (3): e50.

Cairns, John. 1975. “Mutation Selection and the Natural History of Cancer.” *Nature*. https://doi.org/10.1038/255197a0.

Chang, Jiang, Wenle Tan, Zhiqiang Ling, Ruibin Xi, Mingming Shao, Mengjie Chen, Yingying Luo, et al. 2017. “Genomic Analysis of Oesophageal Squamous-Cell Carcinoma Identifies Alcohol Drinking-Related Mutation Signature and Genomic Alterations.” *Nature Communications* 8 (May): 15290.

Chen, Xi-Xi, Qian Zhong, Yang Liu, Shu-Mei Yan, Zhang-Hua Chen, Shan-Zhao Jin, Tian-Liang Xia,

et al. 2017. “Genomic Comparison of Esophageal Squamous Cell Carcinoma and Its Precursor Lesions by Multi-Region Whole-Exome Sequencing.” *Nature Communications* 8 (1): 524.
Ciążyńska, Magdalena, Grażyna Kamińska-Winciorek, Dariusz Lange, Bogumił Lewandowski, Adam Reich, Martyna Sławińska, Marta Pabianek, et al. 2021. “The Incidence and Clinical Analysis of Non-Melanoma Skin Cancer.” *Scientific Reports* 11 (1): 4337.
Coelho, Sergio G., Wonseon Choi, Michaela Brenner, Yoshinori Miyamura, Yuji Yamaguchi, Rainer Wolber, Christoph Smuda, et al. 2009. “Short- and Long-Term Effects of UV Radiation on the Pigmentation of Human Skin.” *The Journal of Investigative Dermatology. Symposium Proceedings / the Society for Investigative Dermatology, Inc. [and] European Society for Dermatological Research* 14 (1): 32–35.
Consortium, The Icgc/tcga Pan-Cancer Analysis of Whole Genomes, and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature*. https://doi.org/10.1038/s41586-020-1969-6.
Crawford, Nicholas G., Derek E. Kelly, Matthew E. B. Hansen, Marcia H. Beltrame, Shaohua Fan, Shanna L. Bowman, Ethan Jewett, et al. 2017. “Loci Associated with Skin Pigmentation Identified in African Populations.” *Science* 358 (6365). https://doi.org/10.1126/science.aan8433.
Cui, Ri, Yoichiro Kamatani, Atsushi Takahashi, Masayuki Usami, Naoya Hosono, Takahisa Kawaguchi, Tatsuhiko Tsunoda, et al. 2009. “Functional Variants in ADH1B and ALDH2 Coupled with Alcohol and Smoking Synergistically Enhance Esophageal Cancer Risk.” *Gastroenterology* 137 (5): 1768–75.
Dar, N. A., G. A. Bhat, I. A. Shah, B. Iqbal, M. A. Makhdoomi, I. Nisar, R. Rafiq, et al. 2012. “Hookah Smoking, Nass Chewing, and Oesophageal Squamous Cell Carcinoma in Kashmir, India.” *British Journal of Cancer* 107 (9): 1618–23.
Davoli, Teresa, Andrew Wei Xu, Kristen E. Mengwasser, Laura M. Sack, John C. Yoon, Peter J. Park, and Stephen J. Elledge. 2013. “Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome.” *Cell* 155 (4): 948–62.
Ding, Qiliang, Ya Hu, Shuhua Xu, Jiucun Wang, and Li Jin. 2014. “Neanderthal Introgression at Chromosome 3p21.31 Was under Positive Natural Selection in East Asians.” *Molecular Biology and Evolution* 31 (3): 683–95.
Druesne-Pecollo, Nathalie, Bertrand Tehard, Yann Mallet, Mariette Gerber, Teresa Norat, Serge Hercberg, and Paule Latino-Martel. 2009. “Alcohol and Genetic Polymorphisms: Effect on Risk of Alcohol-Related Cancer.” *The Lancet Oncology*. https://doi.org/10.1016/s1470-2045(09)70019-1.
Ellis, Peter, Luiza Moore, Mathijs A. Sanders, Timothy M. Butler, Simon F. Brunner, Henry Lee-Six, Robert Osborne, et al. 2021. “Reliable Detection of Somatic Mutations in Solid Tissues by Laser-Capture Microdissection and Low-Input DNA Sequencing.” *Nature Protocols* 16 (2): 841–71.
Farmery, James H. R., Mike L. Smith, NIHR BioResource - Rare Diseases, and Andy G. Lynch. 2018. “Telomerecat: A Ploidy-Agnostic Method for Estimating Telomere Length from Whole Genome Sequencing Data.” *Scientific Reports* 8 (1): 1300.
Ferrucci, Leah M., Brenda Cartmel, Annette M. Molinaro, David J. Leffell, Allen E. Bale, and Susan T. Mayne. 2012. “Indoor Tanning and Risk of Early-Onset Basal Cell Carcinoma.” *Journal of the American Academy of Dermatology* 67 (4): 552–62.
Fioletov, Vitali, James B. Kerr, and Angus Fergusson. 2010. “The UV Index: Definition, Distribution and Factors Affecting It.” *Canadian Journal of Public Health. Revue Canadienne de Sante Publique* 101 (4): I5–9.
Fowler, Joanna C., Charlotte King, Christopher Bryant, Michael W. J. Hall, Roshan Sood, Swee Hoe Ong, Eleanor Earp, et al. 2021. “Selection of Oncogenic Mutant Clones in Normal Human Skin Varies with Body Site.” *Cancer Discovery* 11 (2): 340–61.
Frede, Julia, Philip Greulich, Tibor Nagy, Benjamin D. Simons, and Philip H. Jones. 2016. “A Single Dividing Cell Population with Imbalanced Fate Drives Oesophageal Tumour Growth.” *Nature Cell Biology*. https://doi.org/10.1038/ncb3400.
Gallagher, Richard P., Tim K. Lee, Chris D. Bajdik, and Marilyn Borugian. 2010. “Ultraviolet Radiation.” *Chronic Diseases and Injuries in Canada*. https://doi.org/10.24095/hpcdp.29.s1.04.
Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, et al. 2020. “The Evolutionary History of 2,658 Cancers.” *Nature* 578 (7793): 122–28.
Gerstung, Moritz, Elli Papaemmanuil, and Peter J. Campbell. 2014. “Subclonal Variant Calling with Multiple Samples and Prior Knowledge.” *Bioinformatics* 30 (9): 1198–1204.
Giglia-Mari, Giuseppina, and Alain Sarasin. 2003. “TP53 Mutations in Human Skin Cancers.” *Human Mutation* 21 (3): 217–28.

Green, A. C., and C. M. Olsen. 2017. “Cutaneous Squamous Cell Carcinoma: An Epidemiological Review.” *The British Journal of Dermatology* 177 (2): 373–81.
Grossmann, Sebastian, Yvette Hooks, Laura Wilson, Luiza Moore, Laura O’Neill, Iñigo Martincorena, Thierry Voet, Michael R. Stratton, Rakesh Heer, and Peter J. Campbell. 2021. “Development, Maturation, and Maintenance of Human Prostate Inferred from Somatic Mutations.” *Cell Stem Cell* 28 (7): 1262–74.e5.
Hartevelt, M. M., J. N. Bavinck, A. M. Kootte, B. J. Vermeer, and J. P. Vandenbroucke. 1990. “Incidence of Skin Cancer after Renal Transplantation in The Netherlands.” *Transplantation* 49 (3): 506–9.
Hoang, Diep Thi, Le Sy Vinh, Tomáš Flouri, Alexandros Stamatakis, Arndt von Haeseler, and Bui Quang Minh. 2018. “MPBoot: Fast Phylogenetic Maximum Parsimony Tree Inference and Bootstrap Approximation.” *BMC Evolutionary Biology* 18 (1): 11.
Inman, Gareth J., Jun Wang, Ai Nagano, Ludmil B. Alexandrov, Karin J. Purdie, Richard G. Taylor, Victoria Sherwood, et al. 2018. “The Genomic Landscape of Cutaneous SCC Reveals Drivers and a Novel Azathioprine Associated Mutational Signature.” *Nature Communications*. https://doi.org/10.1038/s41467-018-06027-1.
International Cancer Genome Consortium, Thomas J. Hudson, Warwick Anderson, Axel Artez, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, et al. 2010. “International Network of Cancer Genome Projects.” *Nature* 464 (7291): 993–98.
Jonason, A. S., S. Kunala, G. J. Price, R. J. Restifo, H. M. Spinelli, J. A. Persing, D. J. Leffell, R. E. Tarone, and D. E. Brash. 1996. “Frequent Clones of p53-Mutated Keratinocytes in Normal Human Skin.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (24): 14025–29.
Jones, David, Keiran M. Raine, Helen Davies, Patrick S. Tarpey, Adam P. Butler, Jon W. Teague, Serena Nik-Zainal, and Peter J. Campbell. 2016. “cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 56 (December): 15.10.1–15.10.18.
Kartha, Vinay K., Khalid A. Alamoud, Khikmet Sadykov, Bach-Cuc Nguyen, Fabrice Laroche, Hui Feng, Jina Lee, et al. 2018. “Functional and Genomic Analyses Reveal Therapeutic Potential of Targeting β-catenin/CBP Activity in Head and Neck Cancer.” *Genome Medicine* 10 (1): 54.
Kgomo, Mpho, Ali A. Elnagar, Jaco Nagel, and Taole Mokoena. 2017. “Prevalence of Squamous Cell Carcinoma of the Esophagus in a Single Tertiary Center of South Africa: A Cross Sectional Analytic Study.” *Journal of Public Health in Africa* 8 (1). https://doi.org/10.4081/jphia.2017.563.
Kim, Grace K., James Q. Del Rosso, and Susun Bellew. 2009. “Skin Cancer in Asians: Part 1: Nonmelanoma Skin Cancer.” *The Journal of Clinical and Aesthetic Dermatology* 2 (8): 39–42.
Koh, D., H. Wang, J. Lee, K. S. Chia, H. P. Lee, and C. L. Goh. 2003. “Basal Cell Carcinoma, Squamous Cell Carcinoma and Melanoma of the Skin: Analysis of the Singapore Cancer Registry Data 1968-97.” *British Journal of Dermatology*. https://doi.org/10.1046/j.1365-2133.2003.05223.x.
Kryazhimskiy, Sergey, and Joshua B. Plotkin. 2008. “The Population Genetics of dN/dS.” *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1000304.
Lao-Sirieix, Pierre, Carlos Caldas, and Rebecca C. Fitzgerald. 2010. “Genetic Predisposition to Gastro-Oesophageal Cancer.” *Current Opinion in Genetics & Development* 20 (3): 210–17.
Lee-Six, Henry, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, et al. 2018. “Population Dynamics of Normal Human Blood Inferred from Somatic Mutations.” *Nature* 561 (7724): 473–78.
Lee-Six, Henry, Sigurgeir Olafsson, Peter Ellis, Robert J. Osborne, Mathijs A. Sanders, Luiza Moore, Nikitas Georgakopoulos, et al. 2019. “The Landscape of Somatic Mutation in Normal Colorectal Epithelial Cells.” *Nature* 574 (7779): 532–37.
Leisenring, Wendy, Debra L. Friedman, Mary E. D. Flowers, Jeffrey L. Schwartz, and H. Joachim Deeg. 2006. “Nonmelanoma Skin and Mucosal Cancers after Hematopoietic Cell Transplantation.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 24 (7): 1119–26.
Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.
Lin, De-Chen, Jia-Jie Hao, Yasunobu Nagata, Liang Xu, Li Shang, Xuan Meng, Yusuke Sato, et al. 2014. “Genomic and Molecular Characterization of Esophageal Squamous Cell Carcinoma.” *Nature Genetics* 46 (5): 467–73.
Liu, Xi, Min Zhang, Songmin Ying, Chong Zhang, Runhua Lin, Jiaxuan Zheng, Guohong Zhang, et al.

2017. “Genetic Alterations in Esophageal Tissues From Squamous Dysplasia to Carcinoma.” *Gastroenterology* 153 (1): 166–77.
Long, Millie D., Christopher F. Martin, Clare A. Pipkin, Hans H. Herfarth, Robert S. Sandler, and Michael D. Kappelman. 2012. “Risk of Melanoma and Nonmelanoma Skin Cancer among Patients with Inflammatory Bowel Disease.” *Gastroenterology* 143 (2): 390–99.e1.
Maloney, M. E. 1996. “Arsenic in Dermatology.” *Dermatologic Surgery: Official Publication for American Society for Dermatologic Surgery [et Al.]* 22 (3): 301–4.
Martincorena, Iñigo, and Peter J. Campbell. 2015. “Somatic Mutation in Cancer and Normal Cells.” *Science* 349 (6255): 1483–89.
Martincorena, Iñigo, Joanna C. Fowler, Agnieszka Wabik, Andrew R. J. Lawson, Federico Abascal, Michael W. J. Hall, Alex Cagan, et al. 2018. “Somatic Mutant Clones Colonize the Human Esophagus with Age.” *Science* 362 (6417): 911–17.
Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2017. “Universal Patterns of Selection in Cancer and Somatic Tissues.” *Cell*. https://doi.org/10.1016/j.cell.2017.09.042
Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C. Wedge, et al. 2015. “Tumor Evolution. High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin.” *Science* 348 (6237): 880–86.
Menzies, Andy, Jon W. Teague, Adam P. Butler, Helen Davies, Patrick Tarpey, Serena Nik-Zainal, and Peter J. Campbell. 2015. “VAGrENT: Variation Annotation Generator.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 52 (December): 15.8.1–15.8.11.
Moody, Sarah, Sergey Senkin, S. M. Ashiqul Islam, Jingwei Wang, Dariush Nasrollahzadeh, Ricardo Cortez Cardoso Penha, Stephen Fitzgerald, et al. 2021. “Mutational Signatures in Esophageal Squamous Cell Carcinoma from Eight Countries with Varying Incidence.” *Nature Genetics* 53 (11): 1553–63.
Murai, Kasumi, Greta Skrupskelyte, Gabriel Piedrafita, Michael Hall, Vasiliki Kostiou, Swee Hoe Ong, Tibor Nagy, et al. 2018. “Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations.” *Cell Stem Cell* 23 (5): 687–99.e8.
Murphy, G., V. McCormack, B. Abedi-Ardekani, M. Arnold, M. C. Camargo, N. A. Dar, S. M. Dawsey, et al. 2017. “International Cancer Seminars: A Focus on Esophageal Squamous Cell Carcinoma.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 28 (9): 2086–93.
Muzic, John G., Adam R. Schmitt, Adam C. Wright, Dema T. Alniemi, Adeel S. Zubair, Jeannette M. Olazagasti Lourido, Ivette M. Sosa Seda, Amy L. Weaver, and Christian L. Baum. 2017. “Incidence and Trends of Basal Cell Carcinoma and Cutaneous Squamous Cell Carcinoma: A Population-Based Study in Olmsted County, Minnesota, 2000 to 2010.” *Mayo Clinic Proceedings. Mayo Clinic* 92 (6): 890–98.
Perry, David M., Virginia Barton, and Anthony J. Alberg. 2017. “Epidemiology of Keratinocyte Carcinoma.” *Current Dermatology Reports* 6 (3): 161–68.
Que, Syril Keena T., Fiona O. Zwald, and Chrysalyne D. Schmults. 2018. “Cutaneous Squamous Cell Carcinoma: Incidence, Risk Factors, Diagnosis, and Staging.” *Journal of the American Academy of Dermatology* 78 (2): 237–47.
Raine, Keiran M., Jonathan Hinton, Adam P. Butler, Jon W. Teague, Helen Davies, Patrick Tarpey, Serena Nik-Zainal, and Peter J. Campbell. 2015. “cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 52 (December): 15.7.1–12.
Raine, Keiran M., Peter Van Loo, David C. Wedge, David Jones, Andrew Menzies, Adam P. Butler, Jon W. Teague, Patrick Tarpey, Serena Nik-Zainal, and Peter J. Campbell. 2016. “ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 56 (December): 15.9.1–15.9.17.
Riet, P. van der, D. Karp, E. Farmer, Q. Wei, L. Grossman, K. Tokino, J. M. Ruppert, and D. Sidransky. 1994. “Progression of Basal Cell Carcinoma through Loss of Chromosome 9q and Inactivation of a Single p53 Allele.” *Cancer Research* 54 (1): 25–27.
Riva, Laura, Arun R. Pandiri, Yun Rose Li, Alastair Droop, James Hewinson, Michael A. Quail, Vivek Iyer, et al. 2020. “The Mutational Signature Profile of Known and Suspected Human Carcinogens in Mice.” *Nature Genetics* 52 (11): 1189–97.
Roberts, M. R., M. M. Asgari, and A. E. Toland. 2019. “Genome-Wide Association Studies and Polygenic Risk Scores for Skin Cancer: Clinically Useful Yet?” *The British Journal of Dermatology* 181 (6): 1146–55.

Schmitt, J., A. Seidler, T. L. Diepgen, and A. Bauer. 2011. “Occupational Ultraviolet Light Exposure Increases the Risk for the Development of Cutaneous Squamous Cell Carcinoma: A Systematic Review and Meta-Analysis.” *The British Journal of Dermatology* 164 (2): 291–307.
Schroeff, J. G. van der, L. M. Evers, A. J. Boot, and J. L. Bos. 1990. “Ras Oncogene Mutations in Basal Cell Carcinomas and Squamous Cell Carcinomas of Human Skin.” *The Journal of Investigative Dermatology* 94 (4): 423–25.
Shuck, Sarah C., Emily A. Short, and John J. Turchi. 2008. “Eukaryotic Nucleotide Excision Repair: From Understanding Mechanisms to Influencing Biology.” *Cell Research* 18 (1): 64–72.
Simons, Benjamin D. 2016. “Deep Sequencing as a Probe of Normal Stem Cell Fate and Preneoplasia in Human Epidermis.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (1): 128–33.
Sng, Judy, David Koh, Wong Chia Siong, and Tai Bee Choo. 2009. “Skin Cancer Trends among Asians Living in Singapore from 1968 to 2006.” *Journal of the American Academy of Dermatology* 61 (3): 426–32.
Song, Hoseok, Monica Hollstein, and Yang Xu. 2007. “p53 Gain-of-Function Cancer Mutants Induce Genetic Instability by Inactivating ATM.” *Nature Cell Biology* 9 (5): 573–80.
Song, Yongmei, Lin Li, Yunwei Ou, Zhibo Gao, Enmin Li, Xiangchun Li, Weimin Zhang, et al. 2014. “Identification of Genomic Alterations in Oesophageal Squamous Cell Cancer.” *Nature* 509 (7498): 91–95.
South, Andrew P., Karin J. Purdie, Stephen A. Watt, Sam Haldenby, Nicoline den Breems, Michelle Dimon, Sarah T. Arron, et al. 2014. “NOTCH1 Mutations Occur Early during Cutaneous Squamous Cell Carcinogenesis.” *The Journal of Investigative Dermatology* 134 (10): 2630–38.
Subramaniam, Padmini, Catherine M. Olsen, Bridie S. Thompson, David C. Whiteman, Rachel E. Neale, and for the QSkin Sun and Health Study Investigators. 2017. “Anatomical Distributions of Basal Cell Carcinoma and Squamous Cell Carcinoma in a Population-Based Study in Queensland, Australia.” *JAMA Dermatology*. https://doi.org/10.1001/jamadermatol.2016.4070.
Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21660.
Trakatelli, M., C. Ulrich, V. del Marmol, S. Euvrard, E. Stockfleth, and D. Abeni. 2007. “Epidemiology of Nonmelanoma Skin Cancer (NMSC) in Europe: Accurate and Comparable Data Are Needed for Effective Public Health Monitoring and Interventions.” *British Journal of Dermatology*. https://doi.org/10.1111/j.1365-2133.2007.07861.x.
Urabe, Yuji, Kenichi Kagemoto, C. Nelson Hayes, Koki Nakamura, Kazuhiko Masuda, Atsushi Ono, Shinji Tanaka, Koji Arihiro, and Kazuaki Chayama. 2019. “Genomic Characterization of Early-Stage Esophageal Squamous Cell Carcinoma in a Japanese Population.” *Oncotarget*. https://doi.org/10.18632/oncotarget.27014.
Veierød, Marit B., Elisabeth Couto, Eiliv Lund, Hans-Olov Adami, and Elisabete Weiderpass. 2014. “Host Characteristics, Sun Exposure, Indoor Tanning and Risk of Squamous Cell Carcinoma of the Skin.” *International Journal of Cancer. Journal International Du Cancer* 135 (2): 413–22.
Venables, Z. C., T. Nijsten, K. F. Wong, P. Autier, J. Broggio, A. Deas, C. A. Harwood, et al. 2019. “Epidemiology of Basal and Cutaneous Squamous Cell Carcinoma in the U.K. 2013-15: A Cohort Study.” *The British Journal of Dermatology* 181 (3): 474–82.
Verkouteren, J. A. C., K. H. R. Ramdas, M. Wakkee, and T. Nijsten. 2017. “Epidemiology of Basal Cell Carcinoma: Scholarly Review.” *The British Journal of Dermatology* 177 (2): 359–72.
Von Hoff, Daniel D., Patricia M. LoRusso, Charles M. Rudin, Josina C. Reddy, Robert L. Yauch, Raoul Tibes, Glen J. Weiss, et al. 2009. “Inhibition of the Hedgehog Pathway in Advanced Basal-Cell Carcinoma.” *The New England Journal of Medicine* 361 (12): 1164–72.
Watt, Tanya C., Peter D. Inskip, Kayla Stratton, Susan A. Smith, Stephen F. Kry, Alice J. Sigurdson, Marilyn Stovall, Wendy Leisenring, Leslie L. Robison, and Ann C. Mertens. 2012. “Radiation-Related Risk of Basal Cell Carcinoma: A Report from the Childhood Cancer Survivor Study.” *Journal of the National Cancer Institute* 104 (16): 1240–50.
Weinstein, John N., The Cancer Genome Atlas Research Network, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics*. https://doi.org/10.1038/ng.2764.
Wood, Richard D., Michael Mitchell, John Sgouros, and Tomas Lindahl. 2001. “Human DNA Repair Genes.” *Science*. https://doi.org/10.1126/science.1056154.
Wright, Caradee Y., D. Jean du Preez, Danielle A. Millar, and Mary Norval. 2020. “The Epidemiology

of Skin Cancer and Public Health Strategies for Its Prevention in Southern Africa." *International Journal of Environmental Research and Public Health* 17 (3). https://doi.org/10.3390/ijerph17031017.
Wu, Degang, Jinzhuang Dou, Xiaoran Chai, Claire Bellis, Andreas Wilm, Chih Chuan Shih, Wendy Wei Jia Soon, et al. 2019. "Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore." *Cell* 179 (3): 736–49.e15.
Xiao, Ta, Yuhang Xiao, Wenxiang Wang, Yan Yan Tang, Zhiqiang Xiao, and Min Su. 2020. "Targeting EphA2 in Cancer." *Journal of Hematology & Oncology*. https://doi.org/10.1186/s13045-020-00944-9.
Yang, Xiaorong, Xingdong Chen, Maoqiang Zhuang, Ziyu Yuan, Shuping Nie, Ming Lu, Li Jin, and Weimin Ye. 2017. "Smoking and Alcohol Drinking in Relation to the Risk of Esophageal Squamous Cell Carcinoma: A Population-Based Case-Control Study in China." *Scientific Reports*. https://doi.org/10.1038/s41598-017-17617-2.
Yokoyama, Akira, Nobuyuki Kakiuchi, Tetsuichi Yoshizato, Yasuhito Nannya, Hiromichi Suzuki, Yasuhide Takeuchi, Yusuke Shiozawa, et al. 2019. "Age-Related Remodelling of Oesophageal Epithelia by Mutated Cancer Drivers." *Nature* 565 (7739): 312–17.
Yoshida, Kenichi, Kate H. C. Gowers, Henry Lee-Six, Deepak P. Chandrasekharan, Tim Coorens, Elizabeth F. Maughan, Kathryn Beal, et al. 2020. "Tobacco Smoking and Somatic Mutations in Human Bronchial Epithelium." *Nature* 578 (7794): 266–72.
Yousef, Hani, Mandy Alhajj, and Sandeep Sharma. 2021. "Anatomy, Skin (Integument), Epidermis." In *StatPearls*. Treasure Island (FL): StatPearls Publishing.
Zhang, Y., S. V. Coillie, J-Y Fang, and J. Xu. 2016. "Gain of Function of Mutant p53: R282W on the Peak?" *Oncogenesis* 5 (February): e196.