University of Southampton

Faculty of Medicine

Clinical and Experimental Sciences

Analysis of cancerous and pre-cancerous skin lesions to generate a repository for studying the mutation burden of UV

by

Noeline Dharini Nadarajah

Thesis for the degree of Doctor of Philosophy

September 2022

University of Southampton

Abstract

Faculty of Medicine

Clinical and Experimental Sciences

Doctor of Philosophy

Analysis of cancerous and pre-cancerous skin lesions to generate a repository for studying the mutation burden of UV

by

Noeline Dharini Nadarajah

Exposure of skin to ultraviolet radiation (UVR) can cause DNA mutations in skin and subsequent development of skin cancer, but UVR is also used is as a treatment for skin diseases. However, it is not known how many courses of UVR for skin disease a patient can receive over their lifetime without being at high risk of developing skin cancer. One way to approach this is to identify all the genetic mutations in skin cancers, and to determine which mutated genes are driver genes, i.e. genes in which mutations promote the development of cancer. Following that, one could use this information to look at mutations in the skin before and after a course of UVR treatment for skin disease to assess the amount of mutational damage in driver genes from that UVR course, in order to estimate the number of courses of UVR that patients with skin disease could safely have in their lifetime.

This thesis used a bioinformatics approach to document the genetic mutations in skin cancer, including cutaneous squamous cell cancer (cSCC), basal cell cancer (BCC) and melanoma in order to identify driver genes in these cancers. Along the way, the genetic mutations in squamous cell cancers (SCCs) of four other organs (lung, oesophagus, oropharynx and cervix) were documented to allow comparison of the driver genes in cSCC with SCCs of these other organs.

Whole genome and whole exome sequencing data were identified from online genetic databases and literature searches. Driver genes and mutation signatures were extracted from this data for all the aforementioned cancers. Linux was used for data manipulation and R was used for data analysis. The results of this bioinformatic analysis identified driver genes in each of the three

types of skin cancer and that most of the driver genes in cSCC, BCC and melanoma differed between these cancers. However, some driver genes were common to more than one type of skin cancer, including *TP53* as a driver gene in the three different types of skin cancer, *CDKN2A* as a driver gene in both cSCC and melanoma, *PPP6C* as a driver gene in BCC and melanoma, and *CDC27* and *TMEM222* as driver genes in cSCC and BCC. In the comparison of cSCC with SCCs of the other organs, six driver genes (*TP53*, *CDKN2A*, *FAT1*, *HRAS*, *NOTCH1* and *NOTCH2*) in cSCC were noted as driver genes in one or more of the other cSCC types.

Whole exome sequencing data from precancerous skin lesions and targeted sequencing data from chronically sun exposed skin and chronically sun exposed normal melanocytes were also analysed. This data was used to identify driver genes, that were present in cSCC, BCC and melanoma, in these precancerous skin lesions as well as in the chronically sun exposed skin/melanocytes. This allowed the generation of a repository or "adjunct" to a future pipeline for assessing the carcinogenicity of UVR treatment for skin disease. Specifically, this repository will assist in assessing whether mutated genes in skin after a course of UVR treatment, in comparison with skin prior to that course of UVR, are likely to be driver genes (i.e. promoting skin cancer development). By comparing the number of driver genes mutated in skin after one course of UVR therapy with the number of driver genes mutated in non-cancerous skin of people with skin cancer, it is hoped that one can estimate the number of courses of UVR therapy for skin disease that patients can safely have without significantly increasing their risk of skin cancer development. In this way, the data in this thesis could be used to help inform clinical practice on the maximum number of UVR courses for skin disease that dermatology patients should have in their lifetime.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Noeline Dharini Nadarajah

Title of thesis: Analysis of cancerous and pre-cancerous skin lesions to generate a repository for studying the mutation burden of UV

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature: ....Date: 05/09/2022...........................

Acknowledgements

I would like to thank Professor Eugene Healy for giving me the opportunity to work under his supervision. I would also like to thank Dr Mat Rose-Zerilli for his guidance during the bioinformatics component of this project and Professor John Holloway for his advice during this PhD. I am also grateful to Dr Jane Gibson for her bioinformatics input. I would also like to thank Dr Chester Lai and Dr George Coltart for their support and assistance in this study.

I would like to express my utmost gratitude to my parents, Alex and Malini Nadarajah, who have always highlighted the importance of education. I would like to especially thank my mother, who has been my best friend and emotional support throughout my academic journey.

Finally, I would like to thank the Psoriasis Association for funding this PhD.

Definitions and abbreviations

AK Actinic Keratosis

BCC Basal Cell Carcinoma

BD Bowen's Disease

bp base pair

BRAF Proto-Oncogene, Serine/Threonine kinase gene

cm^2 Square centimetre

COSMIC Catalogue Of Somatic Mutation In Cancer

cSCC cutaneous Squamous Cell Carcinoma

DNA DeoxyriboNucleic Acid

EGFR Epidermal Growth Factor Receptor protein

Gb Gigabase

GRCh37 Genome Reference Consortium 37

HPV Human Papilloma Virus

HRAS Harvey Rat Sarcoma viral oncogene homolog gene

ICGC International Cancer Genome Consortium

Kb Kilobase

KRAS Kirsten Rat Sarcoma viral oncogene homolog gene

LOH Loss Of Heterozygosity

MAPK Mitogen-Activated Protein Kinase protein

Mb Megabase

MC1R MelanoCortin 1 Receptor gene

mRNA messenger RNA

µm Micrometre

NCBI National Center for Biotechnology Information

NF1 NeuroFibromin 1

NGS Next Generation Sequencing

NMSC Non Melanoma Skin Cancer

NRAS Neuroblastoma RAS viral (v-ras) oncogene homolog gene

PCR Polymerase Chain Reaction

PI3K PhosphatIdylinositide 3-Kinases

PIP P53 Immunopositive Patch

RNA RiboNucleic Acid

SNP Single Nucleotide Polymorphism

TCGA The Cancer Genome Atlas

TERT Telomerase Reverse Transcriptase gene

TP53 Tumour Protein 53

Tregs Regulatory T-cells

UCSC University of California, Santa Cruz genome browser

UV UltraViolet

UVA UltraViolet A

UVB UltraViolet B

UVR UltraViolet Radiation

VAF Variant Allele Frequency

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

XP Xeroderma Pigmentosum

1. Introduction

1.1 Structure and function of the skin

Skin is composed of three main layers: epidermis, dermis and the subcutaneous tissue (Rees, 2004). The outermost layer is the epidermis which provides a waterproof barrier (Alberts, 2002). The dermis lies beneath the epidermis and protects the body from mechanical injury and controls thermal regulation. This layer contains collagen, elastic fibres and most of the skin's structures such as blood vessels, lymph vessels, hair follicles and sweat glands (Haake, 2001). The subcutaneous layer is the innermost layer of the skin and consists of a network of fat cells, where fat is stored as an energy reserve for the body. The blood vessels, nerves, and lymph vessels also cross through this layer (Figure 1-1).

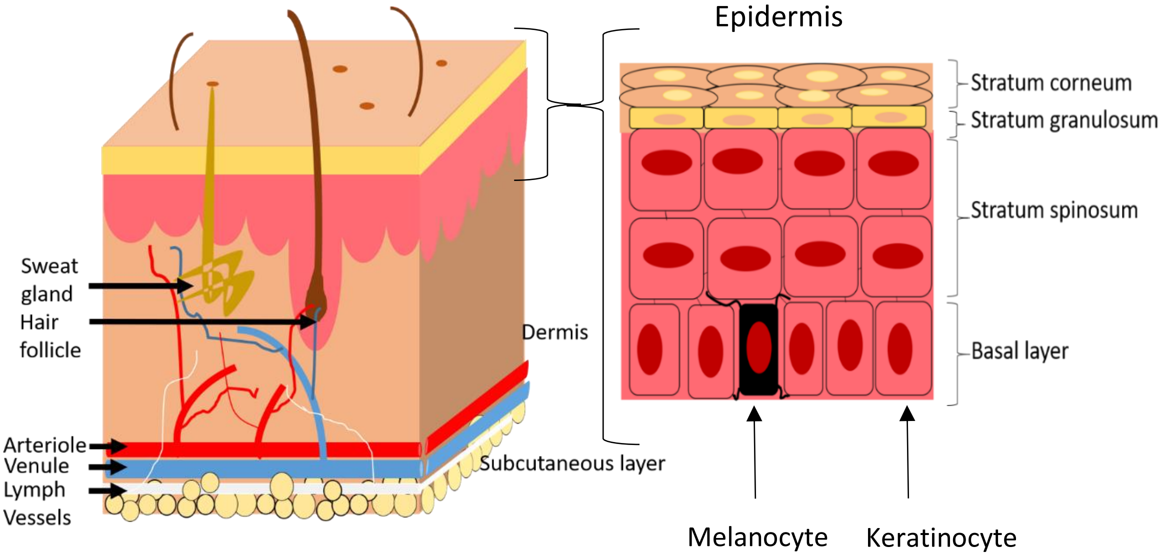


__Figure 1-1: Cross-section of the skin showing the four different layers of the epidermis.__ The left image shows a cross-section of skin and the image on the right shows a labelled diagram of the epidermis.

The epidermis can vary in thickness in different areas of the body, for example the epidermis on the palms and soles is thicker than that on the eyelids. 90% of the epidermal layer consists of keratinocytes, which produce keratin intermediate filaments to protect against trauma (Haake, 2001).

The other main types of cells in the epidermis are melanocytes, Langerhans cells and Merkel cells (Gleason et al., 2008). Langerhans cells are dendritic cells which have a role in immunity (Clayton et al., 2017) and Merkel cells have been associated with neural development and tactile sensation (Moll et al., 1990, Abraham and Mathew, 2019).

Melanocytes produce melanin and are responsible for skin pigmentation (Lin and Fisher, 2007). Melanin is synthesised within melanosomes and ultraviolet radiation (UV) exposure increases melanogenesis (Park et al., 2009). There are two types of melanin; brown-black eumelanin and red-yellow phaeomelanin (Ito and Wakamatsu, 2003). Eumelanin, absorbs UV and helps to prevent DNA photodamage (Brenner and Hearing, 2008) however the function of phaeomelanin is not fully understood (Nasti and Timares, 2015). The amount of melanin pigment in the skin (determined by skin colour) and the effect UV has on the skin, is used to predict the skin type of an individual. The Fitzpatrick classification system is the universal means of skin type characterisation (Fitzpatrick, 1988). Lighter skinned individuals' sunburn easily and tan poorly whereas darker skinned individuals' sunburn less and tan more easily.

The epidermis usually has four sublayers, which are the stratum basale, stratum spinosum, stratum granulosum and stratum corneum (Figure 1-1). The epidermis on the palms and soles has an additional layer called the stratum lucidum which is situated between the stratum granulosum and stratum corneum (Narayan, 2009).

The stratum basale is the innermost layer of the epidermis and consists of melanocytes that rarely undergo mitosis (Cichorek et al., 2013) and are surrounded by keratinocytes which reproduce to provide daughter cells that differentiate as they move towards the stratum corneum where they are eventually shed (Matsui and Amagai, 2015).

The stratum spinosum is located above the basal layer and is characterised by spiny projections, called desmosomes, which hold adjacent cells together (Simpson et al., 2011). It is the thickest layer of the epidermis and contains the former basal cells which have been pushed upwards.

The stratum granulosum has a granular appearance under light microscopy and is composed of keratinocytes that have been pushed up from the stratum spinosum (Simpson et al., 2011). At this stage, the cells become flatter as they differentiate further.

The stratum corneum is the outermost layer of the epidermis consisting of 50 – 150 µm of continually shedding dead keratinocytes. This layer forms a barrier between the skin and the outside world and prevents the passage of water and electrolytes outwards from the body and protects against the entry of microbes and chemicals into the skin (Haake, 2001, Micali, 2001).

The duration for the keratinocytes to travel up from the basal layer to the stratum corneum is, on average, 28 days (Micali, 2001).

1.2 Inflammatory skin conditions

Atopic dermatitis (atopic eczema) is a chronic relapsing inflammatory skin condition which is characterised by itchy papules and vesicles and usually develops in childhood (Williams and Strachan, 1998, Thomsen, 2014). The condition affects up to 30% of children and 3% of adults (Bin et al., 2014). Topical agents such as moisturisers/emollients, corticosteroids and calcineurin inhibitors are used to treat eczema (Chong and Fonacier, 2016). Atopic eczema skin has a less effective outer skin barrier that allows the entry of allergens through the stratum corneum, thus causing an inflammatory response (Hara et al., 2000, Imokawa, 2001). The use of emollients helps restore the skin barrier, reduce the penetration of allergens into the skin and prevent the subsequent development of inflammation (Arkwright et al., 2013). Corticosteroids are medicines that suppress immune responses and thus treat inflammation, whereas calcineurin inhibitors are an alternative class or medications that are used to inhibit the immune system. When these treatment measures have failed to control atopic eczema, the second line of intervention used is phototherapy (Chong and Fonacier, 2016).

Psoriasis is another inflammatory skin condition which is described as presence of red scaly plaques on the skin and arises due to immune activation, inflammation and uncontrolled keratinocyte proliferation and dysfunctional differentiation. It affects 2-3% of the world population (Zhang and Wu, 2018) with 70 – 80% of patients suffering from mild psoriasis (Boehncke and Schon, 2015). The choice of therapy depends on the severity of the condition. Mild psoriasis is treated using a combination of topical agents such as glucocorticoids, vitamin D analogues, tar-based therapies etc. (Rendon and Schakel, 2019).

Vitamin D analogues are a helpful treatment because they regulate immunity and control the proliferation and differentiation of keratinocytes (Barrea et al., 2017). Calcineurin inhibitors are used for psoriasis on the face and in flexures, and corticosteroids are used on a short-term basis for psoriasis on most body sites. UV therapy (termed phototherapy) is used when the psoriasis is more widespread and when topical treatments have failed to improve the skin condition. Systemic drugs such as methotrexate, cyclosporin and biologics are used for moderate to severe psoriasis but can have significant toxicities (Boehncke and Schon, 2015).

Phototherapy was first used as a treatment for the treatment of psoriasis in 1925 (Goeckerman, 1925). In the 1920s, they also observed that eczema improved during the summer, but it was subsequently in 1948 that UV (from carbon arc lamps) was used as an effective treatment for eczema (Rodenbeck et al., 2016).

1.3 UV

UV is a form of electromagnetic radiation and includes UVA (315 nm – 400 nm), UVB (280-315 nm), and UVC (100 – 280 nm) (Narayanan et al., 2010) as displayed in Figure 1-2. The sun emits these three types of UV, however, UVC does not reach the earth's surface due to its absorption by the ozone layer located in the earth's stratosphere (D'Orazio et al., 2013).

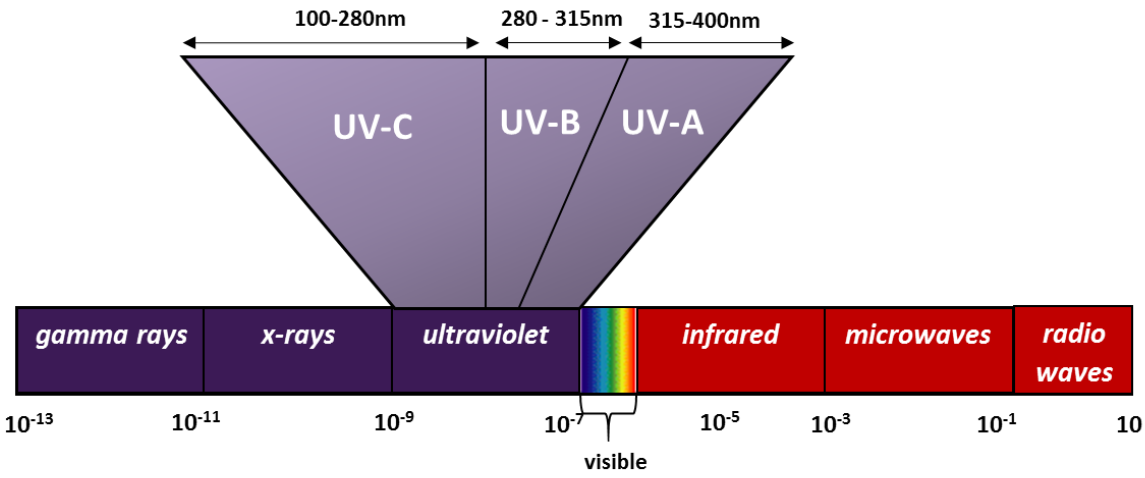


***Figure 1-2**: **Ultraviolet radiation (UV) is part of the electromagnetic spectrum, and includes subtypes UV-C, UV-B and UV-A, which are named according to their wavelengths (nm).** UVC has a wavelength spanning 100 –280 nm, UV-B is 280 – 315 nm and UV-A is 315-400 nm.*

1.3.1 UV exposure

The amount of UV that passes through the earth's atmosphere (which absorbs UV) and that a person's skin is exposed to from sunlight can vary according to certain factors as follows.

The quantity of UV reaching the earth's surface is at its highest when the sun is directly overhead because the pathway through the atmosphere is at its shortest. The distance that UV must pass from the outer atmosphere to reach the earth's surface varies throughout the day and results in diurnal variation of UV during daylight hours (Monteith and Unsworth, 1990). Cloud cover can also affect UV exposure. While clouds can reflect UV from sunlight back into space, thus reducing the terrestrial UV dose, UV can sometimes be increased at the earth's surface in situations when clouds refract UV passing from the sun to the earth (Calbo et al., 2005).

There is also seasonal variation of solar radiation which reaches the earth's surface (Kumar et al., 1997) with most UV present during the summer months. This relates to the fact that the distance that UV travels from the sun through the atmosphere varies throughout the year. Latitude can affect the amount of terrestrial UV, with higher doses closer to the equator and lower doses further away from the equator. This is due to the longer pathway that UV has to pass through the atmosphere at higher latitudes (Madronich et al., 1998).

UV levels also increase with gains in altitude because of the reduced amount of atmosphere which the UV has to pass through (Blumthaler and Ambach, 1988). The physical features of the ground surface can also affect the amount of UV exposure, for example glass, sand and water can reflect UV, thus increasing the dose of UV in the close vicinity (Chadyšiene, 2010).

The amount of UV which penetrates the layers of the skin also varies according to the wavelength of UV (Figure 1-3). UVA, which is of longer wavelength, can penetrate through the epidermis and into the dermis, whereas most UVB is absorbed by the epidermis of the skin with a minimal amount reaching the dermis (D'Orazio et al., 2013). UV has many effects on the skin, including DNA damage, inflammation (sunburn), immunosuppression, photoaging / photodegradation of collagen and skin cancer development (Meeran et al., 2008). UVB also causes vitamin D synthesis in skin; this occurs when 7-dehydrocholesterol absorbs UVB and is converted into pre-vitamin D3 which then isomerizes into vitamin D3 (i.e. cholecalciferol) (Wacker and Holick, 2013).

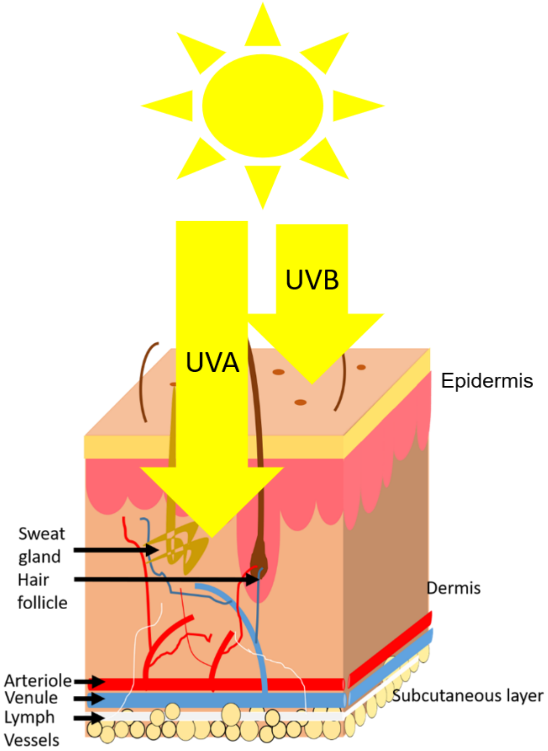


Figure 1-3:Labelled cross-section of skin showing UVA and UVB penetration. *UVB only penetrates the epidermis of the skin and UVA penetrates the epidermis and the dermis.*

1.3.2 Effect of UV on DNA

UV, especially UVB, can damage DNA in cells within the skin (Brash, 1988, Sage, 1993). This damage from UV can be via the production of cyclobutane pyrimidine dimers (CPD), 6-4 photoproducts (also known as a 6-4 pyrimidine-pyrimidone photoproduct; 6-4PP), 8-hydroxy-2'-deoxyguanosine and double strand breaks (Jackson and Bartek, 2009). UVB predominantly promotes the formation of CPDs and 6-4PPs, with CPDs 3 to 4 times more common than 6-4PPs (You et al., 2001). These UV-induced damaging products can affect DNA conformation and regulatory functions (Rastogi et al., 2010).

Formation of CPD occurs when the energy from a photon of UVB splits the double bond between the 5th and 6th carbon in pyrimidine bases which are adjacent to one another on a strand of DNA. As a result, a covalent bond is formed linking the 5th carbons of the adjacent pyrimidines and a second covalent bond is formed linking the 6th carbons of the adjacent pyrimidine bases, producing the CPD (Freeman, 1988, Parrish et al., 1982). Dimers can form between two thymine bases producing a thymine dimer as shown in Figure 1-4 but can also form between a cytosine and thymine or two cytosine bases.

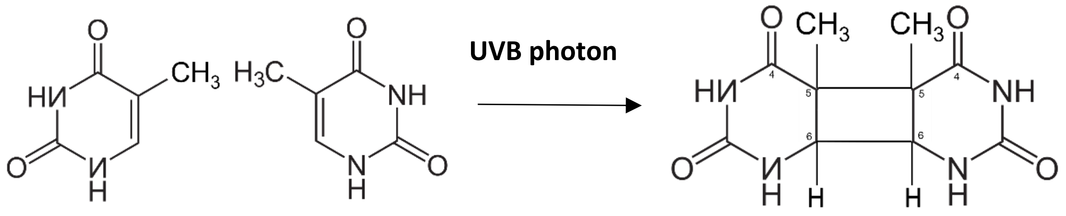


Figure 1-4: The formation of thymine dimer after UVB exposure. *UVB exposure causes splitting of the double bond between the 5th and 6th carbon of two adjacent thymine bases, resulting in the formation of a covalent bond linking the two thymine bases, thus producing a thymine dimer.*

Alternatively, a 6-4PP (Figure 1-5) can form as a result of the splitting of the double bond between the 5th and 6th carbons of one pyrimidine base and the subsequent formation of a bond between the 6th carbon of that pyrimidine and the 4th carbon of the adjacent pyrimidine (Freeman, 1988).

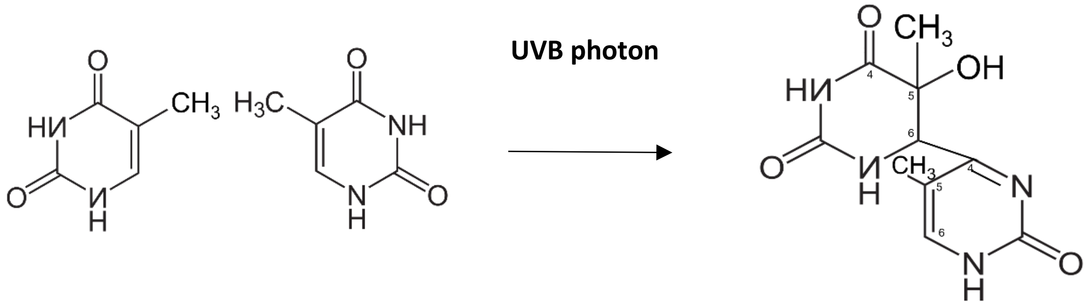


Figure 1-5: The formation of a 6-4 pyrimidine-pyrimidone after UV exposure of two adjacent thymine bases. *The UVB photon breaks the bond between the 5th and 6th carbon atoms of one thymine base which allows the formation of a bond between the 6th carbon of that thymine base and the 4th carbon of an adjacent thymine base, resulting in a 6-4 photoproduct.*

Evidence of the production of UV-induced CPDs and 6-4PPs in DNA leading to subsequent mutations can be identified by the presence of TC to TT or CC to TT transitions at bipyrimidine sites in genes, including in the *TP53* gene (Mouret et al., 2006).

DNA double strand breaks have been identified in UV-irradiated cells, especially in replicating DNA, where it has been observed in the replication of damaged DNA (Dunn et al., 2006). These double strand breaks have been identified in cells exposed to UVB radiation, where DNA lesions such as CPDs and 6-4 photoproducts can result in double strand breaks due to these lesions causing replication blockage. Therefore, it has been suggested that double strand breaks are produced because of an attempt to replicate DNA at the site of unrepaired DNA lesions (Dunn et al., 2006). Greinert et al., 2012 has proposed that UVA exposure can result in a replication-independent induction of double strand breaks which has also been proposed by earlier studies (Greinert et al., 2012).

Increased levels of 8-hydroxy-2'-deoxyguanosine (8-OH-dG), also known as 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxodG) (Valavanidis et al., 2009) have been identified in hairless mice and in human skin after UVB exposure (Hattori et al., 1996, Ahmed et al., 1999) and after UVA exposure (Griffiths et al., 1998). The presence of 8-hydroxy-2'-deoxyguanosine residues in DNA results in GC to TA transversions (Guo et al., 2016).

In response to all the above photoproducts/DNA changes, progression through the cell cycle is usually stalled until DNA is repaired. DNA damage in human skin is thought to be repaired mainly via nucleotide excision repair (NER) and base excision repair (BER).

NER is usually deployed in response to UVB damage in skin and is the principle pathway for repair of CPDs and 6-4PPs (Strachan, 2011). In this process, the damaged area of the DNA strand containing the photoproduct is cleaved and then exonucleases remove the surrounding DNA and DNA polymerase catalyses the re-synthesis of the correct sequence (using the opposite DNA strand as template) and is sealed by DNA ligase. Patients with Xeroderma Pigmentosum (XP), which is an autosomal recessive hereditary disease, have mutations in genes which code for proteins that are responsible for nucleotide excision repair (Cleaver, 2005). These patients are classified into eight subclinical types based on the genes affected, there are seven genetic complementation groups (XPA to XPG) from A to G and XPV which represents deficiency in trans-lesion synthesis (DiGiovanna and Kraemer, 2012). Therefore, these patients are unable to remove the dipyrimidine photoproducts and have DNA repair deficiency which is associated with a massively increased risk of skin cancer.

BER occurs in response to base lesions because of deamination, alkylation and oxidation and is the main repair pathway for removal of 8-OH-dG (Fortini et al., 1999, Kunisada et al., 2005, Javeri et al., 2008). Oxidation can occur indirectly via UV and the production of reactive oxygen species. This type of repair corrects small base lesions that does not affect DNA conformation. DNA glycosylases are used to remove the damaged base (Krokan and Bjoras, 2013).

UV irradiation can affect energy supply in skin cells and DNA repair requires energy. 1α, 25-Dihydroxyvitamin D_3 is produced in the skin and reduces UV-induced DNA damage (Gordon-

Thomson et al., 2012, Dixon et al., 2011). After UV exposure, energy availability is limited in keratinocytes, however in the presence of 1α, 25-Dihydroxyvitamin D_3, there is increased energy availability due to increased glycolysis. This provides energy for NER and repair of CPDs with decreased oxidative DNA damage (Rybchyn et al., 2018).

Unrepaired DNA can result in the generation of mutations or can trigger apoptosis (Rastogi et al., 2010). Cells which survive UV-induced DNA damage frequently harbour UV mutation signatures, for example unrepaired CPDs and 6-4PPs resulting in C to T or CC to TT changes (Brash, 1988, Patrick, 1977, Sage, 1993, Sinha and Hader, 2002, Wang et al., 2008, Wikonkal and Brash, 1999, Brash et al., 1996). These mutations affect the protein products of affected genes and can cause alterations to the cell cycle and in cellular behaviour, resulting in the development of skin cancer.

1.3.3 UV and skin cancer

Skin cancer is the most common cancer in Caucasians, with the three main types of skin cancer comprising basal cell cancer (BCC), cutaneous squamous cell cancer (cSCC) and melanoma. BCC is the most common type of skin cancer and accounts for around 80% of all skin cancers (Berking et al., 2014); BCCs are very slow growing and do not usually spread to other parts of the body. In 2015 there were 166,448 cases of basal cell carcinoma (BCC) in the UK (Venables et al., 2019). cSCC is the second most common type of skin cancer and 44,672 cases of cSCC occur annually in the UK (Venables et al., 2019). The five year rate of recurrence of primary cSCCs is 8% and the rate of metastasis is 5% (Alam and Ratner, 2001). BCC and cSCC arise from keratinocytes, whereas melanoma arises from melanocytes. Melanoma is less common than BCC and cSCC but is more likely to metastasise. There were 232,100 cases of global melanoma cases annually and 55,500 cancer deaths (Schadendorf et al., 2018). Melanoma metastasis has been recorded to occur in 5 – 15% of all melanoma cases (Kalady et al., 2003, Sandru et al., 2014, Meier et al., 2002).

There are a number of risk factors for development of skin cancer, including endogenous factors (e.g. skin pigmentation, genetic alterations that predispose to fairer skin) and exogenous factors (e.g. lots of UV exposure, various medications, etc.). There is an inverse correlation between skin pigmentation and the incidence of skin cancer, which is due to the photoprotective role of melanin in darker skinned people preventing DNA damage and subsequent skin cancer (Gilchrest et al., 1999). Individuals with light skin are 70 times more likely to develop skin cancer compared to individuals with dark skin (Halder and Bang, 1988). UVB can also induce erythema (sunburn) in fairer skinned people, and this usually occurs 4 hours after UV exposure and peaks between 8 and 24 hours which then fades over a day. In lighter skinned individuals, erythema can last for weeks.

Erythema is associated with apoptotic keratinocytes which are sunburn cells (Brenner and Hearing, 2008).

In humans, *MC1R* variants are the cause of red hair and fair skin and are a risk factor for skin cancer (Valverde et al., 1995, Robles-Espinoza et al., 2016, Tagliabue et al., 2015). Due to the pleiotropic nature of *MC1R* variants, it has been identified that variants in *MC1R* can also determine sun sensitivity in individuals without red hair (Healy et al., 2000). MC1R signalling controls the production of eumelanin by inducing expression of proteins responsible for eumelanin synthesis (Rees and Harding, 2012). Inactivating mutations in *MC1R*, which are present in individuals with red hair can reduce the synthesis of eumelanin and increase the risk of skin cancer (Nasti and Timares, 2015).

Drugs such as azathioprine which is an immunosuppressant used to prevent organ transplant rejection can increase the likelihood of developing cutaneous squamous cell carcinoma. Azathioprine has been linked with photosensitivity to UVA suggesting this might be a cause for the increased incidence of skin SCC (Inman et al., 2018). Psoralen is a medication used with UVA for the treatment of psoriasis and can intercalate with DNA to inhibit DNA synthesis and cell division. UVA exposure can activate psoralen to form covalent bonds with double bonds of thymines and the production of monoadducts which can lead to skin cancer (Derheimer et al., 2009).

1.3.4 UV and skin ageing

UV can affect the connective tissue in the dermis and cause skin ageing. Skin ageing can be seen visibly as wrinkles and, histologically, via the loss of mature dermal collagen. There is evidence which suggests that there are higher concentrations of reactive oxygen species generated *in vitro* and *in vivo* after UVA and UVB irradiation. These reactive oxygen species can affect collagen metabolism which results in destroying interstitial collagen (Wlaschek et al., 2001). Although people develop wrinkles as they age, and skin cancer is more common in elderly people, one study has suggested that people who develop UV-induced wrinkling are less likely to develop BCC (Brooke et al., 2001).

1.3.5 UV therapy for skin disease

Due to its ability to immunosuppress, UV is used as a treatment (i.e. phototherapy) for common inflammatory skin conditions such as eczema and psoriasis (Ibbotson et al., 2004, Menter et al., 2010, Pathirana et al., 2009). There are three main types of phototherapies: broadband UVB (BB-UVB), narrowband UVB (NB-UVB), and PUVA (i.e. Psoralen with UVA).

Psoriasis is caused by the hyper-proliferation of keratinocytes mediated by T-cells (Gudjonsson et al., 2004). UV can immunosuppress in a variety of ways such as inhibiting the presentation of antigens, releasing immunosuppressive cytokines and causing the apoptosis of leukocytes such as T-cells (Schwarz, 2005). It has been found that T-cells are 10 times more sensitive than keratinocytes to UVB induced apoptosis; therefore the long remission periods of patients could reflect the apoptosis of psoriasis-specific T cells which have been exposed to UVB (Krueger et al., 1995). However, it has also been noted that UVB induces regulatory T cells (also known as Tregs), and it is likely that this also plays a role in the mechanism of clearance of psoriasis by narrowband UVB (Schwarz, 2008).

NB-UVB has been identified as the most beneficial component of UV for the treatment of psoriasis (Chen et al., 2013). 'Narrowband' refers to a specific wavelength of UVB which is 311-312 nm, as opposed to 'broadband' UVB which is a wavelength of 280-320 nm (and some UVA wavelengths also) (Ibbotson et al., 2004). BB-UVB has been identified as less effective for psoriasis than NB-UVB (Kirke et al., 2007). NB-UVB and PUVA are used as treatments for psoriasis and eczema (Ibbotson, 2018). PUVA has been used since 1974 and this phototherapy is carcinogenic and mutagenic. Studies have shown that patients exposed to high doses of PUVA have an increased risk of skin SCC and melanoma (Stern and Study, 2001) and studies have also shown a small significant increased risk in BCC (Nijsten and Stern, 2003).

1.3.6 UV therapy

NB-UVB was first reported to treat patients with psoriasis in 1988 (Green et al., 1988) and although it has been used for the past few decades as treatment, there is limited information about the long term likelihood of developing skin cancer in patients who have received this form of phototherapy. There are some studies which have reported on previous NB-UVB exposure which have not shown an increased risk of skin cancer when compared to an age and sex matched control population (Hearn et al., 2008, Archier et al., 2012, Man et al., 2005). There has also been a Taiwanese population-based cohort study examining the long-term safety of NB-UVB treatment which concluded that long-term NB-UVB phototherapy does not increase risk for skin cancer compared to short-term phototherapy in individuals with skin phototypes III – V (Lin et al., 2019). However, it is known that individuals which have a skin phototype III – V are considered to rarely or never sunburn and also to tan easily (Fitzpatrick, 1988) therefore these people are at lower risk of skin cancer than more fair-skinned subjects.

By contrast, some reports have suggested that NB-UVB is more carcinogenic than BB-UVB (van Weelden et al., 1988, Flindt-Hansen et al., 1991, Wulf et al., 1994, Gibbs et al., 1995). A study in mice identified that more CPD were present in mice after NB-UVB treatment compared to BB-UVB suggesting higher carcinogenesis because of NB-UVB (Kunisada et al., 2007). Another study was conducted to follow up on this finding to identify the underlying genetic cause for the association between NB-UVB and carcinogenesis and it identified that 1 minimal erythemal dose of NB-UVB produced more UV mutation signatures in p53 than BB-UVB, which supported the original observation that there is more CPD formation and 6-4 photoproducts after NB-UVB exposure (Yogianti et al., 2012). A cohort human study was conducted which reported that cases of skin SCC and BCC were seen in patients after NB-UVB treatment but there were no cases of melanoma (Raone et al., 2018). The keratinocyte skin cancers were more frequently identified in patients who were at an older age (mean 68.8 years) during their first NB-UVB course. This suggests that patients are more likely to develop keratinocyte skin cancer after NB-UVB treatment if they have prior risk factors such as old-age. However, this study by Raone *et al*. only included 375 patients and had a short follow-up time (mean 6.9 years) which means that we still do not really know the long-term risk of NB-UVB treatment in relation to whether it significantly increases the risk of subsequent development of skin cancer.

1.4 Cancer

Cancer is the uncontrollable proliferation of cells (Hanahan and Weinberg, 2011). There are several hallmarks that are acquired by a cancer cell to survive, proliferate, and disseminate. These hallmarks are sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction (Hanahan and Weinberg, 2011). These hallmarks are acquired via enabling characteristics such as the development of genomic instability in precancerous and cancerous cells and the inflammatory state of premalignant and malignant lesions.

Genomic instability includes mutations and chromosomal rearrangements which occur in two broad classes of genes, proto-oncogenes and tumour suppressor genes (Weinstein, 2002). Proto-oncogenes encode proteins which stimulate cell growth, inhibit cell differentiation, and inhibit cell death. Gain of function mutations transforms proto-oncogenes into oncogenes. As a result, there is increased cell growth and abnormal cell proliferation (Carbone and Levine, 1990). Tumour suppressor genes encode proteins which inhibit cell proliferation or play a role in other important cell functions and these genes are prone to loss of function mutations which can prevent processes such as apoptosis and lead to the development of cancer (Lee and Muller, 2010).

Two of the aforementioned hallmarks were identified as emerging hallmarks, as they were not in the original six hallmarks of cancer, but were added subsequently (Hanahan and Weinberg, 2011). These two emerging hallmarks are deregulating cellular energetics to fuel cell growth and avoiding immune destruction. The tumour associated inflammatory response can lead to a dysfunctional immune system in the cancer microenvironment, thus promoting tumorigenesis (DeNardo et al., 2010, Grivennikov et al., 2010, Qian and Pollard, 2010, Colotta et al., 2009). Cancer cells can avoid immune destruction via 'immunoediting', which occurs in three phases: elimination, equilibrium, and escape. In the elimination phase, cells which have become precancerous or cancerous are recognised and killed by the innate and adaptive immune system. Cells which survive this process proceed to the equilibrium phase, where the tumour cells continue to proliferate but the immune system continues to kill a proportion of cancer cells, thus net tumour growth is reduced and can even be stalled. Tumour subclones with reduced immunogenicity, e.g., via loss of tumour antigens, can develop due to a combination of genomic instability and pressure from the adaptive immune system, whereby these immuno-edited tumours are now in the escape phase. In this phase the growth is not restrained, and the cancer becomes clinically apparent and/or metastasises (O'Donnell et al., 2019).

1.4.1 Models/Understanding of cancer development

In 1889, Paget examined the post-mortem data of 735 women with breast cancer and proposed the 'seed and soil' hypothesis (Paget, 1989). He noticed that the organ distribution of metastases was not random and suggested that tumour cells were the 'seeds' which grew in specific 'soils', the latter referring to the tumour micro-environments in these organs.

In 1954, Armitage and Doll suggested through a statistical model, that the development of cancer is a multistage process. They produced a formula which aimed to weight the strength of the carcinogenic factors. The carcinogenic factors were responsible for cellular changes. The changes that were of greatest importance varied according to where the cell was considered to be in the multistage process leading to the development of cancer (Armitage and Doll, 1954).

Subsequently, a two-hit hypothesis was proposed by Knudson in 1971, which was based on certain cancers occurring more frequently in families than in the general population, and the theory suggested that retinoblastoma was caused as a result of two mutational events (Knudson, 1971). In patients who dominantly inherited the risk of developing this cancer, the tumour arose due to the first mutation being inherited in the germline and a subsequent second somatic mutation arising in the cell which developed into the cancer. In individuals where there was not a family history of predisposition to the cancer, both mutations were somatically acquired. This hypothesis was based on 48 patients with retinoblastoma and was a key advance in understanding that most tumour suppressor genes require biallelic inactivation for cancer development and/or progression.

In 1976, Nowell suggested that cancer is a stepwise, sequential evolutionary process consisting of an accumulating series of mutations, with the initial mutation resulting in a growth advantage permitting clonal expansion, and subsequent genetic instability/mutations leading to sub-clonal populations as highlighted in figure 1-6 (Nowell, 1976). The environment of these sub-clones can affect their survival and their potential to evolve (Bierie and Moses, 2006). Restraints (e.g. metabolic or immunologic disadvantages) or selective pressures such as clonal interference and tissue ecosystems allow some sub-clones to expand, become extinct or remain dormant (Gatenby and Gillies, 2008). Clonal interference is the process where two or more different beneficial mutations arise in two individual cells and these cells compete against each other resulting in the loss of the less-fit genotype.

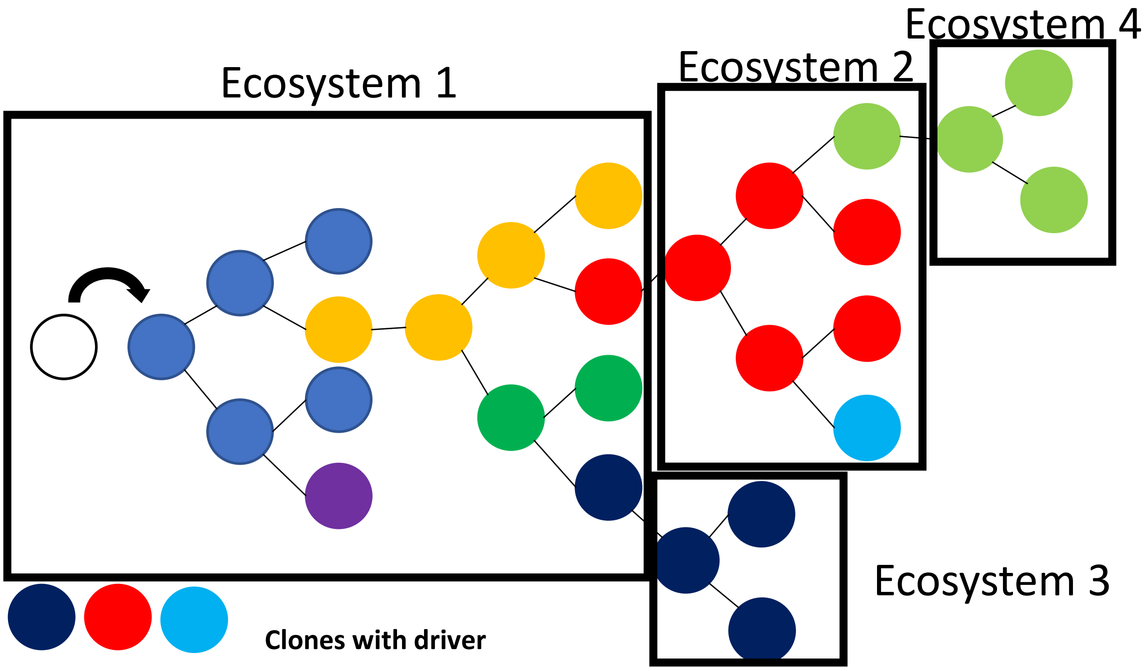


*Figure 1-6: **The clonal evolution of cancer.** The diagram shows how a cell divides in different ecosystems and how some cells stop dividing and others continue dividing and evolve into cancer. Adapted from Greaves and Maley 2012.*

Later, Vogelstein and Fearon proposed their theory for the genetic basis of colorectal cancers and presented four processes required for the formation of a tumour (Fearon and Vogelstein, 1990).

The theory stated that there should be the activation of oncogenes and the inactivation of tumour suppressor genes. Other requirements were that there should be the presence of mutations in 4 or 5 genes, the total accumulation of mutations is more important than the order of mutations and mutant tumour suppressor genes can have a phenotypic effect even at a heterozygous level.

The expansion of clones can be caused by a 'driver' mutation that confers a cell with a growth advantage and have been positively selected for during the evolution of the cancer. For example, a driver mutation may provide advantage in their tissue ecosystem, for example by altering the cell's phenotype resulting in increased cell proliferation (Greaves and Maley, 2012). Cells with driver mutations also have other 'passenger' mutations which are considered not to alter the cell's phenotype (Stratton et al., 2009). Since there are many passenger mutations, these may be detected more frequently than the driver mutations. Next generation sequencing (NGS) has revolutionised our understanding of the evolution of cancer, allowing the detection of numerous mutations within cancers. However, bioinformatics approaches are needed to distinguish driver mutations from passenger mutations using statistical analysis and this approach can be supported by downstream functional assays to identify if these driver mutations alter a cell's phenotype. Sequencing all the cells in a tumour can reveal the clonal architecture of the tumour. This information can also be used to identify the founder genotype involved in the clonal evolution of these neoplasms (Stratton, 2011).

There are three fundamental forces which govern the evolution of cancer: mutation, drift and selection. Mutation and drift are stochastic processes whereas selection is deterministic (Sottoriva et al., 2017). Drift is the stochastic change in allele frequencies in a population of cells due to random birth and death events. Drift and selection can change the frequency of alleles in a population. Selection is when there is a new mutation that increases the ability of the cell to survive and reproduce and has escaped genetic drift. For example, this could alter the rate of cell proliferation due to a certain allelic change (Sottoriva et al., 2017) or immunoediting (Kim et al., 2007).

The Cancer Stem Cell theory suggests that a tumour is driven by a rare sub-population of cancer cells that act as stem cells, which reproduce themselves and sustain the cancer. Cancer stem cells were first identified in acute myeloid leukaemia (Bonnet and Dick, 1997), when it was identified that when AML stem cells were transplanted into mice, these cells were able to proliferate and differentiate and re-establish AML. The specific stem cells in this cancer type were those that expressed the CD34+ CD38- marker, which were able to produce tumour cells with a phenotype that were identical to the donor. Support for this theory was also seen in solid tumours such as brain, breast, colon, prostate, pancreas, lung, liver, melanoma and ovarian cancers which are known to have a heterogeneous population of tumour cells (Wang et al., 2013). The establishment of this theory indicated that it might be possible to generate treatments which are tailored to kill cells with specific cancer stem cell traits (Dick, 2003). The skin stem cell hypothesis is based on the fact that stem cells for the epidermal keratinocytes in the interfollicular epidermis basal layer are constantly regenerated in the basal layer and the cells leave the basal layer by differentiating upwards in the epidermis towards the stratum corneum. The hair follicles are constantly in cycles of regeneration and rest which are driven by the stem cells in the bulge and a cluster of cells below the bulge known as the hair germ. The melanocyte stem cells are also distributed in the bulge and the hair germ (Hsu et al., 2014). They identified that in humans there is heterogeneity of clonogenic keratinocytes, and three types of clones are initiated, holoclones, meroclones and paraclones. All three of these proliferate however it was proposed that the holoclone-forming cell is similar to a stem cell and has long-term proliferation potential (Enzo et al., 2021). These holoclones were discovered when treating a child with junctional epidermolysis bullosa (JEB) via transgenic keratinocyte cultures which regenerated a fully functional epidermis for the child (Hirsch et al., 2017). In humans, the keratinocytes are constantly renewed to achieve homeostasis. Three hypotheses were proposed to explain this process of keratinocyte differentiation and homeostasis: asymmetric division, population asymmetry and population asymmetry with stem cells (Li et al., 2013). Using a 3D agent-based model of an epidermis and observing simulated growth and maintenance of the epidermis over three years suggested that the population asymmetry with stem cells hypothesis is most likely responsible for keratinocyte homeostasis in the epidermis (Li et al., 2013). There is a study which has investigated the reason that subsets of tumour-initiating stem cells escape cancer therapies. They identified that TGF-beta signalling in SCC stem cells can fuel heterogeneity in SCC stem cells which can result in drug resistance (Oshimori et al., 2015).

1.5 Next Generation Sequencing (NGS)

The invention of NGS and the decreasing cost of NGS over recent years has led to an increase in the number of studies reporting on DNA sequencing of cancers (Meyerson et al., 2010). The sequencing of cancers can help dissect the mutational landscape of these tumours to understand how these cancers originate and develop. Sequencing all the cells in a tumour can reveal the clonal architecture of the tumour. This information can also be used to identify the founder genotype involved in the clonal evolution of these neoplasms (Stratton, 2011). All regions of DNA, including all those which code for proteins and the regions of DNA that do not code for proteins, are sequenced during whole genome sequencing (WGS) whereas the only the coding sequence of DNA is sequenced during whole exome sequencing (WES). Genes that are expressed by the cancer cell genome can be analysed using transcriptome sequencing (Stratton et al., 2009).

There has also been the development of multiple sequencing platforms to undertake NGS (Quail et al., 2012). The foundation for this type of sequencing relies on the production of an NGS library. The preparation of a high-quality library is crucial to produce high-quality sequencing data. Illumina Inc. has been at the forefront of NGS and uses a “sequencing by synthesis” approach developed by Solexa. Initially, DNA is sheared into smaller fragment lengths which are then ligated to adapters (Bentley et al., 2008). These adapters enable DNA fragments to bind to a glass flow cell. The type of adapter can vary depending on the type of sequencing being conducted. For WES or targeted sequencing, probes or baits can be used to enrich coding regions or regions of interest. These adapter-ligated DNA fragments are then amplified and sequenced on a glass flow cell. Mass parallel sequencing produces data with high sequencing depth which increases reliability of results. High sequencing depth is particularly important in the sequencing of cancers and precancerous lesions for the identification of the rare variants in a single clone which could progress into a cancer. The sequencing data undergo quality control to identify true variants in the samples and the data is filtered to identify pathogenic variants. The variants can then be annotated using databases which predict the protein and biological function of a mutation (Karaoz et al., 2004). The reference human (haploid) genome is ~3.1 billion base pairs (3 Gb) and an exome is approximately 1-2% of the genome, with WES of the protein coding regions spanning around 50 Mb (Nakagawa and Fujita, 2018). A high-quality WGS would have 30 - 50x read depth, so this would produce 90 – 150 Gb of data for a single sample (Nakagawa et al., 2015, Meyerson et al., 2010) .

These second-generation parallel sequencing technologies have dominated the sequencing market for the last 10 years due to the ability to produce a large amount of data at a cheap rate. However, due to the de novo assembly of the genome, this can result in fragmented assemblies which can cause problems for resolving repetitive sequences in the genome (Schatz et al., 2010). A new approach, known as third-generation sequencing technology, was introduced to enable the sequencing of longer read lengths. Second-generation sequencing was dependent on PCR to make multiple copies of template DNA, whereas third generation sequencing directly sequences DNA molecules for analysis, reducing sequencing biases which can be introduced as a result of PCR. The time required for third generation sequencing, first technology was introduced by Pacific

Biosciences (PacBio, http://www.pacb.com/), is less than second-generation sequencing and is reduced from days to hours and can also be reduced to minutes for real-time applications (Lu et al., 2016). This single-molecule long-read sequencing platform also identifies highly repetitive regions, complex structural variation and can estimate the haplotype of an individual. However, this technology has a higher error rate than second-generation sequencing. The errors are dispersed throughout the genome and can be reduced via sequencing a single molecule template strand and the complement strand can be sequenced multiple times by using circular consensus sequencing (Larsen et al., 2014). This technology is relatively high cost and low throughput. PacBio most recent circular consensus called 'Hi-Fi' produces 20kbp read lengths with an error rate of 0.1% (Wenger et al., 2019).

Unlike, the sequencing technologies previously mentioned which function via polymerase-mediated DNA synthesis, the nanopore based technologies sequence DNA by identifying changes in the ionic current across a membrane as a DNA molecule passes through a protein nanopore (Branton et al., 2008).

A study which looked at the efficacy of nanopore sequencing showed that the technology was able to identify and phase two de-novo variants that were from the same paternal haplotype. However limitations to nanopore sequencing were that there were higher errors in single nucleotide variation-calling rates compared to short read sequencing (Bowden et al., 2019). Recently there has been improvements made to this technology which has brought in the development of the ultra-rapid nanopore genome sequencing platform (Goenka et al., 2022) which has been recorded to have a pipeline that is 50% faster than previous pipelines and is capable of providing accurate small and large variant information from the genome.

A study aimed to compare short read Illumina sequencing with Oxford Nanopore Technology (ONT) sequencing and PacBio sequencing to identify which technology and methods are most reliable for detecting large scale variation in cancer genomes (Aganezov et al., 2020). The study showed that the structural variants identified in ONT and PacBio were more than 90% concordant with each other which shows that even though these technologies function differently their results show their reliability to identify true positives. Using these long-read technologies they also identified a large number of structural variants that were not identified via Illumina's short-read technology. This was also supported in another study which stated that long-read sequencing showed key disruption events in cancer such as understanding aberrant genomic structures (Sakamoto et al., 2021). It was also highlighted in a study that short-read technologies can complement long-read technologies by using short-read technologies to analyse allele-specific copy number variants via the detection of alterations in heterozygous germline single nucleotide polymorphisms (Aganezov et al., 2020). Another study supported this finding by suggesting that single nucleotide variants are more reproducible than small insertions and deletions when using short-read sequencing technologies (Pan et al., 2022).

With these increases in speed and accuracy WGS is being considered a good alternative to other diagnostic practices. A study considered WGS to be a timely and more economical alternative to array Comparative Genomic Hybridisation technologies and WES in the detection of CNVs (Coutelier et al., 2022). WGS has also been reported to have a greater diagnostic yield than cytogenetic analysis in myeloid cancers (Duncavage et al., 2021).

Long read sequencing is effective in identifying structural variations and haplotype phasing therefore it is considered useful for cancer genomes. The introduction of long-read sequencing also enabled gaps in the human genome to be completely sequenced which has included centromeric regions and the short arms of five chromosomes (Nurk et al., 2022). Cancer is caused by a variety of genomic aberrations and long-read sequencing can help elucidate transcriptome and epigenome statuses, fusion transcripts, transcript isoforms and DNA methylation phase information which can be missed using short-read technologies (Sakamoto et al., 2020). Short read sequencing posed limitations for sequencing repetitive regions, phasing alleles and distinguishing between homologous genomic regions therefore long read-sequencing enables these to be identified which can help understand disease pathogenesis (Mantere et al., 2019). Oxford Nanopore long-read sequencing generates the largest contiguous sequence but PacBio some of the most accurate long-read data (Logsdon et al., 2020). Therefore, both of these long-read technologies complement each other. The introduction of this sequencing enables the full sequence of many reference genomes to be available so we can understand human genetic diversity, heritability, and mutational processes. This will ensure that we are not aligning sequences to a single reference genome to identify variation but long-read technologies will enable haplotype phasing so we can dissect genetic variation to completely resolve the correct human genome sequence for each individual.

The limitations surrounding long-read sequencing is that for cancer genomes is that it requires micro-gram order DNA (Sakamoto et al., 2021) and this is not always available when sequencing tumour DNA. Short-read sequencing is possible with small volumes of DNA as the DNA is amplified unlike long-read DNA preparation. Scalability is another challenge for long-read sequencing as it is computationally intensive to produce assemblies for large genomes which requires long periods of time (Amarasinghe et al., 2020). This also impacts data generation, storage, and integration. Error correction rates are improving however short-read sequencing is still outperforming the accuracy in long-read sequencing (Amarasinghe et al., 2020). Long-read transcriptomics provides information on transcript-level differential expression. However long-read transcriptomics is in its infancy and provides low replication and modest read counts, therefore an increase in throughput and a decrease in price would enable this technology to be more effective in the future (Amarasinghe et al., 2020).

There have been different approaches of experimental design used in various NGS studies to understand the evolution of cancer and tumour heterogeneity. Tumour multi-sampling is one method, which includes geographical sampling, when multiple samples are taken from a single tumour at the same point in time, or longitudinal sampling when a tumour sample is taken at different points in time (Campbell et al., 2010, Yachida et al., 2010). Single cell sequencing is another method of tracking the evolution of cancer cells as individual mutations within each cell can be identified, however this requires whole genome amplification which can introduce mutations which are not present in the tumour but occur as a result of the amplification (Yates and Campbell, 2012). Mathematical models have also been created and used in conjunction with NGS to help understand clonal expansion in normal or precancerous tissue samples (Lynch et al., 2017). Using WGS on established cancers, it has additionally become possible to dissect the sequence of events for the development of a cancer (Jolly and Van Loo, 2018).

Bottleneck sequencing, which is a type of next generation sequencing, is used to identify rare somatic mutations (Hoang et al., 2016). This type of sequencing uses a barcoded genomic library with a dilution step before amplification. The dilution causes a bottleneck effect as there is a random sampling of double stranded template DNA molecules. Normally, rare mutations would be masked (or hidden) by the abundance of the wildtype DNA at that site using conventional WGS or WES, but due to this dilution step these rare mutations become visible during sequencing. The library, using this approach, also enables the ‘Watson’ and ‘Crick’ strands of the DNA to be sequenced excessively, therefore if a rare mutation is identified repeatedly (across both Watson/Crick duplicate pairs), it is less likely to be due to artefact (Hoang et al., 2016).

Somatic copy number alterations are also used to identify how cancers develop. Copy number is identified using single nucleotide polymorphism (SNP) arrays or via WGS or WES. WGS is expensive so WES is often used to predict copy number variants. Depth of coverage by measuring the number of reads aligned to a base is used to identify the copy number. The presence of highly repetitive sequences can result in limitations to using whole exome data for this purpose by decreasing the power to detect copy number variations (Kadalayil et al., 2015). Copy number has been analysed in exome data for cSCC tumour samples, with 25 of 40 samples showing copy number changes in at least two regions in one study (Inman et al., 2018). This finding was consistent with previous analyses (Hameetman et al., 2013, Sekulic et al., 2010, Salgado et al., 2010, Purdie et al., 2009). The copy number data indicated that there were higher levels of chromosomal instability in moderately and poorly differentiated tumours compared to well differentiated tumours (Inman et al., 2018).

RNA sequencing is another method used in cancer research, drug discovery, cancer diagnosis and prognosis. Two different approaches can be used in RNA sequencing studies; one method is transcriptome analysis which enables the identification of biologically important transcriptional pathways and the other method is to measure differential gene expression. Small RNAs, such as microRNAs (miRNAs), have been identified as non-coding regulators of biological pathways and regulate the expression of protein-coding regions post-transcriptionally (Wang et al., 2020b). RNA and miRNA sequencing have been conducted in the analysis of cSCC and actinic keratosis (AK) (Chitsazzadeh et al., 2016) which is a pre-cancerous lesion. While the mRNA matrix displayed a significant difference between normal skin and cSCC in at least one pairwise comparison, the miRNA matrix showed a better distinction between normal skin, cSCC and AK samples, which highlights that AKs have intermediate expression of all genes between normal skin and cSCC (Chitsazzadeh et al., 2016). RNA sequencing can also be used to identify mutations in cancers, with the limitation that the analysis is only limited to genes that are expressed.

Epigenetic changes are changes that affect gene expression but do not change the DNA sequence. Gene transcription can be regulated by DNA methylation, especially in promoter regions of genes, and alterations of DNA methylation have been identified in tumours. Bisulfite sequencing is used to identify DNA methylation patterns in cancer (Li and Tollefsbol, 2011). Hypermethylation silences gene transcription and results in the loss of gene expression (Das and Singal, 2004). The hypermethylation of tumour suppressor genes can enable the initiation, promotion and progression of cancer. In melanoma, *CDHN2A* encodes a p16 cell cycle inhibitor protein and has been shown to be inactivated by promoter hypermethylation (Penta et al., 2018). Other tumour suppressor genes identified as hypermethylated in melanoma are *RASSF1A*, *PTEN*, *TNFSF10D* and *COL1A2*. Global hypomethylation was also identified in melanomas which showed the activation of retrotransposons (Penta et al., 2018). Retrotransposons are pieces of nucleic acid that can “copy and paste” themselves in different parts of the genome via an RNA intermediate (Boeke et al., 1985, Garfinkel et al., 1985).

Proteomics is another method used to analyse cancers and can be used identify a set of proteins whose level of expression is altered in the relevant cancer and within different subtypes of the same cancer. In a study of cSCC that included cancers which metastasized after surgery and samples which had not metastasized after surgery, it was noted that there was increased expression of ANXA5 and DDOST in primary skin SCCs which was associated with subsequent development of metastasis (Shapanis et al., 2020a). Similarly, proteomic profiling of primary melanoma samples has identified proteins which might be associated with metastasis (Shapanis et al., 2020b). Similar to studies looking at DNA sequence alterations and RNA expression, the use of proteomics can be used to identify potential biomarkers and key processes which are associated with metastasis.

1.5.1 NGS Databases

NGS of multiple samples can yield large amounts of data. While this can provide vast amounts of information about the relevant disease that is being studied, the generation of large quantities of data has certain storage requirements so that all the data is stored, the data is kept securely on a long-term basis without any introduction of errors into the data, and so that appropriate access can be granted for use of the data by the global scientific community. While earlier sequencing studies often focussed (solely or mainly) on reporting their findings, it soon became apparent that the vast quantities of data being produced by the individual NGS studies could be added together to provide larger datasets and thus greater understanding of the relevant disease. There are many databases available online that store sequencing data for analysis, and this is also the case for data obtained from NGS studies on cancer.

COSMIC (Catalogue of Somatic Mutations in Cancer) is the world's largest database of mutations in cancer (Forbes et al., 2016). This dataset is manually curated by experts and contains clinical classification information for each cancer. The genetic variants contained in this database are from articles that have been published and have reported on the mutations which have now been deposited in the COSMIC database. The genetic database contains full mutation information at nucleotide or protein level for each variant. The Genome-wide screen data contained in the COSMIC database is peer reviewed and contains data from other databases such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA).

The National Centre for Biotechnology Information (NCBI) has created the database of Genotype and Phenotype (dbGaP) for authors to deposit genotype and phenotype data which is protected securely and can be accessed via submitting a request through a dbGaP authorized access portal. dbGaP contains data on many human diseases such as diabetic nephropathy, schizophrenia, psoriasis, etc. as well as on human cancer (Mailman et al., 2007). The National Institutes of Health Data Access Committee (DAC) reviews the request and grants access to datasets so they can be downloaded from the dbGaP authorized access portal (Mailman et al., 2007). The European Genome-phenome Archive also contains genetic and phenotypic data which is not publicly available, and data can be accessed by contacting the DAC similarly to dbGaP (Lappalainen et al., 2015). The data available in dbGaP and EGA are in the format the author has chosen to submit to the database.

The Genomic Data Commons (GDC) Portal contains publicly available exome data which has been funded by the National Cancer Institute such as TCGA data (Grossman et al., 2016). The exome data available has been analysed using different variant calling pipelines such as MuTect2, VarScan2 and Pindel and there has also been the development of the Multi-Centre Mutation Calling in Multiple Cancers (MC3) project (mc3.v0.2.8.PUBLIC.maf) which is a harmonized set of mutation calls across all TCGA samples (Gao et al., 2019). The data available from the MC3 project is available in Mutation Annotation Format.

Margaret Dayhoff was named ‘the mother and father of bioinformatics’ as she pioneered the application of computational methods in biochemistry (Gauthier et al., 2019). In 1978, a mathematical algorithm was developed for amino acid substitutions (Barker et al., 1978). With the developments in DNA sequencing, bioinformatics was used for DNA analysis and Unix-like operating systems were used to analyse this data. The mid 1980s saw the introduction of several scripting languages such as Perl and Python that were used in the production of bioinformatics programs (Fourment and Gillings, 2008). The 1990s led to the developments in non-scripting languages such as Fortran C, R and Java which are still used today to analyse genetic data (Gauthier et al., 2019).

1.5.2 Driver genes

A cancer driver gene is a gene that drives the development of cancer. Cancer driver genes have been described in various ways by different authors, for example as “one whose mutations increase net cell growth under the specific microenvironmental conditions that exist in the cell *in vivo* (Tokheim et al., 2016), or as “genes (called ‘cancer driver genes’), the mutant forms of which affect the homeostatic development of a set of key cellular functions”(Martinez-Jimenez et al., 2020). By contrast, passenger mutations have been described as “having no effect on the neoplastic process” (Vogelstein et al., 2013), or “do not provide any proliferative benefit and do not have an effect on the cancer cell” (Stratton et al., 2009). Although the term driver or passenger refers to the biological effect of a particular mutation in cancer development, both terms have been used to describe specific mutations in genes, e.g., driver mutations/passenger mutations, and to describe genes which contain driver or passenger mutations, e.g. driver genes/passenger genes. Furthermore, the terms are often used as nouns “drivers” or “passengers” to signify either genes or mutations.

Various bioinformatics approaches are used to distinguish drivers from passengers and functional assays are also used to determine whether the drivers identified via bioinformatics actually alter a cell’s phenotype. Different algorithms and statistical tests can be applied within the different bioinformatic programs to NGS data to distinguish driver mutations from passenger mutations. There is currently no gold standard method to identify driver genes in a dataset but there have been studies which have combined the use of various bioinformatic programs and/or allowed comparisons to be made between the performances of driver gene programs and the frameworks proposed (Tokheim et al., 2016, Bailey et al., 2018, Rheinbay et al., 2020). In one study, an analysis was conducted comparing eight driver gene programs (MutsigCV, ActiveDriver, MuSiC, OncodriveClust, OncodriveFM, OncodriveFML, Tumor Suppressor and Oncogenes (TUSON) and 20/20+) for somatic variants in exome data for 7916 samples consisting of 34 cancer types (Tokheim et al., 2016). The analysis checked if there was overlap between the Cancer Gene Census (CGC) and driver genes from the various methods, the observed vs. theoretical p value, number of significant genes predicted, and the prediction of consistency based on independent partitions of the dataset. The methods with the strongest support for identifying a set of genes that were shared by other programs and that were also present in the CGC were 20/20+, TUSON, and MutsigCV. The top drop consistency metric (TDC 10 and TDC 100) which was developed by the authors, were used to identify the top *k* genes in a ranked list when the driver gene analysis was applied to randomly partitioned sections of the datasets. The most consistent set of genes during this analysis was identified using MuSiC, 20/20+ and TUSON. The observed and theoretical p values of the driver gene methods were compared, and the observed p values were lower than the theoretical p values for MuSiC, ActiveDriver, OncodriveClust, OncodriveFM, and OncodriveFML. The theoretical p values were lower than the observed p values for TUSON and MutsigCV (Tokheim et al., 2016). When the theoretical p values are underestimated and are lower than expected this can affect the q value which is calculated using the Benjamini-Hochberg multiple testing correction and is used to determine driver genes ($q < 0.1$). This can result in the overestimation of driver genes.

Another study analysed the exomes of 9423 tumour samples and identified 299 driver genes (Bailey et al., 2018). The driver genes were detected in two phases, driver gene discovery and in silico mutation analysis. There were 8 different tools that were used to detect driver genes and these tools were based on mutation frequency (MuSiC2 and MutSig2CV), features of mutations (20/20+, CompositeDriver and OncodriveFML), clustering of mutations (OncodriveCLUST) and externally defined genomic regions (e-Driver and ActiveDriver). The in-silico mutation analysis was conducted with the use of sixteen tools which predicted the effect of the mutations identified by assessing the presence of clinically actionable mutations and how the mutation could affect protein structure and function. Potential driver genes were gathered from all the programs and a consensus score was calculated using the Gene Discovery Weighting Strategy. A Combined Tool Adjusted Total was used to score the in-silico mutation analysis tools.

Various bioinformatics programs were used in the identification of drivers in the Pan Cancer Analysis of Whole Genomes study analysing non-coding somatic drivers in 2,658 cancer whole genomes (Rheinbay et al., 2020). This study consisted of 13 methods for driver discovery (ActiveDriverWGS, CompositeDriver, DriverPower, dndscv, ExInAtor, LARVA, MutSig tools, NBR, ncdDetect, ncDriver, OncodriveFML and regDriver). The p values were combined from all programs and analysed under a framework before applying the Benjamini-Hochberg correction. After identifying driver genes, then the individual mutations within the driver genes were investigated. Stringent filters were applied to detect driver mutations such as at least three mutations in candidate genes should be present in at least three patients. Other factors were also taken into consideration including the presence of structural variants, breakpoints, and gene expression.

Although different approaches can be taken to discover potential driver genes, the power of the study is an important factor to consider (Lawrence et al., 2014). The background mutation rate can affect the ability to identify driver genes that are significantly mutated. Variation in the background mutation rate in cancer can be due to various factors such as the particular type of tumour, the environmental carcinogen that the organ/cells (from which the tumour arises) is/are exposed to, and the age and genetic background of the individual patients (Lawrence et al., 2014). Therefore, the identification of driver genes is particularly difficult in cancers which are caused by environmental carcinogens such as UV due to the high background mutation rate as a result of skin cancers arising more frequently in fairer-skinned, older individuals whose skin would have received multiple exposures to UV during their lifetime. Therefore, cancers with high background rates require a sufficiently high-powered study with a large sample size to detect significantly mutated genes.

1.5.3 Driver gene detection methods

MutSig2CV is a program that identifies genes which are significantly mutated compared to the background mutation rate. The program applies a patient-specific mutation frequency and a gene specific background mutation rate to correct for variation in background mutation rates between individuals and genes (Lawrence et al., 2013). This model expects each base to be mutated by chance, however the probability of these mutations occurring by chance vary between patients and genes. Patient factors considered in this model are the overall mutation rate and the overall mutation spectrum (the type of mutation such as a transversion or transition). The overall mutation spectrum is considered in order to avoid mutation bias because there can be a high rate of C to T changes at CpG dinucleotide sites which can affect the measure of true selection (Greenman et al., 2006, Lawrence et al., 2013, Yang et al., 2003). The 'CV' in MutSig2CV represents 'covariates' and the covariates that are utilised in this program are gene expression levels, DNA replication time and chromatin state. The gene expression covariate was chosen because genes which are highly expressed in the germline were identified to have less mutations due to transcription coupled repair (Fousteri and Mullenders, 2008). The replication time is important because late replicating regions have been shown to have higher mutation rates which could be due to a decrease in the number of available free nucleotides (Stamatoyannopoulos et al., 2009). Repressive histone modifications such as H3K9me3 have been associated with single nucleotide variation in cancers, thus suggesting that chromatin state can also affect mutation rate (Schuster-Bockler and Lehner, 2012). The number of mutations in each individual is considered in order to calculate patient specific scores. Then the number of mutations in each tumour in different individuals are analysed to identify the gene specific scores. These are subsequently combined to predict the total number of mutations per gene. Further research has shown that applying a gene-specific method of measuring background mutation rate could be more reliable than not using this approach, but that this can lead to low sensitivity when used for small sample sizes (Wong et al., 2014).

dN/dS is a program that measures selection for protein coding changes in the DNA (Greenman et al., 2006, Martincorena et al., 2015, Yang et al., 2003). The dN measures the number of nonsynonymous base changes and is compared to the dS which is the number of synonymous base changes per codon. A nonsynonymous base alteration is a mutation in DNA which causes a change to the amino acid whereas a synonymous mutation does not alter the amino acid.

In a gene under neutral drift, we would expect dN/dS to approximately equal 1, i.e., the number of nonsynonymous DNA changes is equal to the number of synonymous DNA changes. Positive selection is characterised as when the value of dN/dS is greater than 1 and suggests that there is selection for the new amino acid coding allele which increases the fitness of the cell. Negative selection is when dN/dS is less than 1 and is thought to result from the nonsynonymous variants decreasing the fitness of the cell, thus it generates fewer daughter cells per unit of time (Greenman et al., 2006). However, the use of dN/dS to identify positive and negative selection uses the assumption that most synonymous DNA changes are neutral (Martincorena et al., 2017).

There are several limitations associated with the use of dN/dS for measuring selection and additional parameters must be considered (Lawrence et al., 2013). Driver gene mutations confer positive selection, so would be expected to be more frequent than the overall average background mutation rate in a cancer. If the background mutational frequency is underestimated, it can lead to an overrepresentation of significant findings, inadvertently suggesting that certain higher frequency mutated genes are drivers. The background mutation frequency could also be overestimated which can result in an underrepresentation of significant findings and thus not identifying potential driver genes. Mutation rate heterogeneity between cancer types can affect the reliability of using a simplistic background mutation rate, for example paediatric cancers have lower mutation frequencies compared to adult cancers which develop after prolonged exposures to UV or smoking (Lawrence et al., 2013). Heterogeneity in the mutation spectrum can also affect these simplistic models when applied to different cancers, for example the majority of mutations in skin cancers are C to T changes (Pleasance et al., 2010a) which differs from the C to A mutation signature identified in lung cancers secondary to tobacco (Pleasance et al., 2010b). There is also heterogeneity in DNA repair and mutation frequency across the genome, which can relate to gene expression levels, and the mutation rate is lower in genes which are expressed in the germline (Lawrence et al., 2013). The late replicating regions of the genome have also been identified as having higher mutation rates (Stamatoyannopoulos et al., 2009).

Due to the limitations in these simplistic dN/dS models, there have been developments in the methods to measure selection in cancer genomes. As an alternative, the dNdScv R package has been developed (Martincorena et al., 2017). This package contains two different models for dN/dS. The dNdSloc model is the traditional dN/dS which measures the number of synonymous mutations in a single gene to infer the local mutation rate without information from other genes.

The dNdSloc model is used for larger datasets. Unlike MutSigCV which applies a gene specific background mutation rate, the dNdScv model combines the synonymous mutation rate in a single gene with the variable mutation rate across all genes. This statistical model can predict the variable mutation rate across the genome and the sample size will not affect the reliability of the model. MutSigCV relies on three covariates to determine mutation rate whereas dNdScv uses many covariates. The covariates include the first 20 principles of 169 chromatin marks from the Roadmap Epigenomics Project (Roadmap Epigenomics et al., 2015). The gene size, gene sequence and the impact of the substitution are accounted for in the model. The number of synonymous mutations per gene is modelled as a negative binomial distribution which is a Gamma-Poisson compound distribution and is used to estimate the background mutation rate to represent the uncertainty in mutation rate. The covariates are then used in the framework to decrease the unexplained variability.

The dNdScv model also uses a full pentanucleotide model to correct for any bias in the analysis. The dN/dS of whole genomes have been calculated across cancer types and the trinucleotide and pentanucleotide models were compared and it showed that there was not a significant difference between dN/dS values except for melanoma; this was due to the C > T changes, induced by UV, in melanoma being affected by nucleotide context beyond the trinucleotide level (Pleasance et al., 2010a).

OncodriveCLUST is a program which aims to identify genes under positive selection (Tamborero et al., 2013) similar to dNdScv. This method is based on the principle that mutation probability is not the same across a gene sequence. It has been identified that clustering of mutations is a marker of positive selection. The OncodriveCLUST program measures the clustering of mutations in a gene and the background mutation rate is based on coding silent mutations which are under no selective pressure and represents the baseline clustering. OncodriveCLUST was compared with MutSig using TCGA dataset and genes were compared to the CGC (Tamborero et al., 2013). The analysis showed that OncodriveCLUST identified genes that were not present in MutSig and could have the possibility to identify novel cancer driver genes that were not present in the CGC.

OncodriveCLUSTL, a later version of the mutation clustering program, has been identified as outperforming OncodriveCLUST and is based on an algorithm that measures the background mutation frequency according to a model created from the tri- and penta-nucleotide substitution frequency (Arnedo-Pac et al., 2019). The program detects genes from CGC in the dataset and identifies clusters of different sizes.

The aforementioned programs produce a list of genes based on different measures of driver gene status. MutSig2CV calculates driver genes based on mutation frequency whereas dNdScv, OncodriveCLUST and OncodriveCLUSTL are based on selection. The background mutation rate is modelled differently in each of the programs however both MutSig2CV and dNdScv use covariates to correct the background mutation rate. OncodriveCLUSTL and dNdScv corrects for context-dependent mutation biases by applying a penta-nucleotide model.

1.5.4 Examples of cancer driver genes

The *TP53* gene, a tumour suppressor gene, has been identified as one of the most frequently altered gene in human cancers and p53 pathway inactivation occurs in the majority of cancers (Olivier et al., 2010). The p53 protein is a 53kDa protein and is a transcription factor initially thought to enhance the rate of transcription of six or seven known genes (Levine, 1997), however, this number has risen to 3661 known target genes (Fischer, 2017). One of the genes regulated by p53 is *MDM2*. MDM2 protein is an E3 ubiquitin ligase and ubiquitinates p53 for proteasome degradation via the ubiquitin-dependent pathway (Moll and Petrenko, 2003). Therefore, p53 is negatively regulated by MDM2. In stressful cellular conditions there is an increase in p53 protein expression, and it is uncoupled from MDM2. The increase in p53 expression results in the regulation of p53 targets and protein-protein interactions (Vousden and Lu, 2002). This can lead to DNA repair, growth arrest, senescence, and apoptosis.

P16-Ink4a is CDK inhibitor and is part of the INK4A family of inhibitors. This tumour suppressor gene inhibits the S phase of the cell cycle (Romagosa et al., 2011). P16-Ink4a expression also inhibits the phosphorylation of Rb and promotes binding of E2F1 which results in G1 cycle arrest (Serrano, 1997). This gene is also highly correlated with HPV infection in head and neck squamous cell carcinoma (Stephen et al., 2013). P16 expression was measured in a study looking at skin biopsies from cSCC, Bowen's disease, BCC, seborrheic keratosis, and normal skin. P16 was overexpressed in 60% of cSCC samples and 50% of BCC samples. It was also identified that 68% of tumours located on sun exposed areas in comparison to non-sun exposed areas showed over expression (Conscience et al., 2006).

The Notch signalling pathway is the one of the most activated signalling pathways in cancer (Yuan et al., 2015). It is activated when a ligand binds to a receptor; there are four receptors, Notch1 to 4, and there are five ligands, delta-like ligand 1 (DLL1), delta-like ligand 3 (DLL3), delta-like ligand 4 (DLL4), Jagged-1 (JAG1) and Jagged-2 (JAG2) (Capaccione and Pine, 2013). The role of Notch signalling varies according to cancer type and can be considered an oncogene or a tumour suppressor gene (Fan et al., 2004). In mice, Notch1 deletion leads to tumour formation suggesting that Notch1 is a tumour suppressor gene in skin (Nicolas et al., 2003). Mice lacking Notch2 and Notch3 in their epidermis also develop skin cancers (Demehri et al., 2009).

Missense gain of function mutations in all three RAS genes are found in 27% of all cancers (Hobbs et al., 2016). The three RAS genes are HRAS, KRAS and NRAS. KRAS is the most mutated RAS gene in human cancers and HRAS is the least mutated RAS gene. These oncogenes activate the downstream targets by binding to GTP. This binding then promotes non proliferating cells to enter the G1 phase of the cell cycle (Taylor and Shalloway, 1996). Mutations in RAS especially HRAS is most common in cutaneous squamous cell carcinoma (Su et al., 2012).

1.6 Potential driver genes in skin cancer

Skin cancer can arise from keratinocytes and melanocytes. Melanoma arises from melanocytes whereas cSCC and BCC arise from keratinocytes (Figure 1-7). AKs and Bowen's disease are precancerous lesions which may develop into skin SCC (Albibas et al., 2017). It has previously been identified that between 0.025 – 20% of AKs progress to skin SCC (Callen et al., 1997, Marks et al., 1988, Quaedvlieg et al., 2006, Glogau, 2000) and 3 – 5 % of Bowen's disease in extragenital regions and 10% in genital regions progress skin squamous cell carcinoma (Cox et al., 1999).

Most BCCs occur sporadically however they can also arise as a result of Gorlin syndrome (also known as basal cell nevus syndrome (BCNS)) where individuals have germline defects in *PTCH1* (Amakye et al., 2013, Epstein, 2008). It was found that BCNS patients have lower mutational loads and lower proportion of UV mutation signatures and fewer mutations in genes involved in DNA checkpoint repair and genome stability (Chiang et al., 2018). The genes most frequently mutated in BCNS in this study were *TP53*, *FANCA* and *BRCA1*. Potential driver genes have been identified in BCC using the InVEx program and a simple binomial distribution has been used to distinguish driver mutations from passenger mutations (Jayaraman et al., 2014). Bonilla et al., 2016 used MutSigCV for their BCC exome sequencing analysis (Bonilla et al., 2016). The genes identified as significantly mutated in these studies include *TP53* and those identified in the Hedgehog pathway such as *PTCH1*, *SMO* and *SUFU* (Jayaraman et al., 2014, Bonilla et al., 2016).

The most common program used to identify driver genes in skin SCC has been from the MutSig suite (Pickering et al., 2014, Inman et al., 2018, Chitsazzadeh et al., 2016, Li et al., 2015). Studies have also used IntOGen (Pickering et al., 2014, Cammareri et al., 2016) OncodriveCLUST and

OncodriveFM (Inman et al., 2018) and their own statistical methods (Pickering et al., 2014, Yilmaz et al., 2017). NGS analysis has identified that the most common genes which are significantly mutated in skin SCC are *TP53*, *NOTCH1*, *NOTCH2* and *CDKN2A* (Albibas et al., 2017, Pickering et al., 2014, Li et al., 2015, Chitsazzadeh et al., 2016).

Melanoma samples have been included in the large Pan Cancer Studies as previously described (Rheinbay et al., 2020) which have used a range of driver gene packages for analysis. *BRAF, NRAS* and *PTEN* genes are mutated in melanoma*, TP53* and *CDKN2A* are also significantly mutated in melanoma similarly to skin SCC (Hodis et al., 2012).

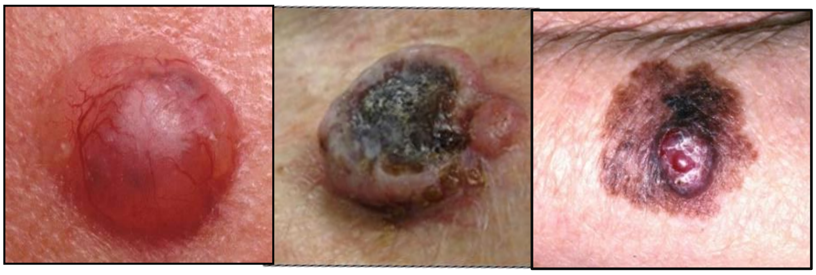


Figure 1-7: Clinical images of skin cancers, showing BCC, cSCC and melanoma (left to right).

1.6.1 Studies in normal skin and precancerous lesions

NGS has also been conducted in chronically sun-exposed 'normal' skin in a study that analysed 234 biopsies of 0.79 – 4.71mm^2 from eyelid skin that had been excised from four individuals and DNA then extracted to sequence the exons of 74 genes (Martincorena et al., 2015). That study identified that the genes which were significantly mutated in sun-exposed normal skin were *NOTCH1*, *NOTCH2*, *FAT1*, *TP53* and noted that these genes are also frequently identified as driver genes for cSCC.

Albibas et al., 2017 investigated gene mutations in p53 immuno-positive patches (PIPs) in chronically sun-exposed 'normal' skin. PIPs are clusters of cells that have accumulated nuclear p53 protein, and thus can be detected immunohistochemically using an anti-p53 antibody, in chronic sun- (or UV-) exposed skin (Berg et al., 1996, Brash et al., 1996, Jonason et al., 1996, Kanjilal et al., 1995, Rehman et al., 1994, Rehman et al., 1996, Ren et al., 1997, Rebel et al., 2005, Tabata et al., 1999, Urano et al., 1995, Ziegler et al., 1994). These clusters represent keratinocytes that frequently harbour *TP53* mutations that lead to clonal proliferation (Jonason et al., 1996), thus the mutated *TP53* keratinocytes can gain a growth advantage over normal keratinocytes (Ziegler et al., 1994). In the Albibas et al. 2017 study, targeted sequencing was conducted on 18 genes with the final analysis reporting on 15 PIPs ranging in size from 0.14–0.27 mm^2. The study identified that PIPs contain mutations in multiple cancer related genes and although some mutations were clonal, there were also subclonal mutations. This finding supported the 'Big Bang' model of cancer where tumours grow as a single expansion where clonal and subclonal alterations occur early in tumour growth (Sottoriva et al., 2015).

In another study, normal skin samples that would have been chronically sun-exposed were taken from ten patients with varying age ranges undergoing Mohs surgery for skin cancer (Lynch et al., 2017). The 16mm^2 epidermal samples were sequenced for 121 genes. This study showed that the distribution of clone sizes in the epidermis was so high that they were unlikely to have occurred by neutral drift according to the time frame of a lifespan. Mathematical modelling and computational simulation showed that the high distribution of clone sizes was likely due to secondary mutations occurring at clone boundaries, such as in clones containing *NOTCH1* mutations, providing a competitive advantage for these cells at the boundary. In addition, the study concluded that the fate of a mutant stem cell in this scenario depends on a combination of neutral drift, cell competition and spatial constraints.

Normal skin samples were taken from different body sites and 2mm^2 (1 x 2mm grid) samples for DNA extraction were collected to sequence 74 genes in another investigation (Fowler et al., 2020). The dN/dS ratio was used to identify genes under selection, and 11 genes were identified under positive selection, including 5 novel genes and 6 genes previously identified in Martincorena et al., 2015 (*NOTCH1*, *TP53*, *NOTCH2*, *NOTCH3*, *FAT1* and *RBM10*). The prevalence of mutant *NOTCH1* and *FAT1* in the normal skin was the same as those observed in cSCC and BCC, which suggests that mutations in these genes may be involved in colonising normal tissues and not driving tumour formation. This has led some researchers to question whether some of these mutated genes are truly driver genes because many of these clones do not give rise to cancer (for example the author of this thesis was asked this question during a presentation of work contained in this thesis), but an alternative view could be that colonisation of normal epidermis by mutant cells helps to drive subsequent skin tumour formation. In the same study, mutant *TP53*, *NOTCH2* and *KMT2D* were enriched more in skin SCC and BCC compared to normal skin (Fowler et al., 2020). This study also identified that competitive selection is not uniform all over the body.

Another approach that involved identifying mutations from a large collection of RNA sequences from the Genotype-Tissue Expression (GTex) project involving 6700 samples from 500 individuals spanning 29 normal tissues (since some mutations in DNA can be found in corresponding RNA) identified that the number of mutations in skin increased with age (Yizhak et al., 2019). That study also identified that the increase in mutations was significantly higher for sun-exposed skin compared to non-sun exposed skin (Yizhak et al., 2019).

In terms of later precancerous skin lesions, WES of AKs was conducted in the Albibas et al., 2017 investigation, as well as in studies by Rodriguez-Paredes et al., 2018 and Chitsazzadeh et al., 2016, but the total number of AKs in which WES has been performed is limited to date. High mutational heterogeneity has been seen in AKs in each of these studies, and although methylation profiles of AKs and cSCCs has shown similar profiles and suggested that there are two corresponding subtypes of AK and cSCC (Rodriguez-Paredes et al., 2018) the exome data did not show any correlation between the methylation profile and the genomic profile to support the two subtypes of skin SCC at a mutation level. The studies by Chitsazzadeh et al., 2016, Albibas et al., 2017 and (Rodriguez-Paredes et al., 2018) also identified that many similar genes were mutated in AKs and cSCCs and, separately in the Albibas et al., 2017 study, in Bowen's and cSCCs. Furthermore, the latter group showed many identical mutations in contiguous lesions where cSCCs had arisen from an adjacent AK or Bowen's, thus proving genetically that AKs and Bowen's can develop into cSCC (Albibas et al., 2017). There had also been previous studies which had conducted single gene analysis in Bowen's disease (Lee et al., 2000) and AKs (Nelson et al., 1994) however the NGS studies have significantly moved the field forward in terms of understanding how mutated AKs are.

As melanoma develops from melanocytes, the normal skin from 19 sites across 6 donors was collected to look for mutations within cutaneous melanocytes (Tang et al., 2020), in a manner similar to the above work on keratinocytes in chronically sun-exposed skin. The number of melanocytes is much lower than the number of keratinocytes in the epidermis, therefore epidermal cells were collected from the tissue samples and the cells were then cultured. Single cells were sorted for clonal expansion, then DNA and RNA were extracted. RNA sequencing was used for cell identification and DNA from the corresponding melanocytes demonstrated loss of function mutations in *NF1*, *CBL* and *RASA2*, which are part of the MAPK pathway (Tang et al., 2020). The study also identified that the mutation burden of melanocytes from sun exposed normal skin varied by several magnitudes.

A study sequenced 293 cancer genes in 150 areas of 37 primary melanomas and their adjacent precursor lesions (Shain et al., 2015b). The results showed that the mutation burden varied significantly according to the anatomical site of excision, age of the patient at diagnosis and the total amount of sun exposure. Patients with BRAF V600E mutations were more common in younger patients who had melanomas in intermittent sun damaged skin (Viros et al., 2008) whereas patients with *NRAS*, *BRAF* V600K and K601E mutations occurred in older patients with chronically sun exposed skin (Long et al., 2011). This study identified that there were different mutation pathways for melanoma progression by comparing the precursor lesions and the primary melanoma. It was found that BRAF V600E mutation were identified in benign naevi whereas BRAF V600K, K601E and NRAS mutations were identified in melanoma in-situ or intermediate lesions which had already accumulated other mutations (Shain et al., 2015b).

Acquired melanocytic nevi can transform into melanoma and this has been suggested to have occurred in around 30% with a range of 4-72% of melanomas and that 70% of melanomas arise de novo (Pampena et al., 2017). Therefore, in one study, WES was conducted in 30 matched nevi and adjacent normal skin to look at genetic factors involved in development of melanocytic naevi (Stark et al., 2018). Mutually exclusive mutations in *BRAF* and *NRAS* were most commonly identified in the nevi, with mutations in *BRAF* approximately five times more common than *NRAS* mutations. Novel genes were identified as mutated in the nevi, in addition to well-known driver genes such as *HDAC9*, *MYH11* and *DCC* (Stark et al., 2018). The study also identified that nevi could be clonal but also that some might not be and could be polyclonal instead. Although *TERT* promoter mutations are common in melanoma (Hayward et al., 2017, Horn et al., 2013, Huang et al., 2013), there were no common *TERT* mutations seen in the acquired nevi (Stark et al., 2018).

WGS was conducted on 14 benign melanocytic nevi, including congenital and acquired nevi, in another study (Colebatch et al., 2019). All the nevi had mutations in genes which activated the MAPK pathway and had mutually exclusive mutations in *BRAF* or *NRAS*. Nevi with high mutational loads showed UV mutation signatures whereas nevi with low mutational loads did not. In this study, *TERT* promoter mutations were detected in nevi from 2 individuals, which suggests that these mutations can arise prior to the development of melanoma.

1.7 Skin cancer prevention and treatments

Different groups of people have varying risks for skin cancer and an individualised approach should be taken to prevent skin cancer. The National Institute of Clinical Excellence guidelines on sunlight and skin; risks and benefits (NICE, 2016) state that children particularly babies, people who tend to burn rather than tan, people with lighter skin, red hair, blue or green eyes, and/or have lots of freckles, people with many moles, immunosuppressed individuals or those who have a history of skin cancer should take extra care to avoid sun damage. People who work outside, have outdoor hobbies and those who take holidays with increased sun exposure are also at a higher risk of skin cancer. Groups who have little or no exposure to sun for their own reasons, are housebound or confined to working indoors for long periods of time are at increased risk of vitamin D deficiency. Therefore, the guidelines promote that there are risks and benefits to sun exposure and a balance needs to be achieved in line with everyone's own risk factors. Protection from the sun can be achieved by wearing suitable clothing, staying in the shade and applying sunscreen (Green et al., 2011). Vitamin D supplements should be taken by those who have minimal sun exposure.

The first line of treatment for cSCC is surgical excision and a low risk cSCC is considered as having a horizontal diameter of less than 2cm and high risk cSCCs have a diameter of more than 2cm (Stratigos et al., 2015). Radiotherapy can be used to treat cSCC and a study has suggested that surgery and adjuvant radiotherapy provide the best treatment for skin SCC (Veness et al., 2005). In advanced/metastatic cSCC, immunotherapy with anti-PD-1 can be employed. First line treatment for most BCCs is surgery but radiotherapy can also be used (Peris et al., 2019). In some countries, locally advanced BCC can be treated by Vismodegib and Sonidegib, which are hedgehog pathway inhibitors targeting the oncogenic smoothened protein (SMO), and hedgehog pathway inhibitors are also approved for metastatic BCC (Peris et al., 2019). Patients with BCNS-BCC are more responsive to SMO inhibitors than sporadic BCC patients and have fewer functionally resistant SMO mutations (Chiang et al., 2018). In melanoma, the treatment can differ depending on the individual features of the tumour such as the location, the tumour stage, and the genetic profile. While the first line treatment for melanoma is surgery, other treatments can include palliative radiotherapy or photodynamic therapy for individual metastases, and chemotherapy, targeted therapy, or immunotherapy for disseminated metastases. The targeted therapy used in melanoma includes BRAF and MEK inhibitors for patients whose melanomas have BRAF mutations (Hamid et al., 2013), whereas immunotherapy includes PD-1 and CTLA-4 inhibitors (Ribas, 2012, Leach et al., 1996).

While melanoma and cSCC are more likely to metastasise than BCC, there are lots of studies that have investigated treatment options for advanced/metastatic melanoma, but treatment options for advanced/metastatic cSCC are more limited. SCC also occurs in other organs, including head and neck (i.e., oropharyngeal), lung, oesophagus, etc., and those SCCs are often more likely to metastasise than cSCC, therefore it might be possible in the future to adopt treatment approaches that are developed to target common mutations that are present in SCCs arising in those organs. Related to this, some studies have revealed that several genes (including *TP53*, *CDKN2A*, *NOTCH1*) mutated in aggressive/metastatic cSCC are also mutated in head and neck SCC, and that certain driver genes (e.g. *TP53*, *CDKN2A*, *HRAS*) in cSCC are shared with lung SCC (Pickering et al., 2014, Li et al., 2015). In addition, Chitsazzadeh et al., 2016 compared mRNA expression profiles of cSCC to SCC in other organs by using gene set enrichment analysis (GSEA) and identified that mRNA profile in skin SCC was most similar to oropharyngeal SCC but least similar to cervical SCC, which is predominantly virally driven by human papillomavirus.

1.8 Mutation rate and UV exposure in skin

NGS databases and publications which have reported on NGS mutations in skin cancers and chronically sun-exposed skin can provide information on the type and number of mutations identified in skin after prolonged UV exposure over many years. However, there have not been any studies identifying how many mutations are generated after a shorter duration of UV exposure (e.g., over several weeks). To do this, it would be necessary to develop a bioinformatics pipeline that would help analyse mutations in epidermal samples from patients who have received UV exposure over several weeks. To relate numbers and types of mutations to the doses of UV received, the UV doses would need to have been measured and documented. NB-UVB is administered over several weeks as a treatment in psoriasis and the UV dose is recorded for each exposure during the course (Bhutani and Liao, 2010). Not only would a bioinformatics pipeline that could analyse the mutation burden from a NB-UVB course be useful to allow one to extrapolate from this the number of NB-UVB courses psoriasis patients could safely receive over their lifetime without being at significantly higher risk of skin cancer, but it might also be helpful to estimate the mutation burden from natural UVB and might also allow better understanding of skin cancer development in relation to UVB exposure. Although not part of this current thesis, skin biopsies from sun exposed and non-sun exposed areas have been collected by colleagues in Dermato-pharmacology, University of Southampton from patients before and after a course of NB-UVB treatment over the past 3 years and the plan is that they will be sequenced in the Wellcome Sanger Institute in Cambridge, and the bioinformatics pipeline generated in this thesis will be used to analyse that data in a future study.

1.9 Aims

To generate a bioinformatic pipeline that could be used in a future study investigating the long-term safety of NB-UVB treatment, this thesis will involve undertaking bioinformatics research on NGS data from skin cancers (cSCC, BCC and melanoma), other SCCs which could inform on common driver genes shared between those SCCs and cSCC, and on precancerous skin lesions and UV exposed normal skin. The initial work in the thesis will focus on cSCC to acquire the relevant bioinformatic skills and then the thesis will address the following aims:

1. To delineate the mutation signatures and driver genes in cSCC

2. To identify mutation signatures and driver genes in oropharyngeal SCC, lung SCC, oesophageal SCC, and cervical SCC, and to compare these between cSCC and the other SCC to examine for common driver genes.

3. To determine whether common driver genes exist between cSCC, BCC and cutaneous melanoma, because each of these arise as a result of UV exposure, despite the former two cancers arising from keratinocytes and the latter arising from melanocytes.

4. To compare the bioinformatics information from cSCC, BCC and melanoma with mutations identified by NGS in chronically sun-exposed normal and precancerous skin lesions to enable the production of a pipeline that would allow future analysis of skin samples from people who receive repeated UV exposure over several weeks (e.g., patients receiving NB-UVB treatment).

2. Methods

2.1 Programming tools

2.1.1 Iridis HPC

Iridis is the University of Southampton high performance compute cluster. Iridis 4 was used for this analysis which was their fourth-generation cluster. Iridis was designed to provide a batch service for users to submit parallel jobs or they could use it to run a multiple resource intensive job. Iridis 5 is currently used by the University of Southampton, and it is four times more powerful than its predecessor. Iridis 5 has 3 login nodes with 40 cores and 384 GB of memory.

Iridis uses the Linux operating system which provides access to bioinformatics tools. Linux operating system relays messages between programs and processes requests using the computer's hardware so one process does not pre-empt another process (Finney, 2001). Linux also has the ability to suspend a process for a limited period of time. Therefore, these characteristics are beneficial when manipulating large-scale genomic data. Iridis was used in this analysis to access a bioinformatics program called Bedtools (version 2.21.0)(Quinlan and Hall, 2010). The Bedtools program used the exome reference file to identify and extract all the variants in the exome from the whole genome mutation file. Bedtools functions with two input files: a reference file which contained all the exome chromosome co-ordinates (hg19 genome positions) and a whole genome mutation file.

2.1.2 Python

Python is a programming language wildly used in genomic analysis (Jenkins, 2004). For local use the, Oncotator bioinformatics program (Ramos et al., 2015) was available as a Python module which was used in this study and Oncotator required Python version 2.7. However, OncodriveCLUSTL, which is another bioinformatics program that was used in this analysis, required Python 3.5. The Miniconda installer (https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html) was used in this instance, as it included Python 3.5 and all the packages OncodriveCLUSTL depended on.

2.1.3 R statistics and packages

R (https://www.r-project.org/) is a language that is used for statistical programming and graphics. R version 3.5.1 was used to run Maftools (v2.4.15). Mutation Annotation Format (MAF) files are the file format required for analysis in Maftools, where they standardise the annotation of mutations. Oncotator was used for annotation and constructed these MAF files. Maftools was used to analyse and visualise mutation data. Mutation signatures for single base substitutions (SBS) was also identified using Maftools as well as manipulating the genomic data to identify the most frequently mutated genes in the dataset and the most common mutation changes.

2.2 Bioinformatic tools

2.2.1 Sigminer

The double base substitution (DBS) mutation signatures were identified using Sigminer 1.0.16 (Wang et al., 2021) as Maftools was only able to identify SBS mutation signatures. A DBS mutation signature shows the proportion of samples where two consecutive reference nucleotides are replaced by another two nucleotides such as adenine, thymine, guanine or cytosine. The mutational pattern was then categorised into one of the reference DBS mutation signatures (Alexandrov et al., 2020). This was important to calculate as UV is known to induce DBS mutations, via direct DNA damage at dipyrimidine sites (CC > TT) (Brash, 2015).

Default settings were used with the addition of the 'add_trans_bias' command to consider transcriptional bias categories. This ensured that the double base substitutions that occur were as a result of DNA damage and not due transcription-coupled nucleotide excision repair in response to DNA damage (Haradhvala et al., 2016). The DBS matrix used was the one which had 78 DBS changes which was in line with the matrix used in the COSMIC database (Alexandrov et al., 2020).

A Nonnegative matrix factorisation (NMF) library was loaded in Sigminer to enable the optimal number of mutation signatures to be estimated, using the Brunet et al., 2004 settings, and the default number of runs were two per each value in the range which was '2:5'. These values were used due to the tumour sample size and the cophenetic plot showed robustness of clusters above 0.9 between the range of 2:5. The range represents the possible number of DBS signatures which can be identified from the matrix.

2.2.2 Oncotator

Oncotator annotates mutation data and was specifically designed for cancer genome annotation (Ramos et al., 2015). It provides information such as the gene name and the functional effect of a point mutation, insertion and deletion. A local version of Oncotator was downloaded as a Python module (version 2.7) due to our large datasets. The annotation directory used was oncotator_v1_ds_April052016 which was the latest version of the annotation directory, and the genome version was hg19 which is the same as GrCh37.

2.2.3 MutSig2CV

MutSig2CV v3.11 (Lawrence et al., 2013) was used to analyse somatic point mutations and to identify genes which have mutations that occur more than expected by chance with regards to the background cellular mutation rate. The MutSig2CV v3.11 package was downloaded and the script for downloading was gifted by Dr Mat Rose-Zerilli and is detailed in appendix 7.7. The download package included a reference folder.

To enable MutSig2CV to function, a mutation table was required and in this case, it was the MAF file produced from the Oncotator output. Then a coverage table was required which contains information about the coverage achieved for each gene in each patient, since this information was not available for this analysis a reference coverage file from the program was available in the reference folder. The coverage was required because this provides information on how many reads were aligned to a base during DNA sequencing. A base change could have differing effects therefore the coverage can provide information on how certain a nonsynonymous change has occurred compared to a synonymous change by analysing the read depth. A covariates table is also used in the analysis and this contains the gene names, the expression level of the gene averaged across many cell lines in the Cancer Cell Line Encyclopedia (Barretina et al., 2012), the replication time of each gene and the chromatin compartment of each gene. This covariate information can affect the background mutation rate of each gene which enables the detection of significantly mutated genes against variable mutation rates.

MutSig2CV uses three independent statistical tests to identify significantly mutated genes. 1) The abundance test infers if the gene is highly mutated relative to the background mutation rate. The program normalises the background mutation rate which can vary across patients, genes and sequence contexts which is accounted for in this analysis. The 'context_and_effect' reference files were used to calculate the abundance score. 2) The clustering of genes is also measured, as genes often harbour hotspots which are regions where there are frequent mutations. Therefore, this enables the program to differentiate between genes which have a uniform distribution and those which have mutation localised to hotspots. Higher significance is given to genes with localised mutations. 3) The conservation test is used to identify the functional significance of a mutated base where higher significance is given to regions of the genome that are highly conserved in vertebrates. The 'conservation' reference files are used for this analysis.

2.2.4 dNdScv

The dNdScv the program was run to identify driver genes via selection using a maximum-likelihood dN/dS method(Martincorena et al., 2017). The program identifies genes under positive selection by comparing the number of non-synonymous mutations to the number of synonymous mutations to compare protein altering mutations to those that do not alter protein structure.

In large datasets, where there are many synonymous mutations, the dNdSloc model is used. This model uses the number of synonymous mutations in a gene to predict the local mutation rate without extracting information from other genes. In this study, the dNdScv model which was used combined dN/dS with a negative binomial regression with many covariates. The dNdScv model combined the local mutation rate with the variation in mutation rate across genes by using epigenetic covariates that included data from 63 cell lines and 10 different epigenetic marks to predict the background mutation rate.

The genomic data file compatible for dNdScv was read into R and then dNdScv was run for the file. The 'dndsout' is the output after dN/dS is run which included a list of objects that show maximum likelihood estimates of the dN/dS ratios for each gene for missense, nonsense, essential splice site mutations, indels; the p and q values for missense, truncation, substitutions, indels and a global value which incorporates all the mutation types. Significant genes were then characterised as genes with a global q value less than 0.1. These genes which satisfied the criteria were then retain for further analysis.

2.2.5 OncodriveCLUST and OncodriveCLUSTL

The OncodriveCLUST program (Tamborero et al., 2013) is present within the Maftools package in R. This program identifies driver genes via spatial clustering of mutations. The parameters used to the run program were that the minimum mutations for any gene was set to five to be included in the analysis. First, the protein altering mutations were identified in the samples and those that are thought to be occurring more than chance are highlighted in the dataset. These positions were grouped to form clusters and scores were assigned. The scores were assigned according to the number of mutations in each and the length of the gene enclosed within the cluster. Then a gene clustering score was calculated by producing the sum of the cluster scores found in a gene. Each gene cluster score was compared to the background mutation rate model. There were different measures of significance in this analysis such as Z score or a p value based on a Poisson model. The method chosen in this analysis was ‘combined’ which meant the lower of the two p values produced from the Z score and Poisson model was chosen and the genes were ranked according to this value.

OncodriveCLUSTL (Arnedo-Pac et al., 2019) identifies driver genes using a sequence-based algorithm that identifies clustering in a genomic region. The coding regions of hg19 were downloaded from the bitbucket repository (https://bitbucket.org/bbglab/oncodriveclust/src/master/). The default settings were used except for concatenate which was changed to ‘TRUE’ for all samples. This was done so the program could detect clusters of two or more single nucleotide variants which span two exons. This was conducted by concatenating the genomic elements by connecting two exons by gluing their consecutive ends together. The program calculated mutation signatures based on mutation frequencies by identifying when the mutated base occurs compared to the reference base and compared this to the total number of substitutions. The input file and the file with the hg19 reference coding regions were used to calculate a background mutation model and by default OncodriveCLUSTL calculated a relative mutation frequency using the methodology published by (Mularoni et al., 2016). The mutational probabilities were calculated from the reference genome and alternate single nucleotide variants were compared. Following this, the program then conducted clustering analysis and the mutational clustering was only conducted on the regions that were annotated in the hg19 reference of genomic coding regions which were referred to as genomic elements. Then mutation clusters were scored according to the number of single nucleotide variants they contain to produce a cluster score. The genomic element score is the sum of the score of the clusters within the element. Analysis of simulated mutations was also conducted in OncodriveCLUSTL where mutations that were observed in a genomic element were simulated at random several times. These simulated mutations were then compared to the observed mutations to calculate clustering probabilities. There were three types of p values which were calculated: empirical, analytical, and top cluster. The empirical p values were calculated based on the fraction of iterations that had a simulated genomic element score that was greater than the observed genomic element score. The analytical p value was calculated by fitting the simulated genomic element score to a gaussian kernel density estimate distribution and then identifying the upper quartile of the observed genomic element score. The third p value was calculated for the top cluster p value and this was when a simulated cluster score was fitted to a gaussian density estimate distribution and the upper quartile of the observed cluster scores were identified. All the p values were adjusted by the Benjamini-Hochberg method to produce q values. The analytical p and q values were used in the analysis in this thesis as per the results in the Arnedo-Pac et al., 2019 publication.

2.3 Comprehensive Literature search

2.3.1 Database resources used

To determine the mutational landscape of SCCs and skin cancer, a comprehensive literature search was conducted to identify all publications containing whole exome, whole genome, or Next Generation Sequencing (NGS) targeted sequencing data using similar guidelines as outlined by PRISMA (Liberati et al., 2009). PUBMED (https://www.ncbi.nlm.nih.gov/pubmed/) is a free database which is maintained by the National Centre for Biotechnology Information (NCBI). It contains citations and abstracts from articles in the fields of medicine, dentistry, veterinary medicine, health sciences and preclinical sciences. The citations provide a record of articles before they are indexed with MeSH and added to MEDLINE. MeSH, which is an acronym for Medical Subject Headings, is the National Library of Medicine's vocabulary thesaurus which is used by MEDLINE indexers in identifying and recording the subject content of a published article. These terms are organised in a hierarchical structure. Therefore, MEDLINE contains a controlled set of MeSH terms which can be personalised in a search enabling a more sensitive and specific literature search (https://www.nlm.nih.gov/bsd/pmresources.html).

According to the MEDLINE Indexing Process: Determining Subject Content, the MEDLINE indexers read the title, introduction and summary/conclusions, and scan the materials and methods, results, abstract, keywords and bibliographic references to generate the MeSH terms for each article (https://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/theindexingprocess/index.html). During the indexing process, terms which mean the same thing but differ slightly in spelling are classified under one MeSH term, therefore 'Bowen Disease' and 'Bowens Disease' would be classified under the MeSH term 'Bowen's Disease'.

2.3.2 Search Terms and searching techniques used

MEDLINE Ovid was used to conduct the literature search. Each search term was searched as an individual word to identify if the term had any MESH terms. If the term did not have a corresponding MESH term, the search was conducted as free text. Free text searches known as multi-purpose searches, scan the title, abstract, unique identifier, keyword heading word, name of substance word, protocol supplementary concept word, rare disease supplementary concept word, subject heading word and synonyms (http://ospguides.ovid.com/OSPguides/medline.htm). A combination of search terms was designed to identify NGS data. Papers were only selected post-2007 to ensure the search was only focused on NGS data and not mutations identified via Sanger Sequencing. The papers identified in this search were compared with those identified from the COSMIC database (See section 2.4.1 below) by using the unique identifiers also known as PUBMED ID (PMID) numbers. COSMIC (Catalogue of Somatic Mutations in Cancer) is the world's largest human-curated database of mutations in cancer (Forbes et al., 2016). This search was conducted for skin SCC, lung SCC, oropharyngeal SCC, oesophageal SCC and cervical SCC to identify if there were any significant papers identified in the literature that were not present in COSMIC. The search was also conducted in skin melanoma and basal cell carcinoma. The search terms used are detailed in appendix 7.9.

2.3.3 Collating results

PMIDs from the literature search were collated and screened to ensure all exome or genome data available in the paper was included. The papers were annotated and links to external databases containing genome data were identified, data was downloaded from supplementary material and data was extracted from tables or figures in the paper. The data which was not available in papers was recorded and the email addresses of the lead author from those papers were noted. The email addresses were checked for each author to ensure it was correct and was changed if they had moved institution or another author was selected in any cases where the author was deceased. An email was sent requesting this data and if there was no response a further email was sent after at least two weeks and the study was discarded if there was no response after a month.

2.3.4 Data Extraction

In the cases of where results from the literature search was stored in external protected databases, the whole genome or exome data was downloaded from dbGaP and EGA database. In dbGaP, the data was available in the dbGaP portal after Data Access Committee approval and was downloaded using Aspera (https://www.ibm.com/products/aspera). A repository key was downloaded and was in the form of an NGC file. The repository key was then used to decrypt the data. For EGA downloads, Miniconda was used to download Python 3.8.3 so the download client would be compatible. The download client pyEGA3 was used to download data from EGA. Once the Data Access Committee approved the data access request, the data was available in the EGA portal.

The data from supplementary material from papers were taken and processed in different ways according to the way the supplementary data was supplied. The supplementary material from the journal articles were in Excel format and contained the exome data or the exome data for each individual tumour was available in separate text files. The chromosome number, start position, end position, reference base, alternate base and sample identification was extracted from the different file types from these journal articles. The files were edited to ensure the chromosome numbers included X and Y chromosomes. Duplicate lines were also deleted using awk and data from each individual paper was processed separately in Oncotator.

2.3.5 Cross-referencing submission to cancer mutation repositories

The data from the literature search was compared to genomic data in the COSMIC database and only genomic data which was not in the COSMIC database was extracted from the literature search. This ensured there were no duplicate datasets. This was done for all exome or genome data from papers about each specific SCC. The data from GDC portal, COSMIC database and journal articles were concatenated together to produce a MAF file with all the data available for each individual cancer.

2.4 Cancer mutation data mining

2.4.1 COSMIC

The data in COSMIC is curated on a per sample basis so mutation or clinical data was only in the database if it had been provided by the author. The release used for analysis was v91 and was downloaded on 21st April 2020. The files downloaded were classification.csv, CosmicMutantExport.tsv.gz, CosmicCodingMuts.vcf.gz and CosmicNoncodingVariants.vcf.gz. Variants were aligned to Genome Reference Consortium Human Build 37 (GrCh37) because Oncotator (Ramos et al., 2015) was used to annotate the genome and this program required variants that are aligned to this build of the genome. The classification.csv file provided a record of the clinical information available from the samples in the database and an example of the information provided for a single comma-separated record (i.e. tumour from one patient) is presented in table 2-1.

***Table 2-1**: **COSMIC classification.csv information provided for a single record.** NS represents 'Not Specified'. Column headings in bold were used in the analysis of the genetic landscape of SCCs and skin cancer.*

Column Headings	**Example Input**
COSMIC_PHENOTYPE_ID	COSO33166091
SITE_PRIMARY	**skin**
SITE_SUBTYPE1	foot
SITE_SUBTYPE2	sole
SITE_SUBTYPE3	NS
HISTOLOGY	**melanoma**
HIST_SUBTYPE1	**nodular**
HIST_SUBTYPE2	arising_in_nevus
HIST_SUBTYPE3	NS
SITE_PRIMARY_COSMIC	skin
SITE_SUBTYPE1_COSMIC	foot
SITE_SUBTYPE2_COSMIC	NS
SITE_SUBTYPE3_COSMIC	NS
HISTOLOGY_COSMIC	malignant_melanoma
HIST_SUBTYPE1_COSMIC	nodular

HIST_SUBTYPE2_COSMIC	NS
HIST_SUBTYPE3_COSMIC	NS
NCI_CODE	C4225
EFO	http://www.ebi.ac.uk/efo/EFO_0008515

The COSMIC mutation data was provided in a CosmicMutantExport.tsv.gz file, this contained all the coding point mutations from targeted and genome-wide screens. A genome-wide screen includes whole genome and whole exome data. The information provided for a single record (i.e., one mutation in one gene from one patient's cancer) is presented in table 2-2. The 'GENOMIC_MUTATION_ID' was also used as a common identifier between the CosmicMutantExport.tsv.gz file and the CosmicCodingMuts.vcf and CosmicNonCodingVariants.vcf files. The sample name was used as a second form of identification of each tumour and the mutation type description was used to organise variants according to their mutation type. The sample type heading was used to ensure the tumour types were only from primary and/or metastatic human tumour samples and not cultured cells.

***Table 2-2**: **COSMIC genome-wide screen data available for a single record**. NS represents 'Not Specified'. Column headings in bold were used in the analysis of the genetic landscape of SCCs and skin cancer.*

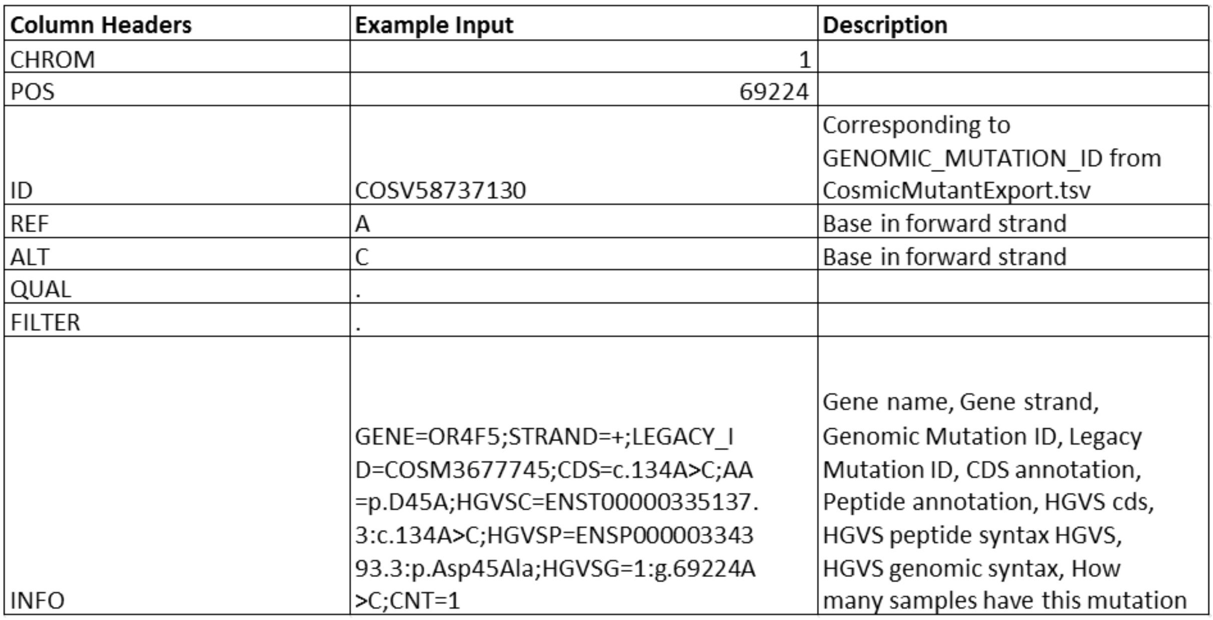
Column Headings	**Example Input**
Gene name	FANCM
Accession Number	ENST00000267430.5
Gene CDS length	6147
HGNC ID	23168
Sample name	**TCGA-D1-A17A-01**
ID_sample	1783523
ID_tumour	1687522
Primary site	**endometrium**
Site subtype 1	NS
Site subtype 2	NS
Site subtype 3	NS
Primary histology	**carcinoma**
Histology subtype 1	**endometrioid_carcinoma**

Histology subtype 2	NS
Histology subtype 3	NS
Genome-wide screen	y
GENOMIC_MUTATION_ID	**COSV57504903**
LEGACY_MUTATION_ID	COSM955832
MUTATION_ID	24750322
Mutation CDS	c.5900G>T
Mutation AA	p.S1967I
Mutation Description	**Substitution - Missense**
Mutation zygosity	
LOH	
GRCh	**37**
Mutation genome position	**14:45668030-45668030**
Mutation strand	+
SNP	n
Resistance Mutation	-
FATHMM prediction	NEUTRAL
FATHMM score	0.17953
Mutation somatic status	Confirmed somatic variant
Pubmed_PMID	
ID_STUDY	419
Sample Type	**fresh/frozen - NOS**
Tumour origin	primary
Age	59
HGVSP	ENSP00000267430.
HGVSC	
HGVSG	

There was also a Variant Call Format (VCF) file of all coding and non-coding mutations in the current release with the file names CosmicCodingMuts.vcf and CosmicNonCodingVariants.vcf.

These VCF files were used to extract the mutations for all the tumour samples in this analysis from the COSMIC database. Table 2-3 shows the information that was available from the COSMIC VCF files.

Table 2-3: An example of the input shown in CosmicCodingMuts.vcf and CosmicNonCodingVariants.vcf. *Details are provided for specific columns.*



Column Headers	Example Input	Description
CHROM	1	
POS	69224	
ID	COSV58737130	Corresponding to GENOMIC_MUTATION_ID from CosmicMutantExport.tsv
REF	A	Base in forward strand
ALT	C	Base in forward strand
QUAL	.	
FILTER	.	
INFO	GENE=OR4F5;STRAND=+;LEGACY_ID=COSM3677745;CDS=c.134A>C;AA=p.D45A;HGVSC=ENST00000335137.3:c.134A>C;HGVSP=ENSP00000334393.3:p.Asp45Ala;HGVSG=1:g.69224A>C;CNT=1	Gene name, Gene strand, Genomic Mutation ID, Legacy Mutation ID, CDS annotation, Peptide annotation, HGVS cds, HGVS peptide syntax HGVS, HGVS genomic syntax, How many samples have this mutation

2.4.2 GDC portal

The Genomic Data Commons (GDC) portal contains The Cancer Genome Atlas (TCGA) somatic mutation data which been called using different mutation calling pipelines such as Somatic Sniper, MuTect, Varscan and MuSE. The Somatic Sniper MAF files were downloaded from GDC portal to extract the sample names of the tumours which were present for each individual cancer type. This file was used because the mc3.v0.2.8.PUBLIC.maf file only contained sample IDs and did not have a description of which organ site the sample originated from. The MC3 pipeline was produced as part of the Multi-Center Mutation Calling in Multiple Cancers project (Ellrott et al., 2018). By combining seven variant calling pipelines: MuTect, MuSE, Radia, Somatic Sniper, Varscan-SNV, Varscan-Indel, Pindel and Indelocator; a mutation calling pipeline was produced that accounted for variance and batch effects (Ellrott et al., 2018). Once the sample identifiers were extracted from the Somatic Sniper file, the variants for those samples were identified in the MC3 MAF file. The variant information from these MAF files were extracted and then run through the Oncotator program which produced MAF files with cancer-specific annotations (Ramos et al., 2015).

2.4.3 Samples used

To identify the genetic landscape of SCC and skin cancer, and to analyse mutations across skin cancers, data was extracted from multiple sources. A summary of the data sources, programs and processes conducted is shown in figure 2-1.

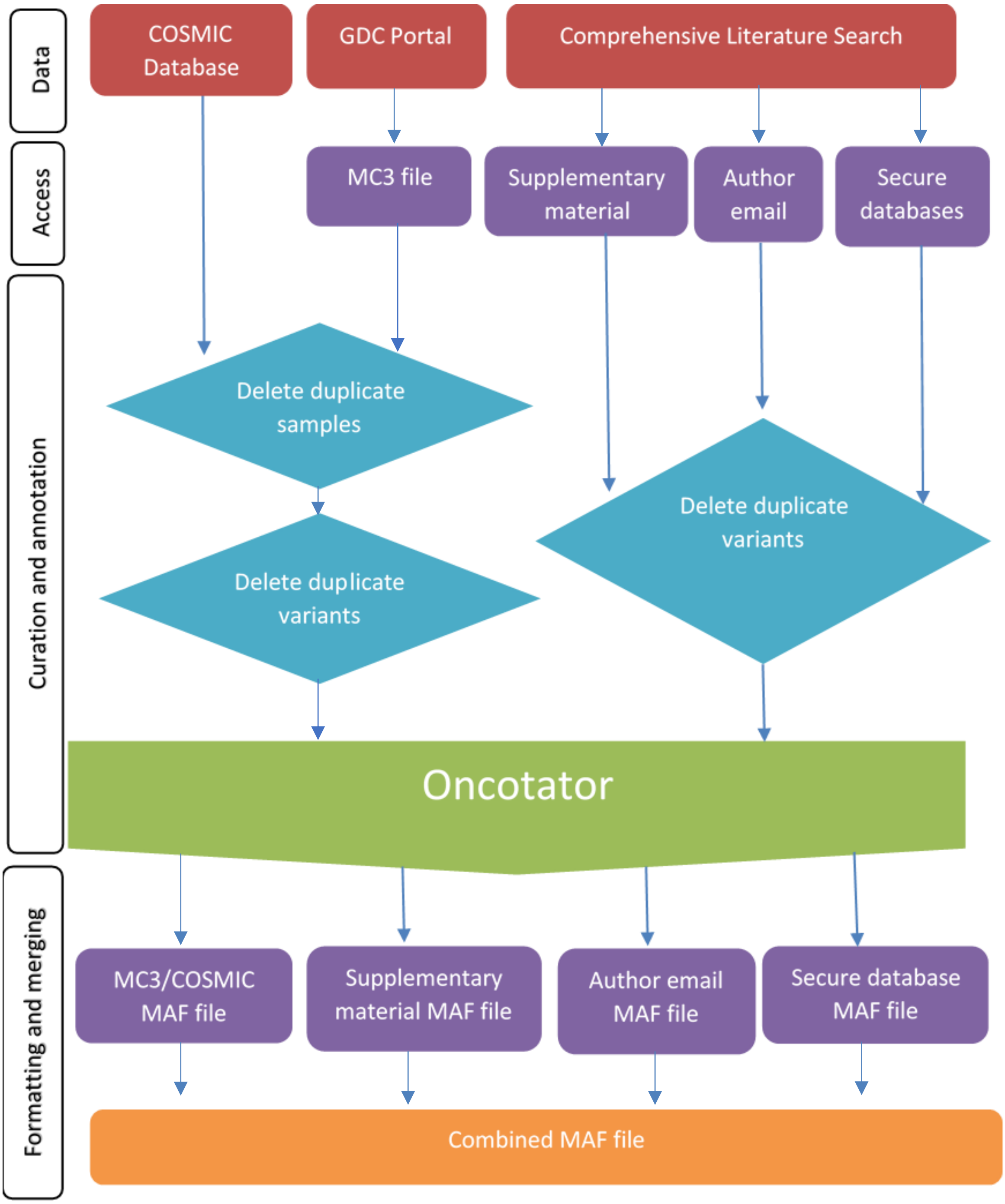


***Figure 2-1: Summary flowchart for extraction and processing of Next Generation Sequencing data**. The flowchart outlines the data sources, programs and processes for analysis of somatic mutations in squamous cell carcinomas (SCCs) and skin cancer.*

2.4.4 Data extraction

In this analysis, data was collated from the three different sources, described above. Since both COSMIC and the MC3 files contained TCGA data, the files were collated, and duplicate samples were deleted. Variant calls detected from the MC3 file were used in the case of duplicates as this pipeline is the most recent analysis of the data (Ellrott et al., 2018). Once the duplicates were deleted, the files were annotated using Oncotator and a MC3/COSMIC MAF file was produced as shown in the 'formatting and merging' section of figure 2-1.

2.5 Skin cancer bioinformatic pipeline

2.5.1 Overview

Mutation data for SCCs and skin cancer was extracted from the COSMIC database (See table 2-1, 2-2 and figure 2-2). Figure 2-3 shows how this data was converted to a format compatible for Oncotator. The bash script for data extraction from the COSMIC database was created in collaboration with Dr Jane Gibson, lecturer in Cancer Bioinformatics and Genomics, University of Southampton, and the script detailing this critical step is in appendix 7.1.

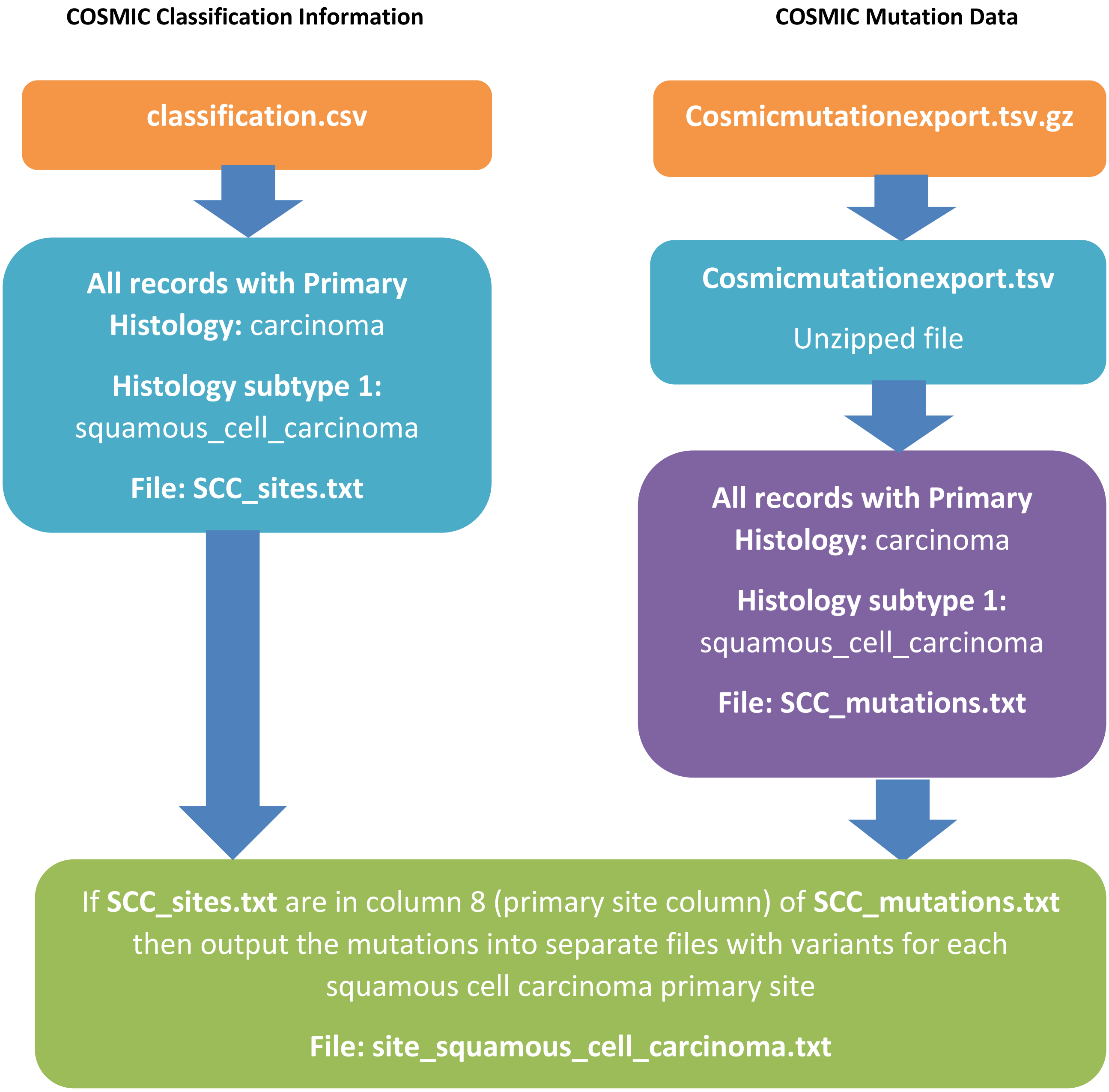

COSMIC Classification Information
COSMIC Mutation Data
classification.csv
Cosmicmutationexport.tsv.gz
All records with Primary Histology: carcinoma
Histology subtype 1: squamous_cell_carcinoma
File: SCC_sites.txt
Cosmicmutationexport.tsv
Unzipped file
All records with Primary Histology: carcinoma
Histology subtype 1: squamous_cell_carcinoma
File: SCC_mutations.txt
If SCC_sites.txt are in column 8 (primary site column) of SCC_mutations.txt then output the mutations into separate files with variants for each squamous cell carcinoma primary site
File: site_squamous_cell_carcinoma.txt

***Figure 2-2**: Flow diagram for extraction of genomic data from COSMIC.* *Data was extracted from classification.csv, the clinical characteristics file and Cosmicmutationexport.tsv.gz containing all cancer mutation data to identify all cancer mutation data associated with SCCs.*

The script takes the COSMIC variant call format file (VCF) for coding and non-coding mutations and merges them to produce an all.vcf file of all the coding and non-coding mutations in the COSMIC database, which was then split by primary histology and histology subtype classification with duplicates removed.

The CosmicMutantExport.tsv file contained all the mutation data for targeted and genome-wide screens and was filtered to only show mutations which were described in the relevant histological types'. The file was also filtered to ensure the mutations were aligned to the GRCh37 so the dataset only included mutations aligned to this version of the genome and only whole exome or whole genome data was included in the analysis.

The script was then used to list and count all the mutation descriptions which show the type of mutation at the amino acid level: substitution, deletion, insertion, complex, fusion or unknown; and then sorted according to their mutation type and only substitutions, deletions and insertions were included in our downstream analyses. All the output files produced were organised according to the body site the cancer originated. The end of these steps are represented the last stage of figure 2-2

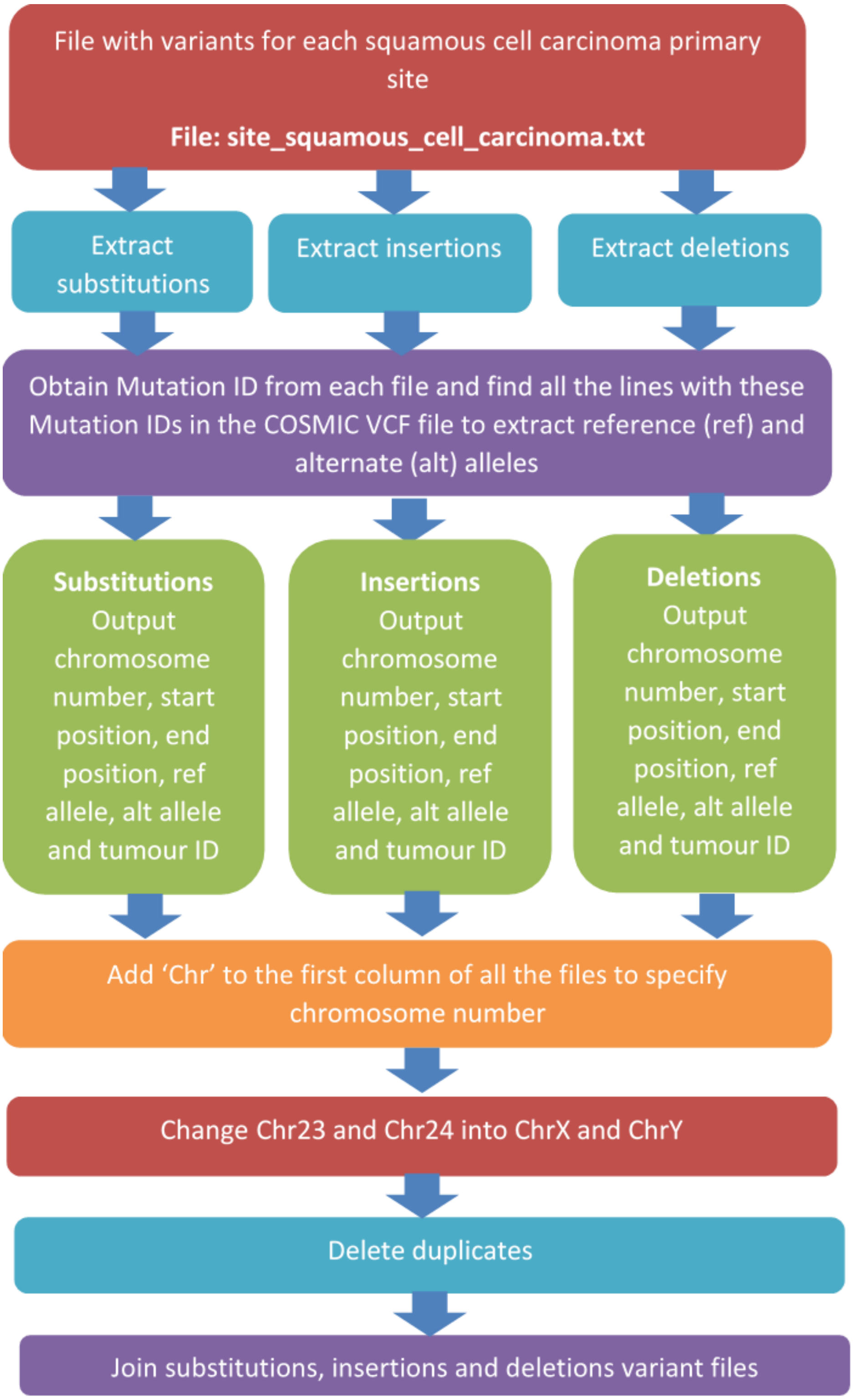

File with variants for each squamous cell carcinoma primary site
File: site_squamous_cell_carcinoma.txt
Extract substitutions
Extract insertions
Extract deletions
Obtain Mutation ID from each file and find all the lines with these Mutation IDs in the COSMIC VCF file to extract reference (ref) and alternate (alt) alleles
Substitutions Output chromosome number, start position, end position, ref allele, alt allele and tumour ID
Insertions Output chromosome number, start position, end position, ref allele, alt allele and tumour ID
Deletions Output chromosome number, start position, end position, ref allele, alt allele and tumour ID
Add 'Chr' to the first column of all the files to specify chromosome number
Change Chr23 and Chr24 into ChrX and ChrY
Delete duplicates
Join substitutions, insertions and deletions variant files

***Figure 2-3*: Flow diagram displaying how mutation data from COSMIC was converted to be compatible for Oncotator.** *Mutation data was cross-checked with COSMIC VCF file to identify all the reference and alternate alleles for insertions, deletions, and substitutions.*

The variant information in each site_squamous_cell_carcinoma.txt file was filtered using the mutation description and the genomic mutation identifiers (COSV) as these were unique numbers for each mutation in the file. Three files were produced, one for substitution mutations, one for insertions and one for deletions mutations with their unique genomic mutation identifiers. The genomic identifers in each of these files were searched for in the all.vcf file and all the lines with the reference and alternate allele for each mutation was identified. This was repeated for the SCCs of each organ site to produce mutation data specific to each SCC and each mutation description. This step produced the three files which are shown in the green boxes in figure 2-3.

To ensure the file was compatible to be run on Oncotator to produce a MAF file, the output files from the bash script were run through another list of commands. This was to ensure that file only included unique variants with chromosome number, chromosome start position, end position, reference base, alternate base and sample name.

The insertion, substitution and deletion files were collated to create one single file with all the COSMIC mutation data for each specific cancer in the study. The file produced at the end of this section of the pipeline is represented in the final purple box in the flowchart in figure 2-3. These files were created for cSCC, oropharyngeal SCC, oesophageal SCC, lung SCC, cervical SCC, BCC and melanoma.

2.5.2 Unifying somatic mutation annotation

The COSMIC database contained data from multiple sources including TCGA data samples. When COSMIC data and MC3 samples were merged together, there were duplicate samples from the MC3 file present. If the same samples were identified in the COSMIC database and the MC3 file, the sample data from the COSMIC database was removed and the data from the MC3 file was used. The MC3 file was considered as better representation of the true variant call as the MC3 pipeline was produced and optimised after comparing different variant calling pipelines such as MuTect2, VarScan2 and Pindel which is outlined in chapter 1, section 5.1.

Any samples which showed variants from cell line or cell culture experiments were not included in the analysis to ensure that only variants from tissue samples remained. A list of sample identifiers, which included mutation information for tumours that were only present in the COSMIC database, were absent from the MC3 file and only originated from tissue samples, was produced. This file was used to extract all the variant information that was specific to this list of sample identifiers.

The variants from samples which were identified under the MC3 pipeline and the variants from samples which were unique to the COSMIC database, were joined together to produce a file with all the variants from the COSMIC database and the MC3 pipeline. This was done separately for each specific cancer type. The COSMIC variant information was extracted from the final file produced in figure 2-3. Therefore each cancer output file contained the chromosome number, start position, end position, reference base, alternate base and tumour barcode for each tumour that was unique to the COSMIC database for that specific tumour type.

The script outlining this process is in chapter 7, section 4 in the appendix.

2.5.3 Duplicate exclusion

Each cancer output file was run through Oncotator to convert the output file into a MAF file. The MAF file was created for analysis in Maftools. When a MAF file is loaded into Maftools, duplicate entries are removed. In Maftools a duplicate entry is when the chromosome number, chromosome start position and the sample ID are the same. To check which duplicate values were removed in the program, duplicate values were removed from the files prior to using Oncotator and Maftools.

Genomic data for each cancer was collated and the Linux operating system was used to remove duplicate entries. In the Linux operating system, a duplicate entry is considered a row which is identical to another row. Therefore, in this file, the chromosome number, start position, end position, reference allele, alternate allele and sample ID should be the same to be considered an identical entry. A file was created with just the chromosome number, start position and sample ID for each cohort to predict the number of duplicates that would be identified in Maftools before uploading the file into Maftools as a method of checking this R package.

2.5.4 Power calculations

To identify the power of the study that was being conducted, the power was calculated using http://www.tumorportal.org/advanced_power. This study aimed to be inclusive to ensure cancer driver genes could be identified. The larger the study, the more likely it would be to identify driver genes which are present in a small subset of samples.

The number of samples available for each cancer was used and the background mutation rate for each cancer type was calculated using published studies. The South et al., 2014 study for skin SCC and the Lawrence et al., 2014 study was used for oropharyngeal SCC, lung SCC, oesophageal SCC and cervical SCC.

2.5.5 False positive genes

Maftools was used to identify the most frequently mutated genes. To ensure the analysis did not include false positive results, false positive genes from two papers (Lawrence et al., 2013, Repana et al., 2019) were compared. A list of false positive genes were collated from these two papers and then compared to COSMIC Cancer Gene Census list, if any genes were identified on this Cancer Gene Census list they were no longer considered false positives. The false positive gene list is in appendix 7.5.1. This list of false positive genes were also referred to when identifying driver genes and any genes which were considered false positives were not considered driver genes.

2.5.6 Driver gene identification

To identify significantly mutated genes, the MAF files were produced for each individual SCC and skin cancer so they could be run on MutSig2CV. The MAF files were reformatted for dNdScv, OncodriveCLUST and OncodriveCLUSTL and these files were run on these programs and the q values were compared to produce a list of potential driver genes. The process is outlined below is figure 2-4.

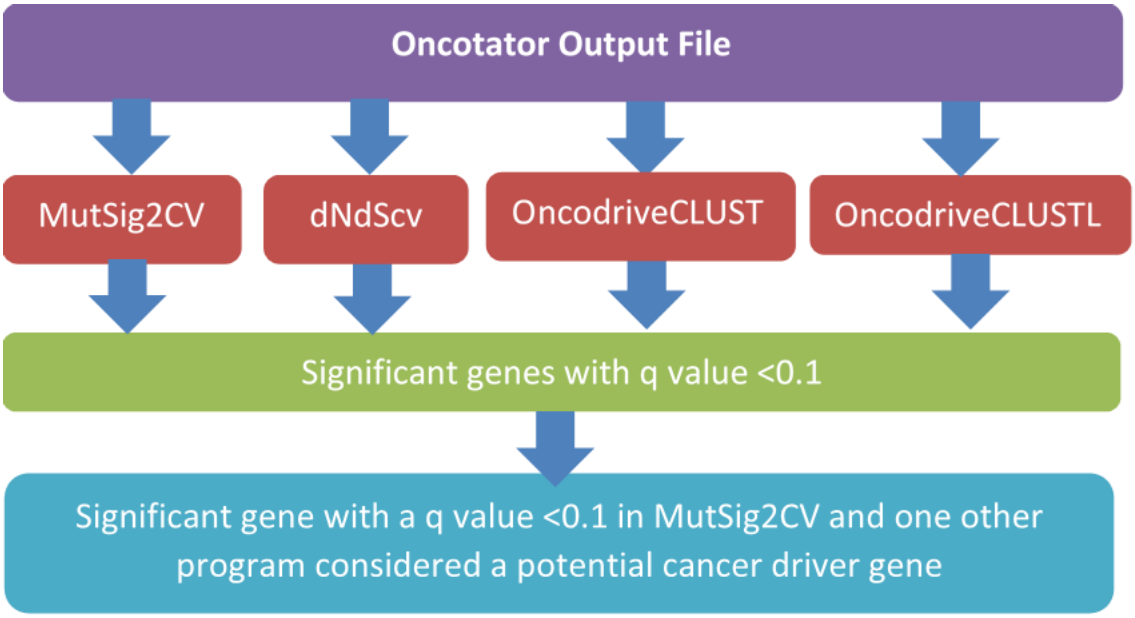

Oncotator Output File
MutSig2CV
dNdScv
OncodriveCLUST
OncodriveCLUSTL
Significant genes with q value <0.1
Significant gene with a q value <0.1 in MutSig2CV and one other program considered a potential cancer driver gene

***Figure 2-4**: **Flowchart outlining the process for identifying potential driver genes.** The four programs used to were MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL.*

A gene was considered significant if it had a q value The potential driver genes identified in cSCC were compared to those identified in oropharyngeal SCC, lung, SCC, oesophageal SCC and cervical SCC. This was done to identify the similarities and differences between cSCC and SCCs at other organ sites. The potential driver genes identified in cSCC were also compared to those identified in BCC and melanoma to analyse the differences between cancers which originate from keratinocytes and cancers which originate from melanocytes.

2.5.7 Data visualisation and analysis

Maftools was used to produce the top frequently mutated genes and the most common base changes and the script to do this is shown in the appendix 7.5. A box and whisker plot was produced to show the most common base changes and an oncoplot was produced to show the most frequently mutated genes. The oncoplot showed the proportion of samples that had a mutation in a specific gene and the specific type of mutation that was identified in each gene.

Nonsynonymous and silent mutations were identified using Maftools. A nonsynonymous mutation included mutations that were classed as frame-shift deletion, frame-shift insertion, in-frame deletion, in-frame insertion, missense mutation, nonstop mutation and splice-site mutation. A silent mutation included mutations that were classed as 3’ untranslated region, 5’ flanking region, 5’ untranslated region, de novo start in frame, de novo start out of frame, intergenic region, intron, RNA, silent, start codon deletion, start codon insertion, start codon single nucleotide polymorphism, stop codon deletion and long non-coding RNA.

Mutation signatures were produced using Maftools (Mayakonda et al., 2018) and Sigminer (Wang et al., 2021). Maftools was used to identify single base substitution (SBS) mutation signatures and Sigminer was used to identify double base substitution (DBS) mutation signatures (Alexandrov et al., 2020).

The SBS mutation signature shows the proportion of samples where the reference nucleotide is replaced by another nucleotide such as adenine, thymine, guanine or cytosine in a specific nucleotide context. The mutational pattern is then categorised into one of the reference SBS mutation signatures (Alexandrov et al., 2020). To identify the single base substitutions in Maftools, the hg19 genome was loaded into R as shown and a trinucleotide matrix was made using the MAF file. The trinucleotide matrix which represents the 96 possible mutated trinucleotides. The 96 possible trinucleotide changes are the six base changes which could occur (C>A, C>G, C>T, T>A, T>C, T>G) multiplied by the 16 different bases which could reside on either side of the base change due to the different combination of the four DNA bases (A, T, C, G).

The Non-negative Matrix Factorisation (NMF) package is used as an unsupervised learning technique and pattern recognition (Gaujoux and Seoighe, 2010). In this case, it was used to identify mutation signatures from the MAF file. This program uses the trinucleotide matrix produced from the MAF file labelled as ‘laml.tnm’ and ran an NMF using Brunet et al’s. (Brunet et al., 2004) estimation of optimal factorisation rank. The default number of NMF runs were 5 per each value in the range which was 2:6. The range represents the possible number of SBS mutation signatures which can be identified from the trinucleotide matrix produced from the MAF file. The result of these runs were saved in the file ‘laml.sign’. A cophenetic plot of ‘laml.sign’ was then produced which was used to measure the robustness of the clusters produced when run at each value of the range.

The NMF was run for each value in the range (2:6) and a cophenetic plot was produced which is shown below (figure 2-5). This is a measure of robustness of the clusters produced which is measured on Y axis and the X axis represents each value in the range, which are the possible number of mutation signatures that can be produced from the MAF file. The cophenetic plot was used to identify the optimum number of SBS mutation signatures which can be produced for each MAF file.

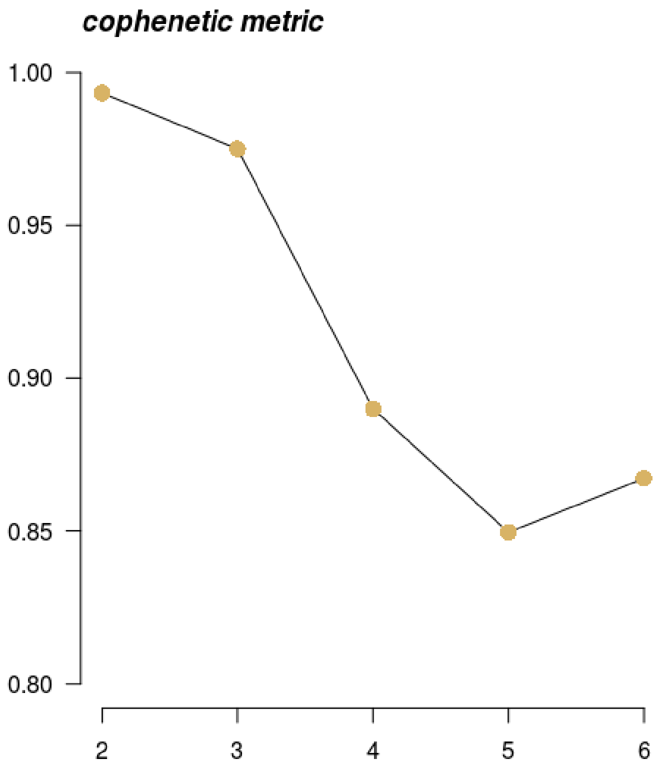
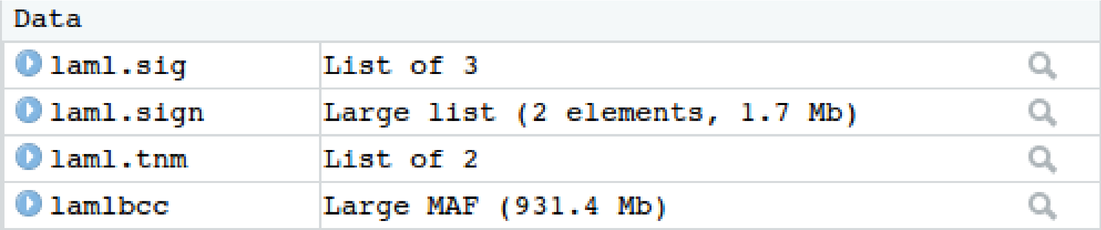


***Figure 2-5**: **Cophenetic plot produced from the genetic data of basal cell carcinoma samples using Maftools.** The genetic data was used to produce a trinucleotide matrix of the single base substitutions and the cophenetic plot was used to determine the optimum number of mutation signatures which are present in the dataset. The X axis represents the number of mutation signatures and the Y axis measures the robustness of the clusters.*

A 'laml.sig' file was created and used the 'extractSignatures' command to decompose the trinucleotide matrix (laml.tnm) into the optimum number of mutation signatures (n). The 'n' was chosen based on the cophenetic plot and optimum number of mutation signatures that can be produced from the trinucleotide matrix produced from the MAF file. The highest point on the y axis of the cophenetic plot which is the robustness of the clusters and the greatest number of mutation signatures that could be produced from the samples (x axis) was identified. A value was chosen for 'n' which related to the value on the x axis of the cophenetic plot.

The screenshot (Figure 2-6) below shows R Studio and all the data files which were stored during the session to produce SBS mutation signatures. The 'laml.sig' represents the mutation signatures produced from the MAF file. The MAF file used in the example below is 'lamlbcc'. The 'laml.sign' are the results of the NMF which was run on the MAF file to identify the optimum number of mutation signatures which could be produced for this dataset. The laml.tnm represents the trinucleotide matrix which was produced from the MAF file (lamlbcc).



Data	
laml.sig	List of 3
laml.sign	Large list (2 elements, 1.7 Mb)
laml.tnm	List of 2
lamlbcc	Large MAF (931.4 Mb)

***Figure 2-6**: A screenshot of the data files produced in R studio during mutation signature analysis using Maftools.* *The 'Data' column shows the name of the file and the second column is a description of the type of file.*

The mutation signatures identified from the dataset were stored in a file which was compared to the COSMIC SBS mutation signatures (Alexandrov et al., 2020). The R terminal was used to present the mutation signatures, their predicted aetiologies and their cosine similarity to the mutation signatures in the COSMIC database. Another command (`maftools::plotSignatures)` was then used to plot the mutation signatures in a graphical format which is shown in figure 2-7.

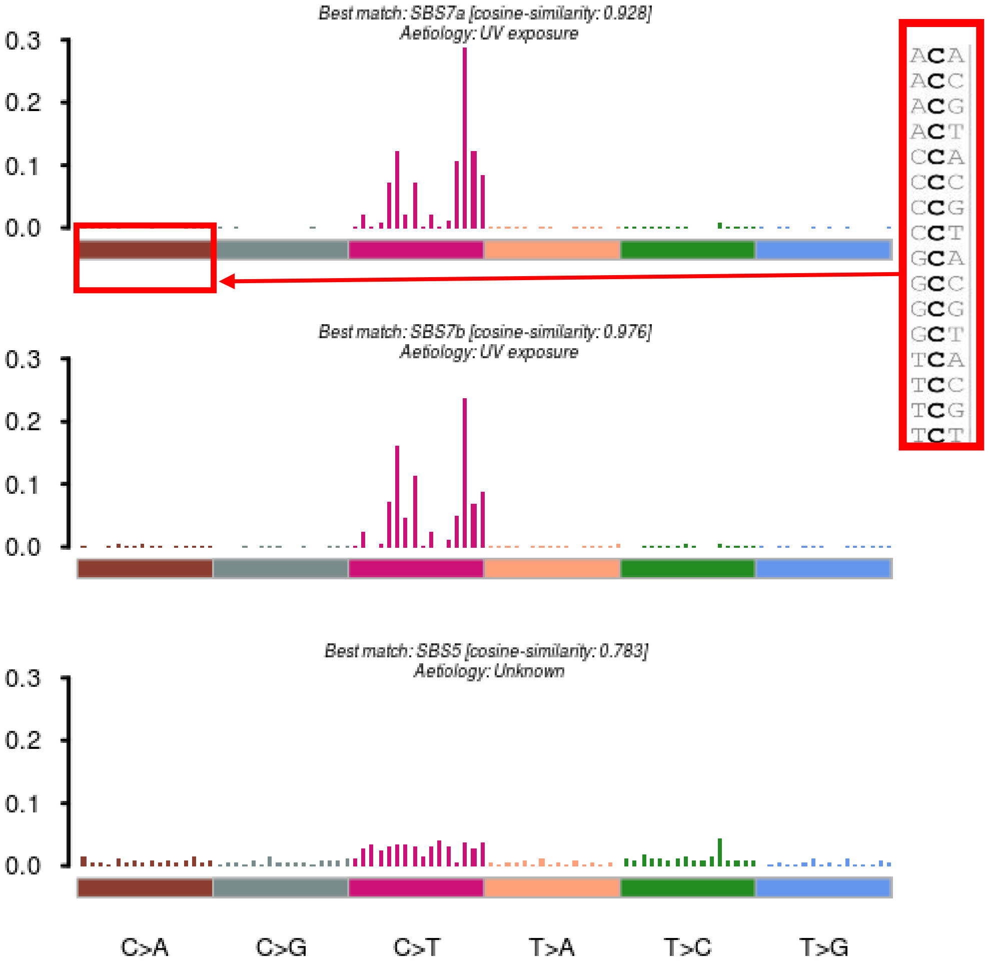


***Figure 2-7**: Example of mutation signature plots produced using Maftools.* *The three graphs represent the three different mutation signatures extracted from the dataset. The text description above each graph represents the single base substitution mutation signature and a measure of its similarity to the COSMIC database and an aetiology is stated.*

Maftools produces figure 2-7 to represent mutation signatures which is shown above. The three graphs are the mutation signatures produced from the MAF file. The X axis of each graph represents the trinucleotide context of each base change and the Y axis represents the proportion of bases with the single base changes from the cohort of samples. The mutation signatures shown in the example below are most similar SBS7a, SBS7b and SBS5.

2.6 Validation of the somatic status of mutations

The COSMIC database is a Catalogue of Somatic Mutations in Cancer, therefore the mutations identified in this study were considered somatic mutations from tumour samples. However, the evidence that these mutations were truly somatically acquired has not been verified for all variants (i.e., by paired sample analysis with matched germline DNA). Therefore, it is possible that this database could contain 'non somatic variants' impacting the validity of our data. The database has a somatic mutation status annotated, which identifies that a 'variant of unknown origin' is when a mutation is known to be somatic but the tumour was sequenced without a matched normal sample. Some mutations are labelled as 'previously observed' which is when the mutation has been reported as somatic previously but has not been identified in the current paper that the genomic data was from. Any mutations labelled as 'confirmed somatic' is when the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient. Therefore, the paired precision of sampling will ensure that the variant is somatic and not a germline mutation. There is expert manual curation conducted when producing this database and papers with incomplete data or insufficient quality are not fully curated but used as additional references for somatic mutations.

All COSMIC variants were included in this study but to ensure the results of this study were reflective of only the confirmed somatic variants, the 'mutation status' of variants identified in COSMIC were recorded and this information is deposited in appendix 7.11.5. Any cancer cohorts which had less than 98% of variants that were considered as confirmed somatic variants were reanalysed. This value of 98% was chosen because it was above the 95% confidence interval for the false positive rate which has been used in other published studies using matched tumour normal paired samples (Shand et al., 2020, Anzar et al., 2019). There were also many studies used in this analysis which used different variant calling algorithms, therefore the false positive rate could vary between different sequencing platforms and variant calling algorithms (Quail et al., 2012).

There were limitations to using this 98% cut-off as variants in driver genes could be present in the 2% of variants which are not confirmed somatic variants. Therefore, by including the 2% of variants that were not included in this analysis, could affect the significance of the driver genes during this reanalysis. In the reanalysis conducted in this study there have been no changes to the driver gene status but if the other cohorts were reanalysed there is the potential for there to be a change in driver gene status. To identify which driver genes could be affected in the other cohorts, the unconfirmed somatic variants of each cohort were extracted and analysed. Variants were analysed separately for each SCC, BCC, and melanoma. The presence of variants in common driver genes that were shared between the SCCs, melanoma and BCC were screened which is shown in appendix 7.11.6. Any studies which had 98% of confirmed variants but included variants in shared driver genes that were not confirmed somatic variants were also reanalysed and the results of this analysis are also in appendix 7.11.5. The appendix 7.11.5 shows a table of the driver genes which were significant in skin SCC and the cancers which were reanalysed and a comparison of their original p and q value after MutSig2CV analysis compared to their new p and q values after reanalysis.

3. Mutational landscape of cutaneous SCC

3.1 Introduction

Squamous cell cancer (SCC) can arise in internal organs as well as in the skin. As there are limited treatments for aggressive cutaneous SCC (cSCC), one important question is whether cSCC has similar somatic genetic changes to SCCs in other organs. One might expect SCCs in different organs to arise in a similar manner, however, the fact that different carcinogens cause them and the fact that they occur in different microenvironments, it is possible that genetic changes may differ between the SCCs in various organs. Using bioinformatics to compare the mutational landscape of cSCC with that of SCCs in internal organs would inform on whether cSCC has common mutations, including driver gene mutations, with those in other SCCs, and thus whether treatments effective for other internal SCCs might have the potential for use in treatment of aggressive cSCC.

Whole exome and whole genome studies have been carried out in cSCC tumours (Inman et al., 2018) and a meta-analysis has been conducted using the raw data from individual studies (Chang and Shain 2021). In a study published in Journal of Investigative Dermatology (South et al., 2014), 20 cSCC tumours were exome sequenced and were used to identify that *NOTCH1* mutations occur early during cutaneous squamous cell carcinoma carcinogenesis. This analysis provided a detailed mutation analysis of *TP53*, *NOTCH1*, *NOTCH2*, *CDKN2A* and members of the *RAS* family of genes. However, it was difficult to identify additional definitive driver genes in cSCC due to the high mutation burden of this cancer and the small sample size. This study suggested that *NOTCH1* mutations arise early in cSCC development as this gene was sequenced in over 170 cSCC samples and ten normal skin samples. Mutations in *NOTCH1* were identified in normal skin samples adjacent to the cSCC tumours suggesting that *NOTCH1* is present in normal skin and has potential for clonal expansion to further initiate tumour formation.

The data from the South et al., 2014 study was included and re-analysed in a subsequent study where the cSCC sample size increased to a total of 40 cSCC tumours (Inman et al., 2018). In this analysis three bioinformatics programs were used to identify significantly mutated genes from whole exome sequencing data, namely MutSigCV, OncodriveFM and OncodriveCLUST. There were 22 genes identified which were identified as significant by at least two bioinformatics programs in this study. Genes which had previously been identified, such as *NOTCH1*, *NOTCH2*, *TP53* and *CDKN2A*, were also replicated in this analysis. However, this analysis also yielded novel significantly mutated genes including *HRAS*, *MAP3K9*, *PTEN*, *SF3B1*, *VPS41* and *WHSC1*. The whole exome sequencing data in this study was also used to identify mutation signatures in cSCC and a novel mutation signature in cSCC associated with Azathioprine exposure was identified.

Mueller *et al*., 2019 used single base substitution mutation signature analysis to identify mutation patterns in metastatic cSCCs. In this study whole genome sequencing was conducted on 15 metastatic cSCC tumours, and the results compared with analysis whole exome sequencing of primary cSCCs by Pickering *et al*., 2014, and showed that mutation signature 7 which is associated with UV is present in both primary and metastatic cSCC tumours. The whole genome sequencing identified that the mutational burden was 171-fold higher in non-coding regions of the genome compared to coding regions. This suggested that mutations in regulatory regions in the non-coding part of the genome might be contributing to tumour progression and metastasis in this tumour. The Mueller *et al*., 2019 study also highlighted that non-coding mutations and mutation signature analysis can be used to distinguish metastatic cSCC from metastases derived from other cancers in cases where a primary tumour cannot be identified.

Due to the high mutation burden of skin SCC, it can be difficult to distinguish driver mutations from passenger mutations using a small sample size, therefore, a meta-analysis was conducted (Chang and Shain, 2021) using published raw data from exome sequencing to identify cSCC driver genes. This study included tumours from individuals with xeroderma pigmentosum, recessive dystrophic epidermolysis bullosa, immunosuppressed patients and sporadic cSCCs. Four bioinformatics programs were used (OncoDriveFML, MutSig, dN/dS and LOFsigrank) to identify driver mutations. The analysis replicated previous documentation of *TP53*, *NOTCH1*, *NOTCH2*, *CDKN2A* and *HRAS* as driver genes. There were 12 novel genes identified in this analysis including *EP300*, *PBRM1*, *USP28* and *CHUK* which were mutated in more than 10% of tumours. The raw data from ten studies were used to call bases with the Mutect2 pipeline which spanned 105 samples. This study was considered the largest analysis of cSCC samples to date. However, this meta-analysis of whole exome samples did not include the whole genome data from the Mueller et al., 2019 study. The Mueller *et al*., 2019 study used the whole genome data to identify mutational signatures and did not use their samples to identify driver genes, whereas Chang and Shain (2021) identified driver genes in their analysis and did not run mutation signature analyses.

The inclusion of underlying genetic conditions such as xeroderma pigmentosum and recessive dystrophic epidermolysis bullosa which increase the chances of a patient developing cSCC (Bradford et al., 2011, Pourreyron et al., 2007) in the above meta-analysis could mean that some of the driver genes identified may not be representative of the wider population, hindering its clinical applicability. In addition, as UV treatment of skin disease is unlikely to be given to patients with xeroderma pigmentosum or recessive dystrophic epidermolysis bullosa, but is used in immunocompetent patients, including patients who later might become immunosuppressed (e.g. if they subsequently required an organ transplant), there is a need to undertake a comprehensive analysis of mutational signatures and driver genes in cSCCs from all immunocompetent and immunosuppressed patients where exome sequencing and/or whole genome sequencing data is publicly available in databases and/or in publications in the literature. The presence of cancer databases such as COSMIC can be utilised with independently curated data to ensure all cSCC data has been identified (Forbes et al., 2016).

In this results chapter, the mutational landscape of cSCC will be dissected by extracting all whole genome and whole exome sequencing data from COSMIC database and a literature search. The collated genomic data will include sporadic cSCC samples which have no underlying genetic conditions to ensure the driver genes identified are pertinent to the general population, including those who go on “sun-holidays” and/or receive treatment with UV for skin disease. The analysis will be used to dissect any single base substitution mutation signatures and double base substitution mutation signatures which can be used to identify the mutational pattern within these skin SCC cohorts. The potential driver genes and mutation signatures from the analysis might also highlight specific regions of the genome that might lead to the identification of potential drug-targets and therapies in skin SCC.

3.2 Methods

To identify data relevant to skin SCC, the COSMIC database v91 (https://cancer.sanger.ac.uk/cosmic) was utilised. The CosmicMutantExport.tsv.gz (containing a tab separated table of all COSMIC coding point mutations from targeted and genome wide screens) from the COSMIC database was downloaded and files with whole genome or exome data for skin SCC were produced as outlined in chapter 2.5. The file name for the skin SCC mutations was called skin_squamous_cell_carcinoma.txt.

The Variant Call Format (VCF) file was downloaded from the COSMIC website for all coding and non-coding variants (VCF/CosmicCodingMuts.vcf.gz). The COSMIC ID was a common identifier in the VCF file and the CosmicMutantExport.tsv.gz files and unique to each mutation in the database. The COSMIC ID from the skin_squamous_cell_carcinoma.txt was then used to extract the reference and mutant bases from the VCF file for each of the skin SCC tumour samples contained in the skin_squamous_cell_carcinoma.txt file. The files were then separated into their mutation type (base substitutions, insertions, and deletions) and three different files were created. To ensure these three files were compatible for the Oncotator program, the data was organised as outlined in chapter 2.5. and the Oncotator program was used to annotate the variants in each file.

A literature search was conducted to identify all whole genome and whole exome skin SCC data until 24th January 2020. The search terms used for skin SCC whole exome files which were identified from literature searches were downloaded from the supplementary data or from datasets sent through email and edited to ensure that the sample name, chromosome number, chromosome start position, end position, reference and alternate base were consistently mined from the files for each study. A whole genome dataset in Strelka VCF format was also downloaded from the EGA database and was done as outlined in chapter 2.3.4 (Mueller et al., 2019). For this Mueller et al., 2019 dataset from EGA, only exonic regions of the whole genome data were used for analysis. The chromosome number, start position and end position of the coding regions of the genome were obtained from Github (https://github.com/Shicheng-Guo/AnnotationDatabase/blob/master/hg19/refGeneExtent.hg19.bed.gz) and the rows which were labelled as being an exon were extracted out of the file. The somatic mutations were identified in each sample from the Mueller et al., 2019 study and Bedtools was used to only extract the coding regions from these cSCC VCF file. The University of Southampton computer cluster, Iridis was used to access bedtools and the bed files produced for each individual tumour were uploaded to the cluster using the script outlined in appendix 7.3.1.

The Mueller et al., 2019 study and the other datasets obtained via the literature search for cSCC were processed through Oncotator for genome annotation to produce a MAF file. Then the COSMIC data and extra datasets were merged together.

The cSCC MAF file was then loaded onto R to be analysed using the Maftools program. This program was used to identify the most common base changes in the skin SCC tumours, the top 25 frequently mutated genes, single base substitution mutation signatures and to investigate which driver mutations co-occurred. The MAF file was also loaded onto another program on R called Sigminer which was used to identify double base substitution mutation signatures.

Driver genes were identified using four different bioinformatics programs. The skin SCC MAF file was analysed using OncodriveCLUST in R to identify driver genes based on mutation clustering. The dNdScv package, which functions on R, was used to generate genes which were selected for in skin SCC tumour samples. The MAF file produced was edited so the format was compatible for the dNdScv package which is outlined in chapter 2.7.2. MutSig2CV was used to identify significantly mutated genes which have a mutation rate higher than the background mutation rate. The OncodriveCLUSTL program was run using Python and due to the high mutation burden of skin SCC the settings used for OncodriveCLUSTL varied from the default (See code chunk below), i.e. the data was concatenated, the smooth window was increased from the default 11 to 15, cluster window was increased from 11 to 15, simulation window was increased from 31 to 35 and the simulation mode was changed to region restricted from the default which is mutation centred. The data was concatenated to ensure two exons in a transcript were joined to identify two or more clusters spanning different exons. By increasing the smooth window and cluster window, it enabled the unsupervised clustering OncodriveCLUSTL to detect larger clusters for observed and simulated windows however a large simulation window would spread out the simulated mutations decreasing the chance of detecting a cluster (Arnedo-Pac et al., 2019). Therefore, to increase the likelihood of detecting clusters in the samples, the smooth window (sw), cluster window(cw) and simulation window (simw) were all increased. The OncodriveCLUSTL code reflects these changes.

```
oncodriveclustl –i /file
location/ONCODRIVECLUSTL/SKIN/skinoncodriveinput.tsv.gz -r /file
location/ONCODRIVECLUSTL/SKIN/cds.hg19.regions.gz -o /file
location/ONCODRIVECLUSTL/SKIN/output_concat_nodefault -sw 15 -cw
15 -simw 35 -sim region_restricted --concatenate
```

Significance testing for the four driver gene programs: If the q value was less than 0.1 in any of the four bioinformatics programs, then the gene was considered as significantly mutated in that program. Lists of genes with q values < 0.1 for each of the four programs were compared, and genes which were significant in MutSig2CV and at least one other program was classed as a driver gene. A summary of the data analysis pathway is outlined in figure 3-1.

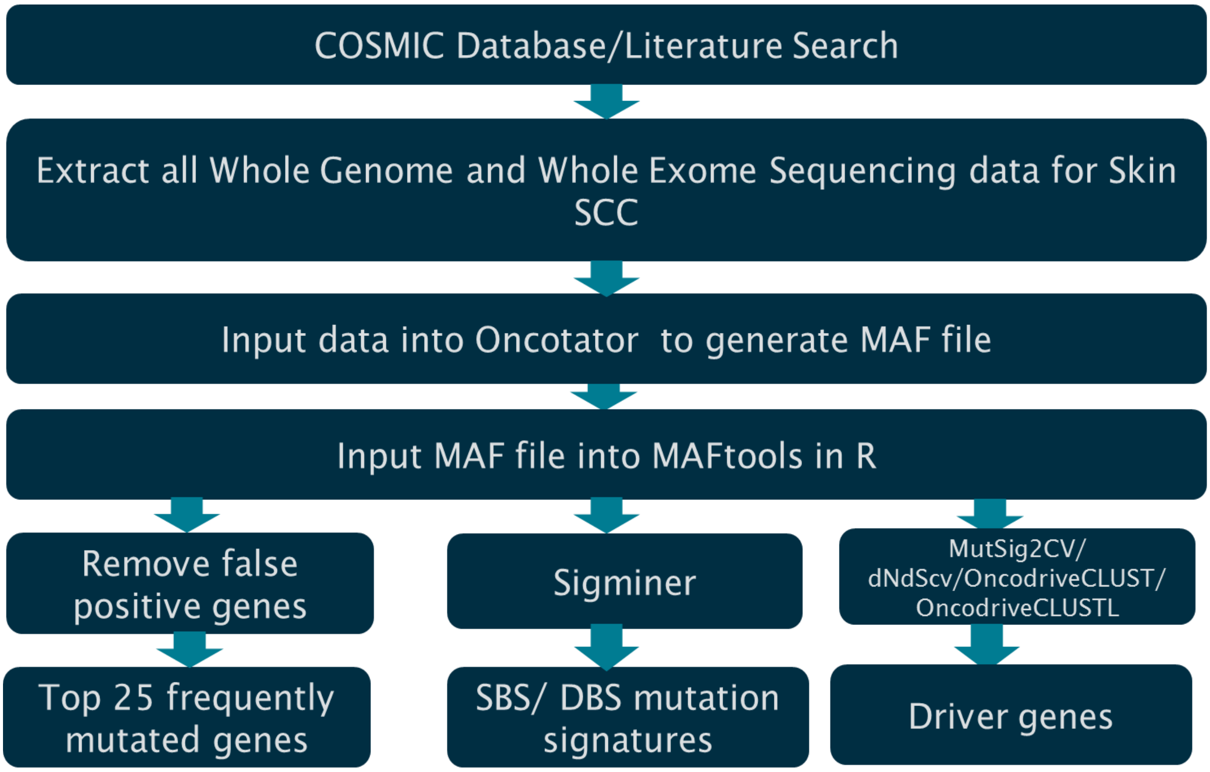

COSMIC Database/Literature Search
Extract all Whole Genome and Whole Exome Sequencing data for Skin SCC
Input data into Oncotator to generate MAF file
Input MAF file into MAFtools in R
Remove false positive genes
Sigminer
MutSig2CV/ dNdScv/OncodriveCLUST/ OncodriveCLUSTL
Top 25 frequently mutated genes
SBS/ DBS mutation signatures
Driver genes

*Figure 3-1: **A flowchart outlining the processing of whole genome and whole exome data for cSCC.** The Mutation Annotation Format (MAF) is the file format that was used in this analysis. This file type is compatible with the programs outlined.*

3.3 Results

The MEDLINE OVID search tool was used to identify whole genome and exome data for skin SCC in the literature search. The COSMIC database was also utilised to extract genomic data and compare with the records identified in MEDLINE. The flowchart in figure 3-2 shows how many records were identified for skin SCC via literature search and from the COSMIC database and how the literature search records were filtered and the reasons that any records were not included in the analysis.

Cutaneous SCC

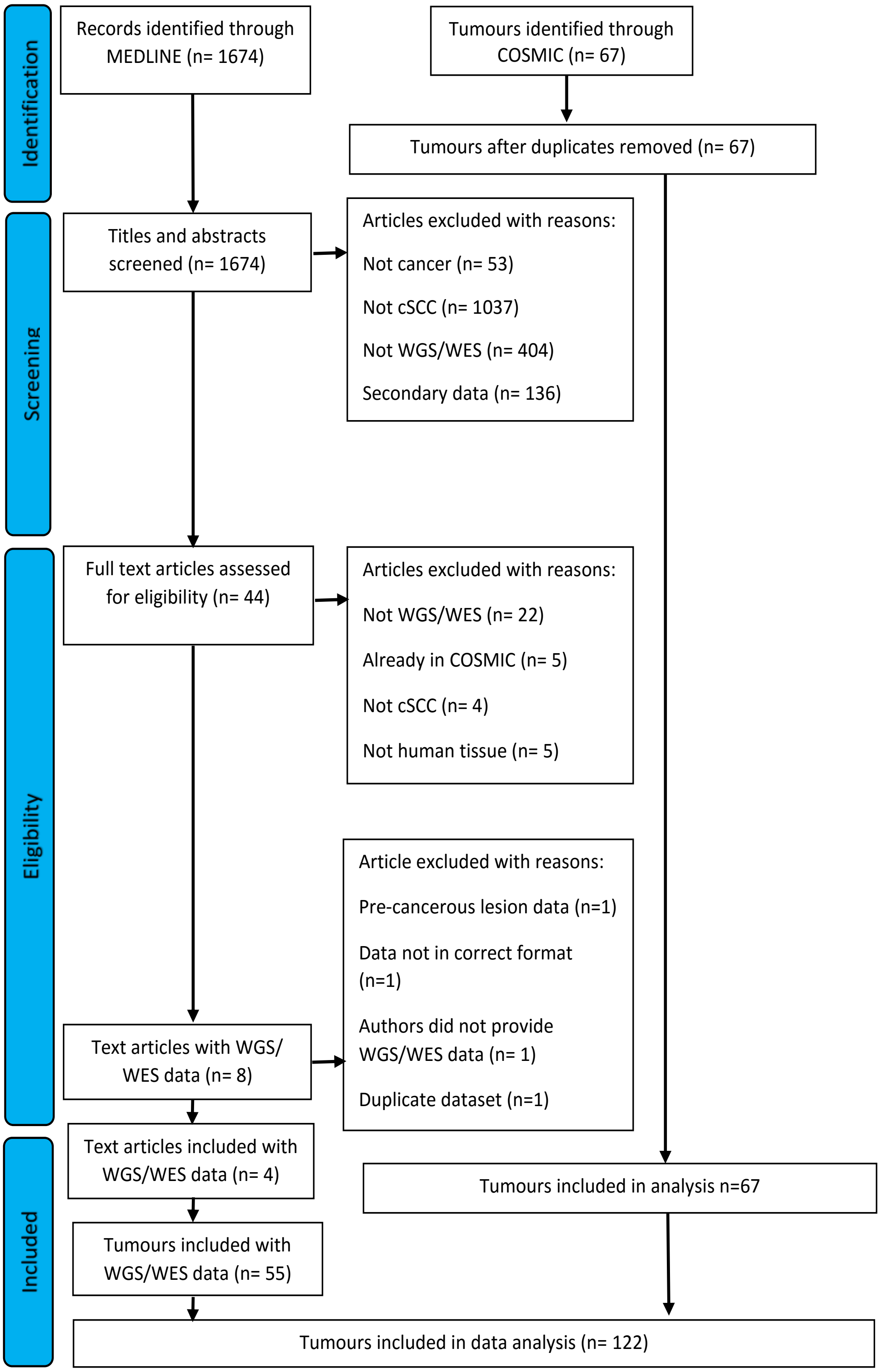

Identification
Screening
Eligibility
Included
Records identified through MEDLINE (n= 1674)
Tumours identified through COSMIC (n= 67)
Tumours after duplicates removed (n= 67)
Titles and abstracts screened (n= 1674)
Articles excluded with reasons:
Not cancer (n= 53)
Not cSCC (n= 1037)
Not WGS/WES (n= 404)
Secondary data (n= 136)
Full text articles assessed for eligibility (n= 44)
Articles excluded with reasons:
Not WGS/WES (n= 22)
Already in COSMIC (n= 5)
Not cSCC (n= 4)
Not human tissue (n= 5)
Article excluded with reasons:
Pre-cancerous lesion data (n=1)
Data not in correct format (n=1)
Authors did not provide WGS/WES data (n= 1)
Duplicate dataset (n=1)
Text articles with WGS/ WES data (n= 8)
Text articles included with WGS/WES data (n= 4)
Tumours included with WGS/WES data (n= 55)
Tumours included in analysis n=67
Tumours included in data analysis (n= 122)

Figure 3-2: Flowchart representing studies identified in a literature search for skin SCC. *Studies in the literature search were compared with those in the COSMIC database and duplicates were removed.*

A Medline OVID search conducted for skin SCC with search terms specified in the appendix 7.9.1 produced a total of 1674 studies. Initially the abstracts of these studies were inspected and 44 studies were identified as containing WGS or WES data. There were 1630 studies which were discarded; 53 of these studies did not include cancer data, 1037 studies were not about skin SCC, 404 studies did not contain WGS or WES data and 136 of the studies were reviews or did not contain primary data. The results of the COSMIC database were filtered as specified in appendix 7.1. The 44 papers retained from the literature search were then read and a further 36 studies were removed. There were 36 studies that were not included in the analysis for the following reasons: 22 studies did not include WGS or WES data, five studies were present in the COSMIC database and were already included in the analysis, four studies did not include cSCC data and five studies included sequencing data for non-human tissue samples. From these eight studies that remained, it was identified that one study included WES for Actinic Keratosis, a pre-cancerous lesion and was not included in the analysis. Data was requested from the dbGaP database for one study and an SRA file was received. This file was not included in the analysis because it was not in the correct data format as a MAF file. An email request was sent to two authors in relation to their publications because their data was not publicly available in their papers and one author did not provide whole exome or whole genome data for analysis. One study's data was included in another paper and was removed as a duplicate. Overall, from the literature search, four studies remained with whole exome data available for 42 tumour samples and whole genome sequencing data for 13 samples. In the COSMIC database 67 tumour samples with whole exome or whole genome sequencing data were identified. In total there were 122 cSCC tumour samples with whole genome or whole exome data available for analysis. All the variants identified from the COSMIC database were confirmed somatic variants and data that was included in the analysis from the literature search were variants that were also confirmed somatic variants.

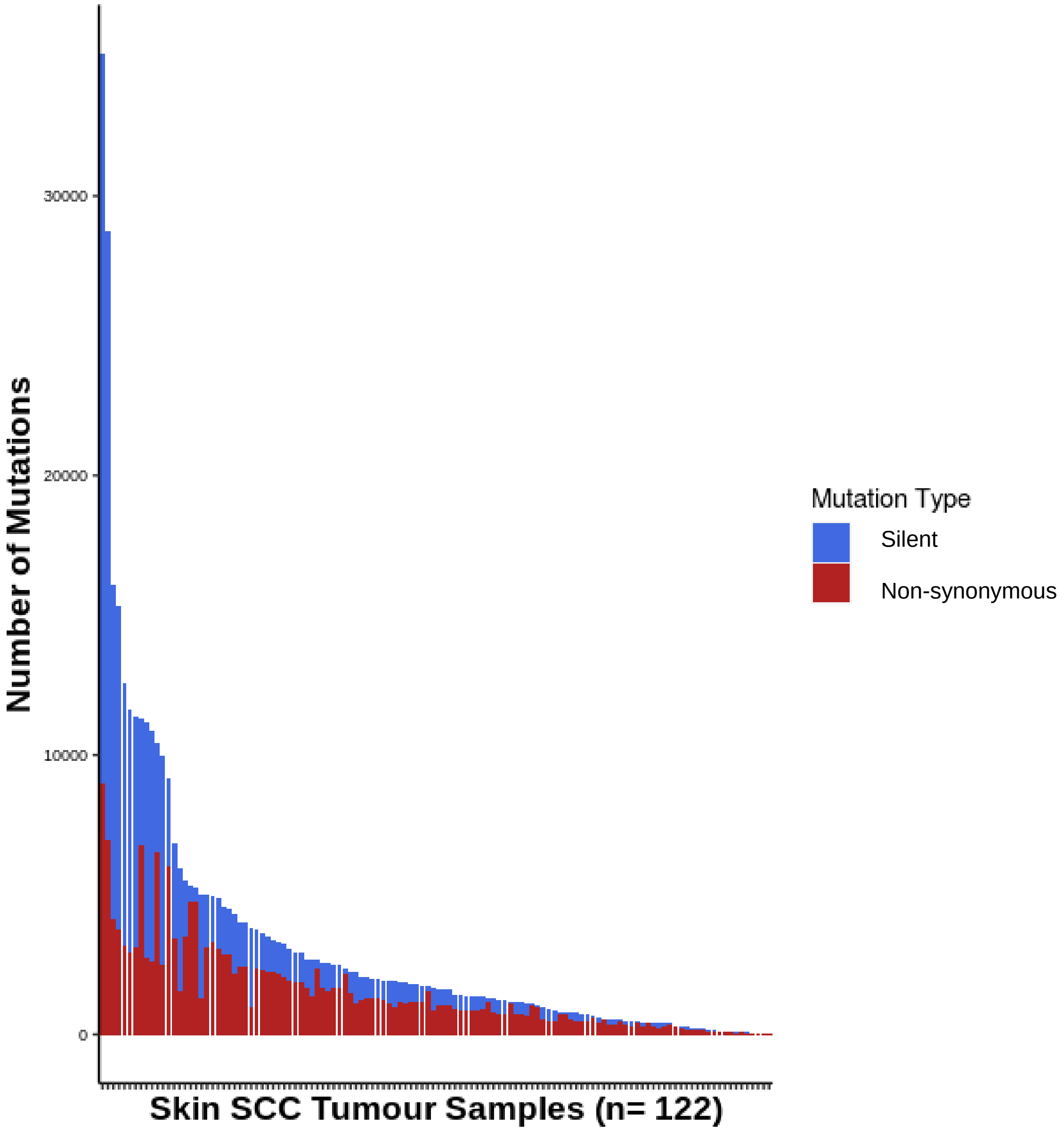


Figure 3-3: A bar graph showing the total number of mutations in each cSCC tumour sample. *The blue coloured bars in the stacked bar chart represent silent mutations and the red coloured bars represent non-synonymous mutations.*

All samples in the skin SCC dataset had a mixture of nonsynonymous and silent mutations and most samples had more nonsynonymous changes than silent changes (figure 3-3). The sample with the highest number of mutations was CSCC_0014_M1 with 35,084 mutations. This sample also has the highest number of silent mutations (26,098). Many cSCC samples had less than 10,000 mutations however there were 12 samples which had more than 10,000 mutations. The sample with the lowest number of silent mutations was sample WD_04 (four mutations) and the sample with the lowest number of nonsynonymous mutations was MD03-Tumor (four mutations).

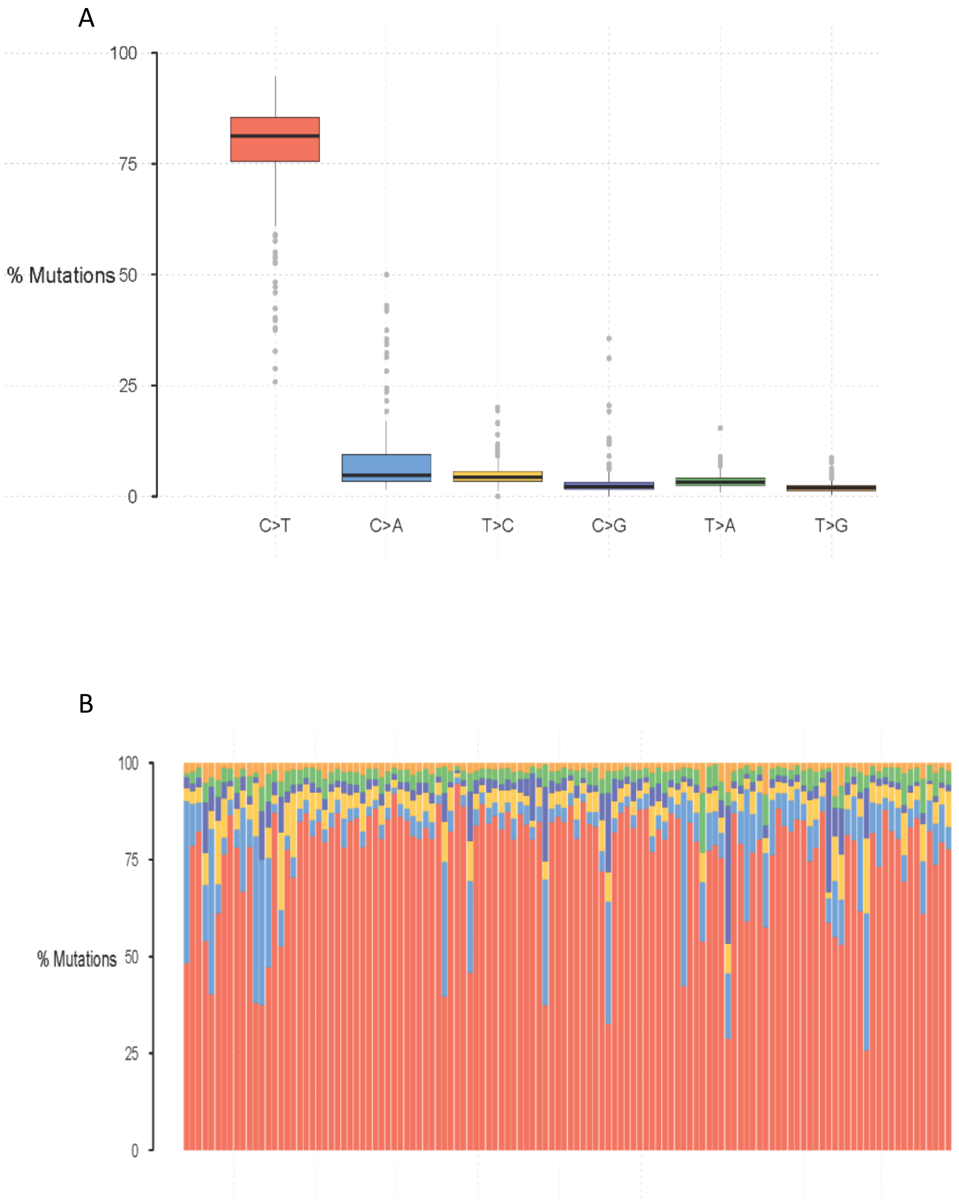


Figure 3-4: A. A box and whisker plot showing the percentage of bases in this cohort of 122 cSCC tumour samples. *The different colours represent the different base changes.* ***B.The relative contributions of the different base changes in each individual skin SCC sample.*** *The proportion of mutations in each sample is measured as a percentage of the total number of mutations in that tumour and the different colours represent the different base changes as denoted in part A of the figure.*

The most common type of base change in the skin SCC dataset was C>T, with more than 75% of mutations in most of the skin SCC samples being C>T changes (figure 3-5). This base change also had the highest interquartile range showing that the number of C>T changes varied the most between samples. Conversely, T>G accounted for the smallest proportion of base changes in the dataset and had the smallest range suggesting that there was more limited variation in this base change across samples. The C>A change was the second most common change, with this alteration almost as frequent as C>T changes in some of the cSCCs.

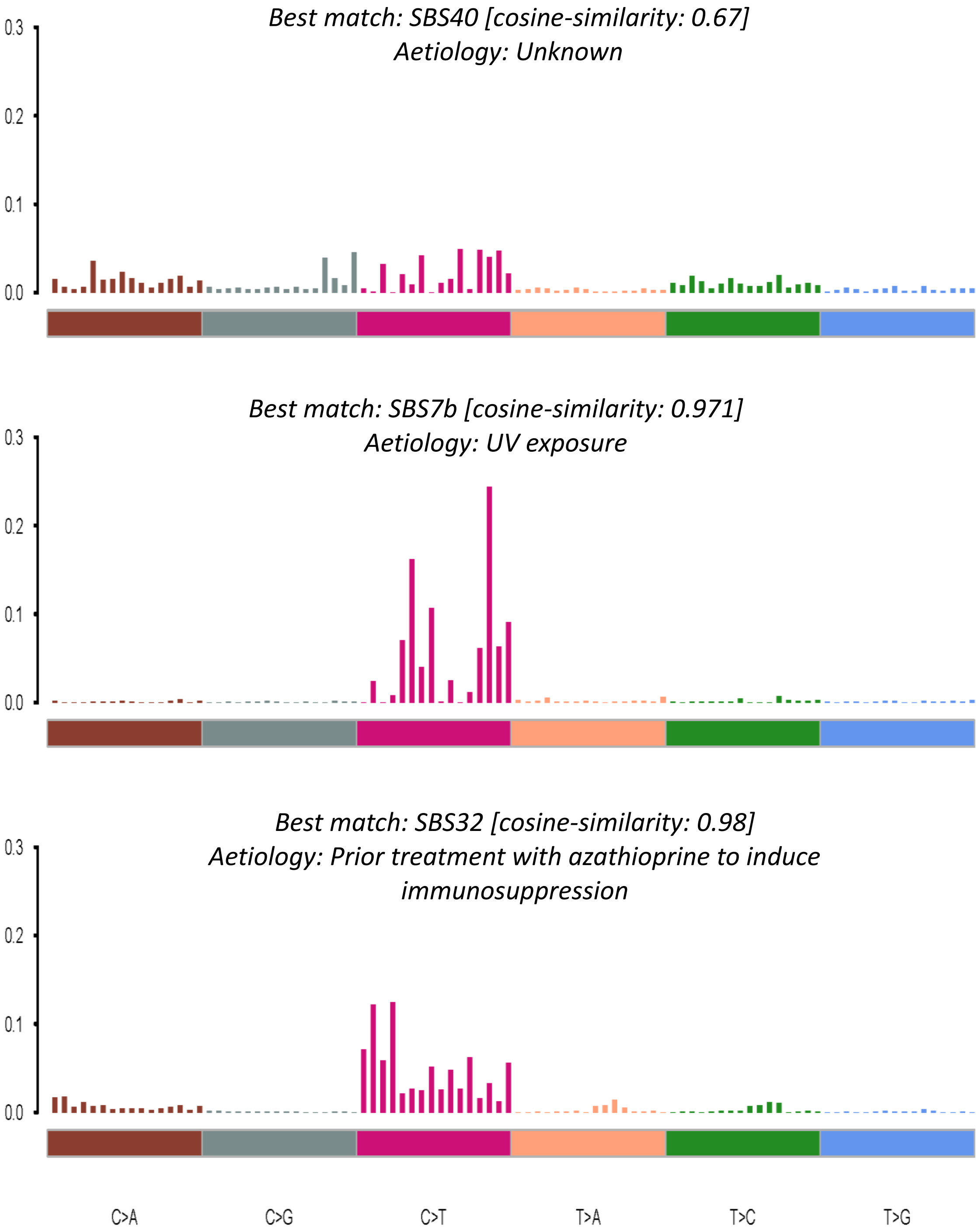


Figure 3-5: Single base subsitution mutation signature plot for cSCC. *The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the skin SCC tumour sample cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

Maftools identified three single base substitution mutation signatures in the cSCC tumour samples. As expected, SBS7b mutation signature which is associated with UV exposure was detected (figure 3-5). This mutation signature contains a varying proportion of C to T base changes with the highest proportion of C to T change noted when a cytosine is between a thymine and cytosine base in the genome. Another mutation signature identified was SBS32, which comprised a mutational pattern with a varying proportion of C to T base changes, including the highest proportion of C to T changes when a cytosine base is between an adenine and thymine base in the genome. This SBS32 mutation signature is associated with the drug Azathioprine, which is used to induce immunosuppression in patients with organ transplants and/or for other diseases affecting the skin (e.g., eczema), bowel (Crohns), etc. The 122 cSCCs included in the analysis in this chapter included cSCC samples from Inman *et al*., 2018 where a proportion of their samples had come from patients who had been documented as having prior Azathioprine exposure. Another mutation signature identified in the current dataset was the SBS40 mutation signature, with a cosine similarity of 0.67; this mutation signature has no known aetiology.

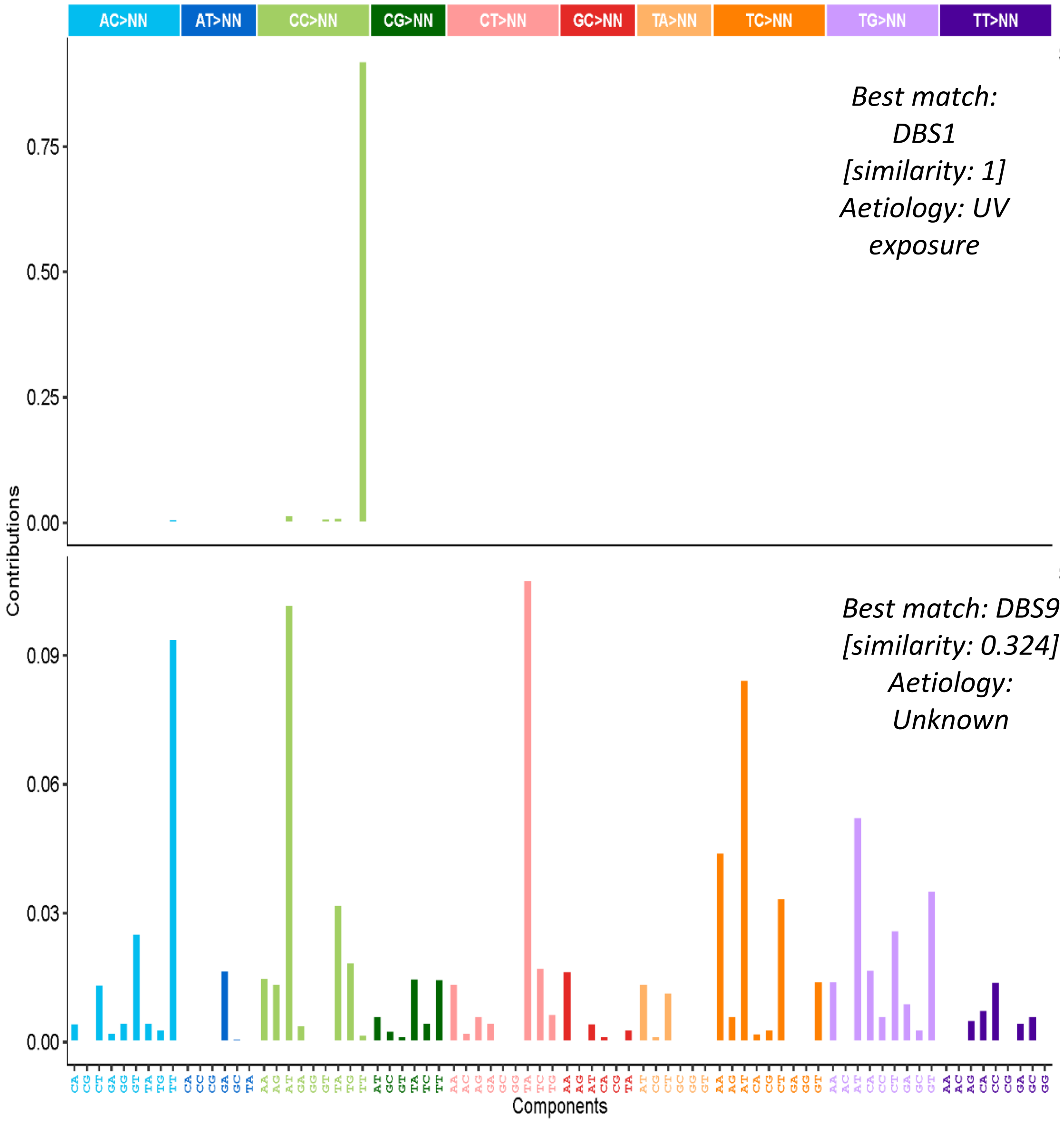


Figure 3-6: The double base substitution mutation signature plots for cSCC. *The x axis represents the base change in its dinucleotide context and the y axis represents the proportion of each base change in the skin SCC tumour sample cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.*

Sigminer identified two double base substitution mutation signatures in the skin SCC tumour samples (figure 3-6). The first mutation signature identified was DBS1 which is associated with UV exposure, with the highest proportion of base changes in this signature being CC to TT. The second mutation signature that was detected had a best match similarity of 0.324 to DBS9, which has an unknown aetiology.

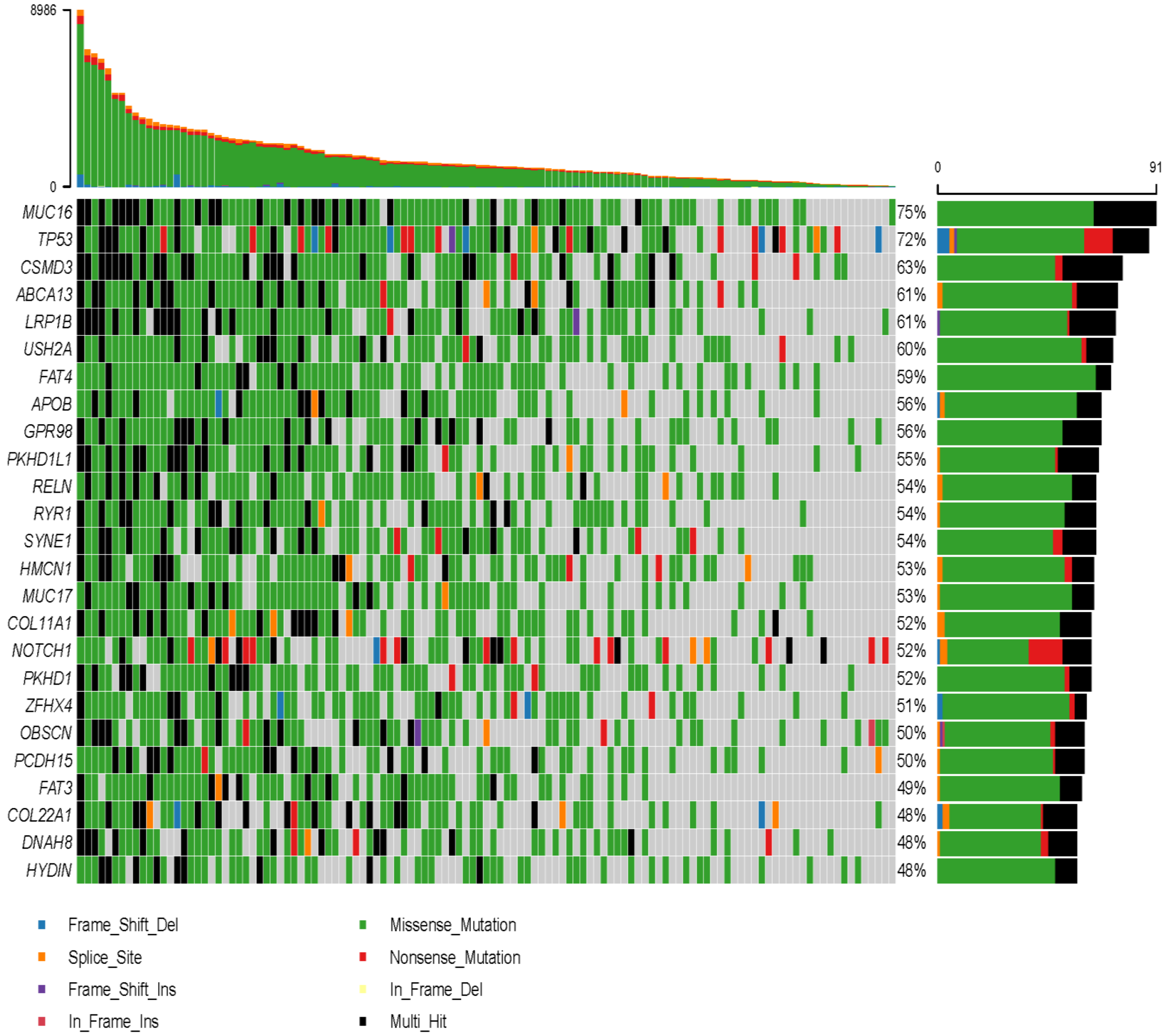


Figure 3-7: Oncoplot with top 25 frequently mutated genes in cSCC tumours identifed in the literature search and COSMIC database. *One or more of these genes was mutated in 119 of 122 samples (97.54%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation that each tumour sample contains with respect to the corresponding gene. The bar chart on the right represents the number of samples which have a mutation in that gene and the colours represent the type of mutation in the gene.*

Analysis of the 122 cSCC samples identified numerous genes that were mutated in these tumours, with 97.54% of the cSCCs having a mutation in at least one of the top 25 frequently mutated genes. Most of the mutations were missense mutations (figure 3-7). As can be seen from the oncoplot of the top 25 frequently mutated genes (figure 3-7), there were numerous tumours with mutations present in many of these genes. Somatic mutations in *MUC16* were most common and were seen in 75% of tumours. *TP53* was the second most highly mutated gene, detected in 72% of tumours, and included a variety of different types of mutations. The proportion of samples which shared mutations in the top 25 frequently mutated genes ranged from 48% to 75%.

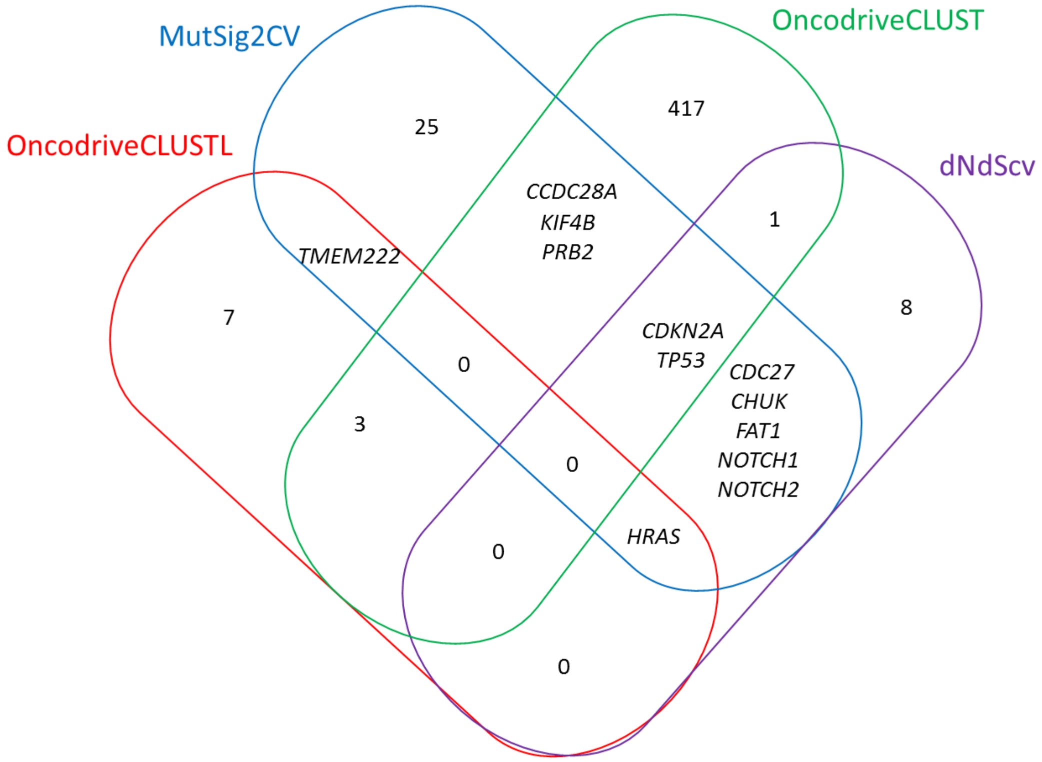


Figure 3-8: Venn diagram of the potential driver genes identified in cSCC using four different bioinformatics programs. *The four different bioinformatics programs are labelled in different colours corresponding to the colour of the outline of the closed curves; OncodriveCLUSTL (red), MutSig2CV (blue), OncodriveCLUST (green), dNdScv (purple). The numbers of potential driver genes, based on a false discovery rate q value <0.1, are highlighted and the names of the specific genes identified by MutSig2CV and one other bioinformatics programe are included in the relevant overlapping sections of the Venn diagram.*

Four bioinformatics programs were used to identify driver genes in skin SCC. In this analysis, a driver gene was characterised as significantly mutated in the MutSig2CV and at least one other of these bioinformatics programs. Using the MutSig2CV program, there were 37 genes with a q value less than 0.1 identified in the cSCC data (figure 3-8). There was a total of 426 genes which had a false discovery rate (q value) less than 0.1 produced from the OncodriveCLUST program using Maftools in R. There were 12 genes that had a q analytical value less than 0.1 in the OncodriveCLUSTL program. In the dNdScv program there were 18 genes with a global q value of less than 0.1. There were 12 genes which were significant in MutSig2CV and at least one other program (i.e., OncodriveCLUST, OncodriveCLUSTL and dNdScv). No genes were significant in all four programs. However, *HRAS* was significant in three programs, OncodriveCLUSTL, MutSig2CV and dNdScv, and the *CDKN2A* and *TP53* genes were also significant in three programs: MutSig2CV, OncodriveCLUST and dNdScv. The *NOTCH1* gene was significant in two programs, MutSig2CV and dNdScv and it was also within the top 25 frequently mutated genes. *TP53* was also present in the top 25 frequently mutated genes.

Table 3-1: The top 25 genes that were only mutated in MutSig2CV for cSCC samples. *Their status in the Cancer Gene Census is highlighted by tier number and N represents that the gene was not present in the Cancer Gene Census tier 1 genes have strong documented evidence relevant to cancer and tier 2 genes strong indications to cancer but less documented evidence. The mutation frequency column represents the frequency of cSCC samples which have a mutation in the associated gene.*

Gene	p	q	Cancer Gene Census Tier Number	Mutation frequency
TP53	1E-16	1.8862E-12	1	72.00%
NOTCH1	1.2977E-12	1.2239E-08	1	52.00%
CDKN2A	1.9971E-11	1.0758E-07	1	25.00%
NOTCH2	2.2814E-11	1.0758E-07	1	47.00%
ZNF750	1.9396E-09	7.3169E-06	N	22.00%
HRAS	3.8247E-09	1.2024E-05	1	16.00%
MOGAT1	2.2206E-07	0.00059836	N	10.00%
FAT1	3.596E-07	0.00084784	1	41.00%
VPS52	5.0009E-06	0.00999101	N	5.00%
COL4A4	5.2969E-06	0.00999101	N	46.00%
CHUK	6.8488E-06	0.01174387	N	11.00%
RPS18	9.0384E-06	0.01420691	N	5.00%
C15orf23	1.1286E-05	0.0163751	N	15.00%
MX2	2.6645E-05	0.03589798	N	16.00%
KCND3	2.9804E-05	0.03747704	N	12.00%
ITGA10	3.2813E-05	0.03868225	N	16.00%
PRB2	4.376E-05	0.04641887	N	37.00%
CDC27	4.4298E-05	0.04641887	N	16.00%
RB1	4.7195E-05	0.04685208	1	15.00%
FAM194A	5.3E-05	0.04998399	N	18.00%
DNAJA2	5.5921E-05	0.05022803	N	7.00%
TMBIM4	8.8798E-05	0.06882943	N	3.00%
SLC15A1	9.0449E-05	0.06882943	N	17.00%
AKAP2	9.1776E-05	0.06882943	N	24.00%
RLF	9.2243E-05	0.06882943	N	20.00%

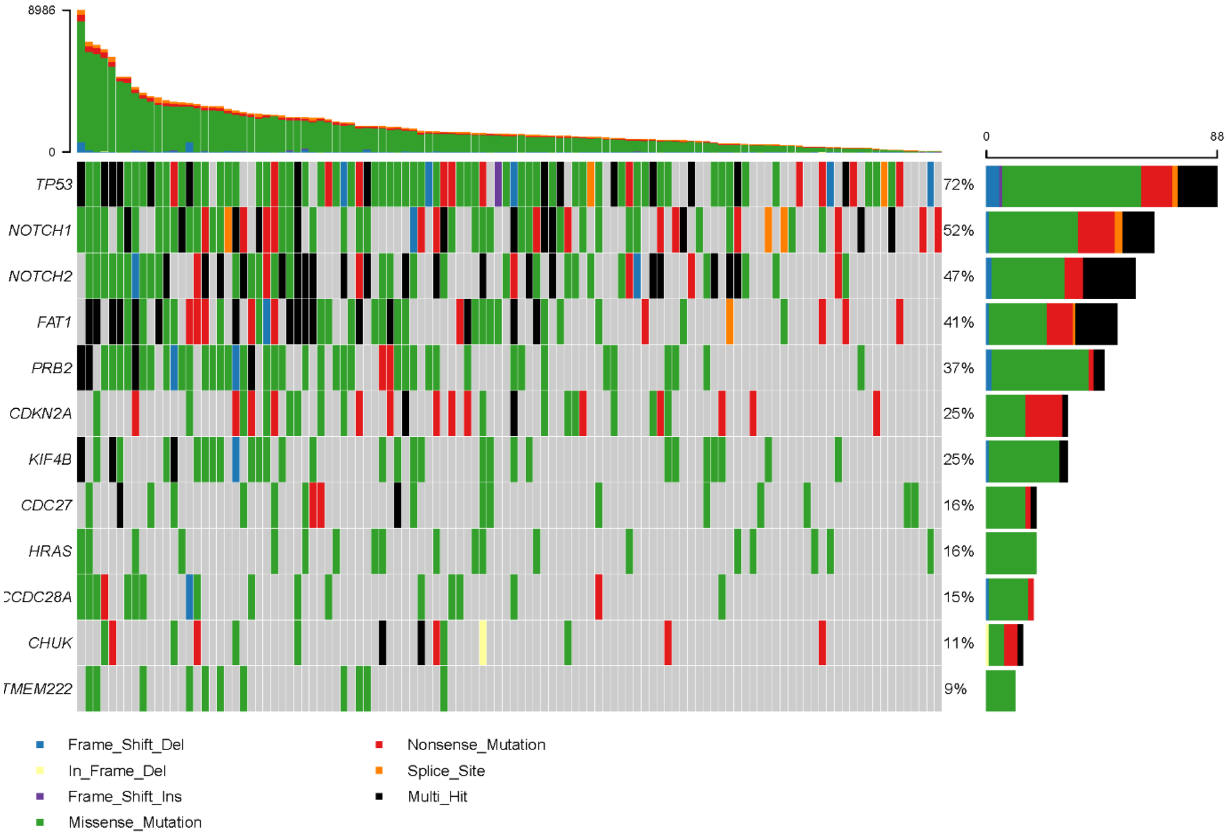


Figure 3-9: Oncoplot of driver genes identified in cSCC tumours using MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL. *One of more of these genes were altered in 112 of 122 samples (91.8%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation each sample contains with the corresponding gene. The bar chart on the right represents the number samples which have a mutation in that gene and the colours represent the type of mutation each of those samples have in the gene.*

The oncoplot of driver genes (figure 3-9) showed that these driver genes were mutated in 112 of 122 samples. The most frequently mutated driver gene was *TP53*, but all tumours which did not have a mutation in *TP53* contained mutations in one or more of the other driver genes. The least frequently mutated driver gene was *TMEM222* which was mutated in only 9% of samples; all *TMEM222* mutations were missense in these cSCCs. Similarly, mutations in *HRAS* were limited to missense mutations in the cSCCs containing *HRAS* mutations. All the other driver genes besides *HRAS* and *TMEM222* contained more than one type of mutation.

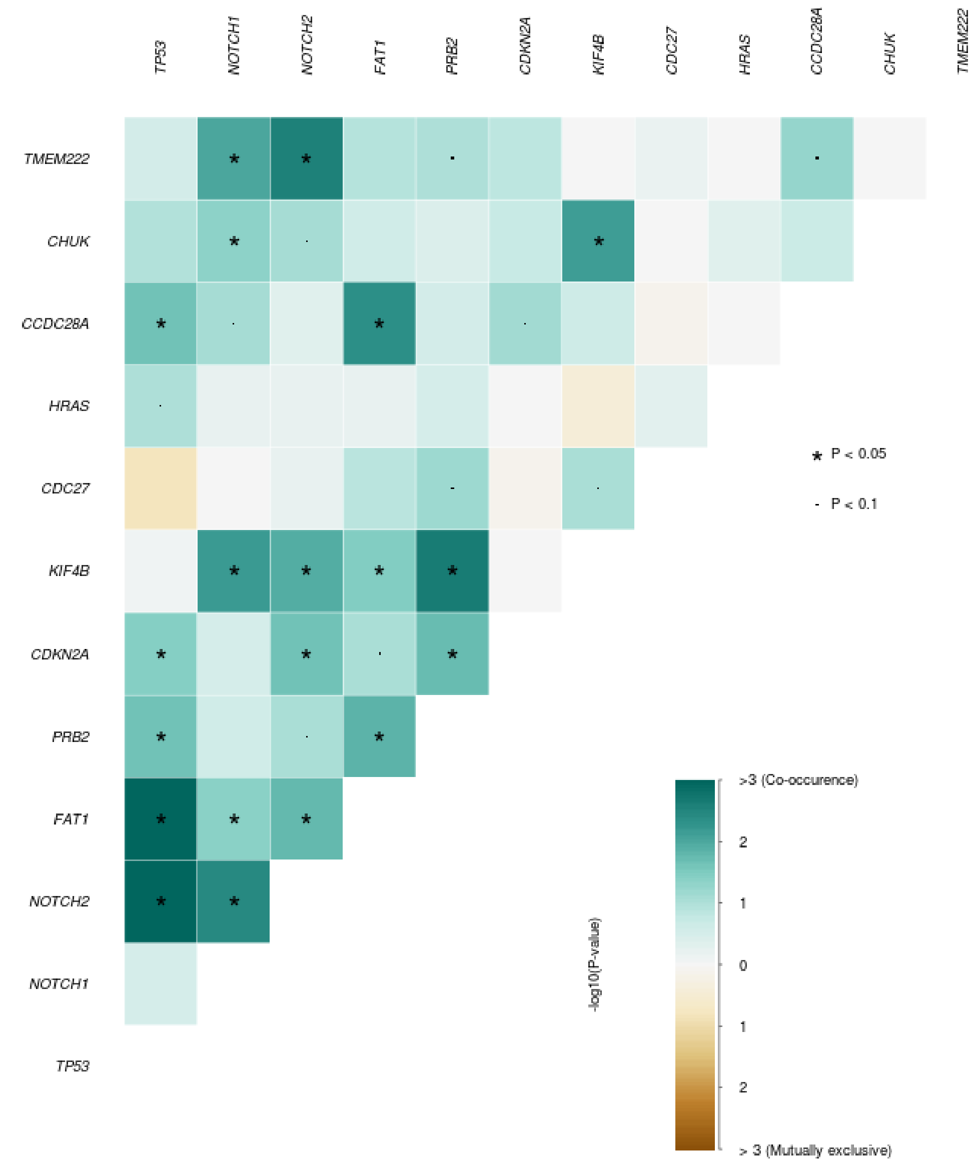


Figure 3-10: Co-occurrence plot for cSCC driver genes for 122 tumour samples from COSMIC database and literature search. *The squares shaded in green show that there is co-occurrence of genes. The squares shaded in yellow suggest that these genes are mutually exclusive. The dot represents that there is a level of significance with a p value of less than 0.1 and the star represents that there is a level of significant with a p value of less than 0.05.*

A co-occurrence plot was produced (figure 3-10), which showed that if a skin SCC tumour sample has a mutation in *TP53* then there is a significant chance that the tumour sample also has a mutation in *CCDC28A, CDKN2A, PRB2, FAT1,* and *NOTCH2* with a p value of less than 0.05. The most significant co-occurrence with *TP53* was *FAT1* and *NOTCH2*. The co-occurrence of *TP53* with *NOTCH2* had p value of 0.000086 and the co-occurrence of *TP53* with *FAT1* had a p value of 0.00020. The co-occurrence plot also showed that if a cSCC tumour sample has a mutation in *TP53*, there is a significant chance the sample also has a mutation in *HRAS* with a p value of less than 0.1. Mutations in *TP53* and *CDC27* seemed mutually exclusive, although this was not statistically significant in this dataset. Another two genes in which mutations looked to be mutually exclusive in the co-occurrence plot were *KIF4B* and *HRAS* however this was also not statistically significant. Where tumour samples had a mutation in *NOTCH1* then there was a significant chance that the tumour samples had a mutation in *TMEM222*, *CHUK*, *KIF4B*, *FAT1* and *NOTCH2*, each with a p value of less than 0.05. Mutations in *NOTCH1* also co-occurred with mutations in *CHUK* with a p value of less than 0.1. Tumours with mutations in both *NOTCH1* and *TP53* genes had co-occurring mutations in *FAT1* and *NOTCH2*. Of note, *NOTCH1*, *TP53*, *NOTCH2* and *FAT1* had also identified as significant driver genes in the MutSig2CV and dNdScv programs in figure 3-10 above.

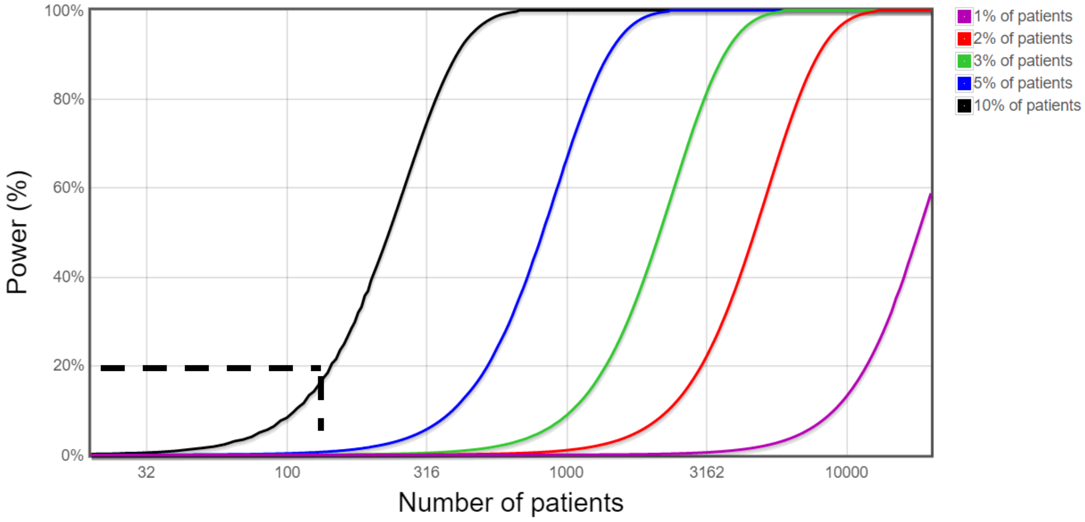


***Figure 3-11**: A power calculation graph showing the number of patients required to identify mutations in genes according to the estimated background mutation rate of skin SCC.* *The graph was produced on http://www.tumorportal.org/advanced_power. The vertical dotted line is showing an estimate of 122 patient samples and the horizontal line is showing the a value for power that corresponds to the sample size.*

The background mutation rate for skin SCC was estimated to be 16 mutations per megabase (Mb). This background mutation rate was calculated using the South et al., 2014 study. The supplementary data was used to identify the total number of silent mutations in each of the 20 cSCC tumour samples. The supplementary data also stated that the Agilent SureSelect Human All Exon 50Mb was used to capture exons for sequencing and there was 69% sequence coverage above 30X. The exon coverage was estimated to be 69% of 50Mb which was 34.5Mb therefore the total number of silent mutations were divided by 34.5 to identify the total number of mutations per Mb. The median mutations per Mb was then identified from the 20 cSCC samples and the figure of 16 mutations per Mb was used. The total mutation burden for cSCC was also estimated using the South et al., 2014 study and was estimated by using collating the total number of mutations for each cSCC divided by 34.5 and then identifying the median from the dataset which was 51 mutations per Mb. In this study, there was a total of 122 cSCC tumour samples retained for analysis. Based on the number of samples and the somatic mutation rate, there was 89% power to detect genes mutated in 20% of patients, 13% power to detect genes mutated in 10% of patients, 1% power to detect mutations in genes in 5% of patients, 0% power to detect mutations in genes for 3% of patients or less above the background mutation rate. The dashed line on figure 3-11 shows that using 122 samples there is 13% power to detect gene mutated in 10% of patients.

3.4 Discussion

This chapter identified driver genes in cSCC by using all the published whole genome and whole exome data available at present. Common cancer driver genes that had been identified in previous analyses were replicated in this study, including *TP53*, *NOTCH1*, *NOTCH2*, *CDKN2A* and *HRAS*. According to the power calculations there was 89% power to detect mutations in genes in 20% of patients. Since the mutation burden of skin SCC is high, many studies are required to identify genes which are significantly mutated compared to the background mutation rate. The top 25 frequently mutated genes show that a high proportion of cSCC samples share mutations in those 25 genes further highlighting the high mutation burden of skin SCC.

The mutation signature plots and the large proportion of C to T changes show that, as expected, UV is an important carcinogen which contributes to this high mutation burden. Other driver genes which were identified in this analysis such as *FAT1* and *CHUK* were also classed as significantly mutated genes in other analyses of cSCC (Chang and Shain, 2021, Pickering et al., 2014, Inman et al., 2018). The Chang and Shain, 2021 study investigated the landscape of driver mutations in cutaneous squamous cell carcinoma and they conducted a meta-analysis with 105 tumours spanning 10 studies. There were six driver genes identified in this chapter that were the same as those identified in the Chang and Shain, 2021 study (*TP53, NOTCH1, NOTCH2, CDKN2A, FAT1 and CHUK*). The Chang and Shain 2021 study used MutSig and dN/dS which were the same programs that were used in this chapter, however they also used two other programs OncodriveFML (Mularoni et al., 2016) and LOFsigrank (Shain et al., 2015a) which could be the reasons there are discrepancies in the driver genes identified. The study also used raw sequencing data from different studies instead of MAF files which could also contribute to the differences in the analyses. Novel genes identified as driver genes in this analysis were *CCDC28A*, *CDC27*, *KIF4B*, *PRB2* and *TMEM222*. However, the program which was consistently used across a wide range of papers was MutSig2CV (South et al., 2014, Inman et al., 2018, Chitsazzadeh et al., 2016) therefore this analysis classed a mutated gene as a driver if it was significant in MutSig2CV and at least one other program.

CCDC28A was mutated in 25% of skin SCC tumour samples in the current analysis and the protein has low expression in skin, and no expression in keratinocytes has been detected in The Human Protein Atlas (https://www.proteinatlas.org/ENSG00000024862-CCDC28A/tissue/skin) (Thul et al., 2017). The function of this gene has not been determined but it has been associated with childhood acute leukaemia (Petit et al., 2012). *CDC27* is a cell cycle protein and is part of the anaphase promoting complex in mitosis (Thornton et al., 2006). The protein has medium expression in keratinocytes (https://www.proteinatlas.org/ENSG00000004897-CDC27/tissue/skin) (Thul et al., 2017)and the gene was mutated in 16% of cSCC tumours analysed in this chapter. *PRB2* is a human salivary glycoprotein (Azen et al., 1992) and there is no data available showing expression in skin in The Human Protein Atlas (Thul et al., 2017) but the gene was mutated in 37% of cSCCs in the current analysis. *KIF4B* is similar to *CDC27* and is also involved in anaphase in mitosis. It is specifically involved in the spindle dynamics in anaphase and cytokinesis stages (Zhu et al., 2005). There is no record of *KIF4B* expression in skin however there has been evidence of amplification of this gene in Kidney renal clear call carcinoma (Chandrasekaran et al., 2015). *TMEM222* has a low expression in skin and is only mutated in 9% of skin SCC tumours. It has been recently reported that *TMEM222* has a role in brain development and is expressed in the parietal and occipital cortex, and that germline biallelic variants in this gene result in an autosomal recessive neurodevelopmental disorder (Polla et al., 2021).

In the co-occurrence plot, mutations in *CDC27* seemed to be mutually exclusive to mutations in *TP53*. However, this analysis did not have a p value of less than 0.1, therefore additional samples would need to be analysed to understand if mutations in these genes do occur exclusively. *CDC27* and *TP53* are both cell cycle proteins however *TP53* is an established tumour suppressor gene whereas *CDC27* is a core component of the anaphase promoting complex/cyclosome (APC/C) (Kazemi-Sefat et al., 2021) and it has been suggested that it plays a tumour suppressor or oncogene role in different neoplasms (Pawar et al., 2010, Qiu et al., 2017). *CDC27* RNA expression has been identified in different tumour types such as cervical SCC (Rajkumar et al., 2005), breast cancer (Talvinen et al., 2013), gastrointestinal cancers (Qiu et al., 2016), lung cancer (Bidkhori et al., 2013) and bladder cancer (Kim et al., 2016). In some of these cancer types elevated RNA expression of *CDC27* can increase cell proliferation, upregulation of *CDC27* can increase stemness in cancer stem cells and downregulation of *CDC27* can increase cancer cell survival. In squamous cell carcinoma of the cervix, *CDC27* downregulation showed poor radio- response and treatment failure (Rajkumar et al., 2005). Furthermore, this study also showed that reduced expression of *CDC27* in an irradiated cervical cancer cell line (SiHa cell line) promoted cell survival. These latter experiments suggest that *CDC27* is radio-sensitive and in this squamous cell carcinoma, *CDC27* is a tumour suppressor where reduced expression is increasing cell survival. Loss of function mutations in tumour suppressor genes and gain of function mutations in oncogenes promote tumorigenesis therefore to further understand the mechanism of *CDC27* in skin SCCs, the mutation types identified in this cancer need to be further investigated. In the MutSig2CV analysis *CDC27* had a q value less than 0.05 in both MutSig2CV and dNdScv, suggesting that it is strong driver gene in these programs. Future work could examine for protein expression of *CDC27* in normal skin and cSCC tumours to identify if there is any differential expression between these two tissues to further support a role for this gene in skin cancer development.

The whole exome and whole genome data in this study was analysed using different pipelines to call bases and tumour samples were also processed using different methods before DNA sequencing. These can affect the mutations identified in all the cSCC tumours. However, the variants were all annotated using the same annotation program to ensure any downstream analysis was consistent. To identify further novel driver genes or mutation signatures in cSCC the number of cSCC samples need to be increased to enable genes which are significantly mutated compared to the high background mutation rate to be identified.

Different bioinformatics programmes identified different driver genes. This was due to the way these programs functioned and dNdScv focused on the dN/dS mutation rate whereas OncodriveCLUST and OncodriveCLUSTL focused on the mutation clustering. All potential driver genes had to be significant using MutSig2CV because this program compared the mutation rate of a gene to the background mutation rate and measures the clustering of gene mutations. The functional significance of mutations was also measured in this program. Therefore, MutSig2CV combines elements of dNdScv, OncodriveCLUST and OncodriveCLUSTL. The dNdScv program would produce a list of genes which are identified as driver genes based on the nonsynonymous mutation rate and considers a wider range of covariates to more accurately estimate the background mutation rate compared to MutSig2CV. However, the dNdScv would not be able to identify genes which are produced with regards to positional clustering patterns. OncodriveCLUST and OncodriveCLUSTL can be used to identify genes which are considered driver genes due to mutational clustering however would not be able to identify genes which are significantly mutated due to the number of nonsynonymous mutations. Therefore, a combination of these programs would provide a realistic list of potential driver genes however the strength of each program to detect driver genes is subjective. By increasing the power of the study could enable the detection of more driver genes using these programs and genes which are present in all programs would have strong evidence of driver gene status.

4. Comparative analysis of genetic mutations in cutaneous SCC and SCCs at other organ sites

4.1 Introduction

SCCs arise in tissues lined with epithelium and the most common types of SCCs are skin SCC, oropharyngeal (also called "head and neck") SCC, lung SCC, oesophageal SCC, and cervical SCC. There are different carcinogens which cause these cancers. The main carcinogen responsible for skin SCC is UV ((Armstrong and Kricker, 2001), (Narayanan et al., 2010). Human Papilloma Virus (HPV) contributes to the development of oropharyngeal and cervical SCC (Doorbar, 2006), (Leemans et al., 2011). HPV is also associated with an increased risk of skin squamous cell carcinoma and an association with oesophageal SCC has been identified, however the link of HPV with oesophageal SCC has not been firmly established (Iannacone et al., 2012, Ludmir et al., 2015). Excessive smoking and alcohol intake is a common cause of HPV –negative oropharyngeal SCCs and oesophageal SCCs (Kobayashi et al., 2018). HPV-negative oropharyngeal SCC predominantly affects elderly males (Zumsteg et al., 2016) and oesophageal SCC is common in Asian populations (Torre et al., 2016).

According to Cancer Research UK, there were 367,167 new cases of cancer diagnosed each year in the UK from 2015 to 2017 (CRUK). Annually during this period there were 47,838 cases of lung cancer, 3,152 cervical cancer cases, 12,238 head and neck cancer, 9,209 oesophageal and 151,739 cases of keratinocyte cancer (formerly called non-melanoma skin cancer). Approximately 25 – 30% of all lung cancers are squamous cell carcinomas (Kenfield et al., 2008), therefore, there were approximately 14,351 cases of lung SCC. Venables et al (2019) documented that there were 44,672 skin SCCs in the UK in 2015, which shows that the largest number of SCC cases are skin SCC (Venables et al., 2019). It is unclear whether SCCs arise from similar or different somatic genetic events in these organs, and whether driver genes that are important in the growth of SCCs in the oropharyngeal region, lung, oesophagus, and cervix play a role in skin SCC development.

Genes that are frequently mutated in skin SCCs include *TP53*, *CDKN2A*, *NOTCH1*, *NOTCH2* (Li et al., 2015). In oropharyngeal SCCs, differences have been reported in the genes mutated in HPV-negative tumour samples and in HPV-positive tumour samples. For example, in HPV-positive oropharyngeal SCCs the most commonly mutated gene is *PIK3CA* and the most commonly altered pathway is the PI3K pathway whereas *TP53* is the most commonly mutated gene in HPV-negative SCCs and with common alterations in genes involved in the cell cycle and PI3K pathways (Chung et al., 2015). Genes which are frequently mutated in lung SCCs include *TP53*, *GRM8*, *BAI3*, *ERBB4*, *RUNX1T1*, *KEAP1*, *FBXW7* and *KRAS* (Kan et al., 2010). In oesophageal SCC, an exome study of 113 tumour-normal pairs reported that the genes frequently altered in 99% of cases were involved in cell cycle and apoptosis regulation, including *TP53*, *CCND1*, *CDKN2A*, *NFE2L2* and *RB1* (Gao et al., 2014).

Some studies have previously investigated the genomic relationship between SCCs at different sites (Campbell et al., 2018, Schwaederle et al., 2015) and the mRNA expression between SCCs at different sites (Chitsazzadeh et al., 2016). A study by Schwaederle and colleagues in 2015 compared the exome sequences of 361 SCC samples (oropharyngeal, lung, cutaneous, gynaecological, gastrointestinal, unknown origin) and 277 non-SCC samples. The study compared all 361 SCC samples and identified that the most frequently altered genes in SCCs were *TP53*, *PIK3CA*, *CDKN2A*, *SOX2* and *CCND1*.The study also investigated if SCC samples had particular set of genes which had an alteration frequency that was statistically different from the non-SCC samples. However, there were only 36 cutaneous SCC samples in the cohort, although the study did highlight that in the 'subset of cutaneous SCCs, NOTCH1 alterations were found in 33% of cases, versus 10% in other types of SCCs'. A "PAN-CANCER" study by another group which analysed the copy number alterations (CNAs) in SCCs (lung, oropharyngeal, oesophageal and bladder cancers) showed that there were 5 clusters corresponding to the number of recurrent amplifications or deletions in each chromosome (Campbell et al., 2018). Each of these clusters included a mixture of SCCs from different organs which suggests that SCCs in these organs can have very similar molecular characteristics in relation to cancer development. A review paper (Dotto and Rustgi, 2016) identified that genes which distinguish SCCs are involved in squamous cell differentiation, including *NOTCH1*, *TP63* and *SOX2* and their interaction with *EGFR* and *RAS* pathways.

To date, there has not been any comprehensive research study comparing the genomic profiles of skin SCCs and SCCs in other organs. Furthermore, identifying whether skin SCCs and SCCs of different organs have similar genetic abnormalities might support the use of targeted therapies in the different types of SCC, including skin SCC, and/or may improve cancer classification systems.

4.2 Methods

Whole genome and exome data were extracted from the COSMIC database v91 (https://cancer.sanger.ac.uk/cosmic) using the script documented in appendix 7.1. The format of the data was changed to ensure it was compatible for the genome annotation program, Oncotator as described in chapter 3. For oropharyngeal SCC, the mutation data from upper_aerodigestive_tract_squamous_cell_carcinoma.txt was used for the analysis. For oesophageal SCC, it was the oesophagus_squamous_cell_carcinoma.txt file. The mutation file that was used for lung SCC was lung_squamous_cell_carcinoma.txt and for cervical SCC it was cervix_squamous_cell_carcinoma.txt. The substitution, insertion and deletion mutation files were created for each type of squamous cell carcinoma and edited to be compatible for Oncotator as outlined in appendix 7.3. To ensure all genomic data were from human SCC samples scripts were created to identify and subsequently discard any genomic data from cell lines. Samples with a description of 'short term' were also identified and discarded because this described cells which had been sequenced after short-term culture.

For the cervix_squamous_cell_carcinoma.txt file, column 5 was extracted which was the sample name and column 35 which included a description of the tumour origin (cell-line). The next part of the script deleted duplicates to produce a file with all the sample names which have originated from cell lines. The same was done for samples which have originated from short-term cultures.

This script was replicated for oropharyngeal SCC, lung SCC and oesophageal SCC.

In oropharyngeal SCC, the sample name (column 5) and site subtype 2 column of the upper_aerodigestive_tract_squamous_cell_carcinoma.txt file was extracted and all the lines with 'lip' were mined from the file. Duplicates were also removed to produce a file with all the sample names which had genomic data for squamous cell carcinomas that originated in the lip. The Linux script that which was used to do this is shown in appendix chapter 7, section 4. To ensure the oropharyngeal SCC tumour samples did not include skin SCC tumour samples, any samples from the lip were identified so they could be discarded in the analysis.

For oropharyngeal SCC, the files with samples that originated from the lip and samples which were not human samples were collated to produce a single file. This file could be used to ensure these sample identifiers and their genomic data were removed from the COSMIC dataset before analysis.

SCC tumour sample names were also downloaded from GDC portal (https://portal.gdc.cancer.gov/) and the mutation data for these samples were taken from the mc3.v0.2.8.PUBLIC.maf which is described in chapter 2, section 4.2 and chapter 2, section 5.2. The sample name, chromosome number, start position, end position, reference base and mutated base were used to produce a file compatible for Oncotator. The samples identified in GDC portal were compared with those from the COSMIC database to ensure there were no duplicates. Then the samples which had been identified as genetic data from cell lines were removed from the analysis. The genetic data which originated from the lip and cell line genomic data were removed from the oropharyngeal SCC analysis.

The SCC genetic data from mc3.v0.2.8.PUBLIC.maf file was merged with the COSMIC SCC genetic data and annotated using Oncotator.

Literature searches for SCCs were conducted to identify WGS and/or WES data for oropharyngeal, oesophageal, lung and cervical SCC. All literature was screened on MEDLINE OVID for oropharyngeal SCC and oesophageal SCC until 30th January 2020. The final literature search for lung SCC was 31st January 2020 and for cervical SCC, the search was completed on 1st February 2020. The data from these literature searches were also annotated using Oncotator and collated with the SCC data from mc3.v0.2.8.PUBLIC.maf and COSMIC. The MAF files produced for each SCC were analysed using Maftools in R. The bioinformatics tools used to identify driver genes were MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL.

4.3 Results

The flowcharts in the following figures show the number of studies that were identified in the literature search for WGS and WES data for SCCs in the different organs. The flowcharts also show how many tumours were identified in the genetic databases, GDC portal and COSMIC for each SCC. Reasons for studies from the literature search not being included in the final analysis and reasons for tumour samples not being included in the final analysis are outlined in the flowcharts.

4.3.1 Comprehensive literature review flowcharts

Oropharyngeal SCC

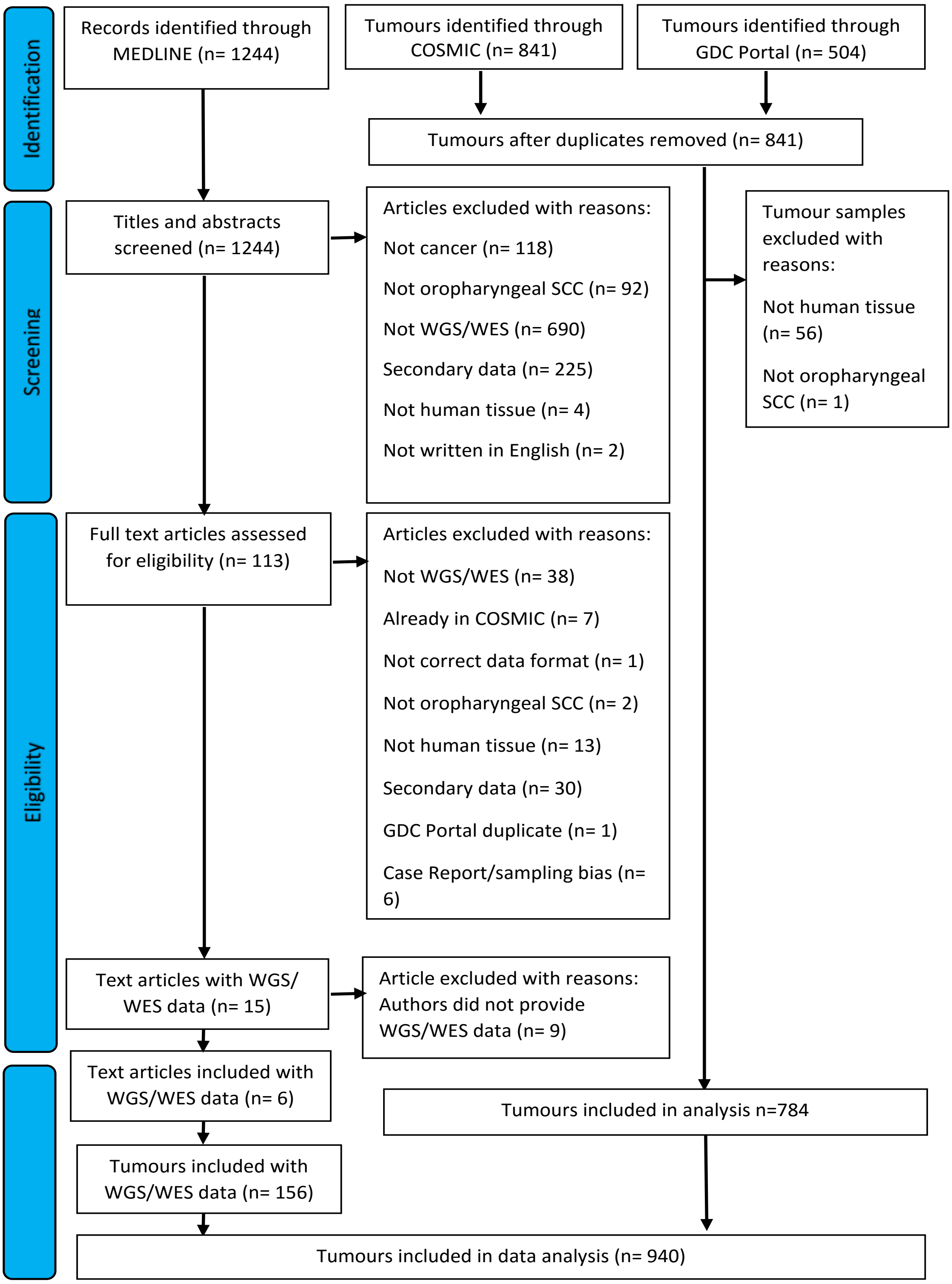

Identification
Screening
Eligibility
Records identified through MEDLINE (n= 1244)
Tumours identified through COSMIC (n= 841)
Tumours identified through GDC Portal (n= 504)
Tumours after duplicates removed (n= 841)
Titles and abstracts screened (n= 1244)
Articles excluded with reasons:
Not cancer (n= 118)
Not oropharyngeal SCC (n= 92)
Not WGS/WES (n= 690)
Secondary data (n= 225)
Not human tissue (n= 4)
Not written in English (n= 2)
Tumour samples excluded with reasons:
Not human tissue (n= 56)
Not oropharyngeal SCC (n= 1)
Full text articles assessed for eligibility (n= 113)
Articles excluded with reasons:
Not WGS/WES (n= 38)
Already in COSMIC (n= 7)
Not correct data format (n= 1)
Not oropharyngeal SCC (n= 2)
Not human tissue (n= 13)
Secondary data (n= 30)
GDC Portal duplicate (n= 1)
Case Report/sampling bias (n= 6)
Text articles with WGS/ WES data (n= 15)
Article excluded with reasons: Authors did not provide WGS/WES data (n= 9)
Text articles included with WGS/WES data (n= 6)
Tumours included in analysis n=784
Tumours included with WGS/WES data (n= 156)
Tumours included in data analysis (n= 940)

Figure 4-1 Flowchart of manuscripts selected for investigation of oropharyngeal SCC. *The number of studies selected for analysis and the reasons for studies being excluded at different stages are specified.*

Medline OVID search was conducted for oropharyngeal SCC with search terms specified in the appendix 7.9.2. The search was designed to identify studies which included WGS or WES data for oropharyngeal SCC and there were a total of 1244 search results (figure 4-1). The titles and abstracts of these studies were read, and 113 studies were identified as relevant to oropharyngeal SCC. There were 1131 studies which were removed, because 118 studies were not about cancer, 92 were not specific to oropharyngeal SCC, 690 studies did not include WGS or WES, 225 studies were reviews and not primary data, four studies included data which were not from human tissue samples and two studies were not written in English. The full text articles were read for the 113 studies and 15 of these studies included WGS or WES data. Of the 113 studies, 38 studies were removed because they were not WGS OR WES data, seven studies were identified as already in the COSMIC database, one study was not in the correct data format (i.e. the data was not in a MAF file and was not in format where it could be converted straightforwardly into a MAF file). In addition, two of the 113 studies were not oropharyngeal SCC studies, 13 studies included data which was not from human tissue samples, 30 studies were reviews which included secondary data, one study included data from GDC portal (thus was a duplicate) and six studies only included an individual sample and therefore were not included due to sampling bias. Of the 15 studies which included WGS or WES data, 11 authors were contacted because their data was not available in the paper and nine authors chose not to share their data. Thus, from the Medline OVID search, supplementary data from four studies were used and two authors shared their data, therefore, six studies remained from the comprehensive literature search, and this included data from 156 tumour samples.

There were 841 tumours that were identified from the COSMIC database and 504 tumours were identified in GDC portal. The tumour barcodes were merged, and duplicates were removed which resulted in 841 tumour samples. Then 56 tumour samples which were not from human tissue were removed and one sample was removed because it was not considered as oropharyngeal SCC because the sample originated from the lip. After the removal of these samples, 784 tumour samples remained which were merged with the 156 tumour samples from the comprehensive literature search. Thus, there were 940 tumour samples with WES data that were analysed for oropharyngeal SCC.

Lung SCC

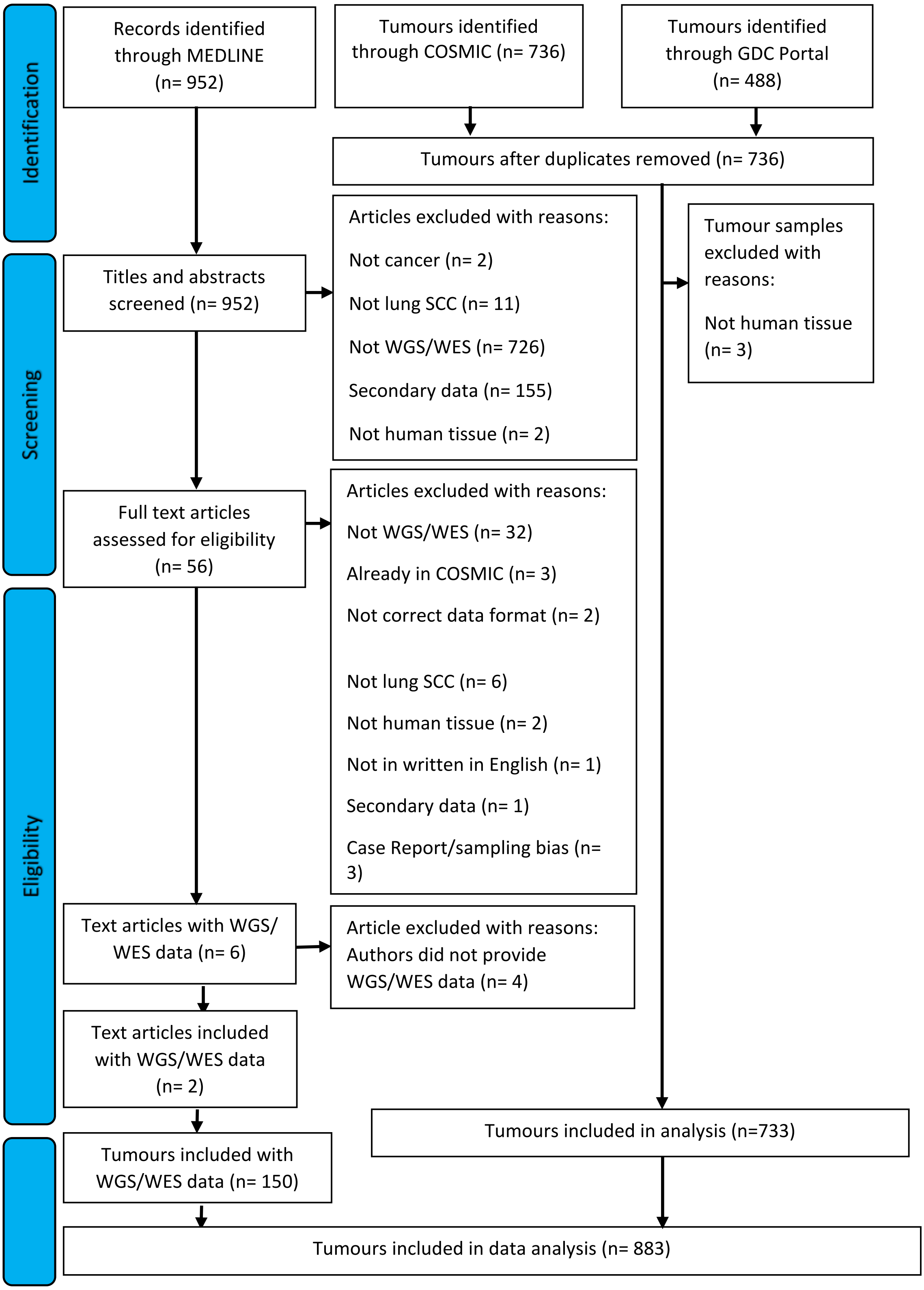

Identification
Screening
Eligibility
Records identified through MEDLINE (n= 952)
Tumours identified through COSMIC (n= 736)
Tumours identified through GDC Portal (n= 488)
Tumours after duplicates removed (n= 736)
Articles excluded with reasons:
Not cancer (n= 2)
Not lung SCC (n= 11)
Not WGS/WES (n= 726)
Secondary data (n= 155)
Not human tissue (n= 2)
Tumour samples excluded with reasons:
Not human tissue (n= 3)
Titles and abstracts screened (n= 952)
Full text articles assessed for eligibility (n= 56)
Articles excluded with reasons:
Not WGS/WES (n= 32)
Already in COSMIC (n= 3)
Not correct data format (n= 2)
Not lung SCC (n= 6)
Not human tissue (n= 2)
Not in written in English (n= 1)
Secondary data (n= 1)
Case Report/sampling bias (n= 3)
Text articles with WGS/ WES data (n= 6)
Article excluded with reasons: Authors did not provide WGS/WES data (n= 4)
Text articles included with WGS/WES data (n= 2)
Tumours included in analysis (n=733)
Tumours included with WGS/WES data (n= 150)
Tumours included in data analysis (n= 883)

*Figure 4-2: **Flowchart representing studies identified in literature search for lung SCC.** The number of studies which have been discarded are specified and their corresponding reasons for removal are stated.*

A Medline OVID search was performed for lung SCC with search terms specified in appendix 7.9.4. The aim of the search was to identify studies which included WGS or WES data for lung SCC. A total of 952 studies were identified via this OVID search (figure 4-2). After the abstracts for these studies were examined, it was ascertained that 56 of these studies were relevant to lung SCC. A total of 896 studies were removed at this first stage of filtering which included two studies because they were not about cancer, 11 studies were not specific to lung SCC, 726 studies did not include WGS or WES for lung SCC, 155 studies were reviews or did not have primary data and two studies included data for samples which were not from human tissue. The full text articles were screened in the remaining 56 studies and 50 of these studies were discarded. Three of these 56 studies were already included in the analysis via the COSMIC database, 32 articles did not include WGS or WES data and two studies were not in the correct format for analysis and could not be converted into a MAF file or was not already in MAF format. Furthermore, there were six of the 56 studies that did not include lung SCC WGS or WES data, two articles contained samples that were not human tissue, one study was not written in English, one study included secondary data and three articles were case reports or only had data on a single patient and was not included due to sampling bias. The six studies that were remained from the 56 publications included analysis of lung SCC WES or WGS data. However, the data was not available from four of the studies, so the data was requested from four authors by email but they did not provide the data for analysis. The other two text articles provided data in the supplementary data of their study, and this included 150 tumour samples, which were then used in the analysis for lung SCC in the current study.

Lung SCC tumour samples which had been whole genome or whole exome sequenced were identified in the COSMIC database and these 736 tumours were merged with 488 lung SCC tumour samples from GDC portal. After duplicate tumour sample barcodes were removed, 736 tumour samples remained. Three of the 736 samples were not from human tissue and were discarded. The 733 lung SCC samples were merged with the 150 lung SCC tumour samples identified from the comprehensive literature search and a total of 883 lung SCC tumour samples were included in the analysis.

Oesophageal SCC

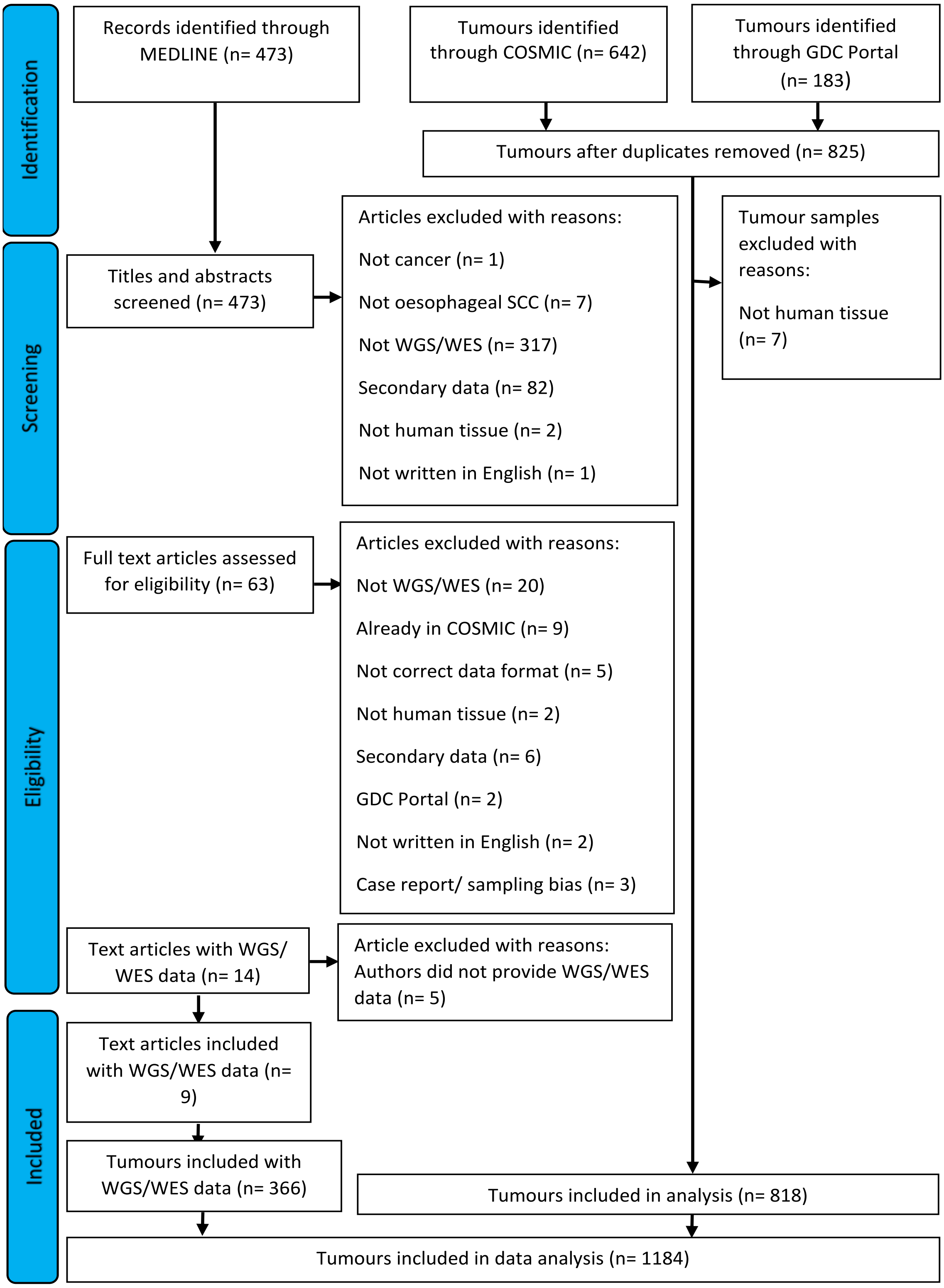

Identification
Screening
Eligibility
Included
Records identified through MEDLINE (n= 473)
Tumours identified through COSMIC (n= 642)
Tumours identified through GDC Portal (n= 183)
Tumours after duplicates removed (n= 825)
Titles and abstracts screened (n= 473)
Articles excluded with reasons:
Not cancer (n= 1)
Not oesophageal SCC (n= 7)
Not WGS/WES (n= 317)
Secondary data (n= 82)
Not human tissue (n= 2)
Not written in English (n= 1)
Tumour samples excluded with reasons:
Not human tissue (n= 7)
Full text articles assessed for eligibility (n= 63)
Articles excluded with reasons:
Not WGS/WES (n= 20)
Already in COSMIC (n= 9)
Not correct data format (n= 5)
Not human tissue (n= 2)
Secondary data (n= 6)
GDC Portal (n= 2)
Not written in English (n= 2)
Case report/ sampling bias (n= 3)
Text articles with WGS/ WES data (n= 14)
Article excluded with reasons: Authors did not provide WGS/WES data (n= 5)
Text articles included with WGS/WES data (n= 9)
Tumours included with WGS/WES data (n= 366)
Tumours included in analysis (n= 818)
Tumours included in data analysis (n= 1184)

Figure 4-3: Flowchart representing the number of studies selected for analysis of oesophageal SCC. *The number of studies which have been discarded and their corresponding reason for removal are specified.*

A Medline OVID search was undertaken to identify studies with WGS or WES data for oesophageal SCC with search terms specified in appendix 7.9.3 and gave a total of 473 search results (figure 4-3). The title and abstracts of these papers were screened, and 63 papers were identified as being relevant to oesophageal SCC. The reason 410 studies were removed was because one paper did not have any association with cancer, seven papers were not specific to oesophageal SCC, 317 papers did not include WGS or WES data, 82 papers were review articles or did not include primary data required for this analysis, two articles included data which were not from human tissue samples and one study was not written in English. Then the full text articles were assessed for the remaining 63 studies and 49 articles were discarded for the following reasons: 20 papers did not include WES or WGS data, nine studies were already included in the COSMIC database, five studies were not in the correct data format for analysis therefore could not be analysed as a MAF file, six articles included secondary data, two studies used data from GDC portal (thus would duplicate data obtained from GDC portal (as below)), two studies were not written in English and three articles were case reports or included data from a single individual and were not included due to sampling bias. The 14 studies which remained included WGS and WES data, however in five of these studies, the data was not available in the paper. The lead authors were contacted for those five studies, but none of the authors provided their WGS or WES data. The remaining nine studies included WES data in the supplementary material and included a total of 366 oesophageal SCC tumour samples.

A search was conducted in the COSMIC database for WGS or WES data for oesophageal SCC samples and 642 samples were identified. The GDC Portal database also included WES data for 183 oesophageal SCC tumour samples. The 642 tumour samples from COSMIC was merged with the 183 samples from GDC portal and after duplicates were removed 825 samples remained. Then, seven samples were discarded because these samples were from non-human tissue. The remaining samples from COSMIC and GDC portal comprised 818 oesophageal tumours. The 818 tumour samples were merged with 366 tumour samples from the comprehensive literature search and a total of 1184 oesophageal tumours were included in the analysis.

Cervical SCC

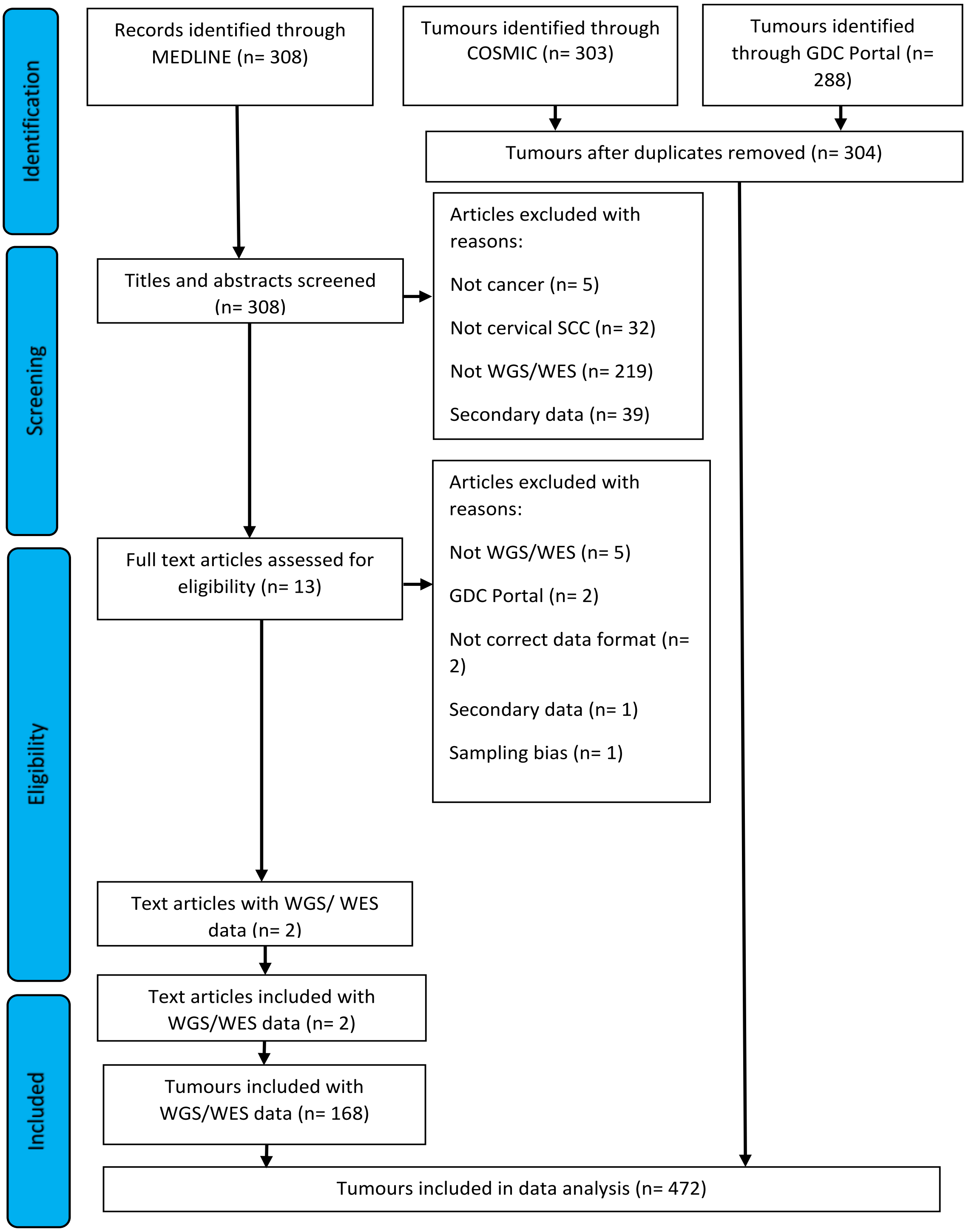

Identification
Screening
Eligibility
Included
Records identified through MEDLINE (n= 308)
Tumours identified through COSMIC (n= 303)
Tumours identified through GDC Portal (n= 288)
Tumours after duplicates removed (n= 304)
Titles and abstracts screened (n= 308)
Articles excluded with reasons:
Not cancer (n= 5)
Not cervical SCC (n= 32)
Not WGS/WES (n= 219)
Secondary data (n= 39)
Full text articles assessed for eligibility (n= 13)
Articles excluded with reasons:
Not WGS/WES (n= 5)
GDC Portal (n= 2)
Not correct data format (n= 2)
Secondary data (n= 1)
Sampling bias (n= 1)
Text articles with WGS/ WES data (n= 2)
Text articles included with WGS/WES data (n= 2)
Tumours included with WGS/WES data (n= 168)
Tumours included in data analysis (n= 472)

__Figure 4-4: Flowchart representing the number of studies selected for analysis of cervical SCC__. The number of studies which have been discarded and their corresponding reason for removal are specified.

A Medline OVID search was conducted for cervical SCC to identify studies with WGS or WES data. The search terms are specified in appendix 7.9.5 and provided a total of 308 search results. The abstracts of these papers were analyzed and 13 papers were identified as being relevant to cervical SCC. The reason 295 studies were removed was because five papers did not have any association with cancer, 32 papers were not specific to cervical SCC, 219 papers did not include WGS, or WES data and 39 papers were review articles or did not include primary data required for this analysis. The whole text articles were screened for the 13 papers that were relevant to cervical SCC and 11 articles were removed for the following reasons: five articles did not include WGS or WES data, two studies used data from GDC portal, two studies were not in the correct format for analysis so did not fulfil the criteria to produce a MAF file, one study included secondary data and one study only had data for a single patient and was not used due to sampling bias. The two studies that contained WGS or WES data included a total of 168 tumour samples. The COSMIC database included WGS or WES data for 303 cervical SCC samples and GDC portal included WES data for 288 cervical SCC tumour samples. Cervical SCC tumour samples from the COSMIC database and GDC portal were merged and following removal of duplicates, 304 tumour samples remained. These 304 cervical SCC tumour samples were combined with 168 samples from the comprehensive literature search and the WES data for 472 cervical SCC tumours were subsequently analyzed.

4.4 Power calculations

Based on a background somatic mutation rate for oropharyngeal SCC of four mutations per megabase (Lawrence et al., 2014) and 940 oropharyngeal tumour samples in the analyses, there was 100% power to detect genes mutated in 5% of patients, 76% power to detect genes mutated in 3% of patients, 24% power to detect genes mutated in 2% of patients and 1% power to detect genes mutated in 1% of patients (figure 4-5).

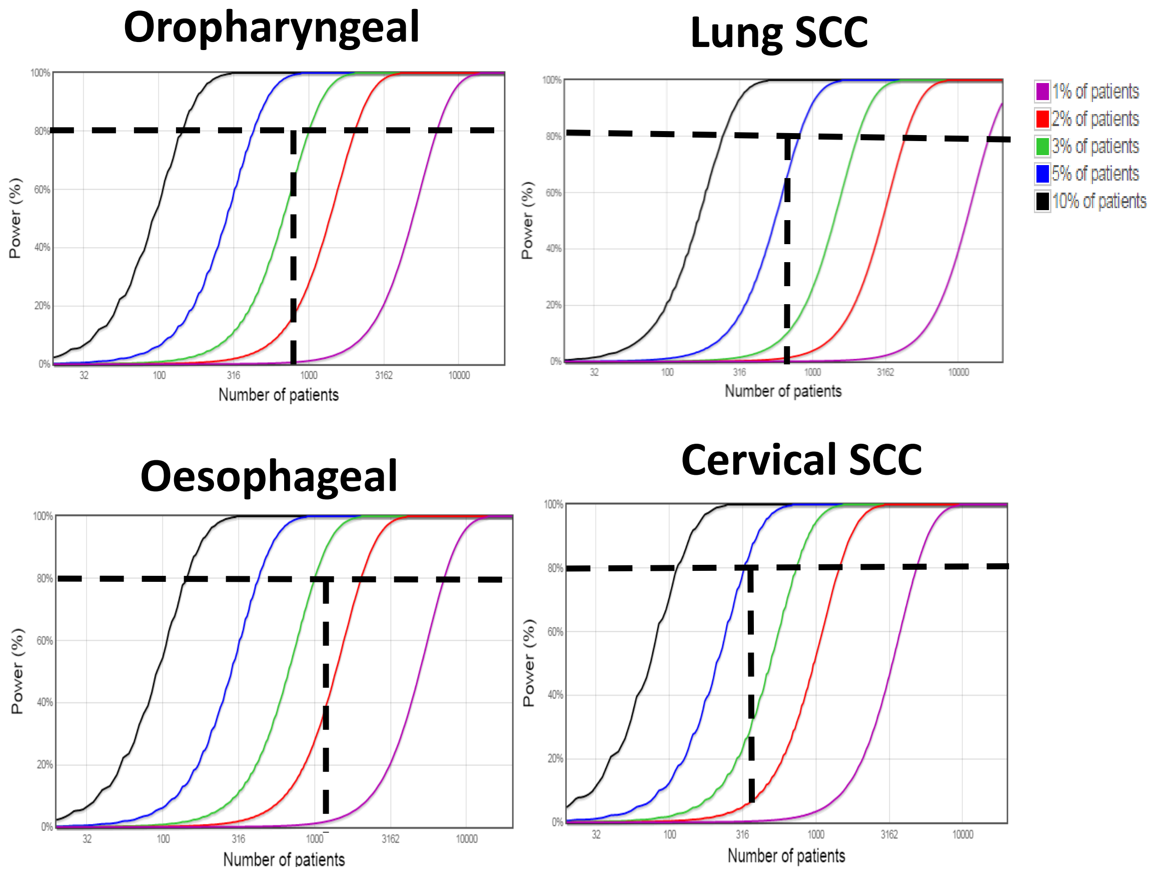


__Figure 4-5: Graph representing the power of the study for whole genome and whole exome sequencing for oropharyngeal, lung, oesophageal and cervical SCC tumour samples identified from COSMIC database, GDC portal and a literature search.__ The x axis shows the number of patients and the y axis shows the power of the study. Each curve represents the percentage of patients in which a mutation can be detected. The vertical dotted line is showing an estimate of the number of patient samples available and the horizontal line is showing the a value for power that corresponds to the sample size.

As the background somatic mutation rate for lung SCC was 10 mutations per megabase (Lawrence et al., 2014) and the current analysis included 883 lung SCC samples, there was 100% power to detect genes mutated in 10% of patients, 85% power to detect genes mutated in 5% of patients, 20% power to detect genes mutated in 3% of patients, 3% power to detect genes mutated in 2% of patients and 0% power to detect genes mutated in 1% of patients.

The background somatic mutation rate for oesophageal SCC was four mutations per megabase (Lawrence et al., 2014) and there were 1184 oesophageal tumour samples, thus the current study had a 100% power to detect genes mutated in 5% of patients, 90% power to detect genes mutated in 3% of patients, 40% power to detect genes mutated in 2% of patients and 2% power to identify genes mutated in 1% of patients.

As the background somatic mutation rate for cervical SCC was 2.5 mutations per megabase (Lawrence et al., 2014), the 472 cervical SCCs gave 100% power to detect genes mutated in 10% of patients, 96% power to detect genes mutated in 5% of patients, 47% power to detect mutated genes in 3% of patients, 12% power to identify genes mutated in 2% of patients and 1% power to detect genes mutated in 1% of patients.

4.5 Oropharyngeal SCC

Silent and non-synonymous mutations in each tumour in Oropharyngeal SCC

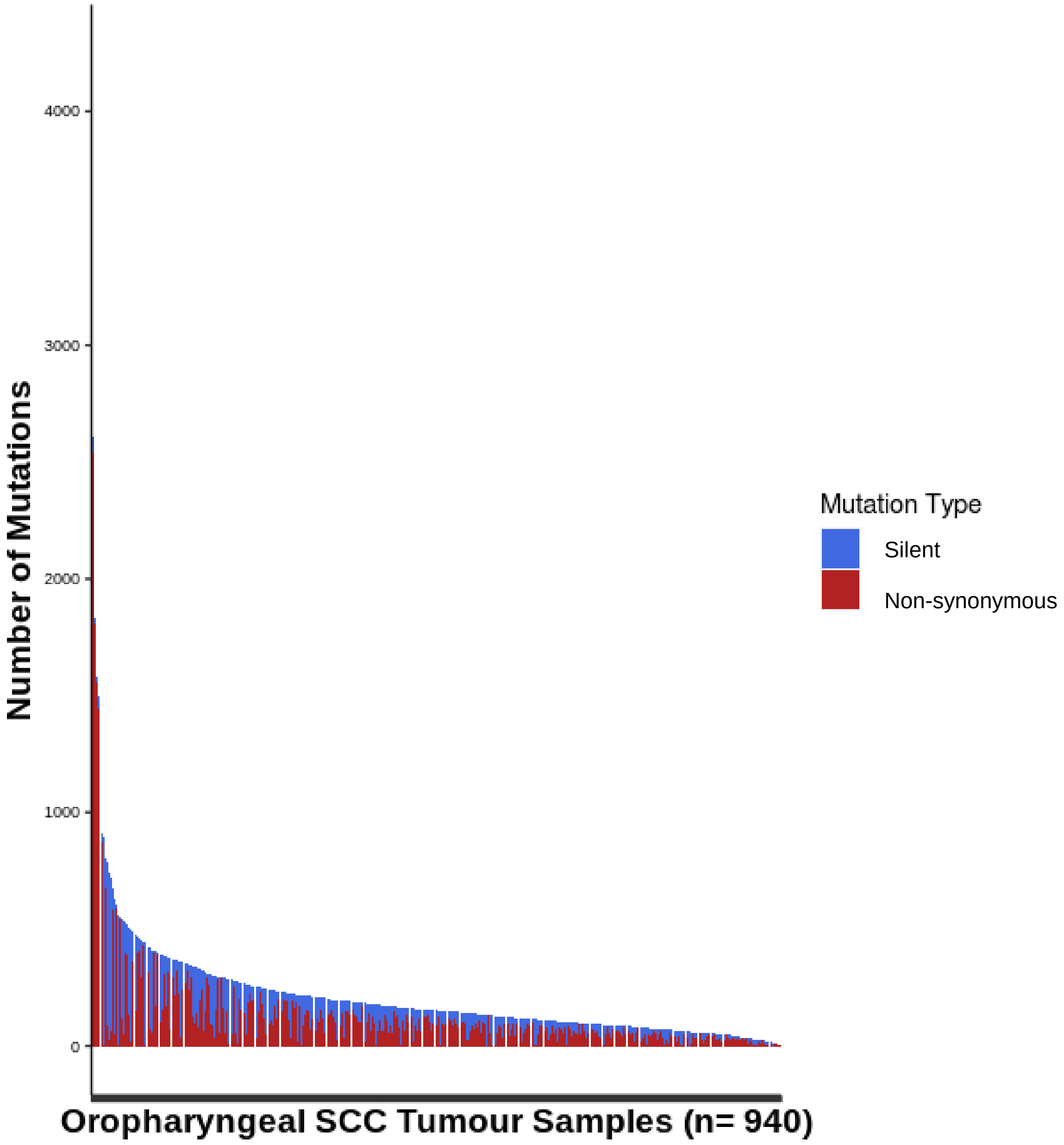


Figure 4-6: The number of synonymous and nonsynonymous mutations in oropharyngeal SCC. The x axis represents each individual tumour identified on COSMIC, GDC portal and literature search.The y axis represents the total number of mutations.

The oropharyngeal SCC tumours showed variation in the proportion of nonsynonymous and silent mutations within each sample (figure 4-6). The majority of oropharyngeal SCC tumour samples had less than 2000 mutations except sample HIPO-HNC1 which had the most (i.e., 4245) mutations. This sample also had the highest proportion of nonsynonymous mutations in the dataset, with 4157 nonsynonymous mutations. There were six samples which had no nonsynonymous mutations and only have synonymous mutations.

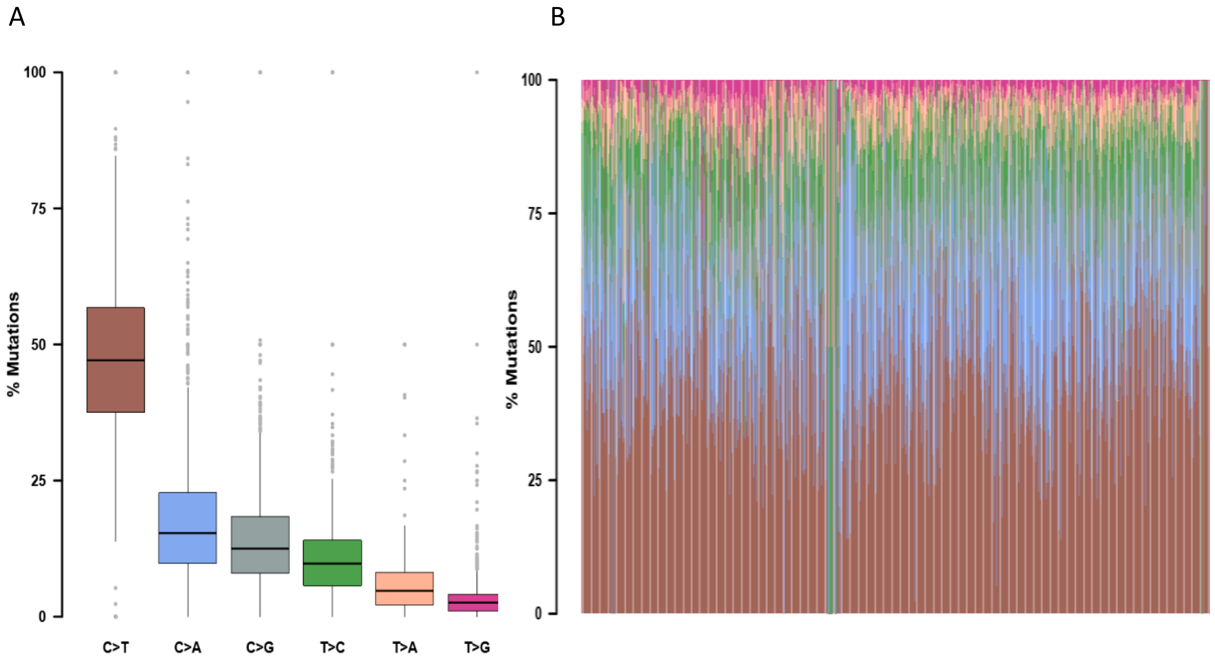


__Figure 4-7: A. The percentage of base changes identified across 940 oropharyngeal SCC samples from COSMIC database, GDC portal and literature search.__ The box and whisker plot illustrates the range of variation represented by the whiskers and the box shows the median and inter-quartile range. The filled circles represent the outliers in the dataset. __B. The relative contributions of the different base changes in each individual oropharyngeal SCC sample.__ The proportion of mutations in each sample is shown as a percentage and the different colours represent the different base changes.

The largest number of base changes in oropharyngeal SCC were C>T mutations (figure 4-7). There were 90,335 C > T changes which was less than skin SCC which had 313,587 C >T changes. The median proportions of the remaining base changes in decreasing order were C>A, C>G, T>C, T>A and T>G changes. The highest interquartile ranges were seen for the C>T and C>A base changes indicating that there was more variation in these base changes in this dataset. The C>A and C>G base changes were positively skewed which implies that the dataset contained a larger range of base changes that were higher rather than lower than the median. The smallest number of base changes observed were T>G.

As expected, there was an amount of variation in the proportions of base changes in each oropharyngeal SCC. Although the highest proportion of base changes in oropharyngeal SCC were C>T, there were some samples which did not have any C>T base changes. The large variation in the percentage mutations that were due to the various base changes indicated that there was high heterogeneity between samples.

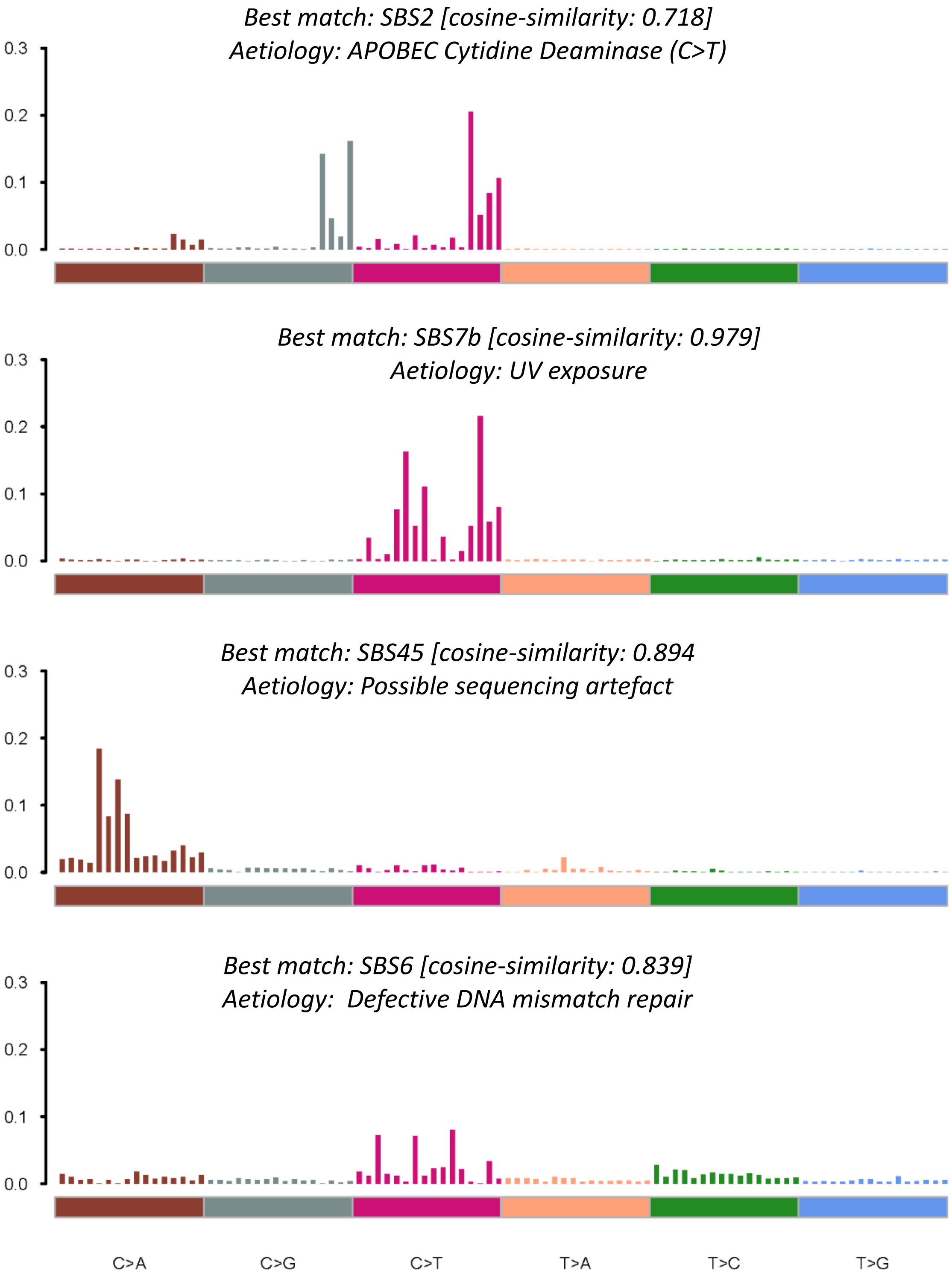


*Figure 4-8: **Single base subsitution mutation signature plot for oropharyngeal SCC.** The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the skin SCC tumour sample cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

Maftools identified four single base mutation signatures associated with oropharyngeal SCC (figure 4-8). One of these was a mutation signature which had a cosine similarity of 0.718 to SBS2, which has been associated with APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like) cytidine deaminases. This enzyme is responsible for converting a cytosine to uracil during RNA editing, however, it can also induce base substitutions in tumour DNA by converting cytosine to uracil, which then becomes converted to thymine during replication of the DNA. APOBECs have mutational specificity for TC motifs therefore the mutation signature plots have peaks at C base changes that are adjacent to T bases (Roberts et al., 2013). Another single base mutation identified was SBS6 which is predicted to be an effect of defective DNA mismatch repair. SBS7b, which is associated with UV exposure, was also noted as a single base mutation signature, however 11 patients seem to have mainly contributed to this mutation signature, which suggested that SBS7b might be an artefact from the analysis or is a signature seen in only a minority of patients. Another mutation signature which had a similarity of 0.894 to SBS45, was identified as a possible sequencing artefact.

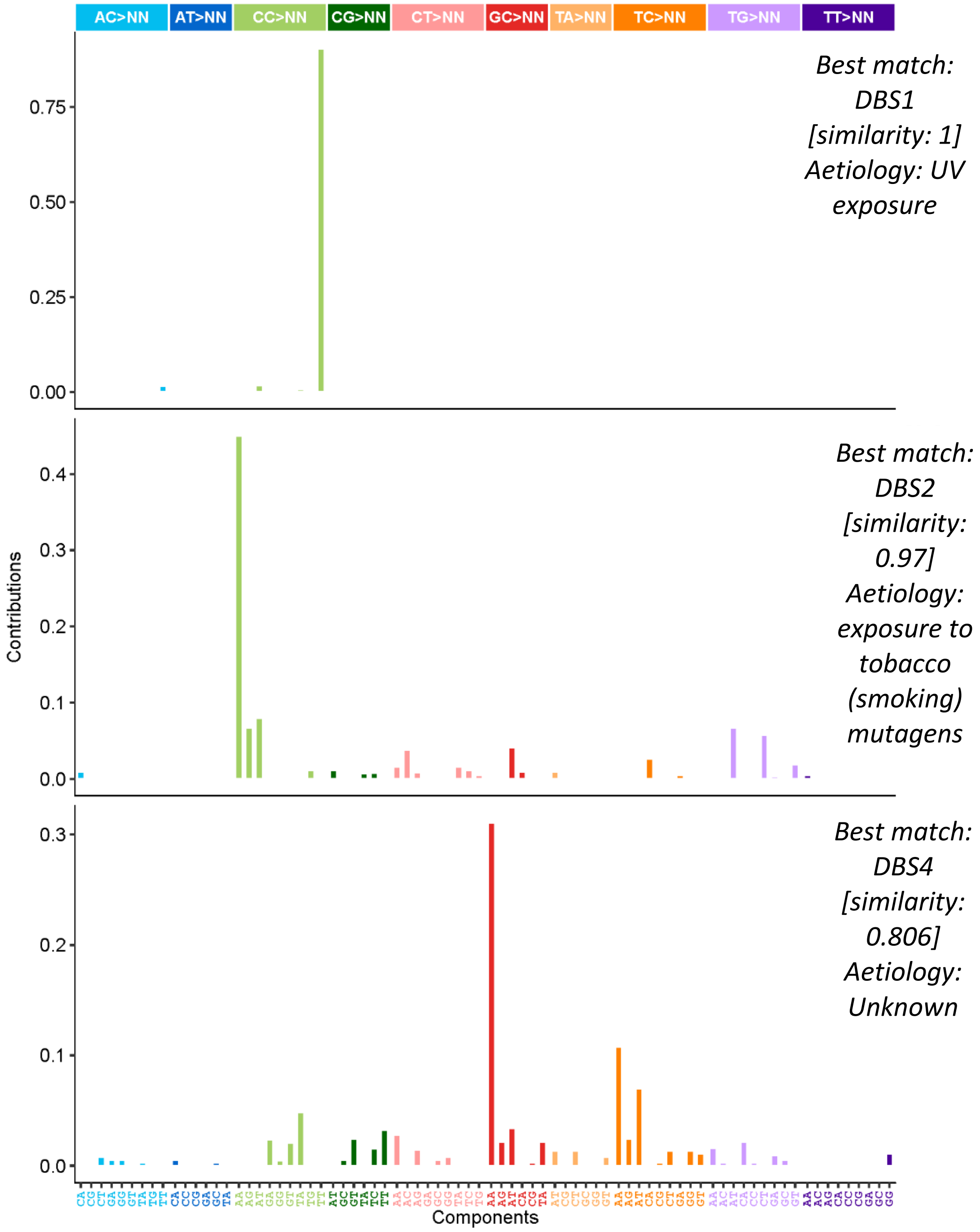


Figure 4-9: The double base substitution mutation signature plots for oropharyngeal SCC. *The x axis represents the base change in its dinucleotide context and the y axis represents the proportion of each base change in the oropharyngeal SCC tumour sample cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.*

The Sigminer program identified a double base substitution signature, DBS1, that is associated with UV in oropharyngeal SCC (figure 4-9). A mutation signature which was 0.97 similar to DBS2 which is associated with smoking was also observed. A DBS4 signature which has an unknown aetiology was also extracted as a signature from the oropharyngeal SCC tumour samples.

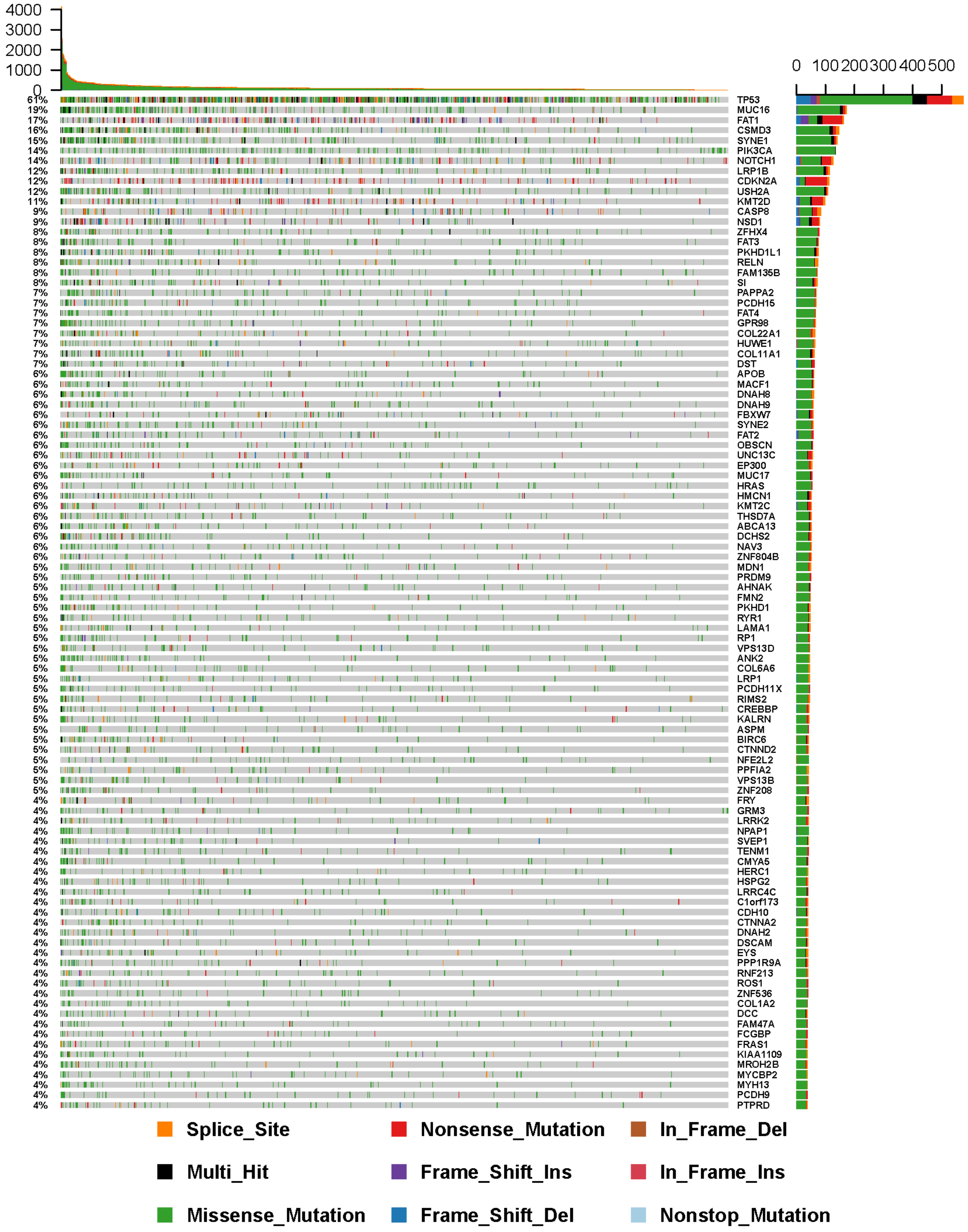


Figure 4-10: Oncoplot with top 100 frequently mutated genes in oropharyngeal SCC tumours from GDC portal, COSMIC database and a literature search. *One or more of these genes was mutated in 888 of 940 samples (94.47%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation that each sample contains within the corresponding gene. The bar chart on the right represents the number of samples which contain a mutation in that gene and the colours represent the type of mutation in the gene.*

An oncoplot showed that 94.47% of oropharyngeal SCC tumours had mutations in one or more of the top 100 frequently mutated genes. However, the proportion of samples which shared mutations in each individual gene varied. A mutation in *TP53* was most common and was shared by 61% of oropharyngeal SCC samples (figure 4-10). The next most frequently mutated gene was *MUC16* but a mutation in this gene was observed in only 19% of oropharyngeal SCCs. The oncoplot also shows that 10 genes had mutations which were shared by 11 – 19% of oropharyngeal SCC samples and that the remaining 89 genes of the top 100 frequently mutated genes had mutations which are shared by less than 11% of oropharyngeal SCCs.

Most mutations in the top 100 frequently mutated genes in oropharyngeal SCCs were missense mutations, however, *FAT1* and *CDKN2A* more commonly contained nonsense mutations than other types of mutations. The number of mutations in each oropharyngeal SCC varied, with most samples had less than 1000 mutations.

4.6 Lung SCC

Synonymous and non-synonymous mutations in each tumour in Lung SCC

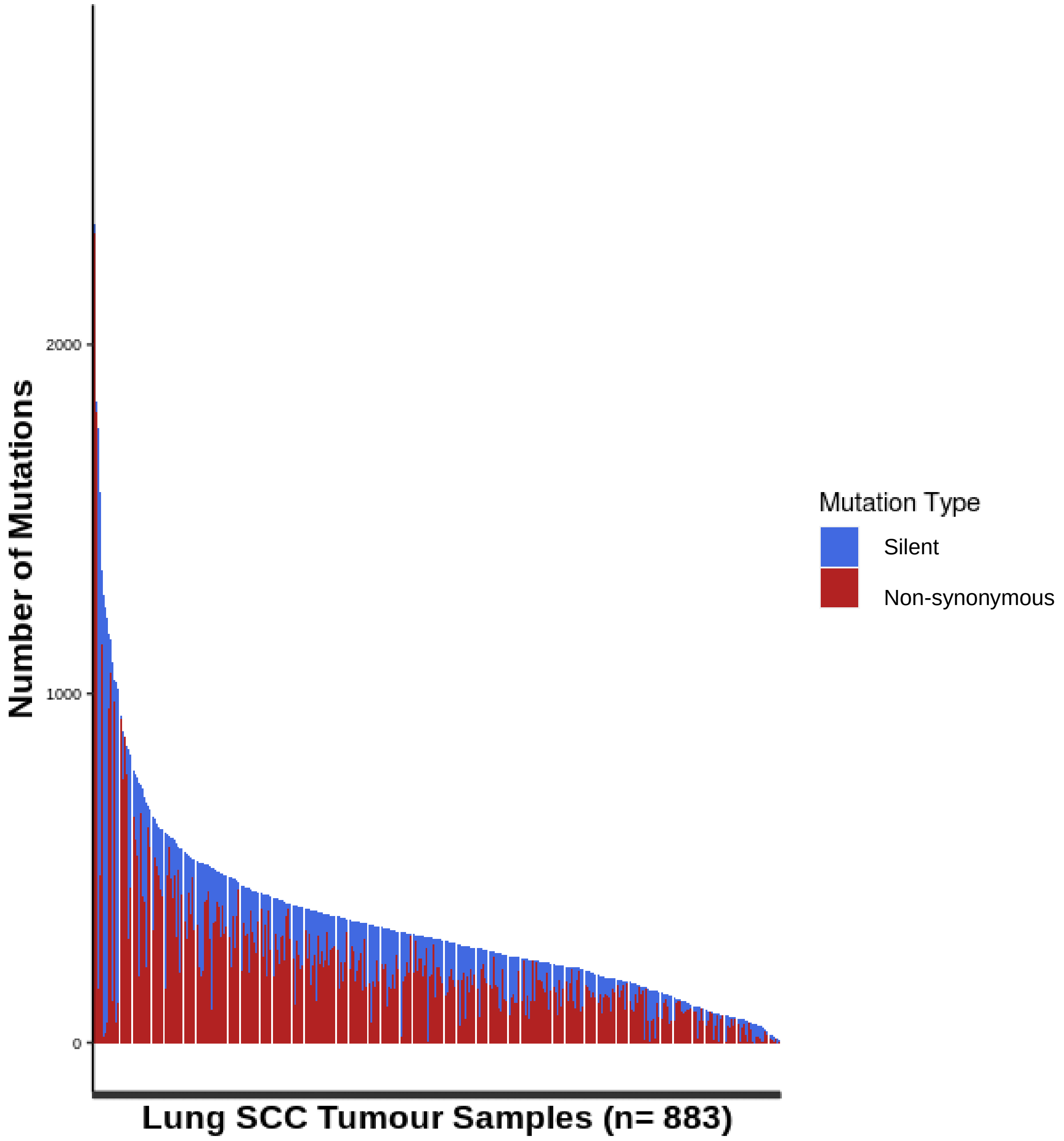


Figure 4-11: The number of silent and nonsynonymous mutations in lung SCC. *The x axis represents each individual tumour identified on COSMIC, GDC portal and the literature search. The y axis represents the total number of mutations.*

The lung SCC tumour samples show that the majority of samples had more nonsynonymous mutations compared to silent mutations in their exomes (figure 4-11). The largest number of mutations were seen in sample WGC665 which had 2832 mutations, with 2627 of those mutations being silent. All the other samples in the lung SCC cohort exhibited less than 2000 mutations, but there was high variation in the number of mutations, including non-synonymous and silent mutation, between samples.

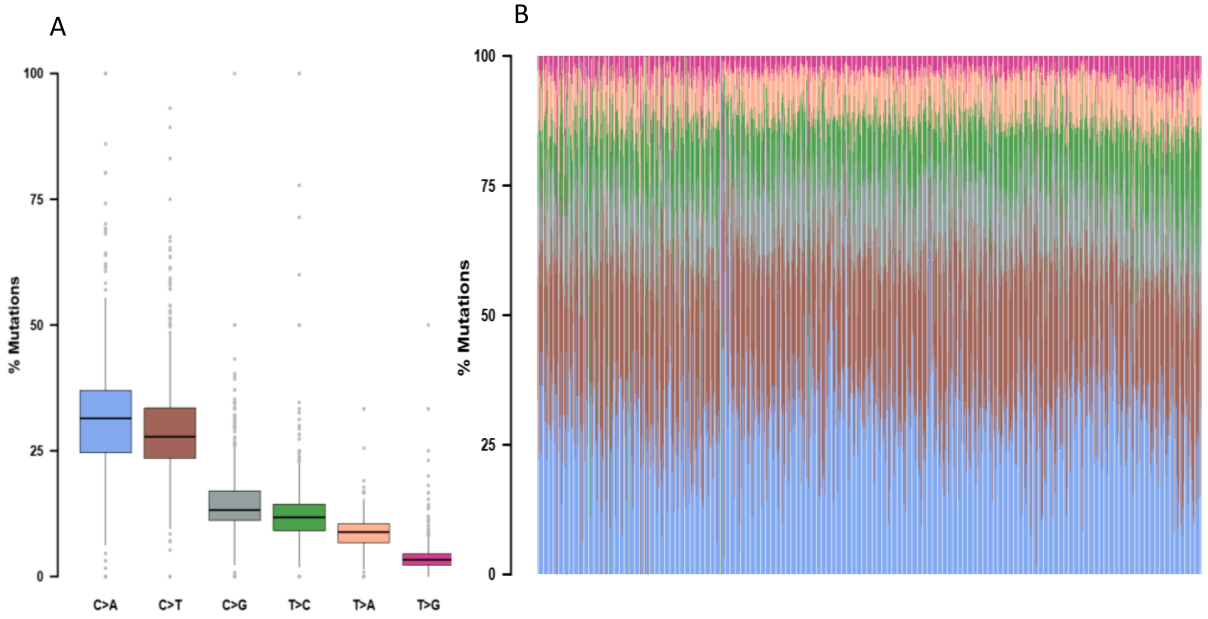


__Figure 4-12: A. A box and whisker plot showing the percentage of different base changes in this cohort of 883 lung SCC tumour samples.__ The different colours represent the individual base changes. __B.The relative contributions of the different base changes in each individual lung SCC sample from COSMIC database, GDC portal and literature search.__ The proportion of mutations in each sample is shown as a percentage of the total mutations and the different colours represent the various base changes.

The highest median number of base changes were C>A, however there were almost as many C>T as there were C>A base changes (figure 4-12). The interquartile range was larger for C>A compared to C>T changes showing that was more variation in C>A changes between samples. The proportion of samples which had C>G, T>C and T>A changes were similar, whereas T>G base change accounted for the lowest proportion of base changes that occurred in the lung SCC samples.

In general, the proportions of mutations within each cancer sample did not show much variation across most samples (figure 4-12B).

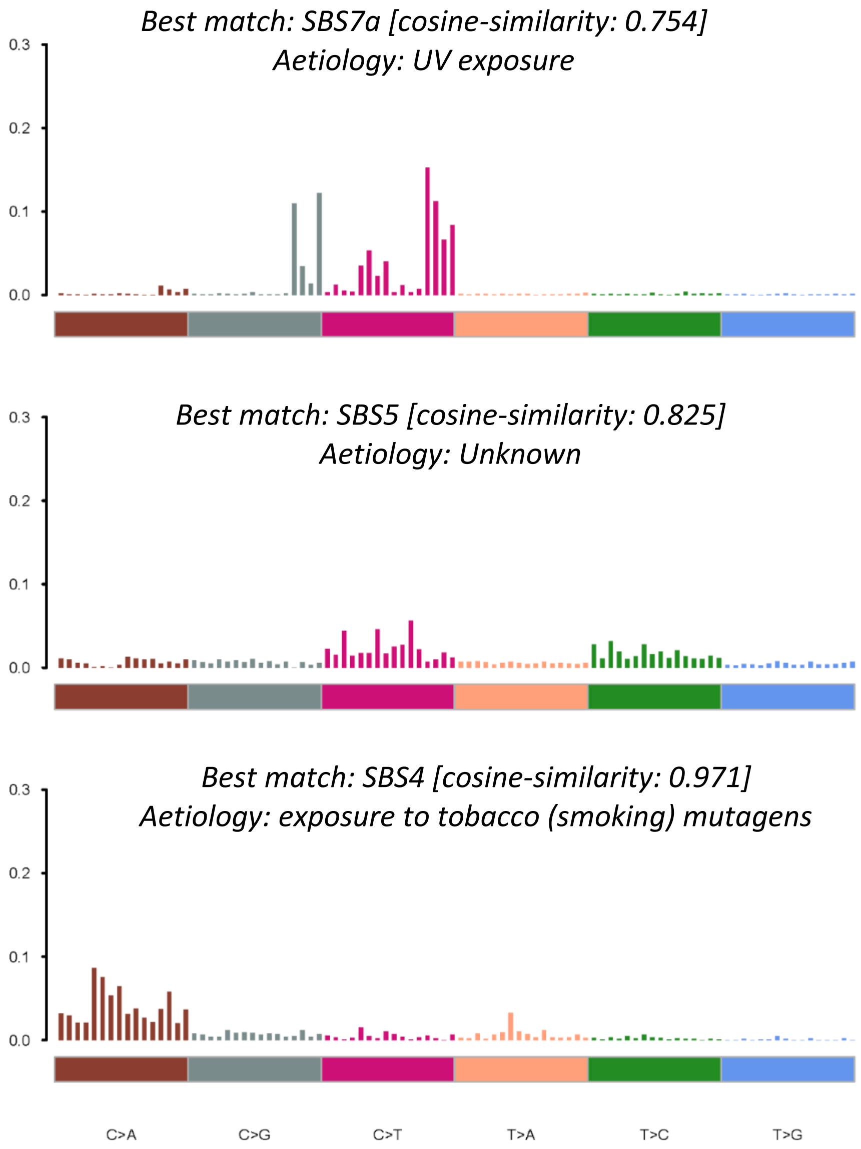


*Figure 4-13: **Single base subsitution mutation signature plot for lung SCC**. The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the lung SCC tumour sample cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

Three single base substitution mutation signatures were identified in the lung SCC cohort using Maftools. The SBS4 mutation signature had a 0.971 cosine similarity to the mutation signature produced (figure 4-13). This signature is associated with tobacco smoking suggesting that these mutational patterns are a direct result of the tobacco carcinogen. Another mutation signature that was detected had 0.825 similarity to SBS5, which has an unknown aetiology. The SBS7a mutation signature associated with UV was identified as having a 0.754 similarity to a mutation signature produced by the lung SCC cohort.

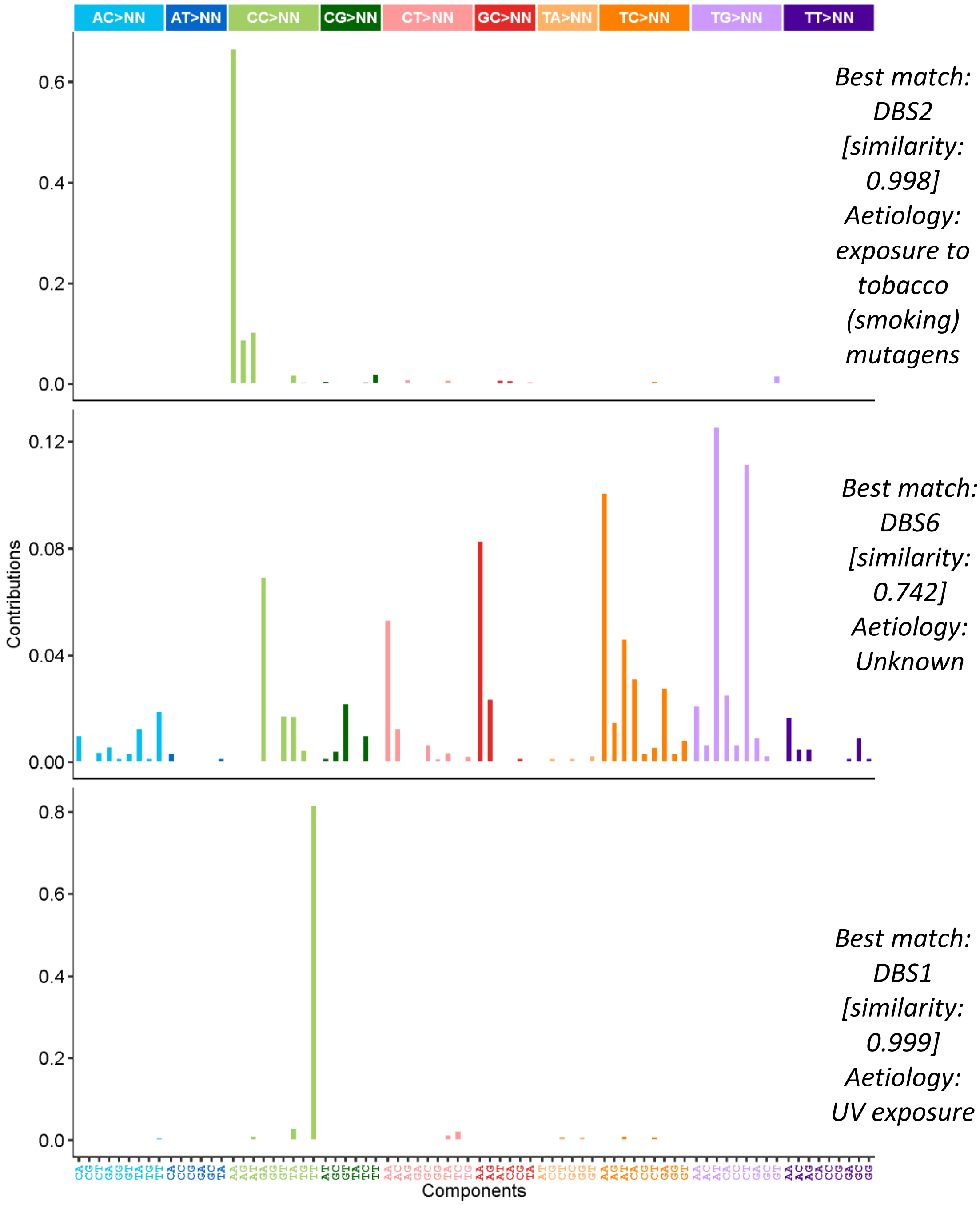


__Figure 4-14: The double base substitution mutation signature plots for lung SCC.__ The x axis represents the base change in its dinucleotide context and the y axis represents the proportion of each base change in the lung SCC tumour sample cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.

Three double base mutation signatures were detected in the lung SCC samples and the DBS2 mutation signature, that is associated with cigarette smoking, had a 0.998 similarity to the mutation signature produced (figure 4-14). The DBS1 mutation signature which is associated with

UV exposure was also identified as a signature in the lung SCC cohort. The analysis of the lung SCC exome data also produced a signature similar to DBS6 (0.742 cosine similarity) however an aetiology for this mutation signature has not yet been identified.

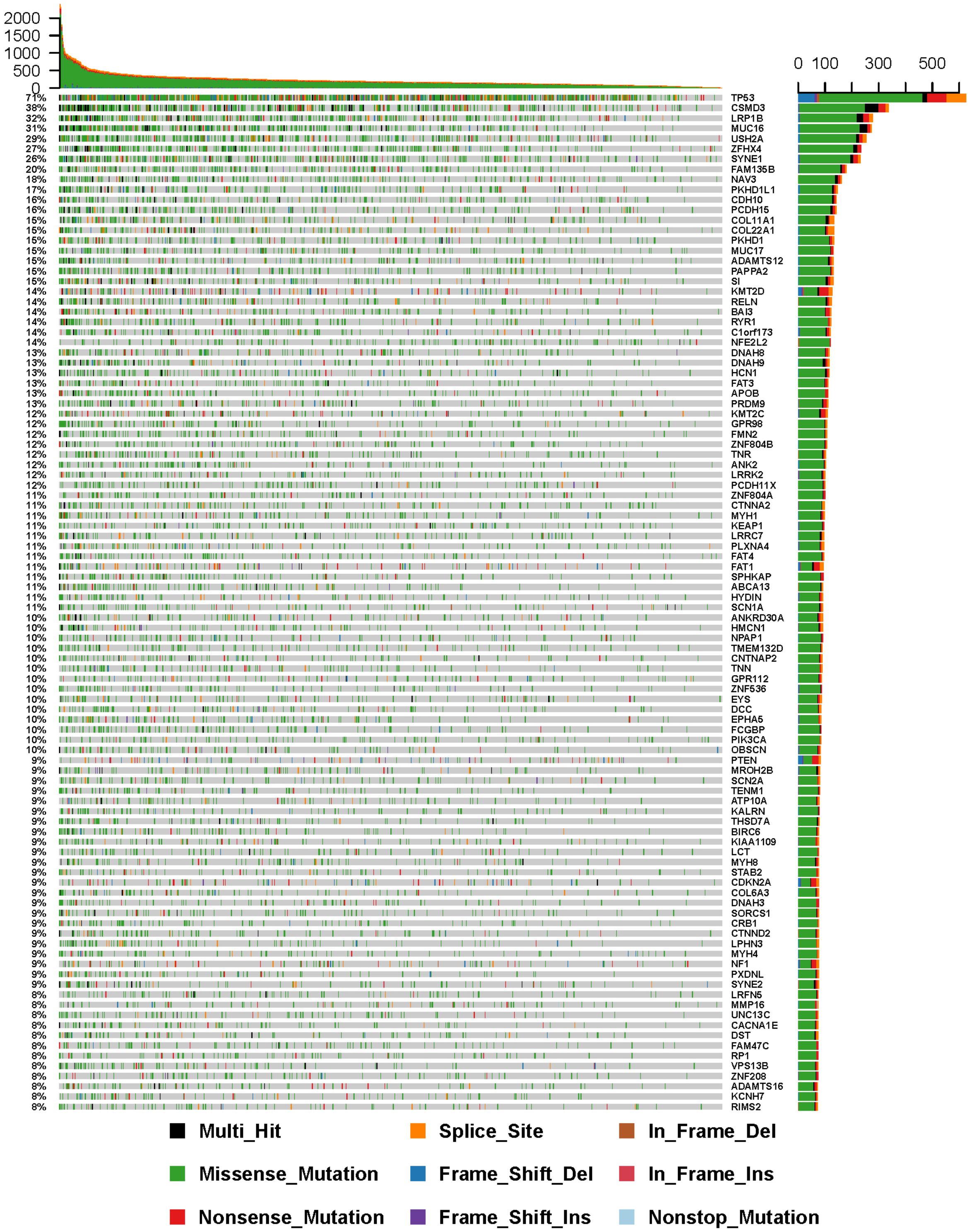


Figure 4-15: Oncoplot with top 100 frequently mutated genes for lung SCC tumours from COSMIC database, GDC portal and the literature search. *One or more of these genes was altered in 843 of 883 (95.47%) samples. The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation that each sample contains within the corresponding gene. The bar chart on the right represents the number of samples with a mutation in that gene and the colours represent the type of mutation within the various genes.*

In lung SCC, 95.47% of samples had a mutation in at least one of the top 100 frequently mutated genes (figure 4-15). The most frequently mutated gene across samples was *TP53* which was mutated in 71% of lung SCCc, with most samples having a missense mutation in the *TP53* gene. The second most frequently mutated gene was *CSMD3*, mutated in 38% of tumour samples. The top eight most frequently mutated genes were mutated in 20% - 71% of tumour samples, with each of the remaining 92 of the top 100 mutated genes exhibiting mutations in less than a fifth of lung SCCs. In general, most mutations in the top 100 most frequently mutated genes were missense, but in some genes, including *KMT2D*, *FAT1*, *PTEN*, *CDKN2A* and *NF1*, a variety of different mutation types including frame shift deletion, frame shift insertion, multi-hit mutations, nonsense and splice site mutations was observed. The number of mutations within in each cSCC sample was less than 1000 mutations in most lung SCC samples.

4.7 Oesophageal SCC

Synonymous and non-synonymous mutations in each tumour in Oesophageal SCC

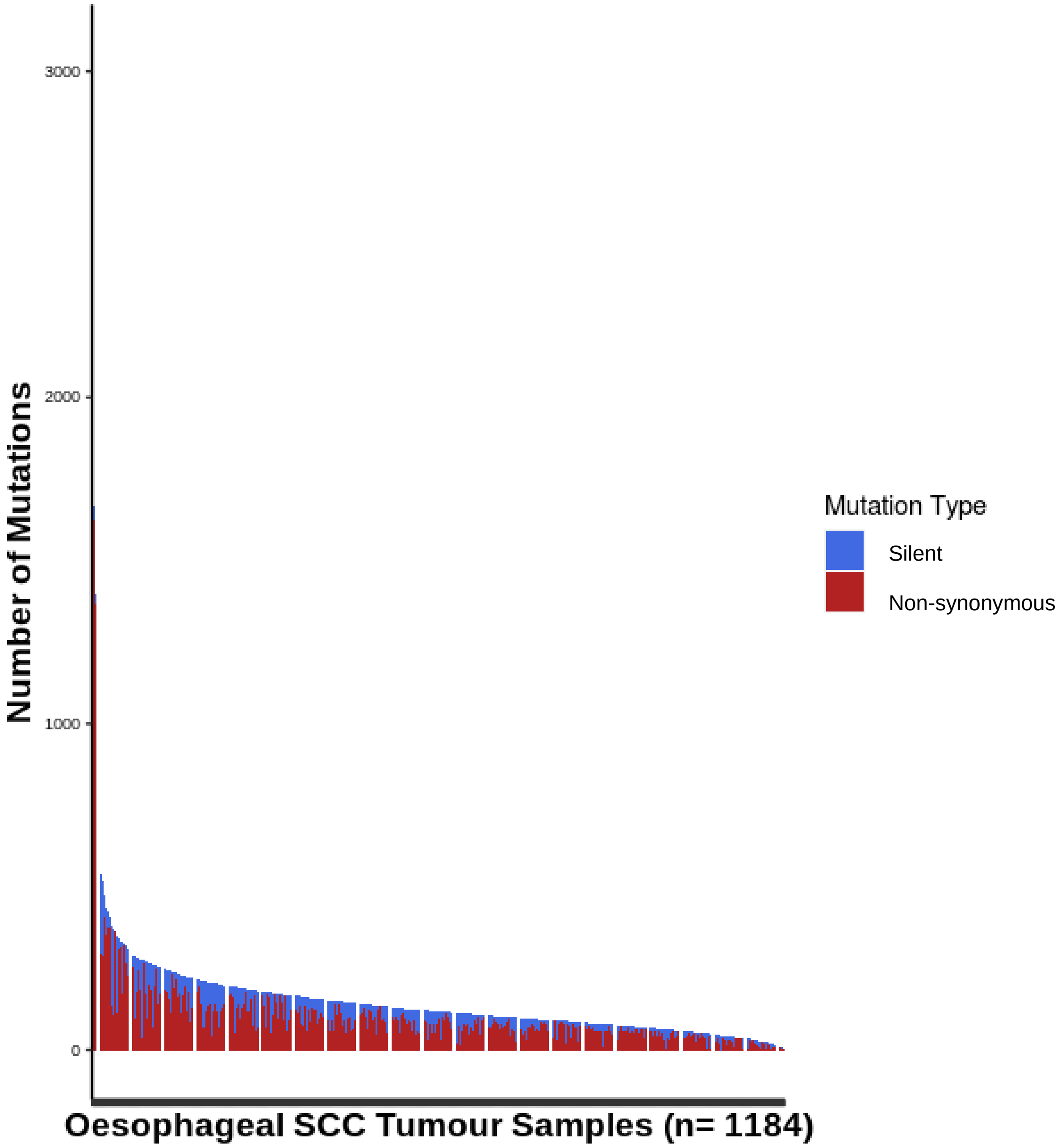


Figure 4-16: The number of synonymous and nonsynonymous mutations in oesophageal SCC. *The x axis represents each individual tumour identified on COSMIC, GDC portal and the literature search. The y axis represents the total number of mutations.*

Most oesophageal SCCs had less than 1000 mutations per tumour, however, as figure 4-16 demonstrates, eight tumour samples had more than 1000 mutations. The sample with the greatest number of mutations (sample ESCC-012T) had a total of 3039 mutations, all of which were nonsynonymous. Five samples had more silent mutations than nonsynonymous mutations.

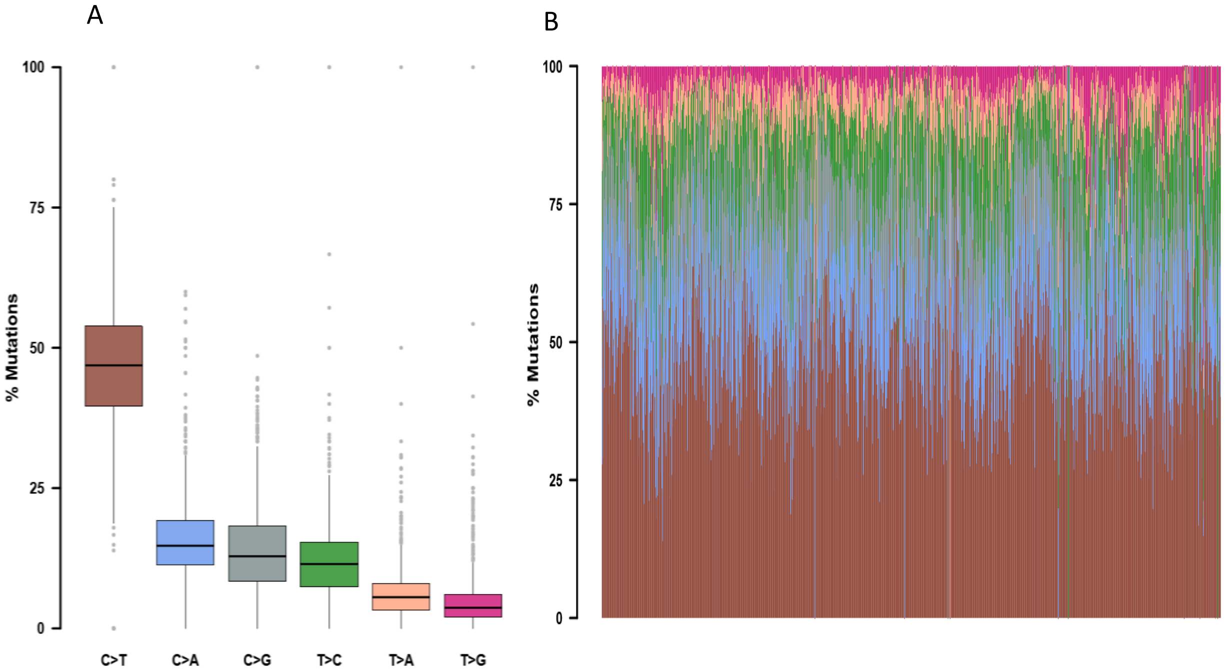


Figure 4-17: ***A. The percentage of base changes identified across all the oesophageal SCC samples from COSMIC database, GDC portal and the literature search.*** *The box and whisker plot shows the range of variation represented by the whiskers and the box shows the median and inter-quartile range. The filled circles represent the outliers in the dataset.* ***B.The relative contributions of the different base changes in each individual oesophageal SCC sample.*** *The proportion of mutations in each sample is depicted as a percentage and the different colours represent the different base changes.*

The highest number of base changes in oesophageal SCC was C>T changes and there were 77,051 of this base change but it also had the largest interquartile range in all oesophageal SCC samples (figure 4-17). The median proportions of C>A, C>G and T>C changes were lower than the C>T alterations but were similar to each other and each of these usually accounted for less than 25% of mutations. The lowest proportion of base changes in the oesophageal SCC samples was T>G changes and, while this change also showed the lowest interquartile range, there were several oesophageal SCCs which are outliers in terms of their proportions of T>G changes.

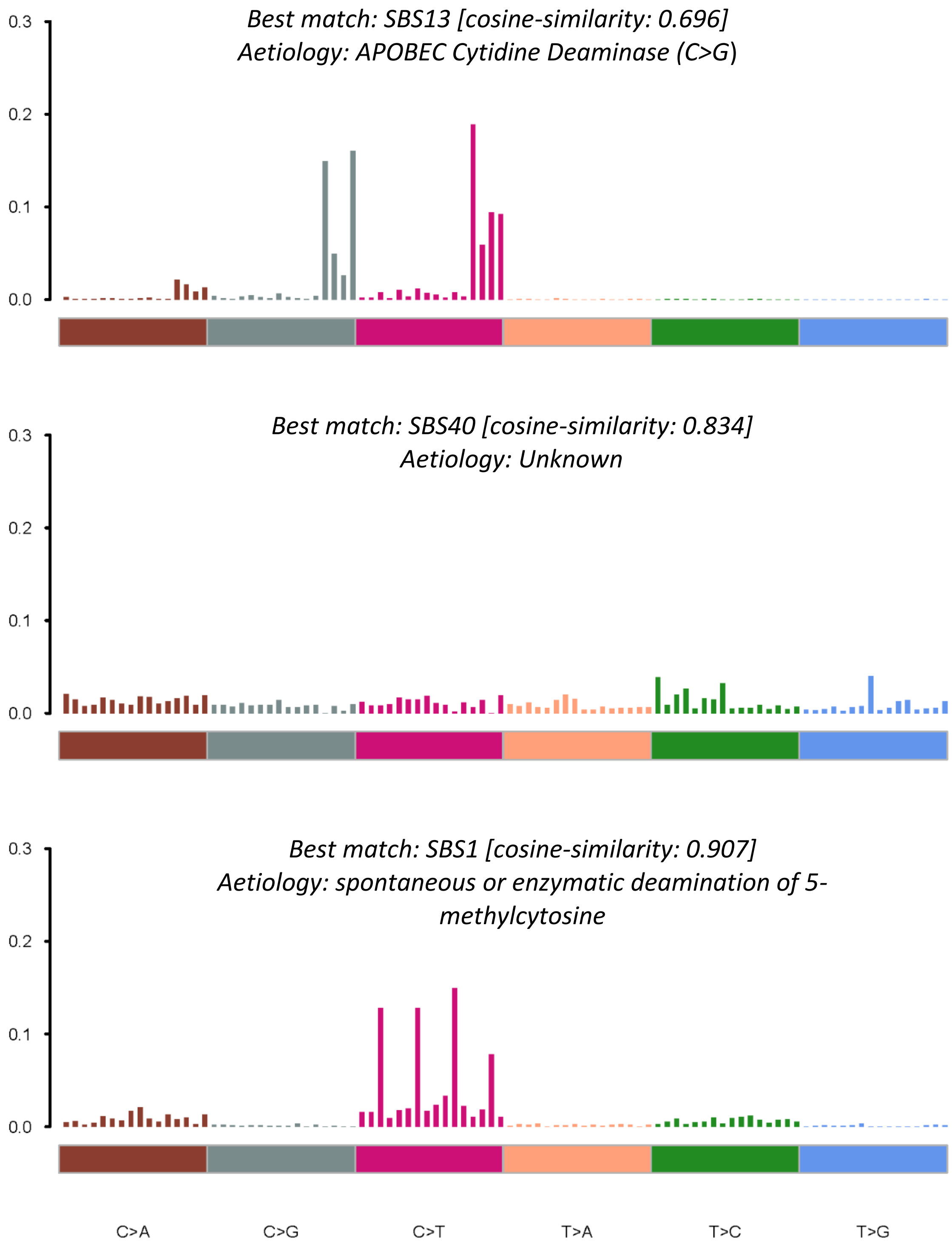
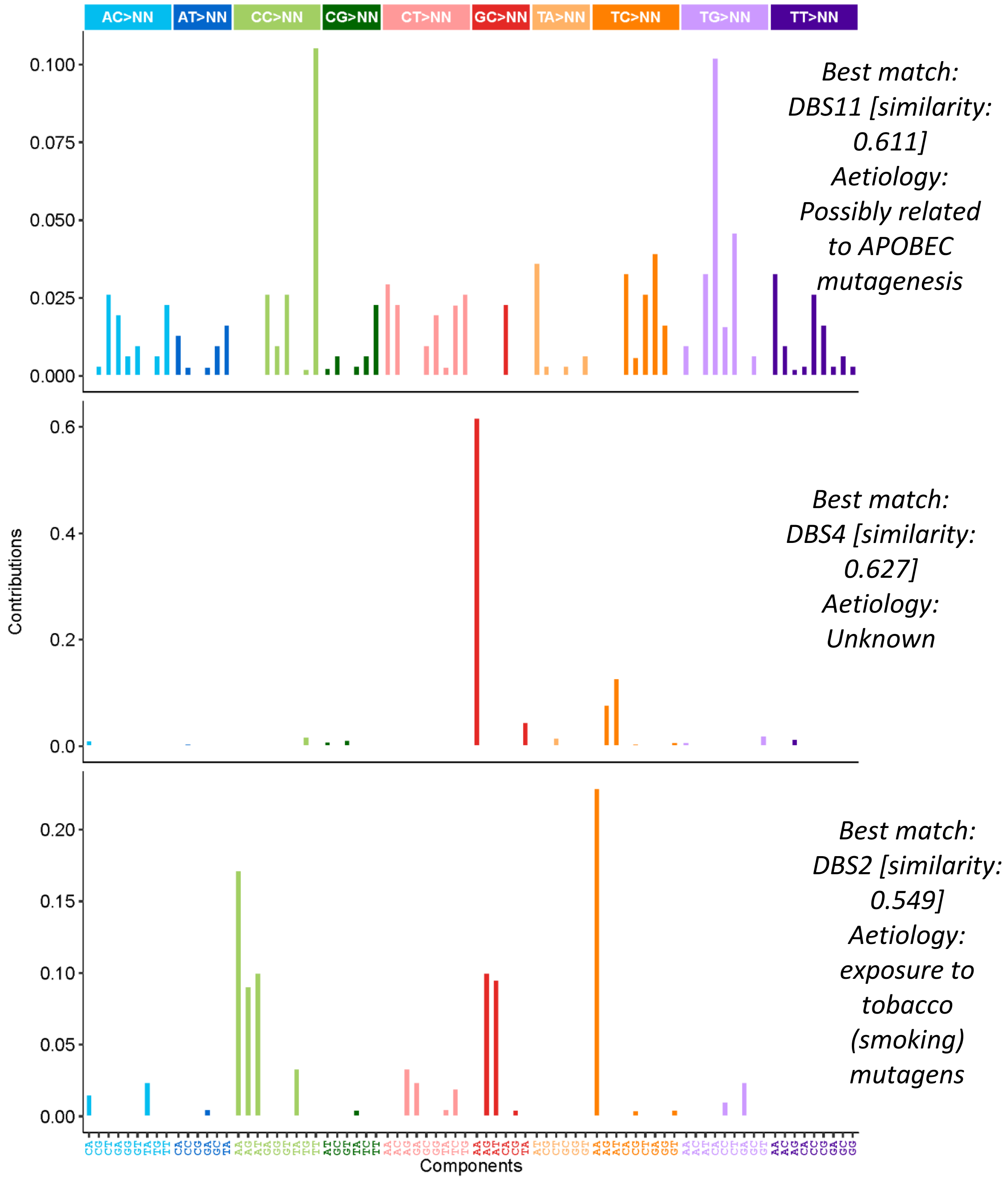


Figure 4-18: Single base subsitution mutation signature plot for oesophageal SCC. *The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the oesophageal SCC cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

The oesophageal SCC tumour samples showed three single base mutation signatures (figure 4-18). One of these mutation signatures had a 0.907 cosine similarity to SBS1, which is associated with spontaneous or enzymatic deamination of 5-methylcytosine to thymine. This deamination results in DNA mismatches in double stranded DNA resulting in guanine bases pairing with thymine. When this DNA mismatch is not replaced before DNA replication, the mismatch remains in the DNA and is replicated. Another single base mutation signature produced from the oesophageal SCC samples had a 0.834 cosine similarity to SBS40, which has an unknown aetiology. SBS13 was another single base substitution mutation signature that was identified and has a similar aetiology to SBS2 (which was present in oropharyngeal SCC (figure 4-8)). However, the APOBEC enzyme functions in a different way in SBS13 and results in a high proportion of a C to G base change. The base change is produced as a result of the error prone polymerases which are generated by base excision remove of the uracil base in RNA (Alexandrov et al., 2020).



Best match: DBS11 [similarity: 0.611] Aetiology: Possibly related to APOBEC mutagenesis

Best match: DBS4 [similarity: 0.627] Aetiology: Unknown

Best match: DBS2 [similarity: 0.549] Aetiology: exposure to tobacco (smoking) mutagens

***Figure 4-19: The double base substitution mutation signature plots for oesophageal SCC.** The x axis represents the base change in its dinucleotide context and the y axis shows the proportion of each base change in the oesophageal SCC cohort.The colour of the bars denotes the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.*

Sigminer identified three double strand mutation signatures in the oesophageal SCC samples (figure 4-19). DBS11 is associated with SBS13 (Alexandrov et al., 2020) and is predicted to have this mutational pattern due to APOBEC mutagenesis. A mutation signature with a 0.627 similarity with DBS4 was identified, however, it is not associated with a known aetiology. The DBS2 mutation signature associated with smoking was also identified in the oesophageal SCC cohort.

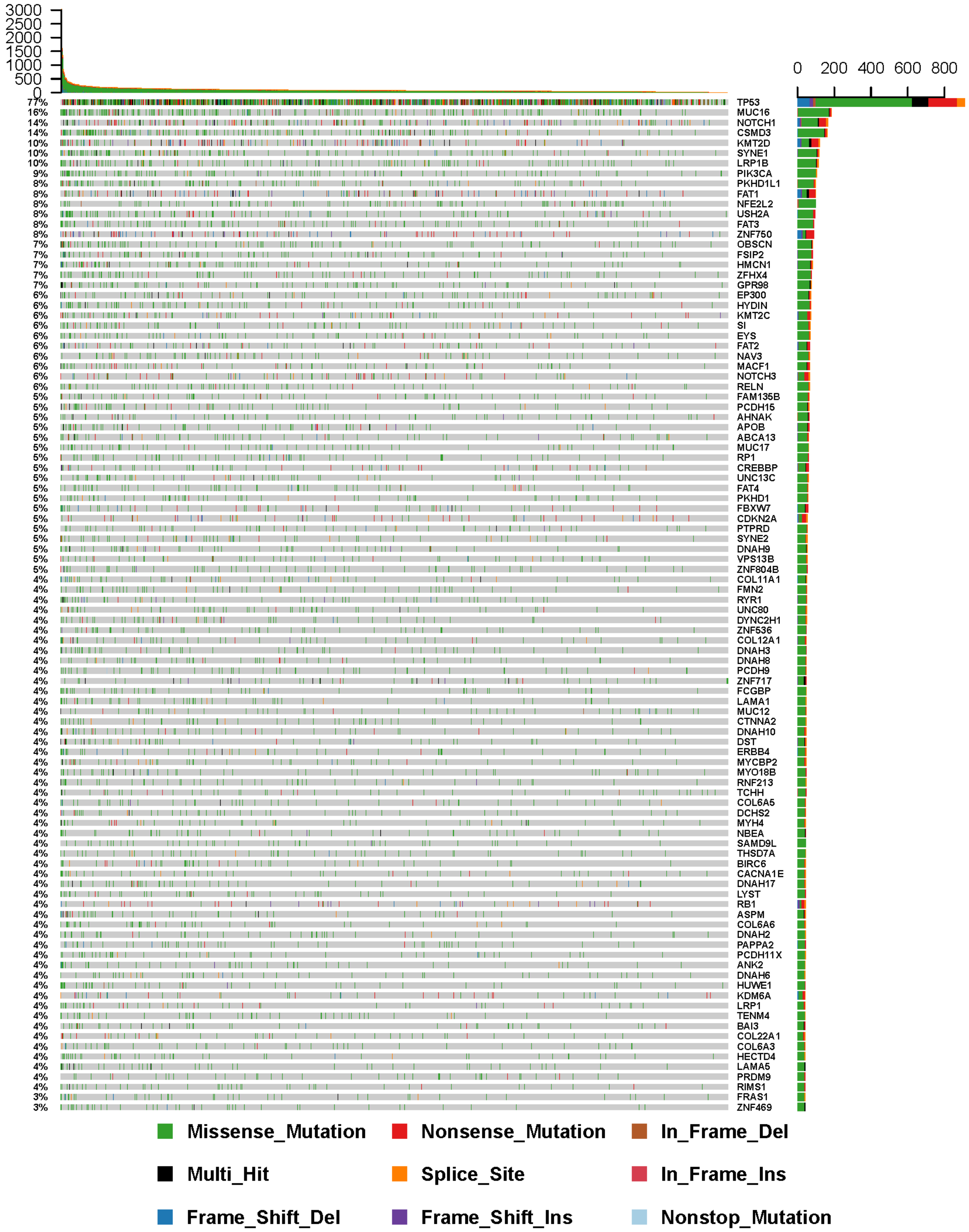


Figure 4-20: Oncoplot with top 100 frequently mutated genes for each oesophageal SCC tumours from COSMIC database, GDC portal and the literature search. *At least one of these genes was mutated in 1141 of 1184 samples (96.37%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation which each sample contains in the corresponding gene. The bar chart on the right represents the number of samples with a mutation in that gene and the colours represent the type of mutation within the gene.*

In oesophageal SCC, 96.37% of samples had mutations in one or more genes that were in the top 100 most frequently mutated genes. *TP53* was the most frequently mutated gene, affecting 77% of samples, however, the second most frequently mutated gene was *MUC16* which was mutated in only 16% of oesophageal SCCs (figure 4-20). As most of the 100 most frequently mutated genes were mutated in <16% of tumours, this suggested that most oesophageal SCCs might share a different combination of mutated genes. Most mutations in the top 100 most frequently genes were missense mutations except in the genes *KMT2D*, *ZNF750*, *CDKN2A* and *RB1*. Most samples with mutations in *KMT2D*, *FAT1* and *ZNF750* were nonsense mutations.

4.8 Cervical SCC

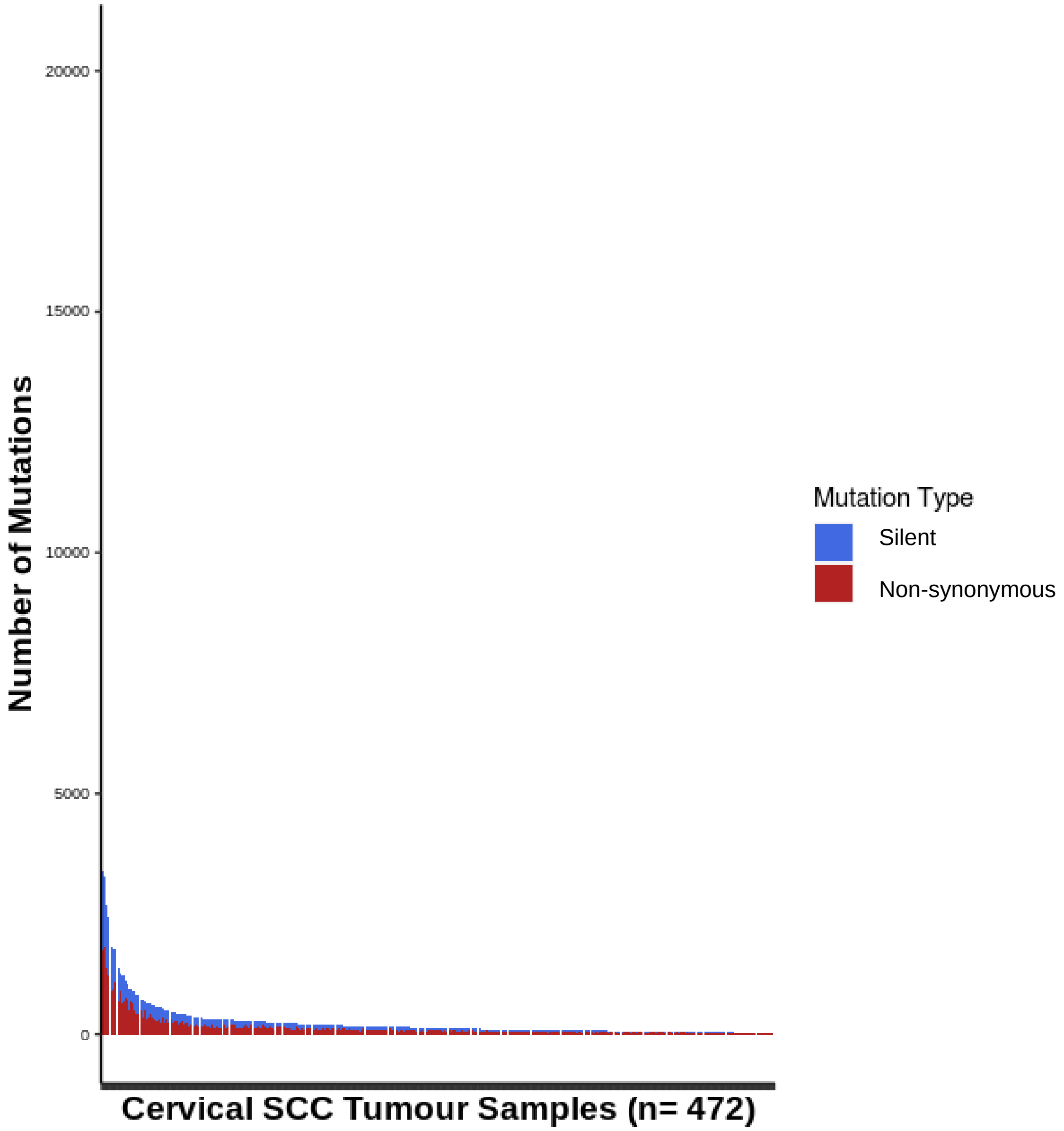


Figure 4-21: A bar graph showing the total number of mutations in each cervical SCC tumour sample. *The blue coloured bars in the stacked bar chart represent silent mutations and the red coloured bars represent non-synonymous mutations.*

Most cervical SCC samples had less than 2000 mutations. However, one sample (TCGA-2W-A8YY-01) had 10691 nonsynonymous mutations and 9662 silent mutations, which was the highest number of mutations in a single tumour in the cervical SCC cohort. Conversely, there was one sample (SGCX-NOR-040_T) which had no nonsynonymous mutations and only one synonymous mutation.

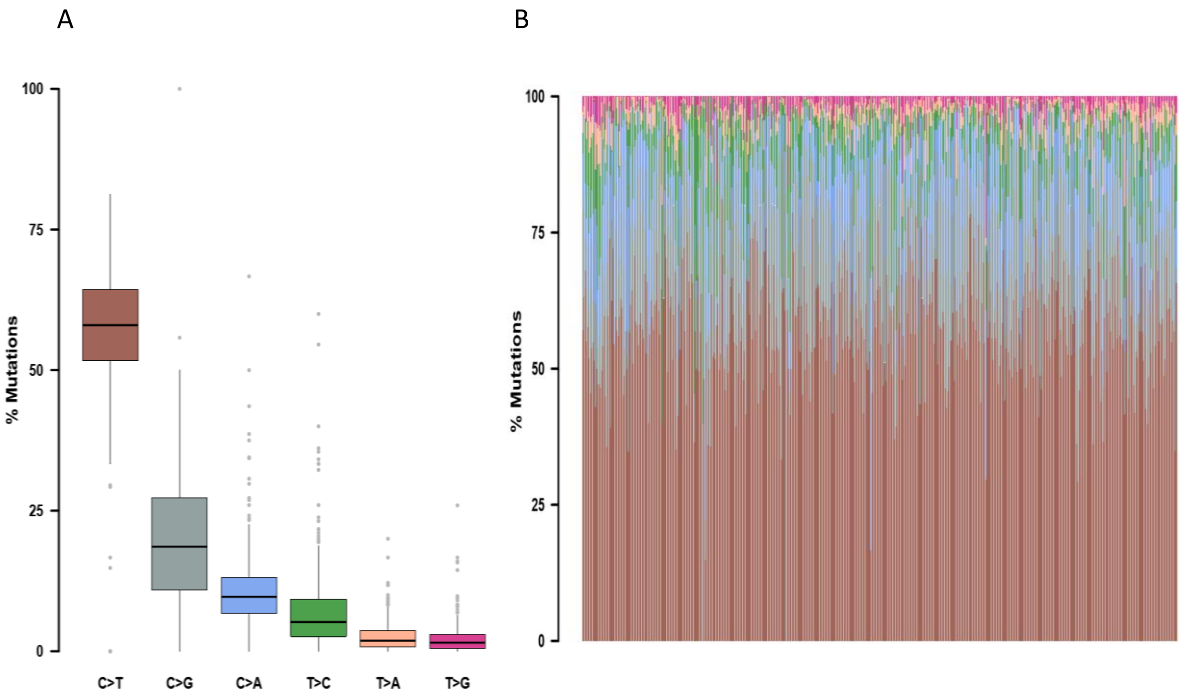


Figure 4-22: A. A box and whisker plot showing the percentage of base changes in this cohort of 472 cervical SCC tumour samples. *The different colours represent the different base changes and the boxes shows the median and inter-quartile ranges, whereas the whisker plot indicates the range.* ***B.The relative contributions of the different base changes in each individual cervical SCC sample.*** *The proportion of mutations in each sample is depicted as a percentage and the different colours represent the different base changes.*

The highest number of base changes in cervical SCC were C>T changes and there were 76,634 C>T base changes (figure 4-22). C>G changes were the next most frequent base change, with lower frequencies of C>A, T>G, T>A and T>G in descending order. In general, the proportion of the different base changes in each cervical SCC sample did not vary appreciably across samples, although there were four samples which had lower proportions of C>T changes and one sample with no C>T changes (figure 4-22B).

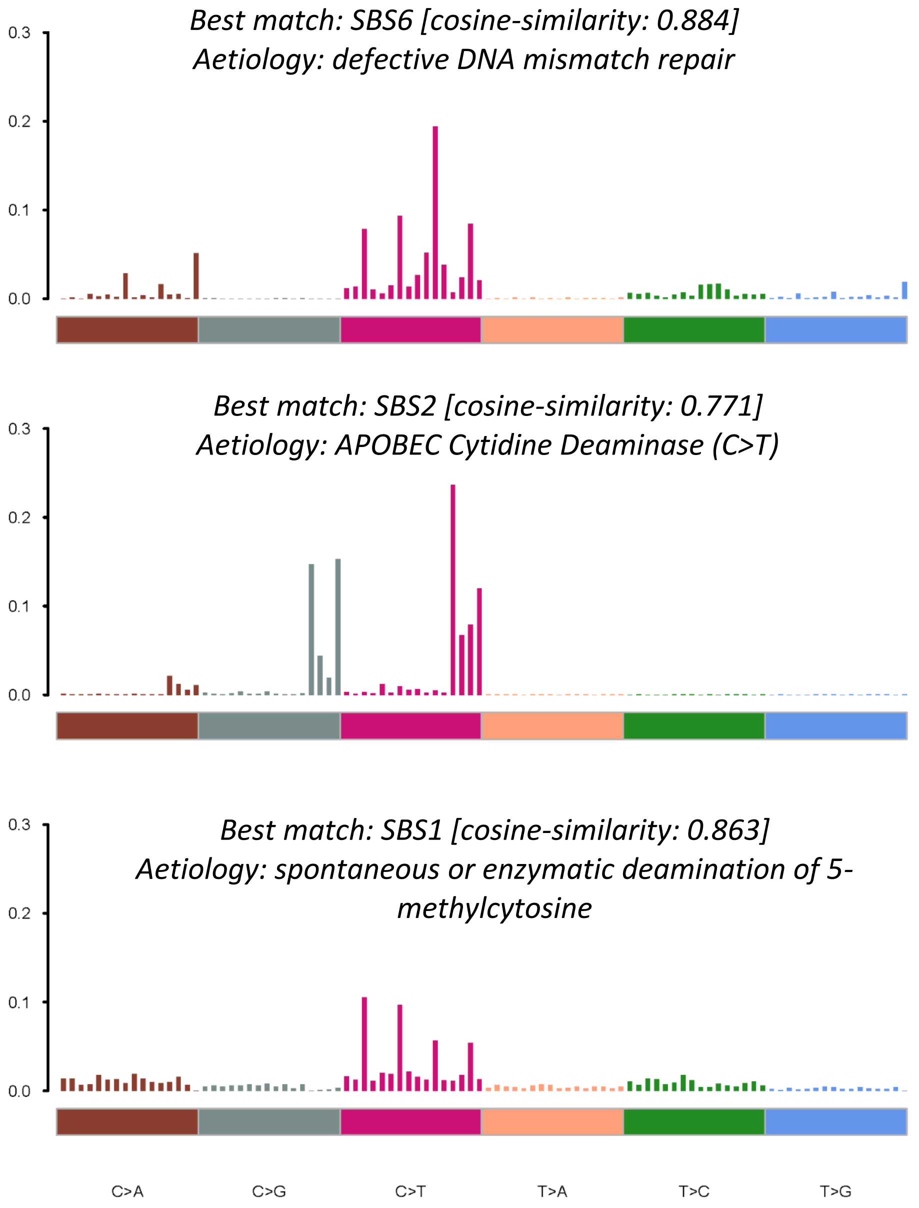


Figure 4-23: Single base subsitution mutation signature plot for cervical SCC. *The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the cervical SCC tumour sample cohort. The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

Maftools identified three single base substitution mutation signatures in the cervical SCC tumour samples (figure 4-23). One of these mutation signatures had a cosine similarity of 0.863 to SBS1, which was also identified in oesophageal SCC, is associated with deamination of 5-methylcytosine resulting in a higher proportion of samples with cytosine to thymine base change compared to other base changes. The SBS2 mutation signature, that had been identified in oropharyngeal SCC cohort, was additionally identified in the cervical SCC cohort with a cosine-similarity of 0.771. The third single base substitution mutation signature seen was 0.884 similar to SBS6, which has been associated with DNA mismatch repair according to previous research conducted in C. elegans (Meier et al., 2018).

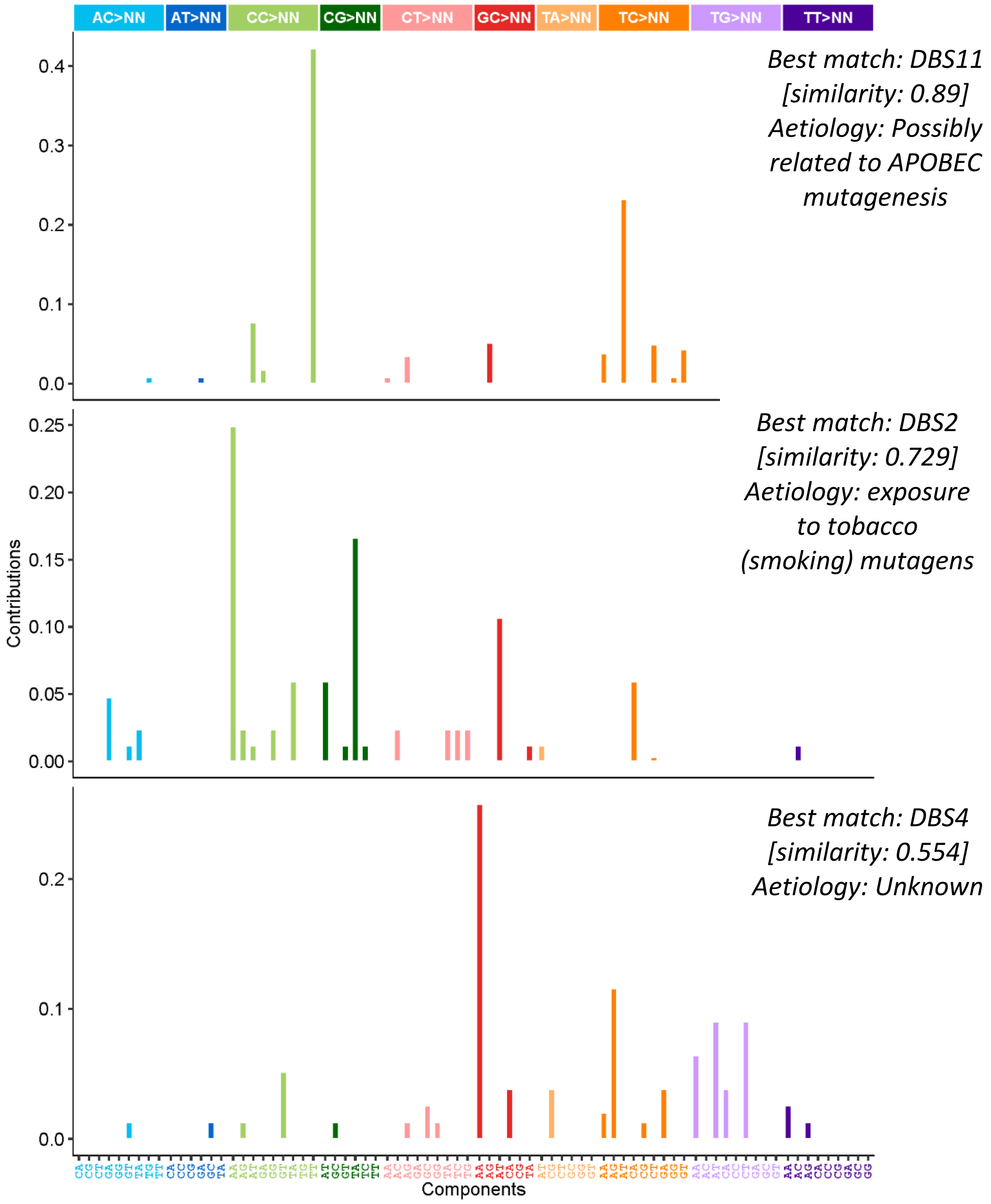


__Figure 4-24: The double base substitution mutation signature plots for cervical SCC.__ The x axis shows the base change in its dinucleotide context and the y axis demonstrates the proportion of each base change in the cervical SCC tumour cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.

The Sigminer program was used to identify double base substitution mutation signatures and showed three double base substitution mutation signatures in cervical SCCs (figure 4-24). DBS11, which is associated with APOBEC mutagenesis, was identified as being 0.89 similar to a mutation signature produced in cervical SCC samples. The DBS2 signature, which is associated with tobacco smoking, was also shown to be present in the exomes of these tumour samples. The third signature categorised in the cervical SCC analysis by Sigminer was 0.554 similar to the DBS4 mutation signature, which has no known aetiology.

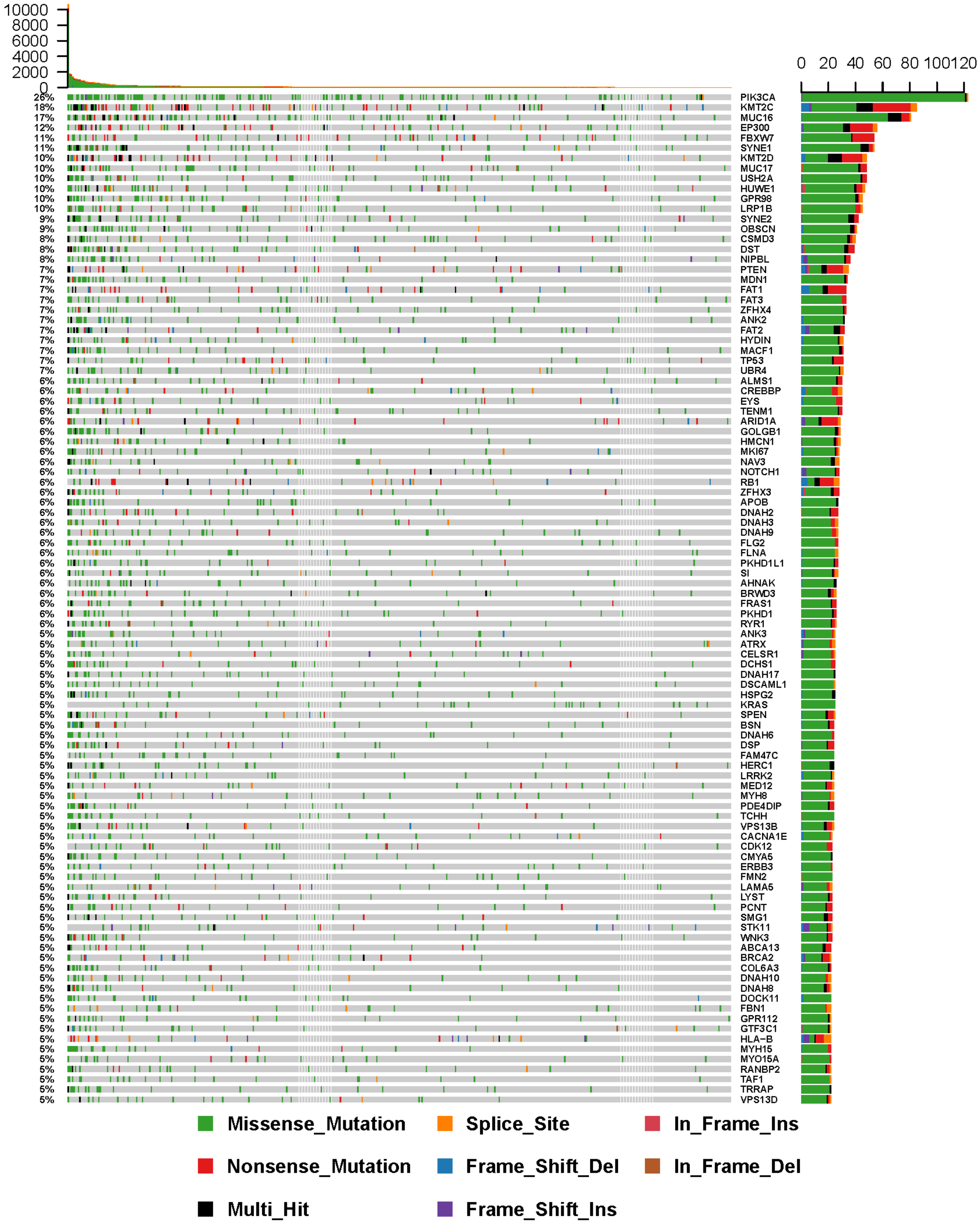


Figure 4-25: Oncoplot of top 100 frequently mutated genes in cervical SCC tumours identifed in GDC portal, a literature search and COSMIC database. *One or more of these genes was altered in 456 of 472 samples (96.61%). The number of mutations identified in each tumour is shown as a bar chart at the top of the figure. Each coloured square in the oncoplot represents the type of mutation in the corresponding gene. The bar chart on the right represents the number samples which have a mutation in that gene and the colours represent the type of mutation in the gene.*

In cervical SCC tumour samples, the most frequently mutated gene was *PIK3CA*, which was mutated in 28% of samples (figure 4-25). The majority of these *PIK3CA* mutations were missense but a minority of samples had a multi-hit or splice site mutation. Within the top 100 most frequently mutated genes, one or more of these genes were mutated in 96.61% of samples. The most common mutation type seen in the samples was missense except in *KMT2C*, *KMT2D*, *PTEN*, *FAT1*, *ARID1A*, *RB1* and *HLA-B*, in which a large proportion of samples contained nonsense or splice site mutations.

4.9 Driver genes and comparison

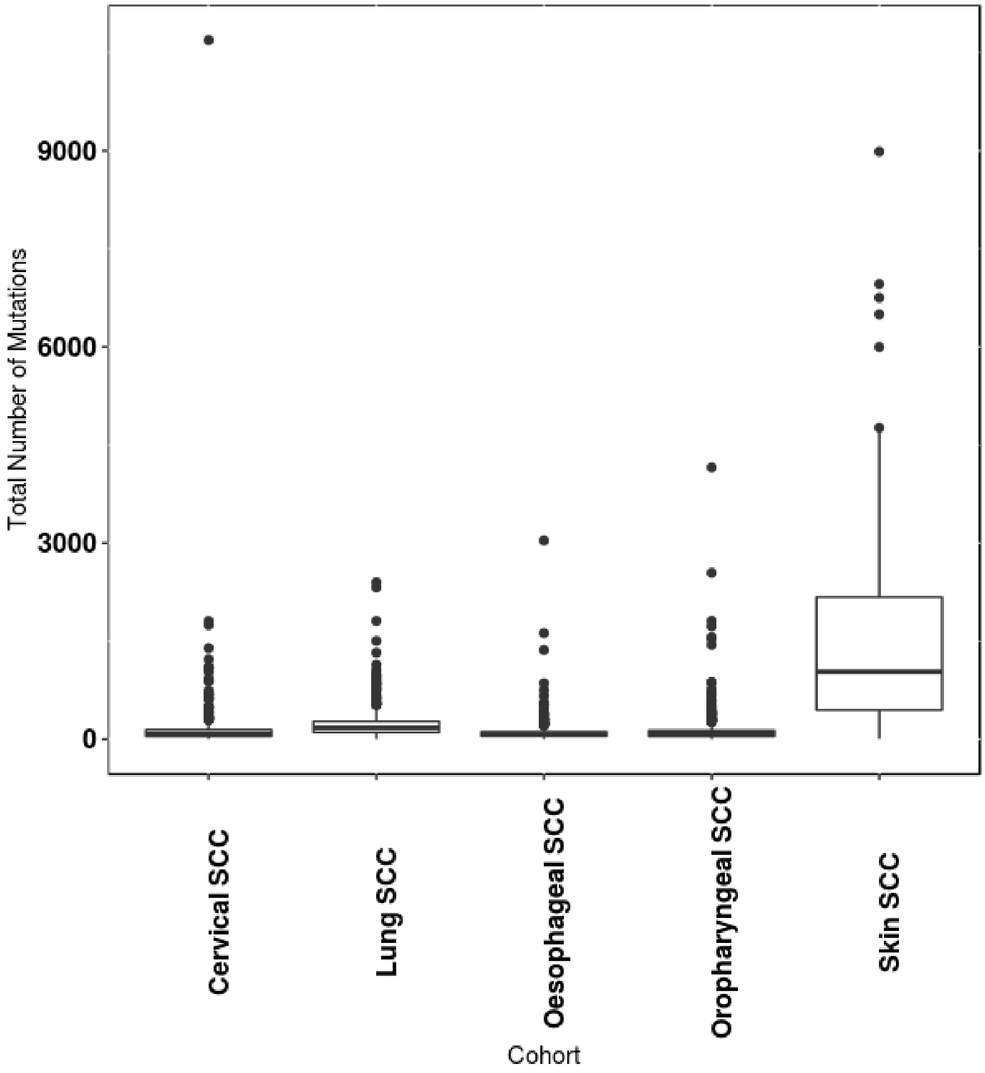


Figure 4-26: Box plot of the mutation burden of SCCs from different organ sites. *Each point represents an individual sample. The x axis shows the name of the SCC cohort and the y axis represents the number of mutations.*

The scatter plot in figure 4-26 shows that the mutation burden of skin SCC is the highest, and with the largest range, compared to the SCCs at the other organ sites. While most cervical SCCs, lung SCCs, oesophageal SCCs and oropharyngeal SCCs have less than 3000 mutations per tumour exome, in skin SCC samples this can range from 0 to 9000 mutations.

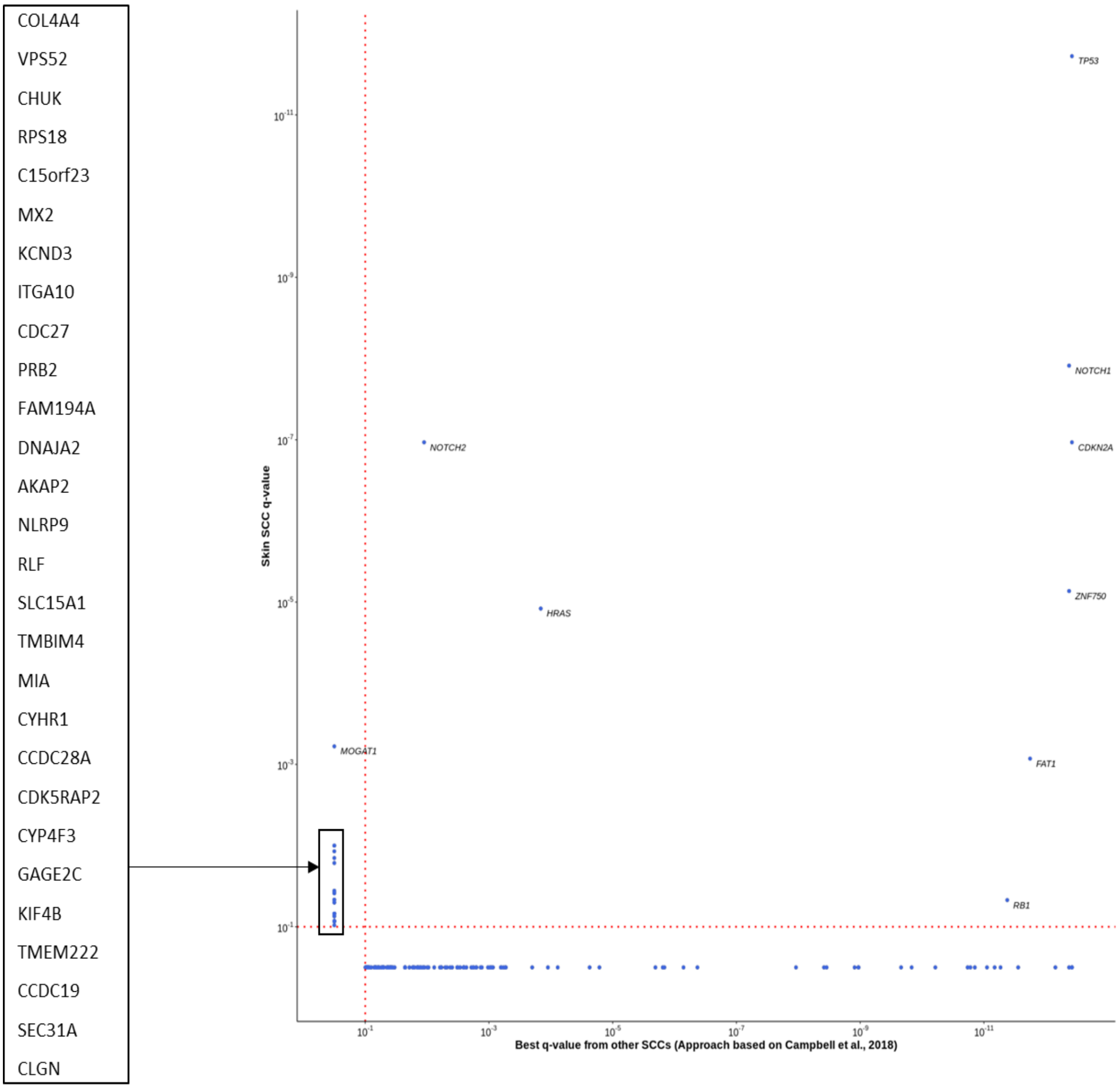


Figure 4-27: Scatter plot which shows the best q values of genes identified as significantly mutated using the Mutsig2CV analysis comparing skin SCC with SCCs at four other internal sites. *The x axis shows genes which were identified as significantly mutated from the MutSigCV analysis in all four internal organ SCCs, comprising oropharyngeal SCC, oesophageal SCC, lung SCC and cervical SCC. The y axis shows the q values for these genes identified in the skin SCC MutSigCV analysis.*

MutSig2CV analysis has been used to identify possible driver genes in cancer, therefore MutSig2CV was conducted on the WES/WGS data from the five different types of SCC. A scatter plot (figure 4-27) was generated to determine whether skin SCC had driver genes that were similar or different to the four other types of SCC. This scatter plot, which compared the significant q values for the genes in these tumour cohorts, showed that the genes which had been identified as significantly mutated (q value less than 0.1) in all the other SCCs as well as in skin SCC by the MutSig2CV analysis were *RB1*, *FAT1*, *HRAS*, *NOTCH2*, *TP53*, *NOTCH1*, *CDKN2A* and *ZNF750*. The gene which was the most significant in skin SCC and in other SCCs was *TP53*. *NOTCH1*, *CDKN2A* and *ZNF750* also had low q values in skin SCC and other SCCs, but not as low as *TP53*. Several genes with significant q values in cSCC but non-significant q values in the other four SCC groups were noted, with *MOGAT1* identified as the most significantly mutated gene in skin SCC in this category.

Table 4-1: Potential driver genes identified in SCCs from different organ sites.

Cervical SCC	Lung SCC	Oesophageal SCC	Oropharyngeal SCC	Skin SCC
PIK3CA	FAT1	PIK3CA	CASP8	CDKN2A
ERBB2	KEAP1	TP53	HRAS	HRAS
FBXW7	TP53	CDKN2A	TP53	TP53
MAPK1	CDKN2A	CREBBP	CDKN2A	CCDC28A
RB1	FBXW7	EP300	CTCF	CDC27
ARID1A	NFE2L2	FAT1	EP300	CHUK
B2M	PABPC3	FAT2	EPHA2	FAT1
ELF3	PIK3CA	FBXW7	FAT1	KIF4B
EP300	PTEN	KEAP1	FBXW7	NOTCH1
HLA-A	RB1	KRAS	NFE2L2	NOTCH2
HLA-B	ARID1A	MUC6	NOTCH2	PRB2
IFNGR1	DDX3X	NFE2L2	PIK3CA	TMEM222
KRAS	ELOVL5	NOTCH1	RAC1	
NF2	HGF	NOTCH3	RIPK4	
NFE2L2	HRAS	PTCH1	TMTC1	
PTEN	IL17F	PTEN	ASXL1	
STK11	MST1	RB1	CYLD	
TP53	NBPF1	SMAD4	FOSL2	
ZNF750	NRAS	ZNF717	HLA-A	
	SYNE1	ZNF750	HLA-B	
	USP13	ARID1A	KEAP1	
	ZNF302	BAP1	MAPK1	
	ZNF814	C10orf76	MLH3	
		CHADL	NEFH	
		ERBB2	PTEN	
		GPR32	RASA1	

		IRF5	RB1	
		L3MBTL4	TGFBR2	
		MYL1	ZNF750	
		NBPF1		
		NEFH		
		NOTCH2		
		OR4L1		
		PARP4		
		RBPJ		
		RGL2		
		SYNJ1		

Footnote: Genes had a q value <0.1 in all four bioinformatics programs (yellow cells), MutSig2CV and two other bioinformatics programs (blue cells), or in Mutsig2CV and one other program (green cells). The genes are in alphabetical order as there are some genes which have a more significant q value in MutSigCV but are not mutated in the other programs such as dNdScv, OncodriveCLUST and OncodriveCLUSTL.

Next, all the genes which were identified as potential driver genes, according to a q value <0.1, in all five cohorts of SCCs using the four bioinformatics programs, MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL were compared (table 4-1). Some genes gave a q value <0.1 in four of these programs, whereas others gave a q value <0.1 in three or fewer programs. It is recognised that these four different programs use different approaches to identify driver genes, and that it would not be expected that all of these programs would identify the same driver genes. However, MutSig2CV has been used frequently in the literature to identify driver genes, but it was thought that the identification of mutations by this program and an additional program would be more robust in designating a gene as highly likely to be a driver gene. Therefore, a gene was considered a driver gene if it was significant at q <0.1 in MutSig2CV and at least one other program.

Table 4-2: Driver genes in skin SCC and SCCs at other organ sites.

Cervical SCC	Lung SCC	Oesophageal SCC	Oropharyngeal SCC	Skin SCC
PIK3CA	FAT1	PIK3CA	CASP8	CDKN2A
ERBB2	KEAP1	TP53	HRAS	HRAS
FBXW7	TP53	CDKN2A	TP53	TP53
MAPK1	CDKN2A	CREBBP	CDKN2A	CCDC28A
RB1	FBXW7	EP300	CTCF	CDC27
ARID1A	NFE2L2	FAT1	EP300	CHUK
B2M	PABPC3	FAT2	EPHA2	FAT1
ELF3	PIK3CA	FBXW7	FAT1	KIF4B
EP300	PTEN	KEAP1	FBXW7	NOTCH1
HLA-A	RB1	KRAS	NFE2L2	NOTCH2
HLA-B	ARID1A	MUC6	NOTCH2	PRB2
IFNGR1	DDX3X	NFE2L2	PIK3CA	TMEM222
KRAS	ELOVL5	NOTCH1	RAC1	
NF2	HGF	NOTCH3	RIPK4	
NFE2L2	HRAS	PTCH1	TMTC1	
PTEN	IL17F	PTEN	ASXL1	

STK11	MST1	RB1	CYLD	
TP53	NBPF1	SMAD4	FOSL2	
ZNF750	NRAS	ZNF717	HLA-A	
	SYNE1	ZNF750	HLA-B	
	USP13	ARID1A	KEAP1	
	ZNF302	BAP1	MAPK1	
	ZNF814	C10orf76	MLH3	
		CHADL	NEFH	
		ERBB2	PTEN	
		GPR32	RASA1	
		IRF5	RB1	
		L3MBTL4	TGFBR2	
		MYL1	ZNF750	
		NBPF1		
		NEFH		
		NOTCH2		
		OR4L1		
		PARP4		
		RBPJ		
		RGL2		
		SYNJ1		

Footnote: Genes designated as driver genes had a q value <0.1 in MutSig2CV and in one of three other bioinformatics programs (dNdScv, OncodriveCLUST and OncodriveCLUSTL). Identical driver genes in skin SCC and in one or more SCCs of other internal organs (cervical, lung, oesophageal, oropharyngeal) are highlighted by identical coloured cells in the table. The genes are in alphabetical order as there are some genes which have a more significant q value in MutSigCV but are not mutated in the other programs such as dNdScv, OncodriveCLUST and OncodriveCLUSTL.

Table 4-2 shows that driver genes which were unique to skin SCC comprised *CCDC28A*, *CDC27*, *CHUK*, *KIF4B*, *PRB2* and *TMEM222*. *TP53* was a driver gene that was shared between all the SCC groups. Two genes, *CDNK2A* and *FAT1*, were seen as driver genes in skin SCC and three other types of SCC, oropharyngeal, oesophageal and lung SCC. Another two genes, *NOTCH2* and *HRAS* were driver genes in skin SCC and SCCs of two other organ sites: *NOTCH2* in skin SCC, oropharyngeal SCC, and oesophageal SCC and *HRAS* in skin SCC, oropharyngeal SCC, and lung SCC. One gene, *NOTCH1*, was only shared as a driver gene between skin SCC and SCC of another organ, i.e., oesophageal SCC.

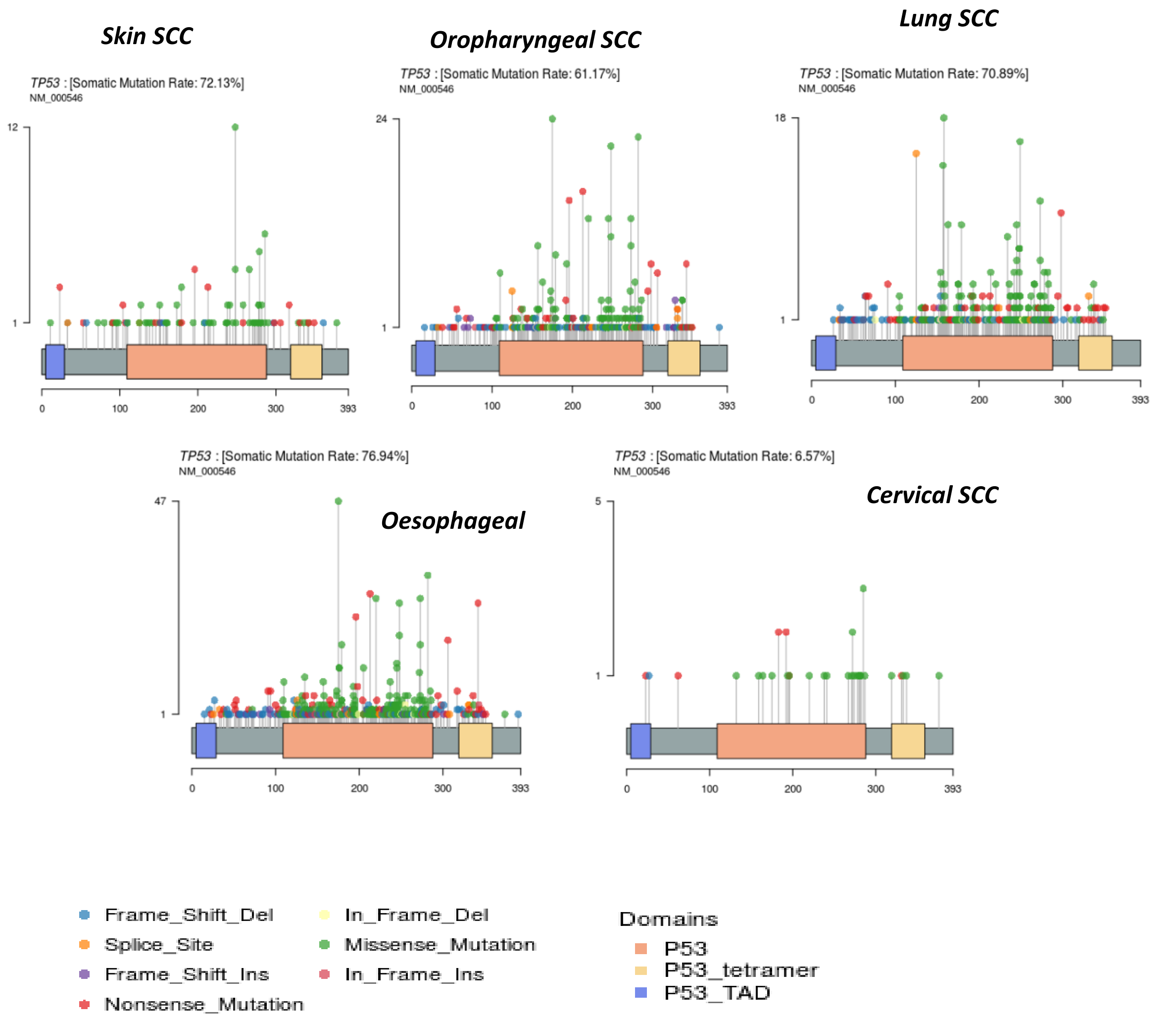


Figure 4-28: Schematic representation of the TP53 protein and the types of mutations in the SCCs from skin, oropharynx, lung, oesophagus and cervix. *This figure was produced from the TP53 mRNA transcript numbers specified by NM 000546 and the TP53 protein's amino acid numbers are stated on the X axis. The respective protein domains have been labelled and were extracted from the Pfam database. The type of mutations are represented by the colours of the circles on the lollipop and show which regions of the protein the gene mutation is affecting. The number of mutations in each region of the gene is represented by the height of the lollipop.*

TP53 was identified as a driver gene in all five SCCs. The mutation rate varied between SCCs with oesophageal SCC having the highest *TP53* mutation rate of 76.94% and cervical SCC with the lowest *TP53* mutation rate of 6.57%. All five types of SCCs contained mutations in all three domains of the protein (figure 4-28). Many mutations in *TP53* were missense mutations which were clustered in the P53 domain of the protein in all the SCCs from the different organs. The skin SCC tumour samples had nonsense mutations and a missense mutation in the P53 TAD domain whereas tumour samples in the other SCCs show tumour samples with frame shift deletions in this part of the protein. Missense and nonsense mutations were noted in the P53 tetramer domain in skin and cervical SCCs but there were more types of mutations seen in this domain in oropharyngeal, oesophageal and lung SCCs.

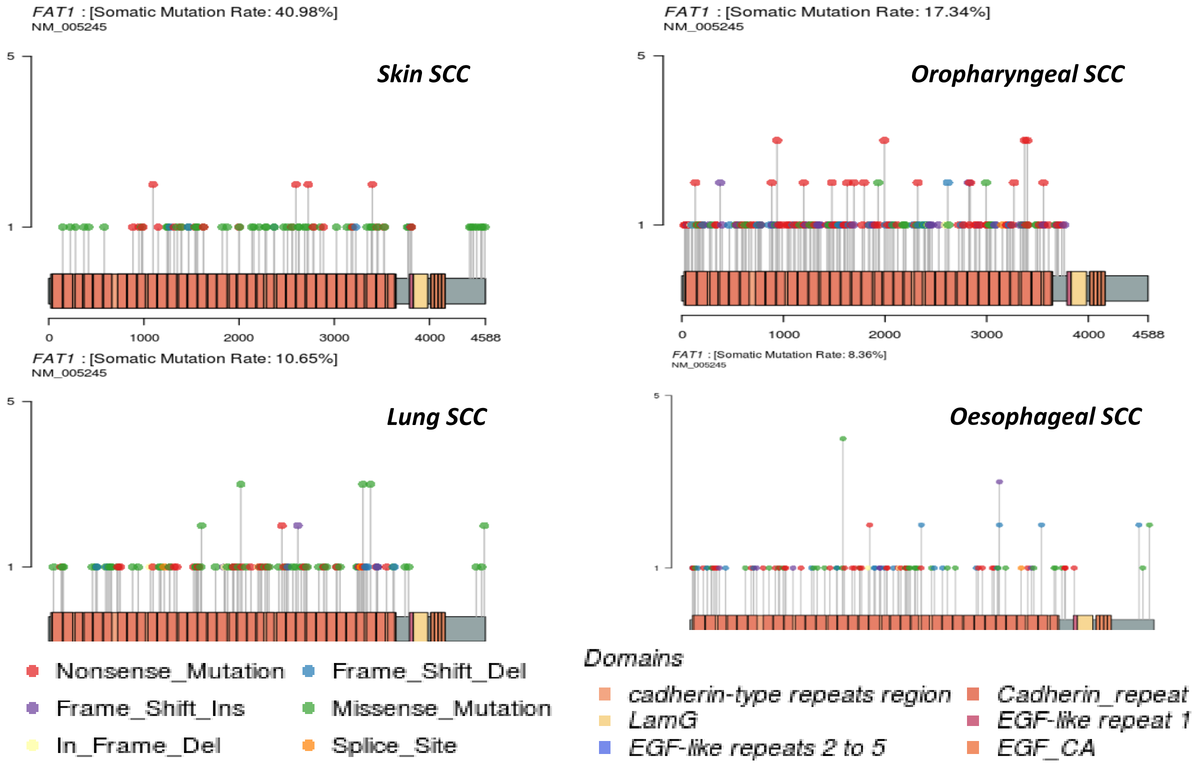


Figure 4-29: Schematic representation of the FAT1 protein and the categories of mutations in SCCs from skin, oropharynx, lung and oesophagus. *The figure uses the FAT1 mRNA transcript numbers specified by NM 005245, with the FAT1 protein's amino acid numbers is shown beneath each graphic. The respective protein domains have been labelled as per the Pfam database. The type of mutations are represented by the colours of the circles on the lollipop and show which regions of the protein the gene mutation is affecting. The number of mutations in each region of the gene is represented by the height of the lollipop.*

FAT1 was a driver gene in skin SCC, oropharyngeal SCC, lung SCC and oesophageal SCC, with skin SCCs having the highest mutation rate (40.98%) and oesophageal SCC having the lowest mutation rate (8.36%) in this gene (Figure 4-29). All four SCCs had mutations in the cadherin repeat domain of *FAT1*, with lots of these being missense or nonsense mutations. None of the SCCs had mutations in the LamG domain of the FAT1 protein, but oesophageal SCC and skin SCC had mutations in the EGF-like repeat 1 domain of the FAT1 protein.

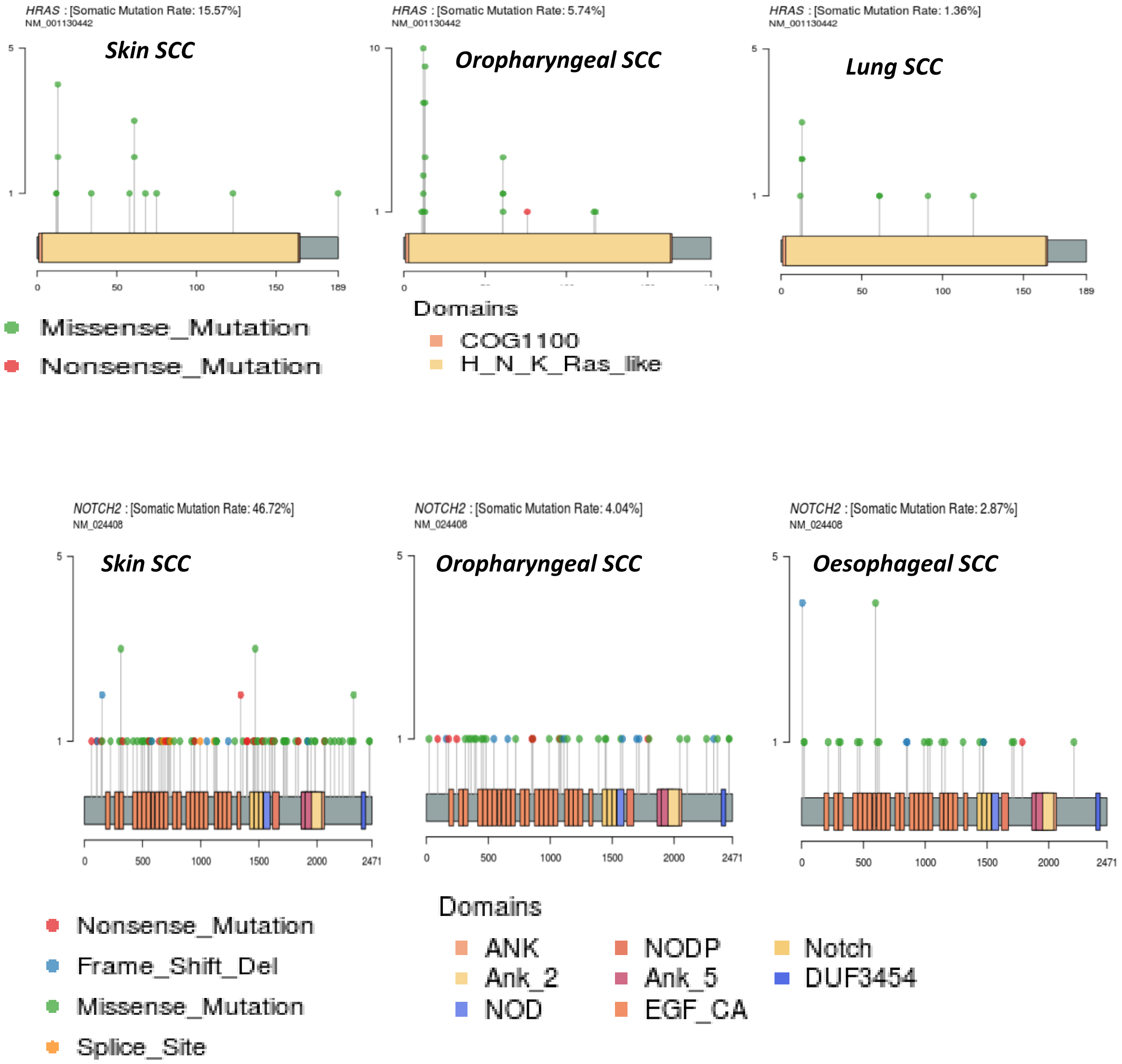


Figure 4-30: Schematic representation of the HRAS and NOTCH2 proteins and the categories of mutations in SCCs from skin, oropharynx and lung (HRAS) and skin, oropharynx and oesophagus (NOTCH2). *The lollipop plots were produced using the transcript numbers specified by NM 001130442 for HRAS and NM 024408 for NOTCH2 and the amino acid numbers are presented beneath each schematic representation. The respective protein domains have been labelled according to the Pfam database. The types of mutation are represented by the colours of the circles on the lollipop and show which region of the protein each gene mutation is affecting. The number of mutations in each region of the gene is represented by the height of the lollipop.*

HRAS and *NOTCH2* were identified as driver genes in skin SCC and two other SCC tumour types (Figure 4-30). As expected, most *HRAS* mutations were at codons 12 and 13 at the start of the HNK Ras-like domain, as well as in codon 61, but some mutations were detected at other exonic sites. Most *HRAS* mutations were missense mutations in skin, oropharyngeal and lung SCCs

NOTCH2 mutations were more common in skin SCC (mutation rate 46.72%) than in oropharyngeal and oesophageal SCCs. NOTCH2 mutations were generally scattered across the exonic regions in each of these three types of SCC, with many of the *NOTCH2* mutations distributed along the ANK domain in all three SCCs.

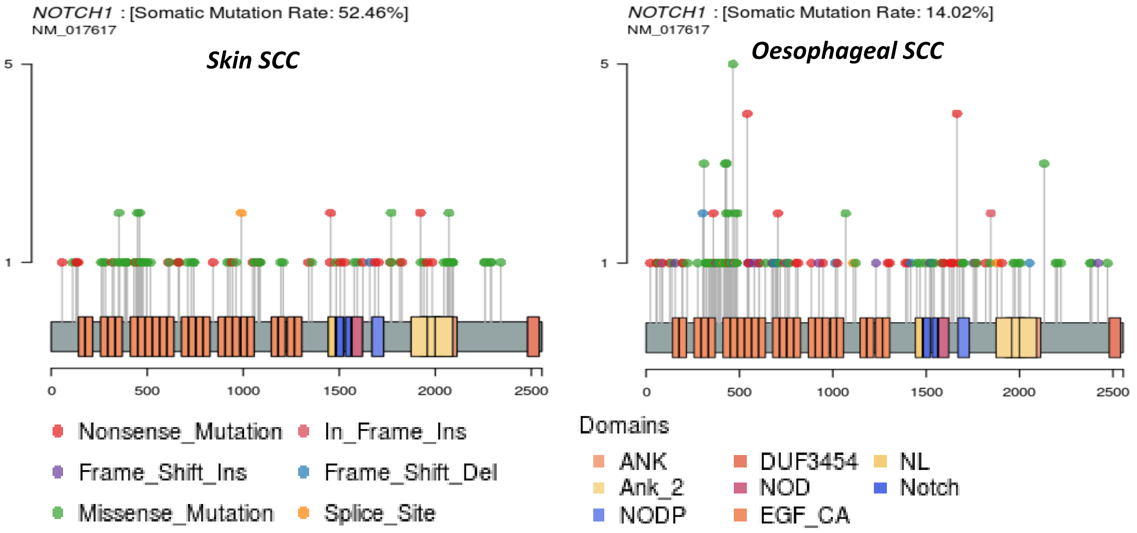


Figure 4-31: Schematic representation of NOTCH1 protein demonstrating the mutation profile in skin and oesophageal SCCs*. The transcript numbers specified by NM 017617 were used to generate the figures and the amino acid numbers are inserted below each figure. The respective protein domains have been labelled as per the Pfam database. The type of mutations are represented by the colours of the circles on the lollipop and show which regions of the protein the gene mutations affect. The number of mutations in each region of the gene is represented by the height of the lollipop.*

NOTCH1 was identified as a driver gene in cSCC and oesophageal SCC (figure 4-31). The mutation rate of *NOTCH1* in skin SCC was 52.14% which was much higher than the *NOTCH1* mutation rate in oesophageal SCC. In general, mutations were scattered throughout the NOTCH1 gene in both these cancer types, with a number of these mutations in the ANK domains in cSCC and oesophageal SCC. There were no mutations in the DUF3454 and EGF CA domain in both SCC tumour types.

Looking at identical mutations seen in skin SCC and SCCs of the other organs in relation to these six genes (*TP53*, *FAT1*, *CDKN2A*, *HRAS*, *NOTCH2* and *NOTCH1*) (table 4-3), identical mutations were frequently shared between oesophageal SCC and skin SCC (n=33 identical mutations detected in both types of SCC). However, there were 1184 oesophageal SCCs, 940 lung SCCs, 883 oropharyngeal SCCs and 472 cervical SCCs in this study, and the corresponding identical mutations (not accounting for how many mutations were present at each of these sites within each tumour type) when comparing skin SCC and other types of SCC amounted to 33 for skin versus oesophageal, 19 for skin versus lung, 21 for skin versus oropharyngeal and 3 for skin versus cervical. This indicates that the proportion of identical mutations between skin and other SCC types were largely similar for oesophageal, lung and oropharyngeal SCC.

Table 4-3: Mutations shared between driver genes in skin SCC and SCCs at other organ sites.

Gene	Mutation	Skin SCC	Oropharyngeal SCC	Lung SCC	Oesophageal SCC	Cervical SCC
TP53	c.1024C>T	Y*	Y	Y	Y	
	c.1045G>T	Y	Y	Y		
	c.159G>A	Y	Y		Y	
	c.272G>A	Y	Y		Y	
	c.310C>T	Y	Y	Y	Y	
	c.373A>C	Y		Y		
	c.380C>T	Y	Y		Y	
	c.413C>T	Y			Y	
	c.437G>A	Y		Y	Y	
	c.476C>T	Y	Y	Y	Y	Y
	c.517G>A	Y	Y	Y	Y	
	c.527G>T	Y	Y	Y	Y	
	c.625_626delAG	Y			Y	
	c.69G>A	Y			Y	Y
	c.713G>A	Y			Y	
	c.796G>A	Y	Y	Y	Y	
	c.808T>A	Y		Y	Y	
	c.815T>G	Y		Y		
	c.824G>T	Y	Y		Y	
	c.833C>T	Y	Y	Y	Y	
	c.836G>A	Y		Y	Y	
	c.856G>A	Y	Y	Y	Y	
	c.949C>T	Y		Y	Y	

	c.98delC	Y		Y	Y	
	c.991C>T	Y	Y	Y	Y	Y
CDKN2A	c.176G>A	Y	Y	Y	Y	
	c.18C>T	Y	Y			
FAT1	c.10198C>T	Y	Y			
	c.2653C>T	Y	Y		Y	
	c.2844G>A	Y			Y	
	c.3286C>T	Y		Y	Y	
	c.440C>T	Y	Y			
	c.4879C>T	Y	Y		Y	
	c.8176C>T	Y	Y			
NOTCH1	c.1363G>A	Y			Y	
	c.1367G>A	Y			Y	
	c.4579C>T	Y			Y	
	c.5308C>T	Y			Y	
NOTCH2	c.4403C>T	Y			Y	
	c.938G>A	Y			Y	

* Y = yes (i.e., mutation is present), blank cells = mutation not present.

4.10 Discussion

The results of this analysis identified a number of significantly mutated genes which were unique to cSCC and which were not mutated in SCCs from the other four internal organs. In addition, the analysis identified several driver genes which were the same in skin SCCs as those identified in SCCs at other organ sites. The comprehensive literature search, COSMIC database and GDC portal showed that the largest amount of WGS or WES data was available for oesophageal SCC and then for lung, oropharyngeal and cervix in descending order, with the lowest amount of WGS or WES data available for skin SCC.

Some mutation signatures were shared between the SCCs of the different organs, whereas some mutation signatures were unique to cSCC. There were a high proportion of C to T changes across SCCs at various organ sites except in lung SCC where the highest number of changes were C to A. This C to A change is associated with tobacco smoking which is the main contributor to the development of lung SCC. The high proportion of C to T changes in cSCC was reflected in the mutation signatures produced. SBS7b and DBS1 were identified as mutation signatures in cSCC and are associated with UV exposure, which also aligns with the high proportion of C to T and CC to TT base changes that were observed in skin SCC. In lung SCC, mutation signatures SBS4 and DBS2 which are associated with exposure to tobacco smoking mutagens were also seen.

However, mutation signatures associated with UV exposure were noted in SCCs in other organ sites. In oropharyngeal SCC two mutation signatures which had a cosine similarity of 0.979 with SBS7b and 1 with DBS1 were identified. As SCCs of the lips had been excluded from this oropharyngeal cohort, and most people do not generally have their mouth wide open when sunbathing or during routine sunshine exposure, this suggests that SBS7b might arise via another mechanism other than UV. There were 11 oropharyngeal samples in the analysis which had the largest contribution to this mutation signature, therefore further study was conducted to identify the exact primary site of these oropharyngeal SCCs to further understand the reason for this mutation signature.

The 11 samples were removed and further reanalysed which produced all the original mutation signatures, SBS2, SBS6 and SBS45, except SBS7b. Which suggests that there are no new mutation signatures identified in the oropharyngeal SCC samples in reanalysis and the SBS7b is originating from the 11 samples. In lung SCC, a mutation signature which had a cosine similarity of 0.754 to SBS7a and a signature which had a similarity of 1 to DBS1 were observed. Both mutation signatures are associated with UV, but it is extremely unlikely that these SCCs arose in the lung as a result of UV, so this provides additional support for these signatures arising via another pathogenetic mechanism in some cases.

In appendix 7.12, the 11 samples which displayed this UV signature were investigated further. Here it is highlighted that some of the samples originated from different data sources such as COSMIC, GDC/MC3 and published literature searches. The samples also had different mutation burdens which varied from 1 mutation to 4157 mutations. The sample with the highest mutation burden originated from the skin of face (B5T_B5N). This suggests that this sample was mislabelled and considered as an oropharyngeal sample but should have been considered as a skin sample. Another sample from the same journal article (Ren et al., 2017) also originated from the left ear which has been mislabelled as oropharyngeal SCC. This suggests that in other samples, mislabelled of samples could have been a contributing factor to the production of a UV signature in this analysis. The mean mutation burden for oropharyngeal SCC was approximately 204 mutations and the median mutation burden was 61 and the majority of the 11 samples did not have mutation burdens that were similar to the median or mean.

During the sequencing of these samples, there could have been index hopping which is when a certain number of sequencing reads are incorrectly assigned from one sample to a different sample in a pool (Guenay-Greunke et al., 2021). The remaining samples originated from different sites of the oropharynx which further suggests this UV signature could also be a result of a sequencing artefact. The Pan Cancer study that looked at the mutation signatures of several cancers (Alexandrov et al., 2020) also showed a low number of samples which displayed mutation signatures 7a and 7b in their Head and Neck SCC samples. TCGA samples which were included in the Alexandrov et al., 2020 study were also used in this study which could also be one of the reasons for a UV signature being displayed in oropharyngeal SCC in these samples.

The COSMIC database was also a source of these 11 samples, COSMIC database is manually curated therefore human error of misclassification of samples could also be contributing to this UV signature. The Maftools program uses the samples to extract mutation signatures and then compares them to known mutation signatures. Therefore, the algorithm which produces the UV-based mutation signature has based this predominantly on these 11 samples instead of identifying a signature which is present in most of the oropharyngeal samples which is producing a bias in the algorithm. Therefore, technical issues in producing the mutation signature computationally could also be contributing to the SBS7b mutation signature in oropharyngeal SCC.

Since this analysis used secondary data from multiple studies, it is most likely that the reason for these 11 samples contributing to this UV signature is due to misclassification as correctly identified by the examiners of this PhD thesis. This is evidenced in appendix 7.12 as a sample considered to be oropharyngeal SCC originated from the skin of the face. Therefore, this is a limitation of using secondary data as misclassification could occur during the curation process in the genomic database which can then produce results which are not reflective of the disease cohort.

Driver genes were compared between skin SCC and SCCs at other organ sites. The driver genes were identified using four different bioinformatics programs which were run using three different computational languages, R, Python and Linux. The MutSig2CV program was previously used to identify driver genes in SCCs by Campbell et al., 2018. A similar approach was initially taken in the current study to identify how similar cSCCs were to SCCs at other organ sites, and this identified a set of driver genes unique to skin SCC. The *MOGAT1* gene was identified as the most significant driver gene unique to cSCC in this MutSig2CV analysis. The *MOGAT1* gene codes for an enzyme that is involved in triglyceride synthesis and storage (Hayashi et al., 2014). There is limited information on the expression of *MOGAT1* in skin, but the Digital Aging Atlas suggests that "differential expression with age was identified in *MOGAT1* in skin" (https://ageing-map.org/atlas/change/DAA4292/) and that MOGAT1 expression in skin decreases with age (Glass et al., 2013) (https://ageing-map.org/atlas/results/?lid=100152&sort=identifier&species%5B%5D=9606&l=tissue&page=56), therefore further research will be required to examine whether *MOGAT1* is indeed a strong driver gene for skin SCC. However, other driver genes such as *CHUK* which were identified as unique to skin SCC in this current analysis have been identified as driver gene in cSCC in the literature using different bioinformatics programs LOFsigrank (Shain et al., 2015a) and OncodriveFML (Mularoni et al., 2016, Chang and Shain, 2021). The *CCDC28A*, *CDC27*, *KIF4B*, *PRB2* and *TMEM222* genes that were highlighted by MutSig2CV as potential driver genes unique to skin SCC (figure 4-27) were also identified as significant in other bioinformatics programs in this analysis (see table 4-2).

The MutSig2CV analysis showed genes which were shared between SCCs of other organ sites and cSCC. *RB1* and *ZNF750* had a q value less than 0.05 for skin SCC however these genes were not identified as significant in cSCC in other bioinformatics programs and therefore were not classed as potential driver genes. In skin SCC, there were no mutated genes which were identified as significant in all four bioinformatics programs (MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL) that were employed in the current study. Admittedly, the skin SCC sample size was the lowest of all five different types of SCC, therefore this could have affected the ability for driver genes to be detected in the skin SCC cohort. In the MutSig2CV analysis eight driver genes were shared between all SCCs: *TP53*, *NOTCH1*, *NOTCH2*, *FAT1*, *HRAS*, *CDKN2A*, *RB1* and *ZNF750*.

In the final part of the analysis, a gene was only classed as a driver if it was significant in MutSig2CV and one other program. *TP53* was the only driver gene that was shared between all five different types of SCCs. *TP53* was also the only driver gene that was shared with cervical SCC, albeit at a lower frequency because most cervical SCC arises from human papilloma virus (HPV), and the E6 protein in HPV types that cause cervical cancer targets the p53 protein for degradation, thus releasing the requirement for TP53 mutation in this cancer (Crook et al., 1991). Five driver genes were shared between cSCC and oropharyngeal SCC and, separately, between cSCC and oesophageal SCC, whereas four driver genes were shared between skin SCC and lung SCC and only one driver gene was shared between skin SCC and cervical SCC. Part of the reason for the low numbers of shared driver genes between skin SCC and SCCs of the other organs is likely to be the limitations due to the lower number of skin SCCs with WES and WGS.

To further understand how similar skin SCC is to SCCs at other organ sites, it will be important to perform this type of analysis with larger numbers of skin SCCs (and ideally more cervical SCCs) to increase the power of the study. However, this study shows that there are similarities in the driver genes identified for skin SCCs and SCCs at other organ sites which might prove useful in future years if targeted therapies which address the altered signalling in the relevant cellular pathways are produces for SCCs of other organs, because these could then be trialled for treatment of aggressive skin SCCs. Furthermore, as the bioinformatics programs used to identify driver genes in this study showed that there is potential for more genes to be shared between cSCC and SCCs at other organ sites, higher powered studies are likely to be helpful in identifying additional driver genes that are shared across these different SCC types.

5. Genomic landscape of skin cancer, precancerous skin lesions and normal skin

5.1 Introduction

To identify the mutagenicity of NB-UVB in normal skin, driver genes and mutation signatures for skin cancer can be identified and compared with precancerous lesions and normal skin to highlight potentially pathogenic variants if these appear in normal skin following NB-UVB exposure. In the initial part of this chapter, there will be an investigation into the genomic landscape of two types of skin cancer types, BCC, and melanoma, to allow comparison of all three common types of skin cancer, namely BCC, cSCC and melanoma. The cSCC and BCC tumours arise from keratinocytes whereas melanomas originate from melanocytes. Although, there have been studies which have identified driver genes in cSCC, BCC and cutaneous melanoma, there has been limited comparison of these three types of skin cancer despite the fact that all of them arise in skin and mainly as a result of UV exposure (South et al., 2014, Chitsazzadeh et al., 2016, Inman et al., 2018, Mueller et al., 2019, Chang and Shain, 2021, Hodis et al., 2012, Jayaraman et al., 2014).

Actinic keratosis is a cSCC premalignancy and studies have shown that up to two thirds of cSCCs may develop from an AK, however, fewer than 0.6% of AKs progress to a cSCC (Thomson et al., 2021). Although many dermatologists consider that BCCs probably arise de novo, and not from any specific precancerous clinical lesion (Madan et al., 2010), Criscione and colleagues reported that 36% of BCCs diagnosed in their study arose in lesions that had been previously diagnosed clinically as AKs and that the risk of progression of AK to primary BCC was 0.48% at 1 year and 1.56% at 4 years (Criscione et al., 2009). There have been studies which have reported on the genetic changes identified using next generation sequencing, including the Thomson et al., 2021, Albibas et al., 2018 and Chitsazzadeh et al., 2016 papers, in which genetic changes identified in AKs were also compared with cSCCs. The Albibas et al., 2018 paper additionally compared p53 immunopositive patches (PIPs) in chronically sun exposed skin with AKs and cSCCs and commented on genetic alterations in Bowen's disease (seen from targeted next generation sequencing) with cSCCs.

In addition to the study by Albibas et al., 2018, a few other studies have used targeted next generation sequencing to look for mutations in chronically sun exposed skin. In the Martincorena et al., 2015 study, targeted sequencing was conducted in chronically sun exposed normal skin from the eyelid of four individuals. The eyelid skin was split into 234 biopsy samples which were 0.79-4.71mm^2 and the proportion of samples which had mutations in 12 genes that are frequently mutated in skin cancers were compared between the eyelid skin, melanoma, BCC and cSCC. In the Lynch et al., 2017 investigation, normal skin samples with an area of 4 x 4mm were obtained from 10 individuals undergoing Mohs surgery and targeted sequencing identified gene mutations in multiple genes, including several genes that had been shown to be mutated in cSCC and genes that were "known drivers of epithelial malignancy". In the Fowler et al., 2020 study, targeted sequencing was conducted on 2 x 2mm normal skin samples in 28 individuals, again demonstrating multiple mutated genes as well as 11 genes under positive selection / driver genes in normal skin (Fowler et al., 2020).

Therefore, this chapter aimed to (i) characterise WES/WGS data in melanoma and BCC to allow comparison of the driver genes in the three common types of skin cancer (i.e., BCC, SCC and melanoma), and (ii) examine WES/WGS and targeted next generation sequencing data in publications on AKs, normal skin and melanocytic naevi and evaluate for the presence of skin cancer driver genes in normal skin and these potentially precancerous lesions.

5.2 Methods

Whole genome and exome data were extracted from the COSMIC database using the script explained in appendix chapter 7, section 1 and literature searches were conducted to identify sources of whole genome and whole exome data for skin melanoma. Melanoma data was also extracted from GDC portal as outlined in chapter 7, section 4. BCC samples were used from the Bonilla et al., 2016 paper.

The bash script in chapter 7, section 1 from the appendix was edited for melanoma to extract all genetic information classified with a primary histology of 'malignant melanoma'. The format of the data was changed to ensure it was compatible for the genome annotation program, Oncotator.. To ensure all genomic data were from human samples, any genomic data from cell lines or samples which were from cultured cells were discarded. This was done by extracting column five (which was the sample name) and column 35 (which included a description of the tumour origin, e.g. cell-line) from the mutation data in skin_cutaneous_melanoma.txt, and then duplicate variant information was discarded.

Sample names which were described as originating from the mucosal region were discarded to ensure that mucosal melanoma samples were not included in the analysis. In the skin_cutaneous_melanoma.txt file, column five was extracted with the sample names and column nine with site subtype 1 'mucosal' was extracted, then duplicate sample names were removed.

Acral lentiginous melanoma is on the palms, soles, and nails (Bradford et al., 2009). This study did not include samples which were classified as acral lentiginous because this type of melanomas is not thought to be caused by UV (Hayward et al., 2017, Newell et al., 2020). The melanoma data included in this study were likely to reflect UV-induced melanomas. The COSMIC classification.csv file (described in chapter 2.4.1) showed that there were only cases of acral lentiginous melanoma classified in the foot in the COSMIC database. These samples of acral lentiginous melanoma in the foot were removed from the melanoma cohort of samples. The other histological types such as superficial spreading, nodular, lentigo maligna of melanoma were included in the analysis.

Melanoma tumour sample names were also downloaded from GDC portal and the mutation data for these samples were taken from the mc3.v0.2.8.PUBLIC.maf. The samples identified in GDC portal were compared with those from the COSMIC database to ensure there were no duplicates.

Literature searches for melanomas were conducted to identify WGS or WES data for BCC and melanoma. All literature was screened on MEDLINE OVID for BCC until 8th February 2020 and for melanoma until 19th May 2020. The data from these literature searches were also annotated using Oncotator and collated with the melanoma data from mc3.v0.2.8.PUBLIC.maf and COSMIC. The MAF files produced for each skin cancer were analysed using Maftools in R. The bioinformatics tools used to identify driver genes were MutSig2CV, dNdScv, OncodriveCLUST and OncodriveCLUSTL.

The cSCC data was compared with AK WES data, which was identified in the literature search, which had which was the same as the one conducted in chapter 3, section 2, any data that was AK data from that search was saved for future analyses (see search terms in appendix 7.9.1). The AK WES data in Thomson et al., 2021, Albibas et al., 2018 and Chitsazzadeh et al., 2016 was examined to identify if there were any cSCC driver genes present in the AK genomic dataset and any AK samples with mutations in the driver genes were recorded.

The number of AK samples with mutations in cSCC driver genes was recorded in R using ggplot2. The number of mutations in skin cancer driver genes in known publications with targeted sequencing data of normal skin was also recorded and compared using R version 3.5.1. The melanoma driver genes identified in melanocytic naevi and normal melanocytes were presented in a table and displayed in R version 3.5.1.

5.3 Results

Medline OVID searches were conducted for BCC and melanoma as highlighted in appendix 7.9.6 and 7.9.7. The flowcharts in this chapter show the number of studies that were identified in the literature search and the genetic databases: GDC portal and COSMIC. These studies were collated and analysed using a variety of graphs, such as oncoplots to identify the most frequently mutated genes in BCC and melanoma.

5.3.1 BCC

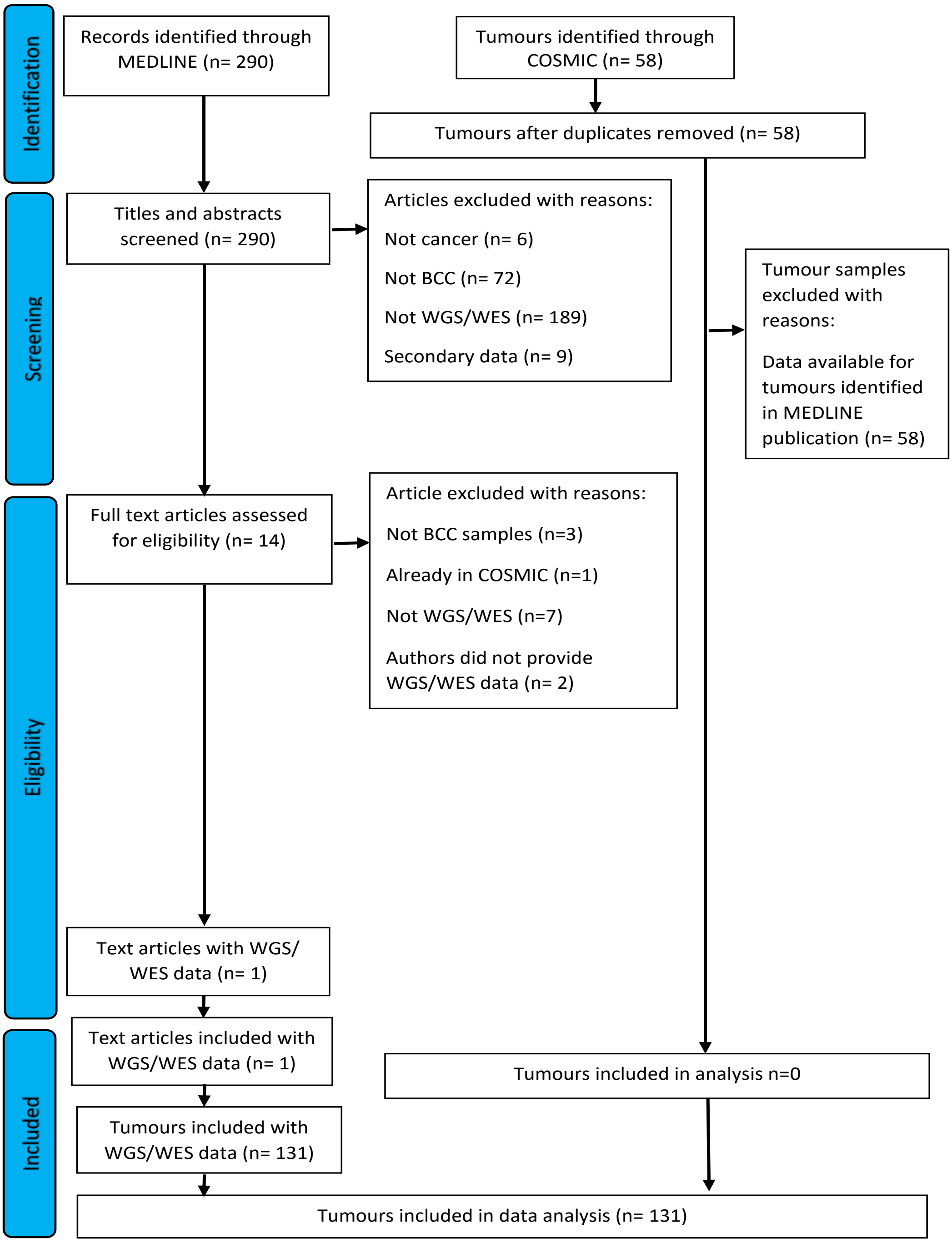

Identification
Screening
Eligibility
Included
Records identified through MEDLINE (n= 290)
Tumours identified through COSMIC (n= 58)
Tumours after duplicates removed (n= 58)
Titles and abstracts screened (n= 290)
Articles excluded with reasons:
Not cancer (n= 6)
Not BCC (n= 72)
Not WGS/WES (n= 189)
Secondary data (n= 9)
Tumour samples excluded with reasons:
Data available for tumours identified in MEDLINE publication (n= 58)
Full text articles assessed for eligibility (n= 14)
Article excluded with reasons:
Not BCC samples (n=3)
Already in COSMIC (n=1)
Not WGS/WES (n=7)
Authors did not provide WGS/WES data (n= 2)
Text articles with WGS/ WES data (n= 1)
Text articles included with WGS/WES data (n= 1)
Tumours included in analysis n=0
Tumours included with WGS/WES data (n= 131)
Tumours included in data analysis (n= 131)

__Figure 5-1: Flowchart representing WGS/WES studies identified from a literature search and WGS/WES search of the COSMIC database for BCC.__ Studies in the literature search were compared with those in the COSMIC database and duplicates were removed.

A Medline OVID search was conducted for BCC with search terms specified in the appendix 7.9.6 and ascertained a total of 290 studies (figure 5-1). The titles and abstracts were inspected for the 290 studies and 276 studies were removed. Of the 276 studies that were removed, six studies were not related to cancer, 72 studies were not specifically about BCC, 189 studies did not include WES or WGS data and nine studies contained secondary data. The publications on the 14 remaining studies were then examined in detail. Thirteen of these 14 studies were removed, including three where the samples in the study were not BCC, one study which had data deposited in the COSMIC database, seven articles that did not include WES or WGS data and two studies where the lead authors were emailed but did not provide their WES or WGS data. The one remaining study from the literature search contained WES data for 131 BCC samples. There were 58 BCC samples identified in the COSMIC database, however, the WES data for these samples had been included and re-analysed in the single study remaining from the literature search. Therefore, the samples identified in the COSMIC database were removed and the 131 samples from the study extracted from the literature search were analysed.

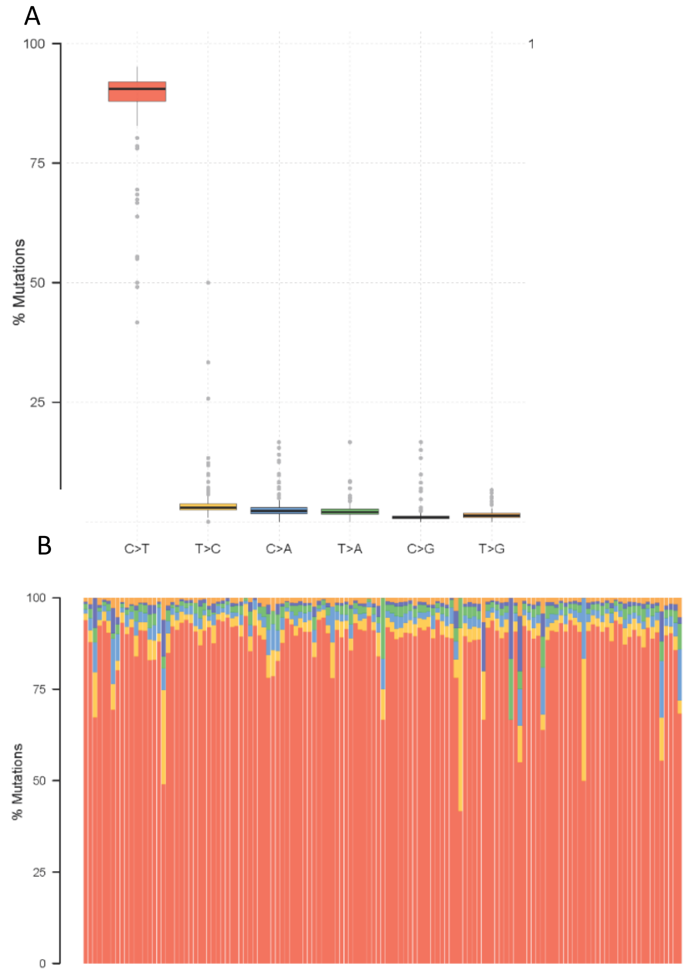
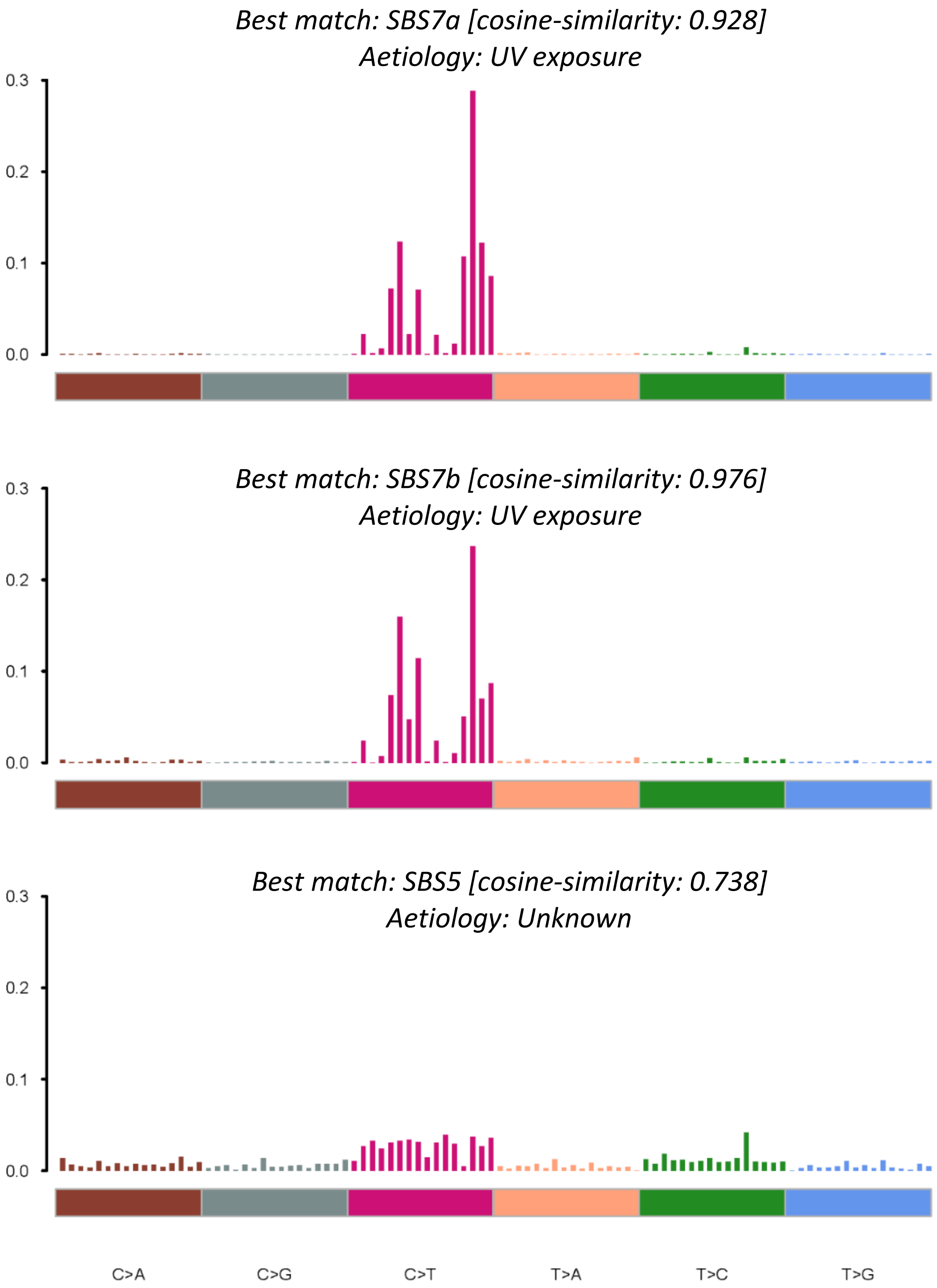


*Figure 5-2: **A. A box and whisker plot showing the percentage of base changes in the cohort of 131 BCC samples.** The different colours represent the different base changes. **B.The relative contributions of the different base changes in each individual BCC sample.** The proportion of mutations in each sample is measured as a percentage of the total number of mutations in that BCC and the different colours represent the different base changes as signified in part A of this figure.*

The BCC samples showed a total of 18,742 C to T base changes (figure 5-2). While the majority of BCCs had a much lower proportion of other base alterations, there were 12 BCCs where the combination of the other base changes amounted to >25% of the total base alterations, including two samples which had less than 50% of mutations that were a C to T base change. There was also one sample which had 50% of mutations that was a T to C base change.



Best match: SBS7a [cosine-similarity: 0.928]
Aetiology: UV exposure

Best match: SBS7b [cosine-similarity: 0.976]
Aetiology: UV exposure

Best match: SBS5 [cosine-similarity: 0.738]
Aetiology: Unknown

Figure 5-3: Single base subsitution mutation signature plot for BCC. *The x axis represents the base change in its trinucleotide context* and the y axis represents the proportion of each base change in the BCC tumour cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

There were three SBS mutation signatures identified from this cohort of BCC samples (figure 5-3). The signatures 7a and 7b, with a high cosine similarity of 0.928 to SBS7a and 0.976 to SBS7b, were identified which are both associated with UV exposure. The SBS5 mutation signature was also identified, however, it has an unknown aetiology.

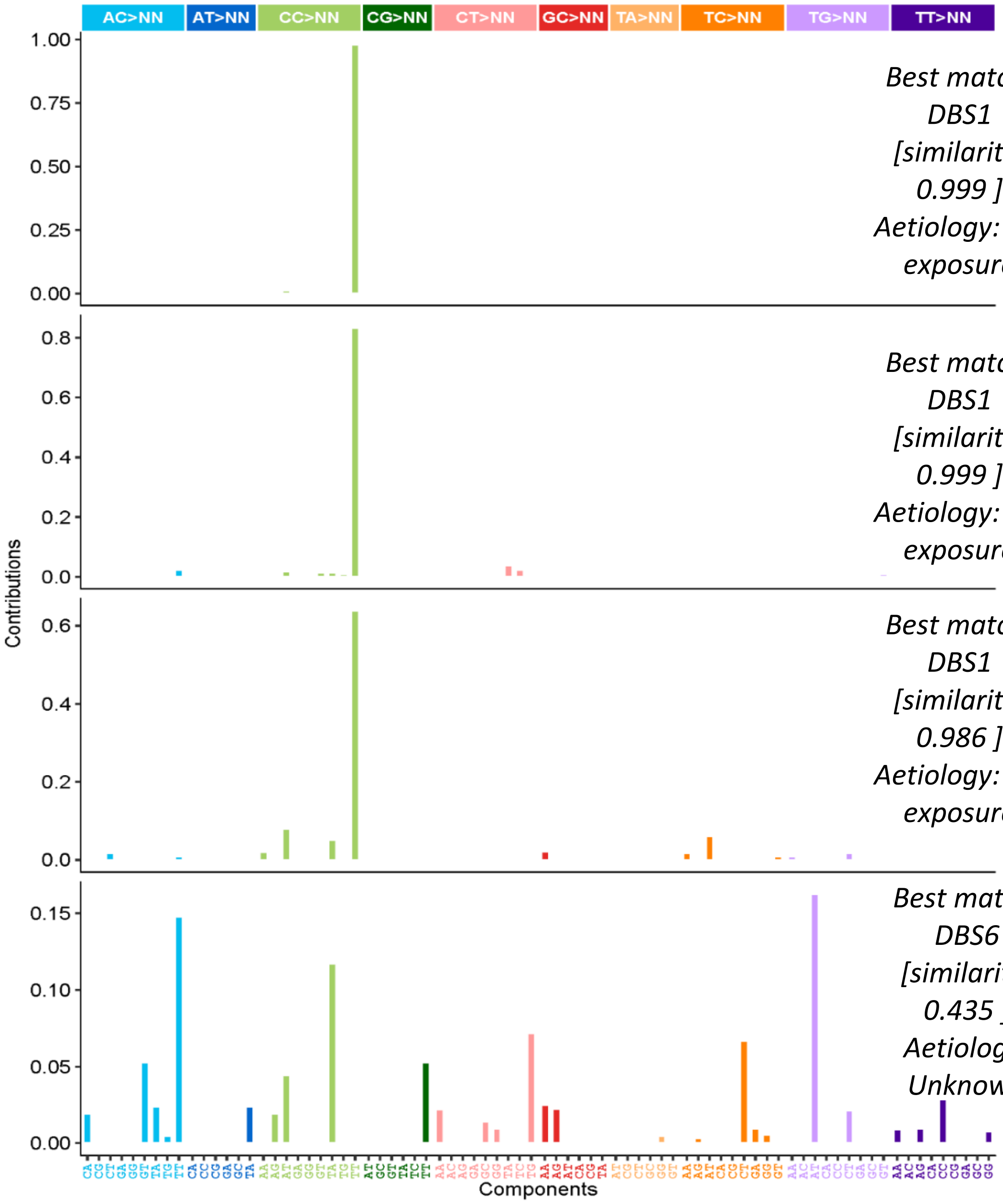


*Figure 5-4: **The double base substitution mutation signature plots for BCC.** The x axis represents the base change in its dinucleotide context and the y axis represents the proportion of each base change in the BCC tumour sample cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.*

There were four double base mutation signatures that were identified in BCC samples (figure 5-4). Three of the mutation signatures had a similarity that was most like DBS1 which is associated with UV exposure. The similarity of these DBS signatures to DBS1 ranged from 0.986 to 0.999. The fourth double base mutation signature identified in the BCC samples was most similar to DBS6 with a similarity of 0.435. DBS6 has an unknown aetiology.

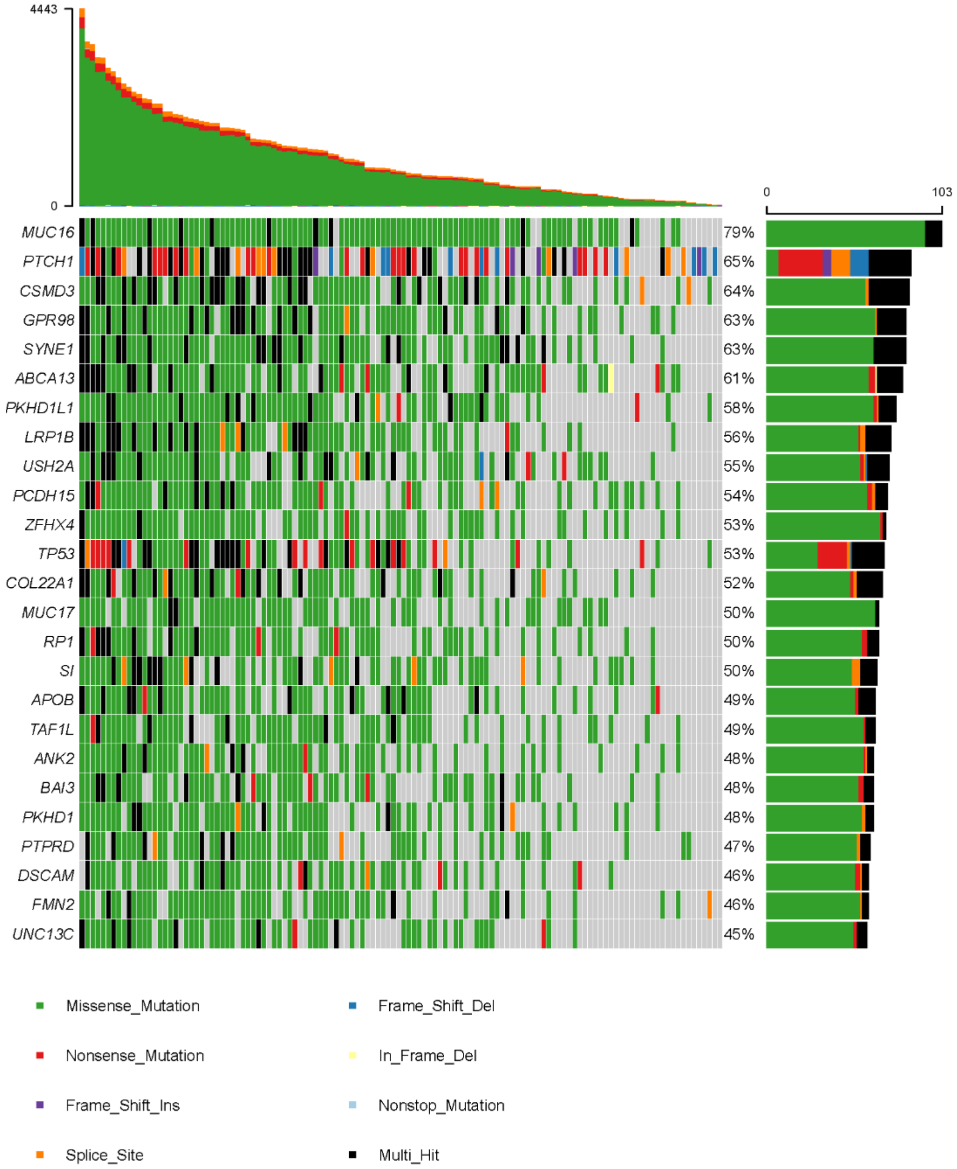


Figure 5-5:Oncoplot with top 25 frequently mutated genes in BCC tumours identifed in the literature search and COSMIC database. *One or more of these genes was mutated in 124 of 131 samples (94.66%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation that each sample contains within the corresponding gene. The bar chart on the right represents the number samples which have a mutation in that gene and the colours represent the type of mutation in the gene.*

At least one of the top 25 frequently mutated genes in BCC was mutated in 124 of 131 samples showing that 94.66% of samples share mutations in these genes (figure 5-5). The highest number of nonsynonymous mutations in an individual sample was 4443 mutations. Mutations in the *MUC16* gene is shared across the highest proportion of samples, i.e., 103 of 131 (79%) BCCs. The second most frequently mutated gene was *PTCH1*, with 65% of BCCs containing mutations in this gene. Missense mutations were the most common type of mutation in most of the top 25 most frequently mutated genes but the *PTCH1* gene had a much lower proportion of missense mutations and contained a mixture of missense, nonsense, frameshift insertion, splice site, frameshift deletion and multi-hit mutations.

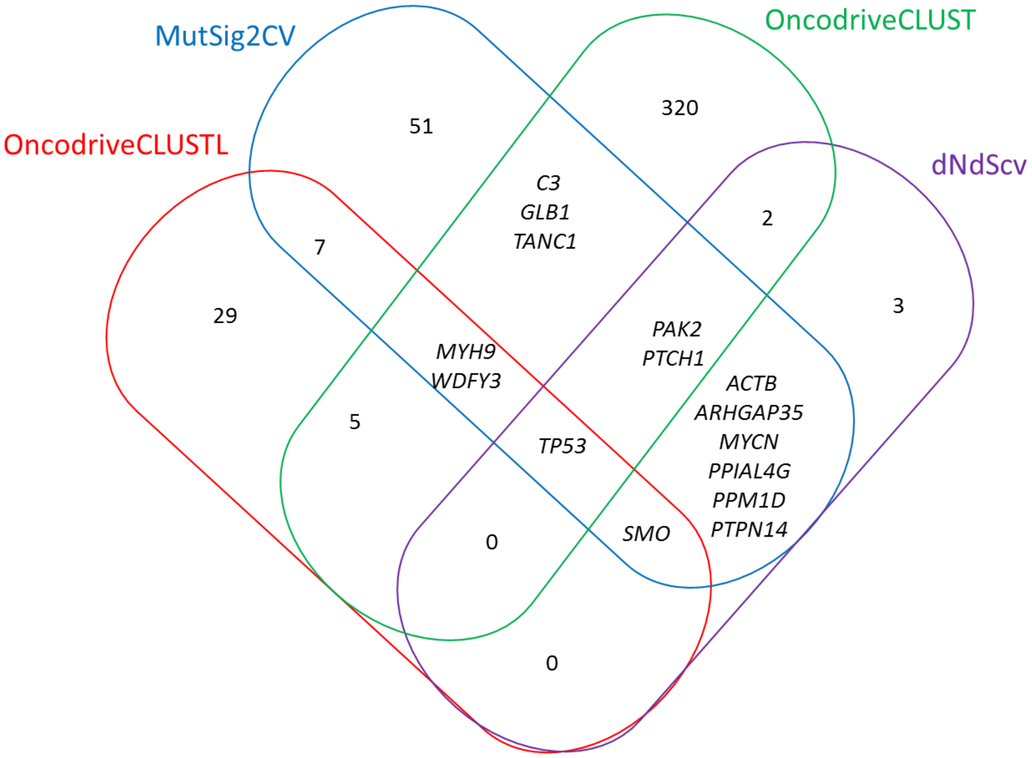


Figure 5-6: Venn-diagram of the potential driver genes identified in BCC using four different bioinformatics programs. *The four different bioinformatics programs are labelled in different colours corresponding to the outline of the closed curves; OncodriveCLUSTL (red), MutSig2CV (blue), OncodriveCLUST (green), dNdScv (purple). The numbers of potential driver genes, based on a false discovery rate q value <0.1, are highlighted and the names of specific genes identified by MutSig2CV and one other bioinformatics programe are included included in the relevant overlapping sections of the Venn diagram.*

The four programs used to identify potential driver genes were OncodriveCLUSTL, MutSig2CV, OncodriveCLUST, and dNdScv. The largest number of potential driver genes were identified using the OncodriveCLUST program as there were 335 genes with an analytical q value of less 0.1. However, in this analysis a driver gene was defined as a gene which has been identified as significantly mutated using the MutSig2CV program and one other program. Using this approach, 22 genes were considered driver genes in BCC (figure 5-6). The *TP53* gene was significant at q<0.1 in all four programs. There were several genes that were significant at this level in three programs; these comprised *MYH9* and *WDFY3* in MutSig2CV, OncodriveCLUSTL and OncodriveCLUST, the *SMO* gene in dNdScv, OncodriveCLUSTL and MutSig2CV, and the *PAK2* and *PTCH1* genes in MutSig2CV, dNdScv and OncodriveCLUST.

5.3.2 Melanoma

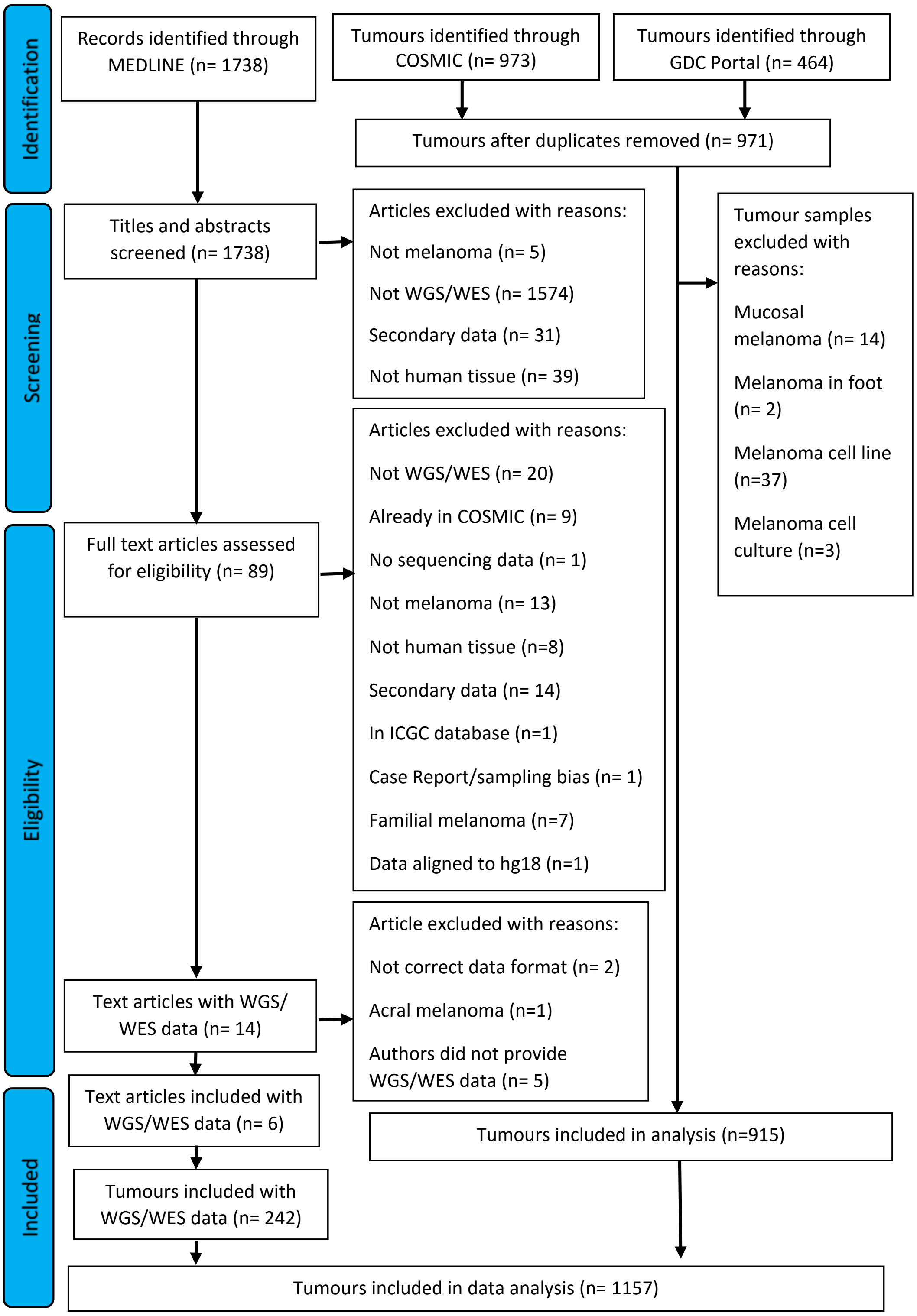

Identification
Screening
Eligibility
Included
Records identified through MEDLINE (n= 1738)
Tumours identified through COSMIC (n= 973)
Tumours identified through GDC Portal (n= 464)
Tumours after duplicates removed (n= 971)
Titles and abstracts screened (n= 1738)
Articles excluded with reasons:
Not melanoma (n= 5)
Not WGS/WES (n= 1574)
Secondary data (n= 31)
Not human tissue (n= 39)
Tumour samples excluded with reasons:
Mucosal melanoma (n= 14)
Melanoma in foot (n= 2)
Melanoma cell line (n=37)
Melanoma cell culture (n=3)
Articles excluded with reasons:
Not WGS/WES (n= 20)
Already in COSMIC (n= 9)
No sequencing data (n= 1)
Not melanoma (n= 13)
Not human tissue (n=8)
Secondary data (n= 14)
In ICGC database (n=1)
Case Report/sampling bias (n= 1)
Familial melanoma (n=7)
Data aligned to hg18 (n=1)
Full text articles assessed for eligibility (n= 89)
Article excluded with reasons:
Not correct data format (n= 2)
Acral melanoma (n=1)
Authors did not provide WGS/WES data (n= 5)
Text articles with WGS/ WES data (n= 14)
Text articles included with WGS/WES data (n= 6)
Tumours included in analysis (n=915)
Tumours included with WGS/WES data (n= 242)
Tumours included in data analysis (n= 1157)

Figure 5-7: Flowchart detailing studies identified in the literature search, COSMIC and GDC Portal databases for melanoma. *Studies in the literature search were compared with those in GDC portal and COSMIC databases and duplicates were removed.*

A Medline OVID search was conducted for melanoma with search terms specified in the appendix 7.9.7 and ascertained a total of 1738 studies (figure 5-7). The titles and abstracts were inspected for the 1738 studies and 1649 studies were removed from the analysis for the following reasons: 1574 studies did not include WGS or WES data, five articles were not about melanoma, 31 studies included secondary data and 39 studies included data which were not from human samples. The remaining 89 studies were analysed and 75 studies were excluded. Of these 75 studies, 20 did not include WGS or WES data, 14 included secondary data, 13 were not about melanoma, nine included data deposited in the COSMIC database, eight were on non-human tissue samples, seven included familial melanoma cases, one did not include sequencing data, another article’s data was in the ICGC (international cancer genome consortium) database which is a controlled access database that was not included in this analysis, one contained a single sample thus leading to potential sampling bias and in one study the sequencing data was aligned to the hg18 genome build. The 14 remaining studies were analysed further, and the lead authors were emailed to request data, however five authors did not provide WGS or WES data, one study included data on acral melanoma, and two studies did not have data available in MAF format. The remaining six articles included WGS or WES data for 242 samples. WGS or WES data for melanoma samples from COSMIC and GDC portal were also extracted. Duplicate samples were removed and any melanoma samples which were described as mucosal melanoma, melanoma which arose from the foot, melanoma which had been cultured from cells and cell lines were removed from the analysis. There 915 samples remaining with WGS or WES data for melanoma from GDC and COSMIC databases were joined with the 242 samples identified from the literature search to produce a file with 1157 human melanoma samples.

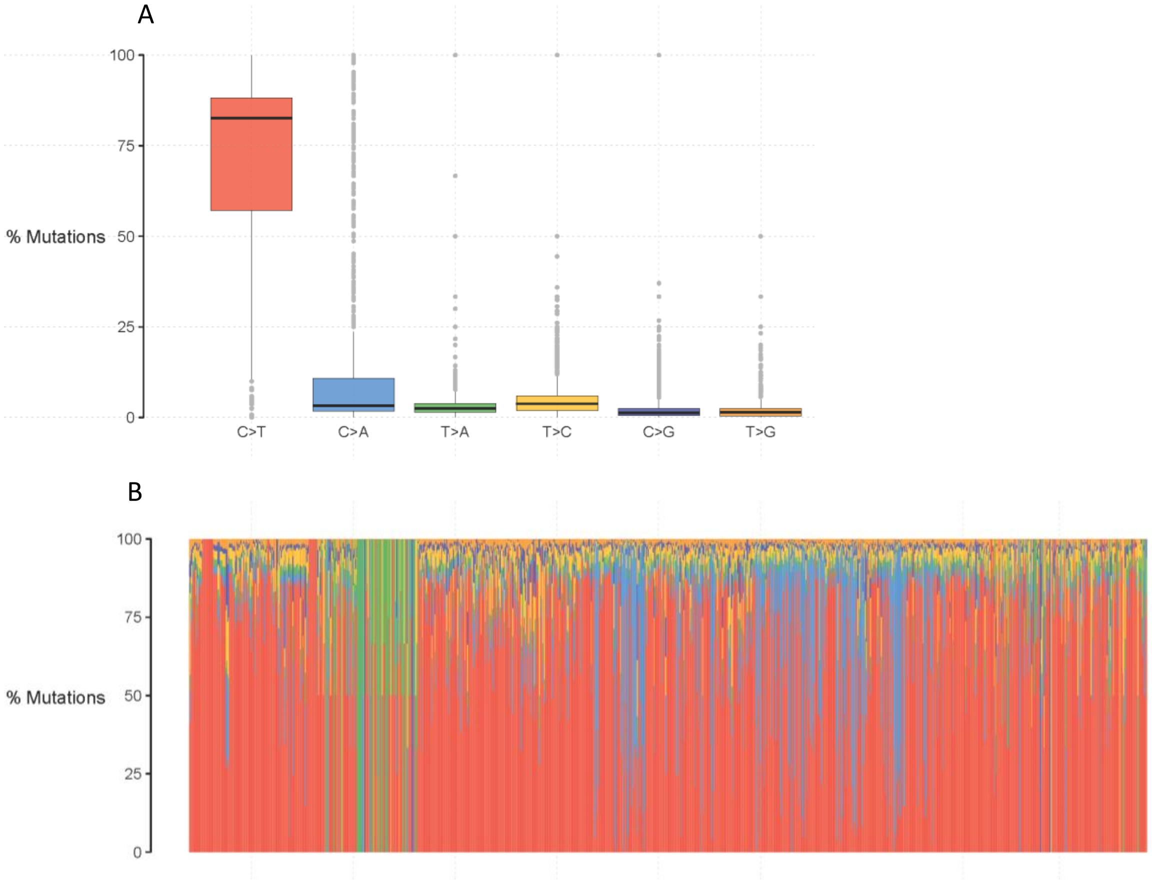


***Figure 5-8**: **A. A box and whisker plot showing the percentage of bases in this cohort of 1157 skin melanoma samples.** The box and whiskers shows the median and inter-quartile range and the total range respectively and the different colours represent the different base changes. **B.The relative contributions of the different base changes in each individual melanoma sample.** The proportion of mutations in each sample is depicted as a percentage of the total mutations and the different colours represent the base changes as delineated in part A of the figure.*

The largest proportion of base changes were C to T whereas the T to G base change was least common (figure 5-8). Despite C to T alterations being most frequent, some melanomas contained predominantly T to A changes, whereas other melanomas had mainly C to A alterations, and some other samples exhibited a mixture of base changes, representing the heterogeneity of this cancer type (figure 5-8B).

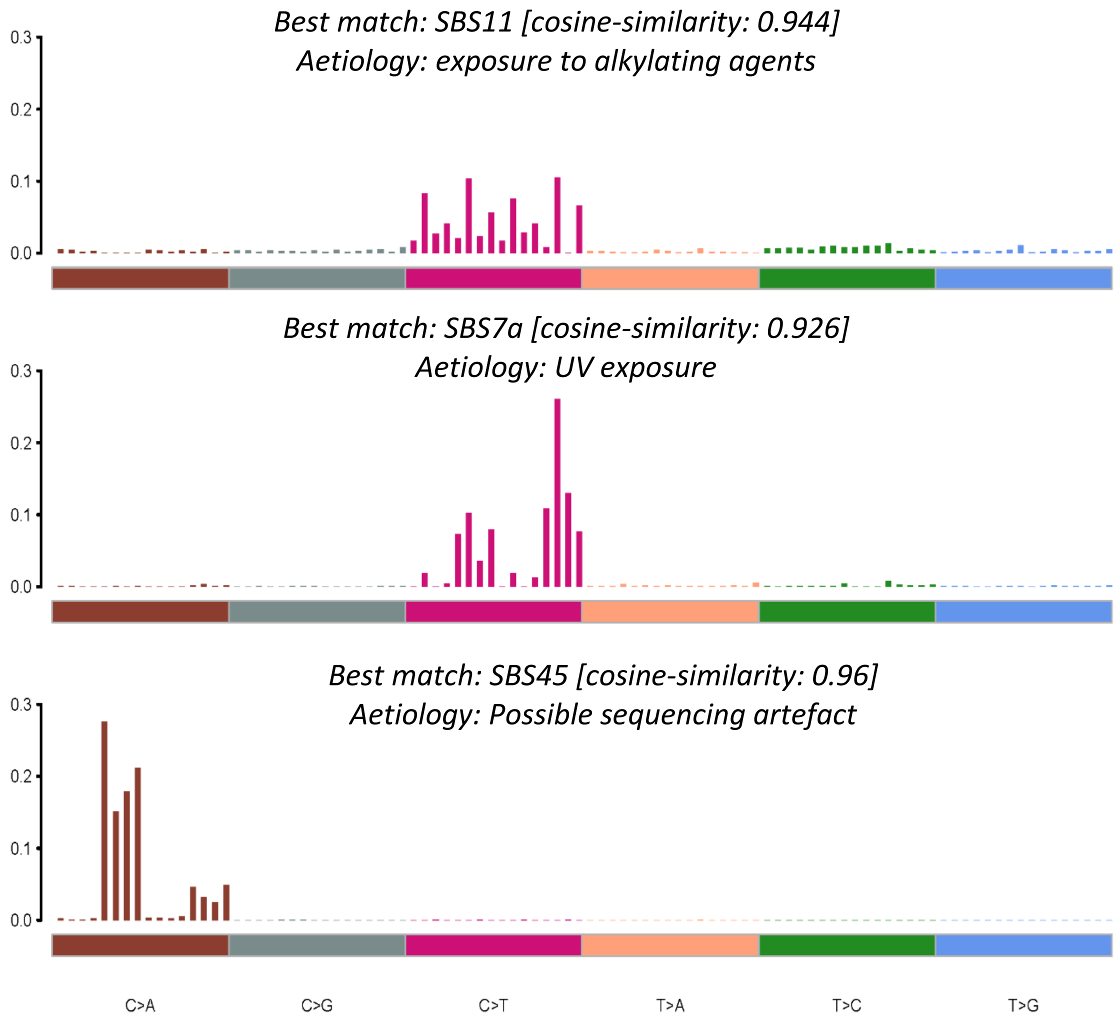


__Figure 5-9: Single base subsitution mutation signature plot for cutaneous melanoma.__ The x axis represents the base change in its trinucleotide context and the y axis represents the proportion of each base change in the melanoma sample cohort.The colour of the bars represents the specific base change described at the bottom of the graph. The mutation signature plot was created using Maftools in R. *Trinucleotide context from left to right for each base change, where N represents the base undergoing the change, is as follows; ANA, ANC, ANG, ANT, CNA, CNC, CNG, CNT, GNA, GNC, GNG, GNT, TNA, TNC, TNG, TNT.*

The melanoma samples showed three single base substitution mutation signatures (figure 5-9). One mutation signature had a cosine similarity of 0.96 to SBS45 which shows a high proportion of C to A base changes and is considered a possible sequencing artefact as a result of oxidative damage to the guanine resulting in 8-oxoguanine being introduced during DNA (Costello et al., 2013). Another mutation signature had a cosine similarity of 0.926 to SBS7a which is associated with UV. This signature reflects the high proportion of C to T base changes in the melanoma samples. A third mutation signature identified in the melanoma samples had a 0.944 cosine similarity to SBS11, which is associated with exposure to alkylating agents and could have occurred as a result of treatment with chemotherapeutic agents (e.g. dacarbazine, temozolomide) (Alexandrov et al., 2020).

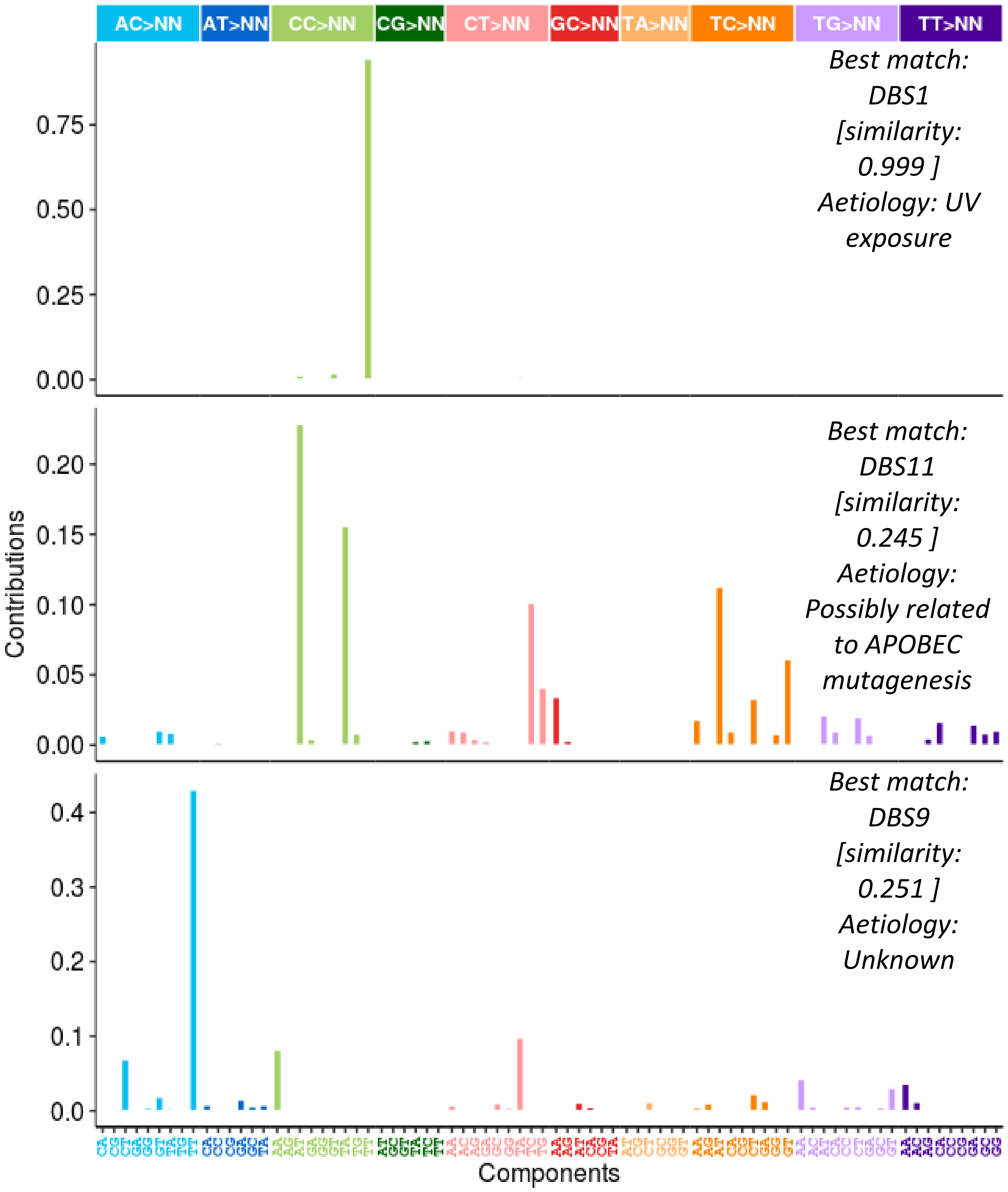


__Figure 5-10: Double base substitution mutation signature plots for cutaneous melanoma.__ The x axis represents the base change in its dinucleotide context and the y axis represents the proportion of each base change in the melanoma sample cohort.The colour of the bars represents the specific base change corresponding to the horizontal bar at the top of the graph. The mutation signature plot was created using Sigminer in R.

Three double base substitution mutation signatures were detected in the melanoma samples using Sigminer (figure 5-10). The first mutation signature has a similarity of 0.999 to DBS1 which is due to UV exposure and had been similarly identified in BCC and cSCC (in this chapter and chapter 3 respectively). However, in melanoma there was also a signature identified with a 0.245 similarity to DBS11, which is related to APOBEC mutagenesis and was not seen in cSCC or BCC. The melanoma tumour samples also contained a mutation signature which had a 0.251 similarity to DBS9 (which has no known aetiology) that was also identified in the cSCC tumour samples (where it had 0.324 similarity to DBS9).

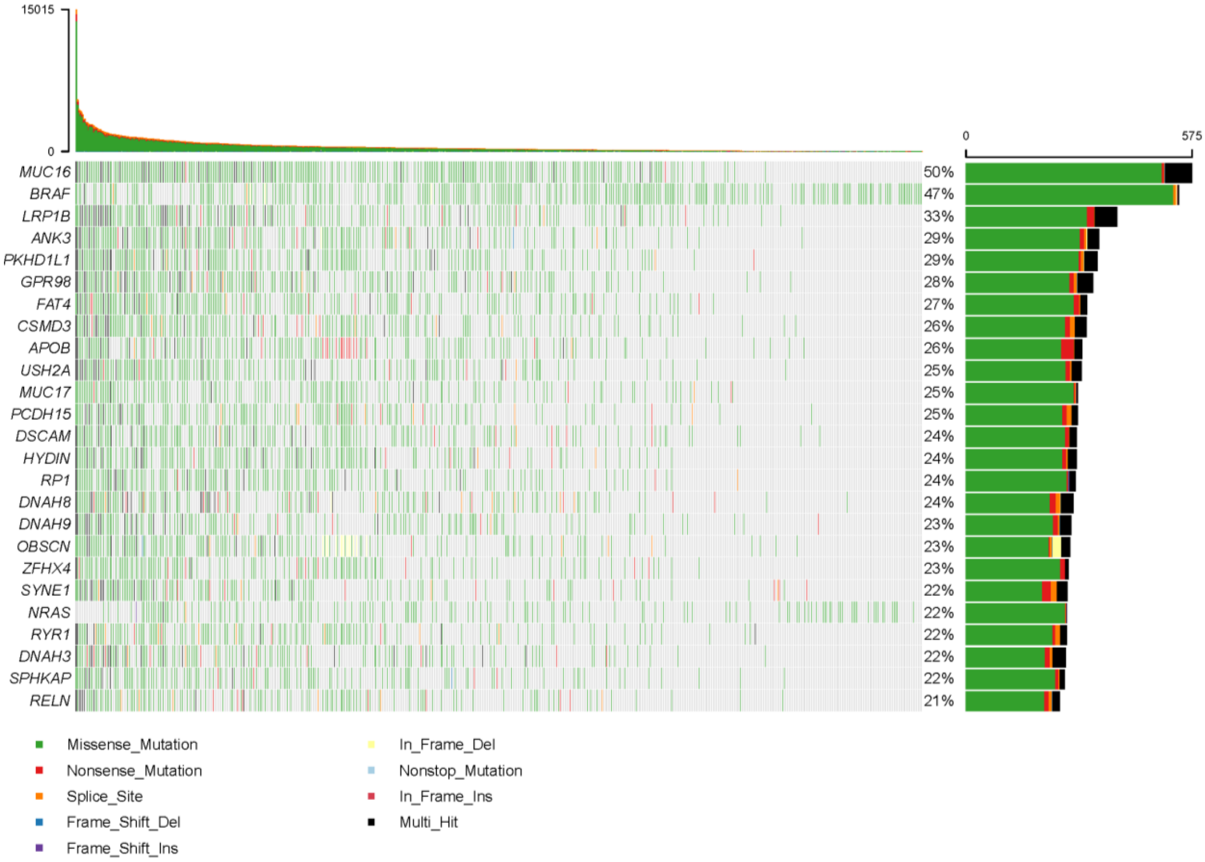


Figure 5-11: Oncoplot of top 25 frequently mutated genes in melanoma tumours identifed from the literature search, COSMIC and GDC portal databases. *At least one of these genes was mutated in 1024 of 1157 samples (88.5%). The number of mutations identified in each tumour is presented as a bar chart at the top of the figure. Each coloured square represents the type of mutation that each sample contains within the corresponding gene. The bar chart on the right represents the number of samples which have a mutation in that gene, with the colours representing the type of mutation in the respective gene.*

The oncoplot of the top 25 most frequently mutated genes in melanoma (figure 5-11) indicates that *MUC16* was mutated in 50% of melanoma samples and *BRAF* was the next most mutated gene, affecting 47% of tumours. The tumour with the highest number of nonsynonymous mutations contained 15,015 mutations. The proportion of mutations shared by each of the samples in the 25 most frequently mutated genes ranged from 21% of samples to 50% of samples, with the most common mutation type shared by all the top 25 frequently mutated genes being a missense mutation. One or more of the top 25 frequently mutated genes was mutated in 88.5% of the 1157 melanoma samples.

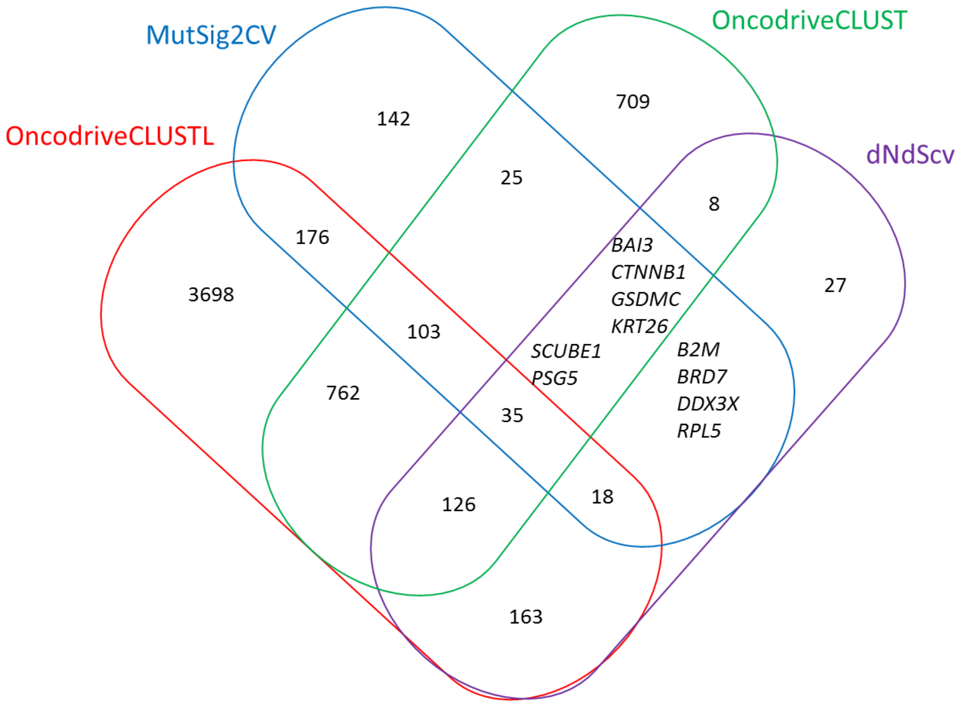


Figure 5-12: Venn-diagram of the potential driver genes identified in skin melanoma using four different bioinformatics programs. *The four different bioinformatics programs are labelled in different colours corresponding to the outline of the closed curves. OncodriveCLUSTL (red), MutSig2CV (blue), OncodriveCLUST (green), dNdScv (purple). The numbers of potential driver genes, based on a false discovery rate q value <0.1, are highlighted are highlighted and, where possible, the names of the specific genes identified by MutSig2CV and one other bioinformatics programe are included in the relevant overlapping sections of the Venn diagram.*

In the melanoma samples, the highest number of genes with a q value of less than 0.1 was recognised using the OncodriveCLUSTL program (figure 5-12). There were 35 genes which were significant at q<0.1 in all four programs (OncodriveCLUST, MutSig2CV, OncodriveCLUST and dNdScv). The lowest number of significant genes with a q value of less than 0.1 were identified using the dNdScv program (i.e. 387 genes). Like previous work in this thesis, a gene was considered a driver gene if it had a q value of less than 0.1 in MutSig2CV and at least one other program. There were 367 driver genes identified this way in the melanoma samples (appendix 7.11.3), which was the highest number of driver genes identified in all three skin cancers investigated in this thesis. Notably, the total number of melanoma samples in the analysis was much higher than the number of BCC and cSCC samples available for analysis in this thesis, which is likely to be the reason why more driver genes were detected in cutaneous melanoma.

5.4 Comparison of skin cancers

As stated earlier in this thesis, UV exposure is the main cause of skin cancer, including for melanoma, BCC and cSCC, and the data on the single base substitution and double base substitution mutations seen in this project supports this view. However, melanoma develops from melanocytes, whereas BCC and cSCC arise from keratinocytes, and even though BCCs and cSCC arise from the same type of cell, there are considerable differences in their clinical behaviour, with cSCC more likely to metastasise than BCC. Therefore, a comparison was conducted in relation to the results of the WES and WGS data for cSCC, BCC and melanoma, focusing on common signalling pathways affected by the mutations and on driver genes.

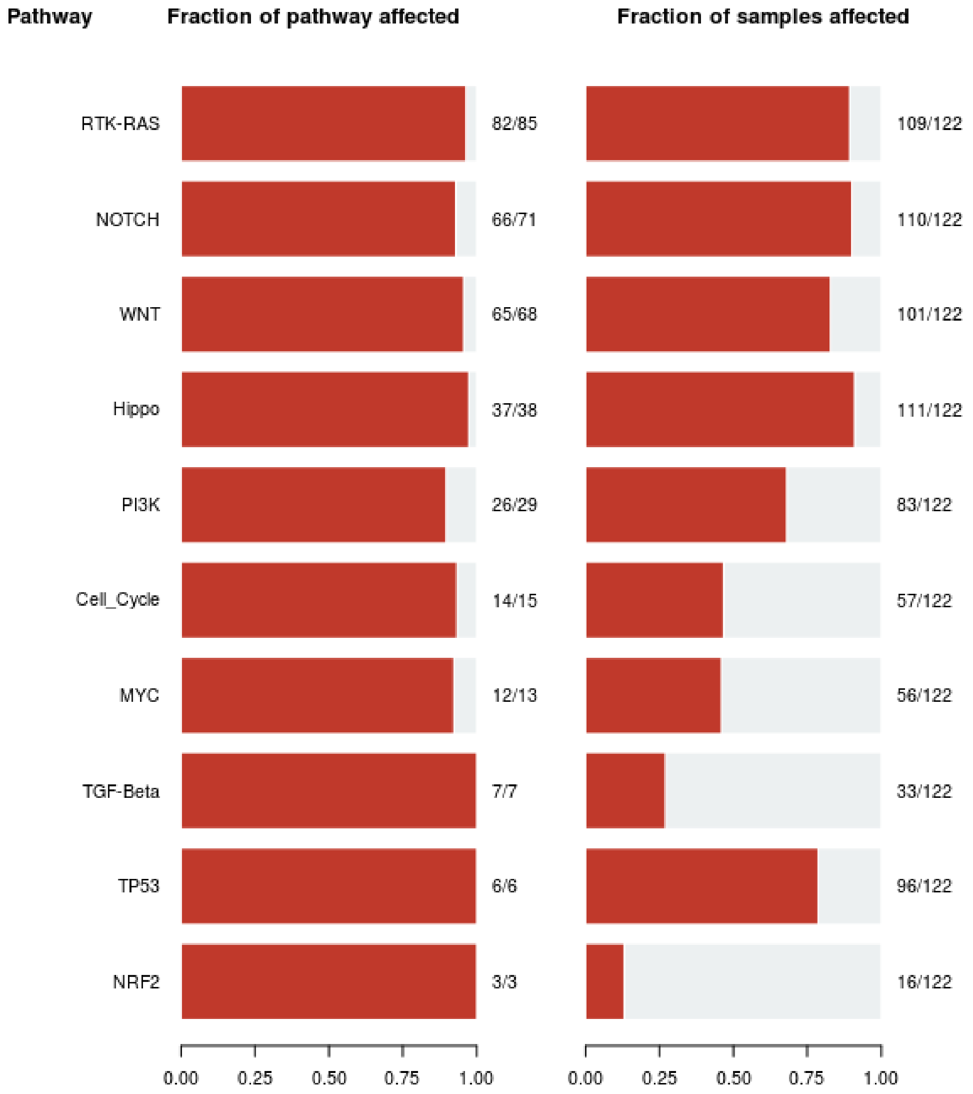


Figure 5-13: The signalling pathways affected in cSCC. *The fraction of genes that are mutated in the relevant signalling pathway is shown on the left and the number of samples which show mutations in these genes is on the right.*

In cSCC, 82 out of 85 genes in the RTK-RAS pathway were mutated in the overall cohort, with mutations in this pathway present in 109 out of a total of 122 samples (figure 5-13). The NOTCH pathway was affected in 110 cSCCs, however mutations affecting this pathway (in 66 of 71 genes) were proportionally slightly less common that those involving the RTK-RAS pathway. The WNT pathway had mutations in 65 of 68 WNT pathway related genes with 101 cSCC samples containing mutations affecting this. Other pathways that were highlighted as being altered due to mutations in relevant genes in the cSCCs included the Hippo, PI3K, MYC, TGF-Beta, TP53 and NRF2 pathways.

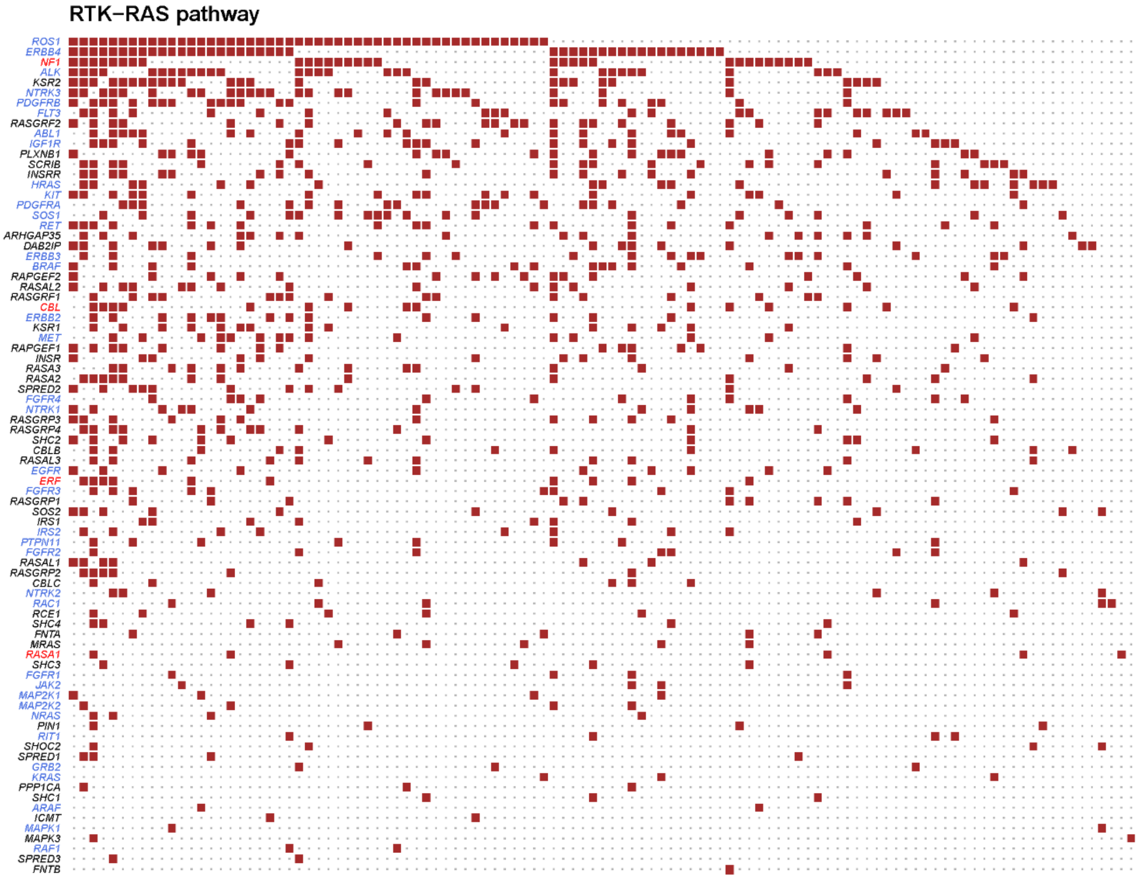


***Figure 5-14: Oncoplot of the RTK-RAS pathway in cSCC**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are depicted in red font, and oncogenes in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

In genes affecting the RTK-RAS pathway, the oncogene *ROS1* was mutated by the greatest number of cSCC samples (figure 5-14). The *ROS1* gene encodes for the receptor tyrosine kinase (RTK) which can activate the RAS pathway and cause cell proliferation (Drosten et al., 2010). There were cSCC samples which did not have mutations in *ROS1* but had mutations in other oncogenes and tumour suppressor genes within the RTK-RAS pathway. The most frequently mutated tumour suppressor gene in the RTK-RAS pathway was *NF1* which, in its wildtype form, converts active RAS-GTP to inactive RAS-GDP thus negatively regulating RAS signalling (Weiss et al., 1999).

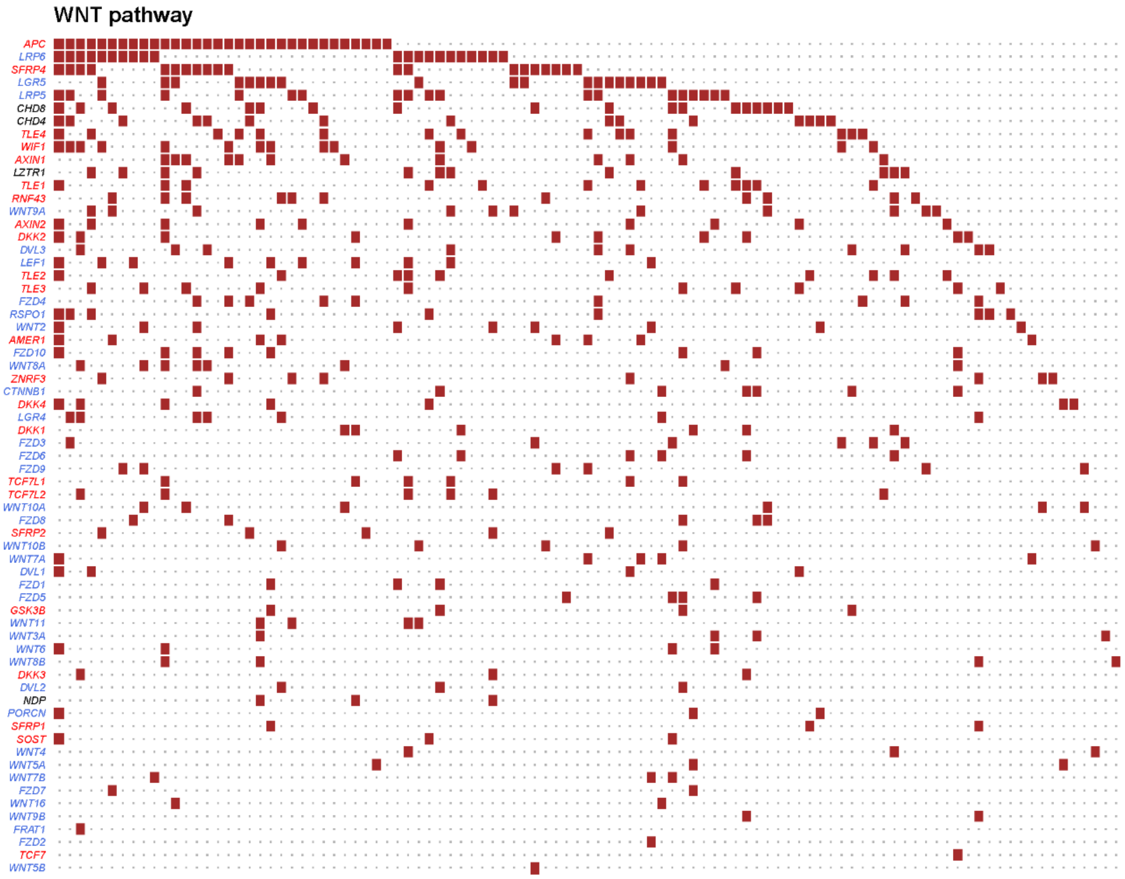


Figure 5-15: Oncoplot of the WNT pathway in cSCC*. The left of the plot shows the genes which have been mutated in the pathway, with tumor suppressor genes highlighted in red font and oncogenes in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

Genes within the WNT signalling pathway were frequently mutated in cSCC, and the gene most mutated in this pathway was the *APC* TSG (mutated in 32 cSCCs, figure 5-15). The *APC* gene encodes for the APC protein which associates with many other proteins, including beta catenin which controls Wnt target gene expression (Li et al., 2012). and cell proliferation; the binding of APC with beta catenin encourages degradation of beta catenin (Eklof Spink et al., 2001). The oncogene which was most often mutated in the cSCC samples was *LRP6*, and there were also 10 samples which contained mutations in both *APC* and *LRP6*. The LRP6 protein is a co-receptor with LRP5 and is responsible for transducing signals thought the WNT pathway. The skin SCC samples contained mutations in more oncogenes than tumour suppressor gene in the WNT pathway as samples share mutations in 23 tumour suppressor gene and 38 oncogenes (figure 5-15).

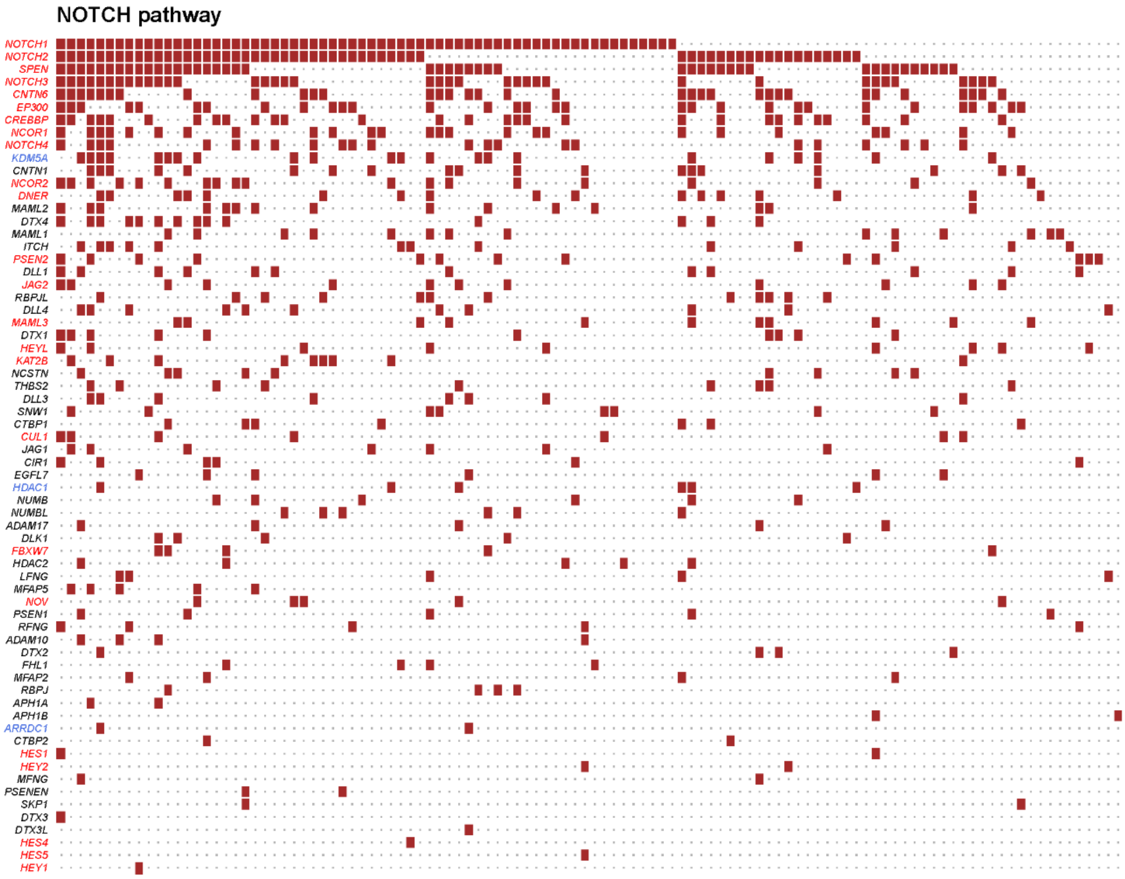


***Figure 5-16: Oncoplot of the NOTCH pathway in cSCC**. The left of the plot shows the genes which have been mutated in the pathway; TSGs are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left. TSG = tumour suppressor gene.*

The nine most highly mutated genes in the NOTCH pathway were tumour suppressor genes (figure 5-16). Of note, *NOTCH1* and *NOTCH2* were the top two mutated genes in the oncoplot of the NOTCH pathway in cSCC and were also identified as driver genes in cSCC in this project (see figure 3-10). There have been studies which consider *NOTCH1* as an oncogene and a tumour suppressor gene, however, in skin, *NOTCH1* has been identified as a tumour suppressor gene (Lobry et al., 2011). In support of this, deletion of *NOTCH1* in a murine study resulted in increased epidermal proliferation and subsequent development of skin tumours (Nicolas et al., 2003). A cluster of samples which did not have mutations in *NOTCH1*, had mutations in either *NOTCH2*, *SPEN* or *NOTCH3*. There were only three oncogenes which were mutated in the NOTCH pathway, namely *KDM5A*, *HDAC1* and *ARRDC1*.

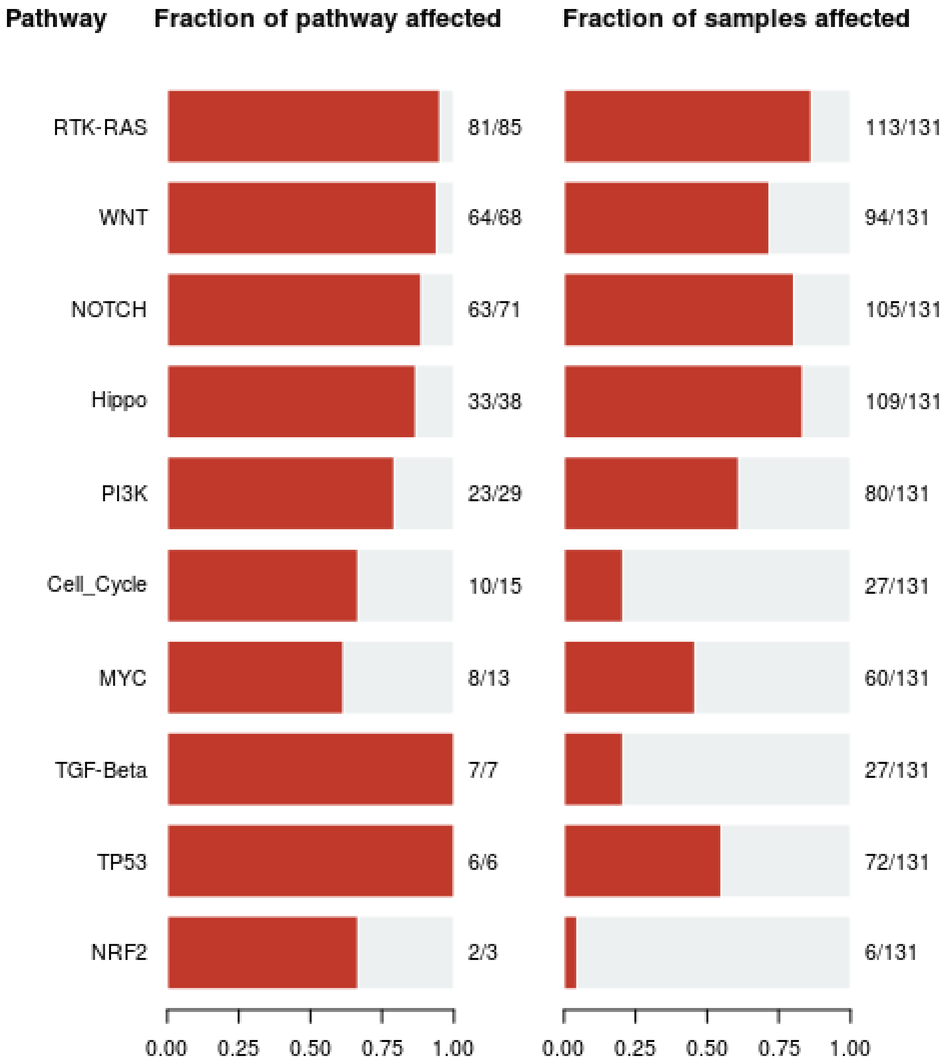


Figure 5-17: The signalling pathways affected in BCC. *The fraction of genes that are mutated in the relevant signalling pathway is shown on the left and the number of samples which show mutations in those genes is on the right.*

Like cSCC, mutations in BCC affected the RTK-RAS, WNT and NOTCH pathways (figure 5-17). The RTK-RAS pathway was impacted in 113 out of 131 BCC tumour samples, with81 of 85 genes in the RTK-RAS pathway mutated in BCC. The WNT pathway was affected in 94 BCC samples with 64 of 68 genes mutated in this pathway and the NOTCH pathway in 105 BCCs with the overall group of BCCs containing mutations in 63 of 71 genes from this pathway. The other pathways which contained mutations in relevant genes in BCC included the Hippo, PI3K, MYC, TGF-Beta, TP53 and NRF2 pathways, which were also affected in the cSCC samples (see figure 5-13).

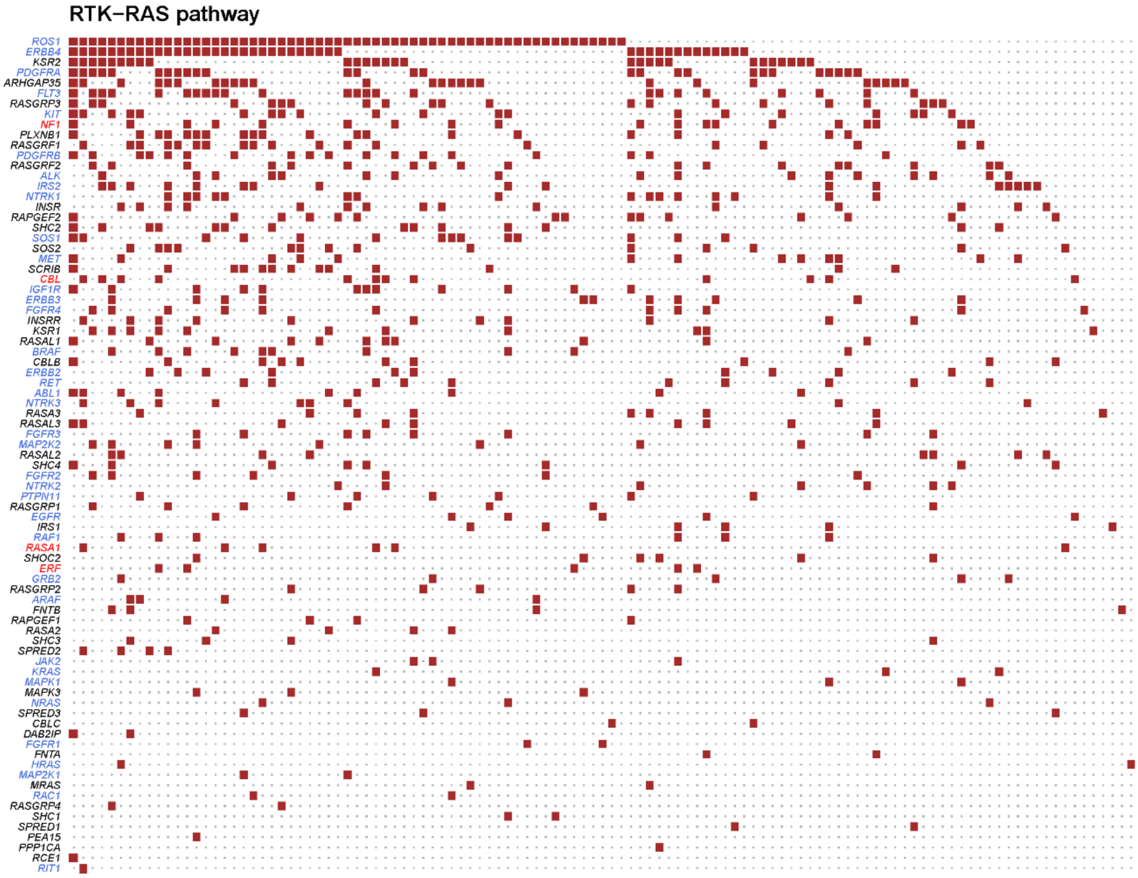


***Figure 5-18: Oncoplot of the RTK-RAS pathway in BCC**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

The two most mutated genes in the RTK-RAS pathway in BCC were the *ROS1* and *ERBB4* oncogenes, which was like that seen in cSCC in relation to this pathway (figure 5-18). Furthermore, as was seen with cSCC, most of the genes in the RTK-RAS pathway in BCC were oncogenes. There were four tumour suppressor genes which were affected in BCC tumour samples in the RTK-RAS pathway, i.e. *NF1*, *CBL*, *RASA1* and *ERF*. Overall, the data showed that there was an amount of heterogeneity in BCCs in relation to the RTK-RAS pathway, with some tumours containing mutations in several/multiple genes whereas other BCCs had mutations in only one or a few of the genes related to this pathway.

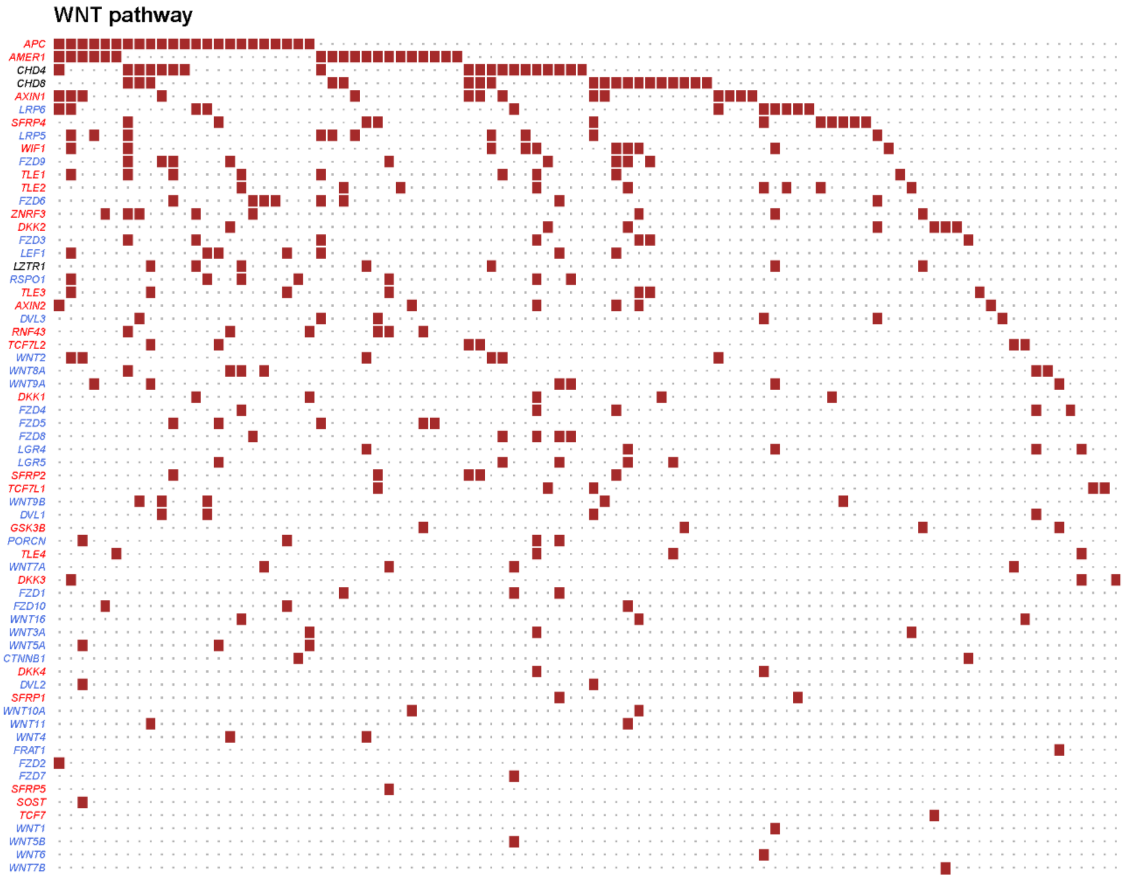


Figure 5-19: Oncoplot of the WNT pathway in BCC. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.

APC was the most commonly mutated gene in the WNT pathway in BCC samples, although this was only in a proportion of BCCs (i.e. 23 tumours), which was similar to cSCC (figure 5-19). However, the second most mutated gene in this pathway in BCC samples was the tumour suppressor gene *AMER1*, in contrast to skin SCC where the second most mutated gene in this pathway was *LRP6*. *AMER1* is an inhibitor of the WNT pathway and induces beta catenin degradation (Tanneberger et al., 2011), therefore, mutations in this gene can result in uncontrolled cell proliferation. There are more oncogenes (37 oncogenes) mutated in the WNT pathway compared to tumour suppressor genes (24 tumour suppressor genes) in BCC tumour samples.

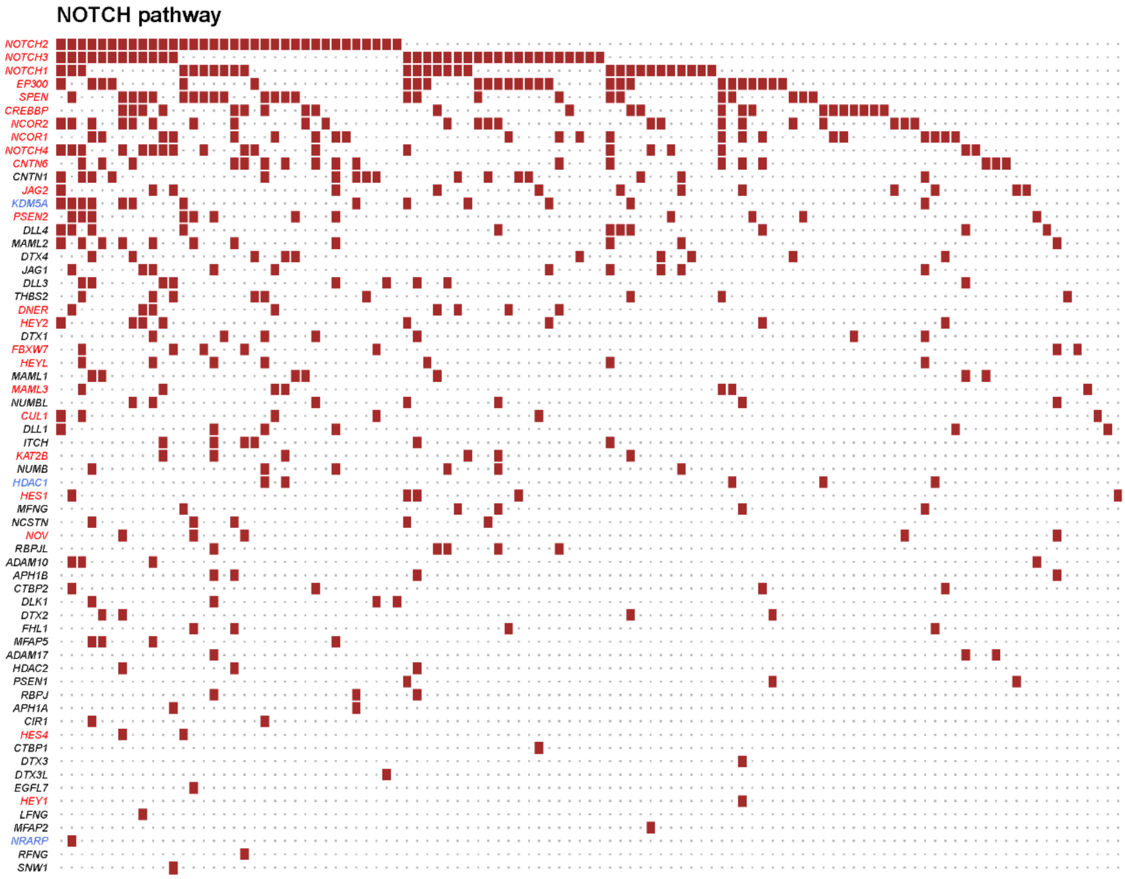


***Figure 5-20**: **Oncoplot of the NOTCH pathway in BCC**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

In relation to the NOTCH signalling pathway, the *NOTCH2* gene was most frequently mutated in the BCC samples, but mutations of NOTCH2, NOTCH3 or NOTCH1 were seen in over half of the cases, and in some tumours two or three of these genes were mutated (figure 5-20). NOTCH signalling has been linked to epidermal cell differentiation in BCC and it has been reported that *NOTCH1*, *NOTCH2* and *NOTCH3* had their highest level of transcription in the basal layer of skin, that is the part of the epidermis where proliferation normally occurs (Thelu et al., 2002). The 10 most highly mutated genes in the NOTCH pathway were tumour suppressor genes, whereas there were only three oncogenes (*KDM5A*, *HDAC1* and *NRARP*) mutated in BCC samples in the NOTCH pathway.

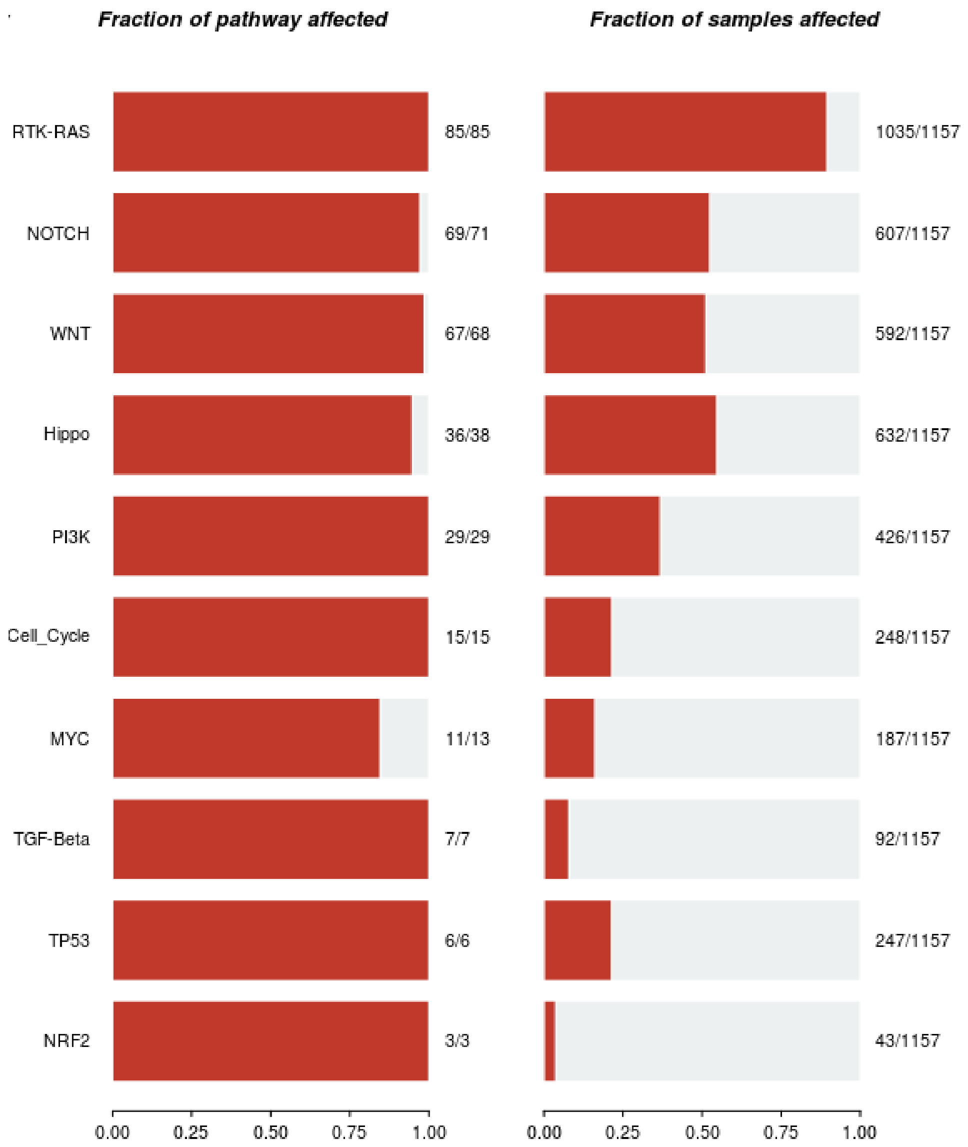


Figure 5-21: The signalling pathways affected in melanoma. *The fraction of genes that are mutated in the relevant signalling pathway is shown as the bar graph on the left and the number of samples which show mutations in those genes is on the right of the figure.*

Interestingly, the signalling pathways that were involved by mutations in the keratinocyte cancers (cSCC and BCC) were also involved in melanoma, with most of the genes in the RTK-RAS, NOTCH and WNT pathways mutated (figure 5-21). However, in melanoma, the RTK-RAS pathway was affected in most cases (1035 of 1157 tumours) whereas the NOTCH and WNT were affected in only 607 and 592 of 1157 melanomas respectively. Although the Hippo, PI3K, MYC, TGF-Beta, TP53 and NRF2 pathways were also affected in some melanomas, the proportion of melanomas with mutations in TP53 (247/1157, figure 5-21) was lower than that seen in cSCC (96/122, figure 5- 13) and BCC (72/131, figure 5-17).

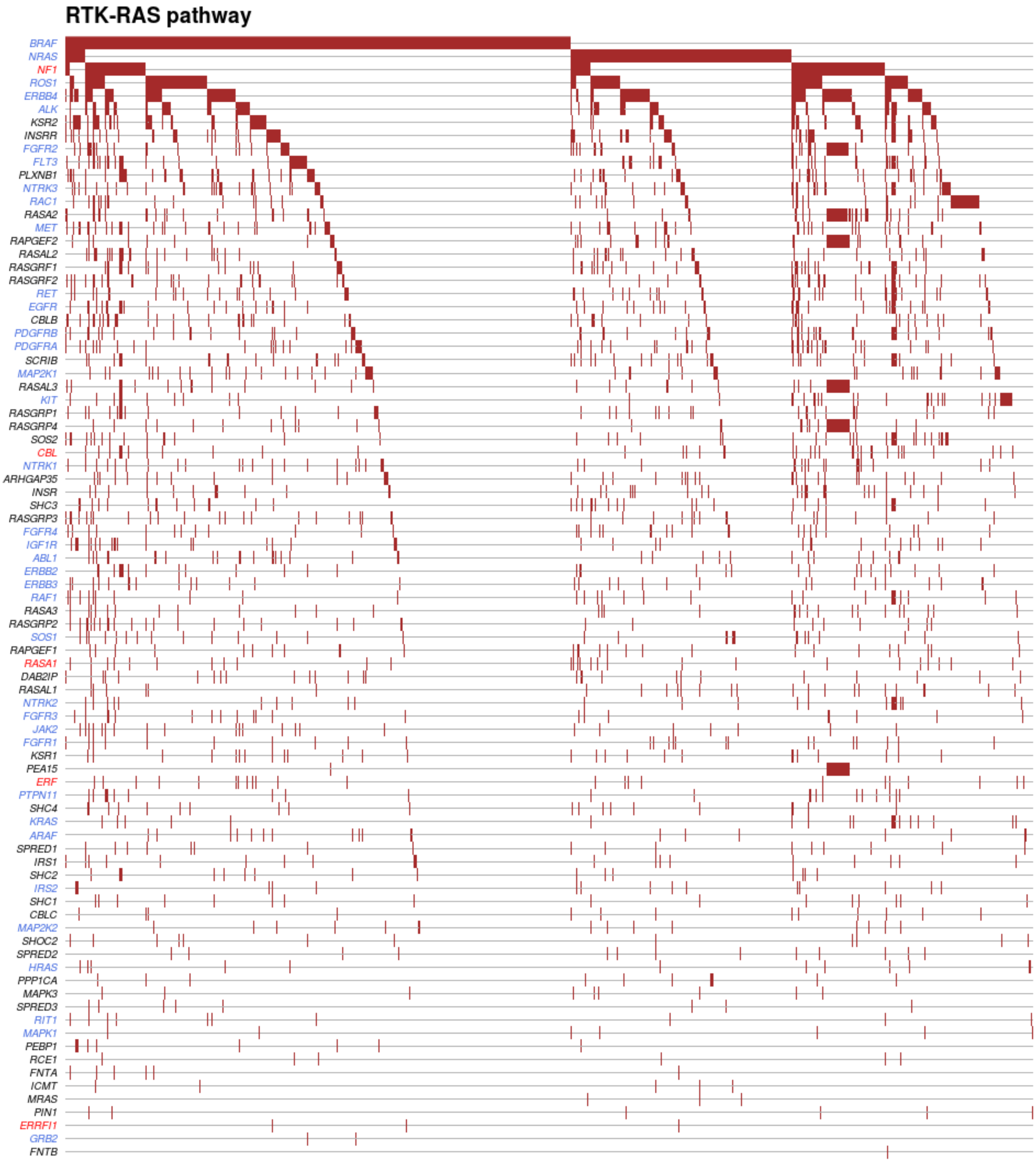


***Figure 5-22**: **Oncoplot of the RTK-RAS pathway in melanoma**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

Genes affecting the RTK-RAS pathway were the most frequently mutated in the melanoma samples (figures 5-21 and 5-22). The most mutated gene in this pathway in melanoma was the *BRAF* gene. Mutations in *BRAF* increase BRAF kinase activity and increases phosphorylation and activates ERK (Yaeger and Corcoran, 2019), thus therapeutic agents have been developed to treat melanoma patients with a *BRAF* mutation (mainly the v600 mutation) (Munoz-Couselo et al., 2015). Many melanomas which did not have mutations in the *BRAF* oncogene had mutations in *ROS1* or *NRAS* oncogenes. The MAPK signalling pathway is activated by RAS proteins such as

NRAS. This RAS activity is controlled by GTPase-activating proteins such as neurofibromin 1 (NF1), which protein converts RAS to its inactive GDP-bound state (Weiss et al., 1999). Perhaps not surprisingly, the most frequently mutated tumour suppressor gene in this pathway in melanoma was NF1

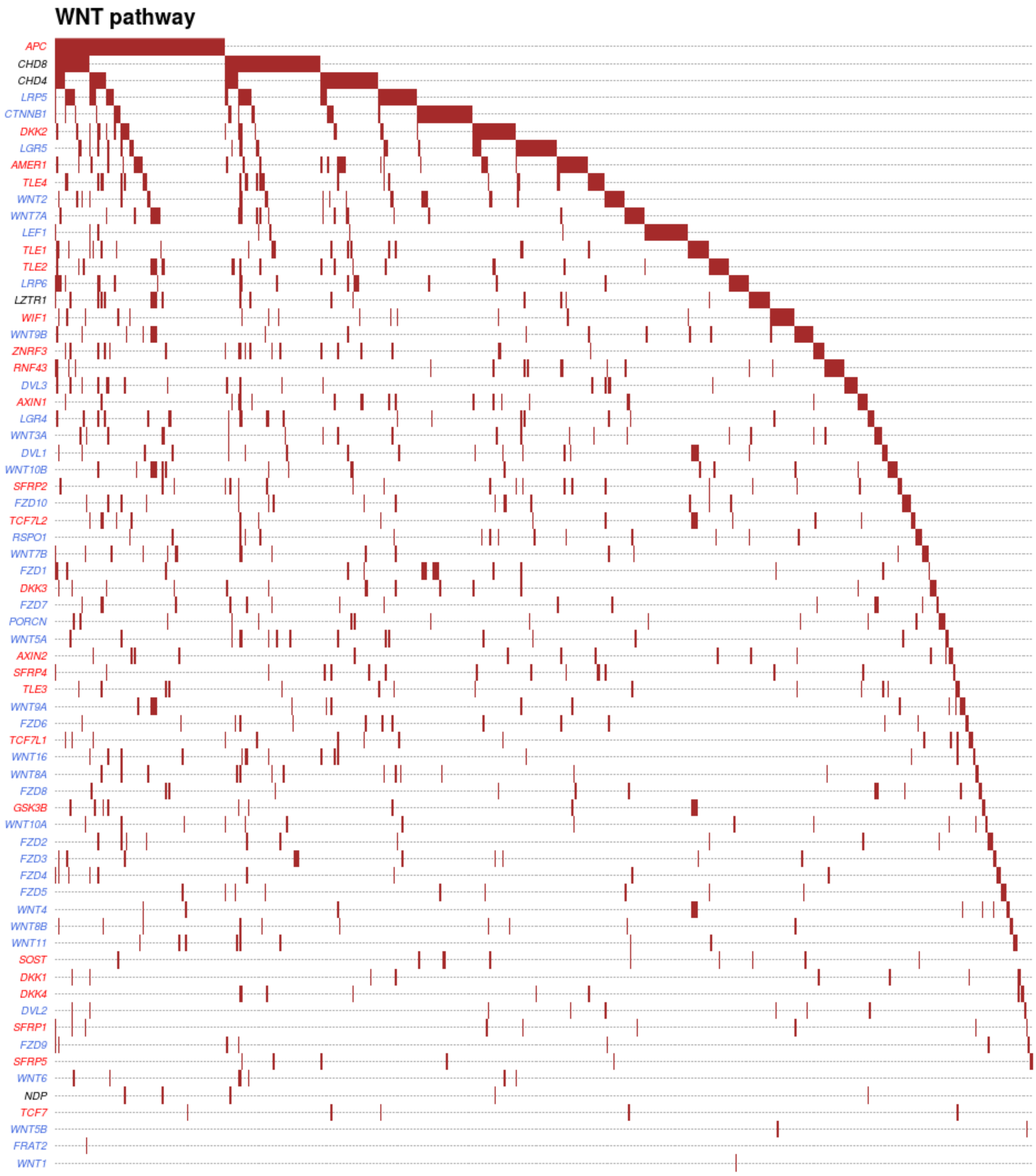


***Figure 5-23: Oncoplot of the WNT pathway in melanoma**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

The *APC* gene was the most mutated gene in the WNT pathway in melanoma, but only in 9% of melanomas (figure 5-23), which is a lower proportion of tumours where *APC* was mutated than in cSCC and BCC. Although the *CHD8* gene was the next most frequently mutated in the WNT pathway in melanoma, in general there were a limited number of tumours with clusters of mutations in any particular gene and most of the tumours exhibited a lot of heterogeneity with respect to mutations in the genes relevant to the WNT pathway. Mutations were frequently observed in oncogenes and tumour suppressor genes affecting this pathway in cutaneous melanoma.

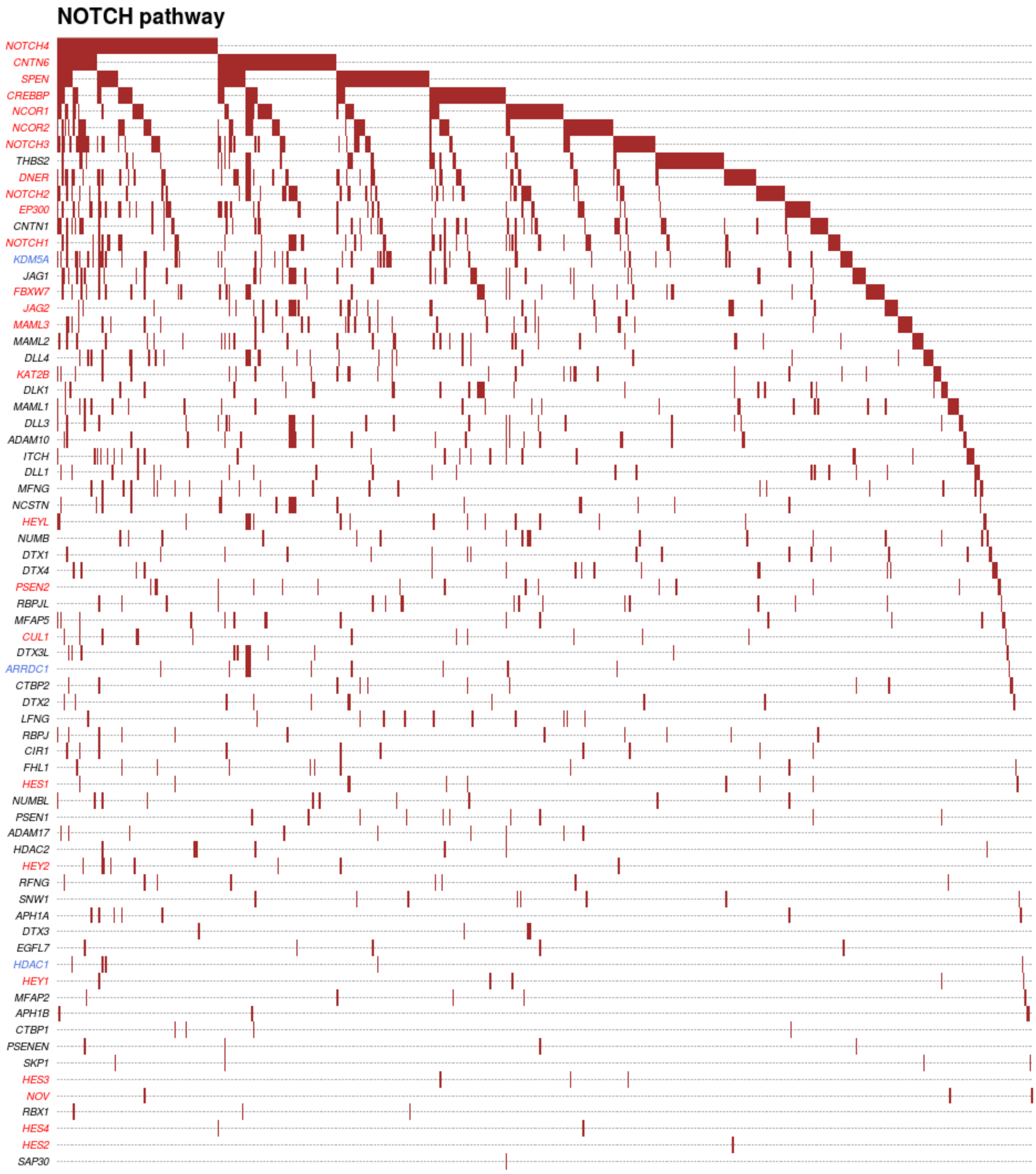


***Figure 5-24: Oncoplot of the NOTCH pathway in melanoma**. The left of the plot shows the genes which have been mutated in the pathway. Tumor suppressor genes are in red, and oncogenes are in blue font. Each coloured square represents a sample which has a mutation in the corresponding gene on the left.*

The most mutated gene in the NOTCH pathway is the tumour suppressor gene *NOTCH4* which is the nineth most mutated gene in skin SCC and BCC. Large clusters of samples in melanoma which do not have mutations in *NOTCH4*, have mutations in other tumour suppressor genes in the NOTCH pathway such as *CNTN6*, *SPEN*, *CREBBP*, *NCOR1*, *NCOR2* and *NOTCH3*. There are only three oncogenes mutated in this pathway for melanoma samples, *KDM5A*, *ARRDC1* and *HDAC1*. *NOTCH4* has shown expression in melanoma cells and triggered a switch from a mesenchymal-like phenotype to an epithelial phenotype and reduced invasive, migratory, and proliferative signalling (Bonyadi Rad et al., 2016). There has also been evidence to suggest that *NOTCH4* exhibited the most significant correlation to *HEY1* expression in head and neck squamous cell carcinoma (Fukusumi et al., 2018). *HEY1* tumour suppressor gene is also mutated in a low proportion of melanomas in figure 5-24.

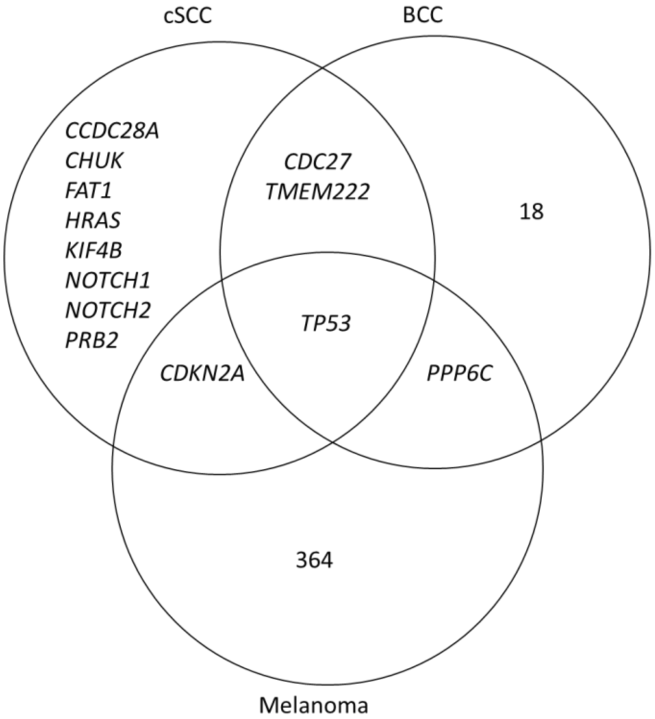


Figure 5-25: Venn diagram of the driver genes identified in the three most common types of skin cancer. *The driver genes or number of driver genes identified in each type of skin cancer are shown and the specific driver genes which are common between skin cancers are shown in the intersecting circles.*

Next, a comparison of the driver genes in the three most common types of skin cancer (BCC, cSCC and melanoma) was conducted. The only driver gene which was common to all three skin cancer types was *TP53* (figure 5-25). There was a much higher number of driver genes identified in the melanoma samples (as documented earlier in this chapter). The *CDKN2A* gene was noted as a driver gene in both melanoma and cSCC, whereas the *PPP6C* gene was a driver gene in melanoma and BCC. In addition to TP53, two other driver genes were common to cSCC and BCC; these were *CDC27* and *TMEM222*. Although BCC and cSCC arise from keratinocytes, the fact that only three driver genes were common to both these types of skin cancer and a greater number of driver genes were not shared between cSCC and BCC, could be seen as part of the reason why these two types of keratinocyte cancer behave different clinically. The limited number of driver genes that were common to melanoma and the other two types of skin cancer was not surprising because they arise from different cell types in the epidermis and the clinical behaviour of melanoma is generally more aggressive than that of keratinocyte cancers.

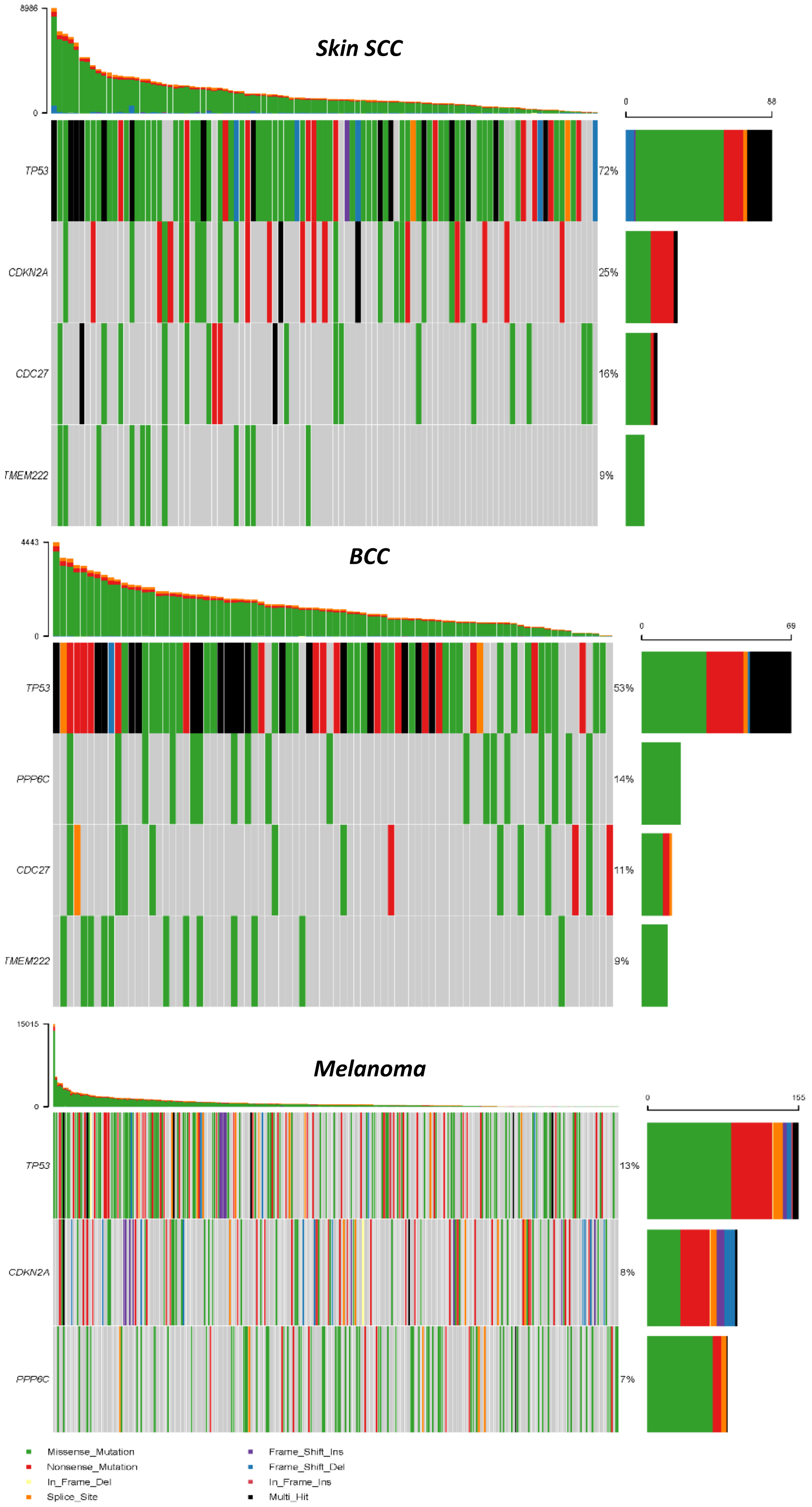


***Figure 5-26: Oncoplot of skin cancer driver genes that were common to at least two of the three different type of skin cancers, i.e. cSCC, BCC and melanoma.**. The number of mutations identified in each tumour across all genes (i.e. not restricted to these driver genes) is presented as a bar chart at the top of type of cancer. Each coloured square represents the type of mutation within the corresponding gene. The bar chart on the right represents the number samples which have a mutation in that gene and the colours represent the type of mutation in the gene.*

In 99 of 122 cSCCs, there were mutations in driver genes that were also identified as driver genes in BCC and/or melanoma. For BCC, mutations in driver genes that overlapped with cSCC and/or melanoma were seen in 82 of 131 samples and for melanoma the driver genes that were common to this tumour and BCC and/or cSCC were mutated in 293 of 1157 samples. While the *TP53* gene was identified as a driver gene in cSCC, BCC and melanoma, and there was some variation in the proportion of the types of mutations (missense, nonsense, multi-hit, etc.) between these tumours, the main difference was that the mutation rate of *TP53* was highest in cSCC, intermediate in BCC and lowest in melanoma (figure 5-26). *TMEM222* was a driver gene in cSCC and BCC and was mutated in 9% of both these cancer types (all of which were missense), but the *CDC27* gene which was a driver gene in cSCC and BCC was mutated slightly more in cSCC than BCC. *CDKN2A* was mutated in 25% of samples in skin SCC and 8% of samples in melanoma was a driver gene common to both these cancers. The *PPP6C* gene, identified as a driver gene in BCC and melanoma, was mutated in 14% of BCCs with all of these missense mutations, *PPP6C* was mutated in a lower proportion of melanomas (i.e., 7%) with some other types of mutations as well as missense. While the mutations in each of the above driver genes often affected different bases and/or codons within and across the different skin cancer types, some of the mutations occurred at identical codons, resulting in identical amino acid alterations, in two or more of the different types of skin cancer (table 5-1).

Table 5-1: Table showing mutations within same codons of individual genes that were identified as driver genes for cSCC and/or BCC and/or cutaneous melanoma.

Gene	Mutation in DNA	Amino acid change in protein	cSCC	BCC	Melanoma
TP53	c.380C>T	p.S127F	Y	Y	Y
TP53	c.476C>T	p.A159V	Y	Y	Y
TP53	c.535C>T	p.H179Y	Y	Y	Y
TP53	c.586C>T	p.R196*	Y	Y	Y
TP53	c.637C>T	p.R213*	Y	Y	Y
TP53	c.722C>T	p.S241F	Y	Y	Y
TP53	c.743G>A	p.R248Q	Y	Y	Y
TP53	c.832C>T	p.P278S	Y	Y	Y
TP53	c.833C>T	p.P278L	Y	Y	Y
TP53	c.844C>T	p.R282W	Y	Y	Y
TP53	c.856G>A	p.E286K	Y	Y	Y
TP53	c.949C>T	p.Q317*	Y	Y	Y
TP53	c.991C>T	p.Q331*	Y	Y	Y
TP53	c.1024C>T	p.R342*	Y	Y	Y
CDKN2A	c.109G>T	p.E37*	Y		Y
CDKN2A	c.152C>T	p.A51V	Y		Y
CDKN2A	c.176G>A	p.W59*	Y		Y
CDKN2A	c.177G>A	p.W59*	Y		Y
CDKN2A	c.188C>T	p.P63L	Y		Y
CDKN2A	c.189C>T	p.P63P	Y		Y
CDKN2A	c.18C>T	p.A6A	Y		Y
CDKN2A	c.19C>T	p.R7*	Y		Y
CDKN2A	c.52G>T	p.E18*	Y		Y
CDKN2A	c.85C>T	p.R29*	Y		Y
CDKN2A	c.97G>A	p.D33N	Y		Y
TMEM222	c.17G>A	p.G6E	Y	Y	
CDC27	c.213C>A	p.C71*	Y	Y	
CDC27	c.332G>A	p.G111D	Y	Y	
CDC27	c.821C>A	p.A274D	Y	Y	
CDC27	c.2072C>T	p.S691L	Y	Y	
PPP6C	c.775C>T	p.P259S		Y	Y
PPP6C	c.790C>T	p.R264C		Y	Y
PPP6C	c.809C>T	p.S270L		Y	Y
PPP6C	c.912C>T	p.F304F		Y	Y
PPP6C	c.913C>T	p.L305F		Y	Y

The sites of mutations for these driver genes in the different types of skin cancer were examined using “lollipop plots”, as this allowed visual comparison of the sites and types of mutations in these genes in the different skin cancer groups (figures 5-27 to 5-31).

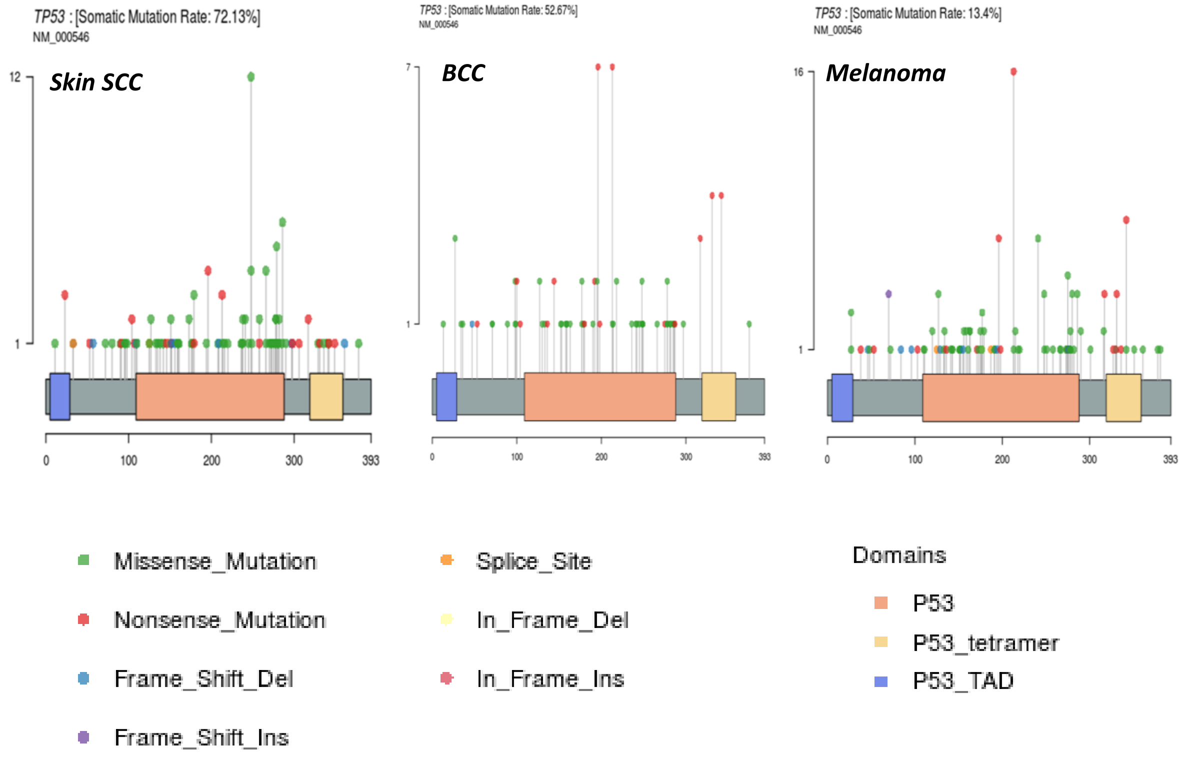


Figure 5-27: Schematic representation of the TP53 protein and the mutations identified in different types of skin cancer. *This has been produced from the transcript numbers specified by NM number 000546 and the amino acid numbers are stated below each schematic representation. The respective protein domains have been labelled as extracted from the Pfam database. The type of mutation are represented by the colours of the circles on the lollipop and show which regions of the protein the gene mutation is affecting. The number of mutations in each region of the gene is represented by the height of the lollipop.*

The mutations in *TP53* gene were scattered across the different protein domains in each of the three types of skin cancer (figure 5-27). Most of the mutations were missense, with nonsense mutations the second most frequent, in the three skin cancer categories. Based on the comparison of the lollipop plots, there were no major differences in the site or type of mutation between cSCC, BCC and melanoma that would suggest that the *TP53* gene was likely to behave differently as a driver gene in any of the three types of skin cancer.

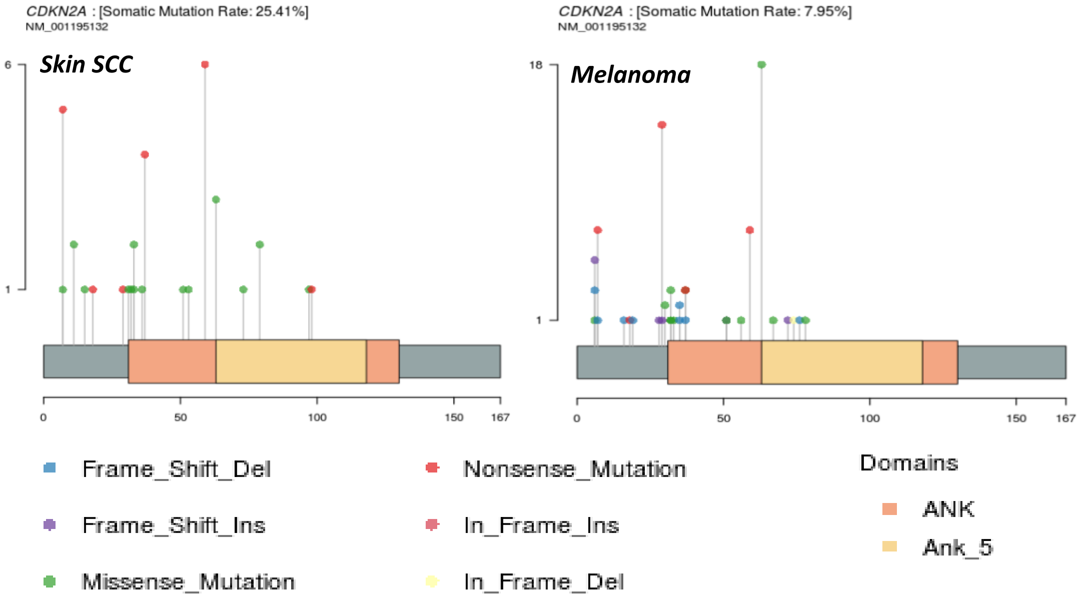


Figure 5-28: Lollipop plot of CDKN2A demonstrating site and type of mutations in skin SCC and melanoma. *Figure was produced from transcript NM 001195132 and the amino acid numbers are shown below the schematic. The respective protein domains have been labelled as per the Pfam database. The type of mutations are represented by the colours of the circles on the lollipop and show which regions of the protein the gene mutation affects. The number of mutations in each region of the gene is represented by the height of the lollipop.*

Most of the mutations in *CDKN2A* were scattered along the first half of the p16 protein from the N-terminus, including in the ANK and Ank_5 domains, in cSCC and in melanoma (figure 5-28). The mutations in cSCC were nonsense or missense whereas in melanoma there were frame-shift deletions, frame-shift insertions, in-frame insertions, and in-frame deletions.

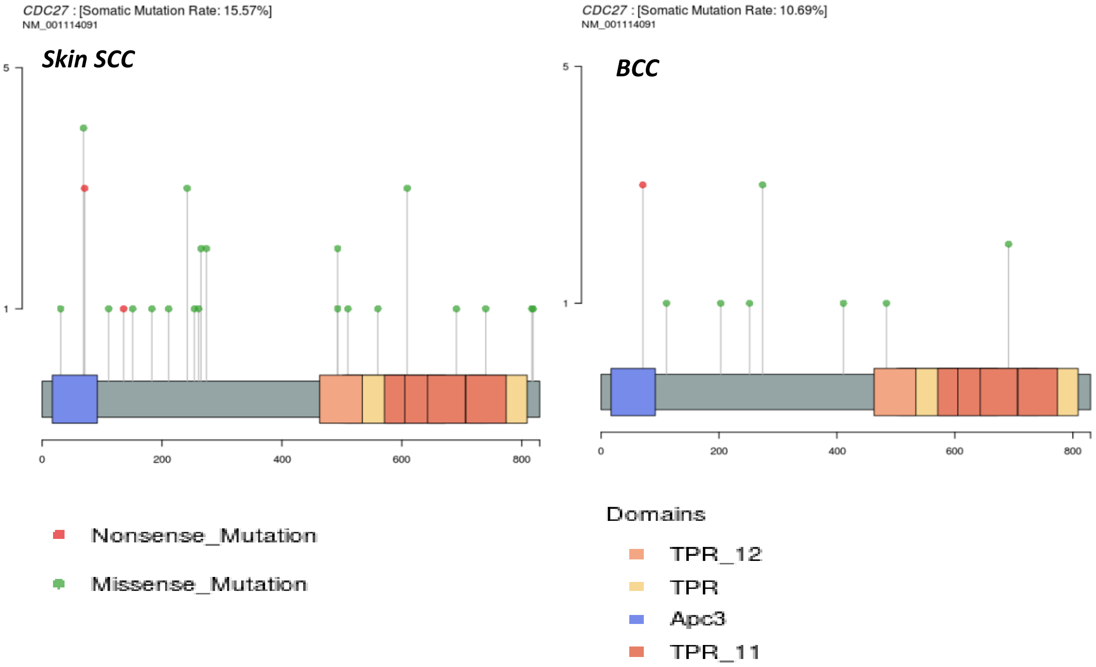


Figure 5-29: Schematic representation of the CDC27 protein and the mutations identified in cSCC and BCC. *Schematic produced from transcript NM 001114091. The amino acid numbers of the protein are stated beneath the protein and the respective protein domains labelled according to data extracted from the Pfam database. The type of mutations are represented by the colours of the circles on the lollipop and show which regions of the protein are affected by the respective gene mutations. The number of mutations in each region of the gene is represented by the height of the lollipop.*

CDC27 was more frequently mutated in cSCC (15.57%) than in BCC (10.69%). Although the mutations were scattered across the gene, the TPR domains were also affected more commonly in cSCC than in BCC.

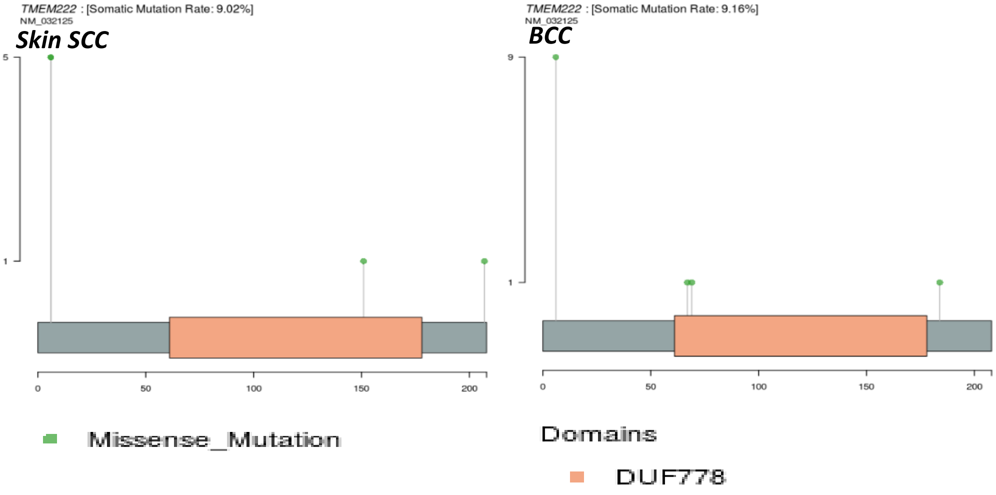


Figure 5-30: Representation of the TMEM222 protein and mutations identified in cSCC and BCC. *Transcript NM 032125 was employed for the schematic representation, with the amino acid numbers written beneath the protein. The respective protein domains have been labelled as per the Pfam database. The type of mutation (all of which were missense) is represented by the green coloured circles on the lollipops and demonstrate the regions of the protein affected by the mutations. The number of mutations at each site are represented by the height of the lollipop.*

The *TMEM222* was mutated in 9% of cSCCs and in a similar percentage of BCCs. All mutations in this gene in both types of keratinocyte cancer were missense, with several of these at the N-terminal region, and some occasionally in the DUF778 domain of the protein (figure 5-30).

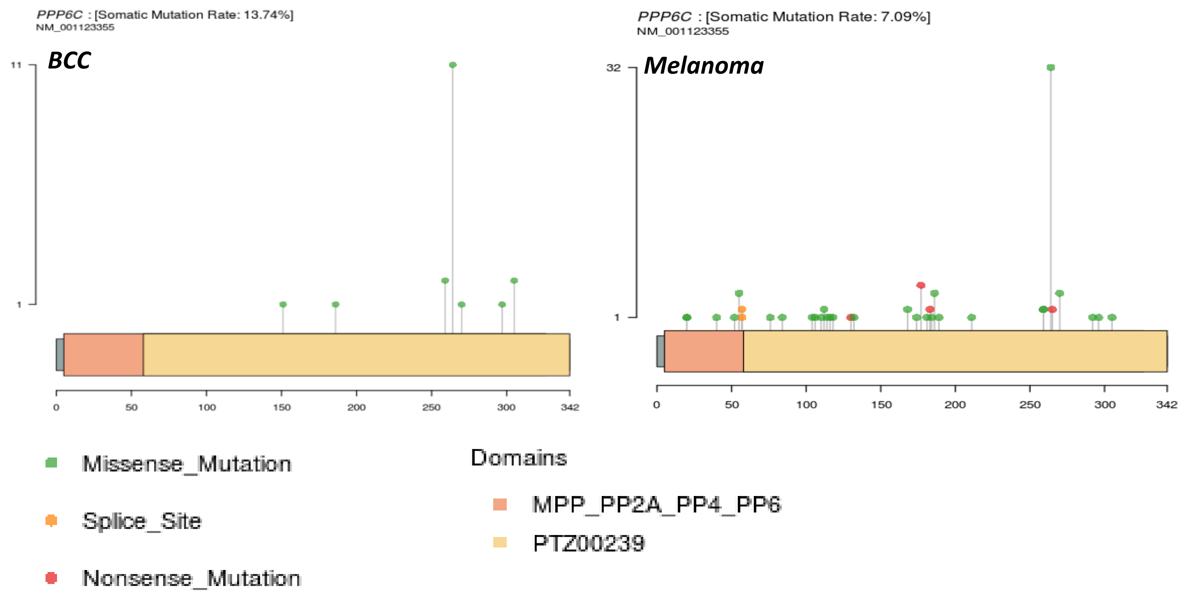


Figure 5-31: Schematic of the PPP6C protein and the mutations in BCC and melanoma. *This was produced from transcript NM 001123355 and shows the amino acid numbers underneath the protein. The protein domains have been labelled according to the Pfam database and the type of mutation are indicated by the colours of the circles on the lollipop. The number of mutations is represented by the height of the lollipop and the site of the lollipops shows the sites of the protein affected by the respective mutation.*

Mutations in *PPP6C* were scattered more across the protein in melanoma than in BCC (figure 5-31). Whereas all the mutations in this gene were missense in BCC, and the majority of mutations in the gene were missense in melaoma, there were some nonsense and splice site mutations in melanoma also. All the mutations in BCC and most of the mutations in melanoma affected the PTZ00239 domain of the PPP6C protein, but some mutations in melanoma modified the MPP P2A PP4 PP6 domain.

While the vast majority of BCC, cSCC and melanoma skin cancers are treated by surgical excision, it is worth looking at whether the mutations identified in the overall WES and WGS data for these skin cancers would offer opportunities for drug treatment of the more limited numbers of skin cancers that could not be adequately excised clinically. Using the Drug Gene Interaction database allows one to estimate whether certain mutated genes might be treatable by medical agents, i.e. termed "druggable". Therefore the "druggable" effects of the mutations in these cancers were assessed using all genes which had variants.

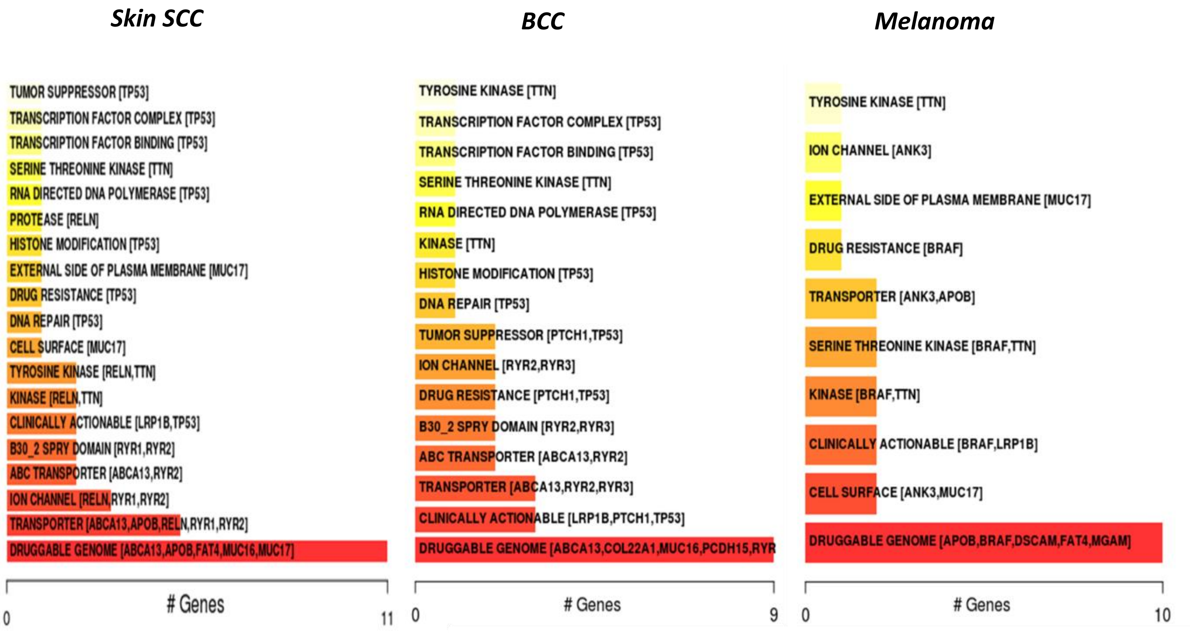


Figure 5-32: Graphs representing the drug-gene interactions of genes which are mutated in skin cancer. *The x axis represents the number of genes which have mutations in the druggable pathways and the most commonly mutated genes in each pathway are shown in the graph. The drug information has been compiled from the Drug Gene Interaction database.*

Figure 5-32 shows the "druggable" gene categories for the mutations in cSCC, BCC and melanoma (Griffith et al., 2013). The graphs show up to the top five genes which are in each druggable gene categories. The *MUC16* gene is mutated in skin SCC and BCC and MUC17 gene is mutated in skin SCC and melanoma and has been identified as a druggable part of the genome. *ABCA13* has been identified as only in skin SCC and BCC and has been categorised as a transporter druggable gene category. *APOB* is a transporter that has been mutated in skin SCC and melanoma. In BCC clinically actionable genes include *LRP1B*, *PTCH1* and *TP53* which is similar to skin SCC which includes *LRP1B* and *TP53*. The least number of genes are involved in the tyrosine kinase druggable pathway in BCC and melanoma.

5.5 Potential pre-malignant skin lesions and chronically sun-exposed skin

To understand which UV-induced mutations (including sun-induced mutations) in human skin might lead to the subsequent development of skin cancer, it is useful to compare the driver gene mutations in cSCC, BCC and melanoma with the mutations that have been reported in potential pre-malignant skin lesions and in chronically sun-exposed skin. While this type of comparison does not mean that the presence of mutated driver genes in skin will lead to skin cancer development, the observation of a mutated driver gene in UV-exposed skin, potential pre-malignant lesions and skin cancers suggests that this type of gene mutation in UV-exposed skin may be more deleterious than a mutation in a gene that is not seen as a driver gene in skin cancer. This, in turn, might allow one to assess how much risk is associated with repeated exposure to NB-UVB (and/or to natural sunshine) by focussing on the number of mutated driver genes that arise in the skin following repeated UV exposure. Therefore, studies in the literature on next-generation sequencing of potential pre-malignant lesions were obtained using search terms as shown in appendix 7.9.1 for actinic keratoses (AKs) and data from known studies for melanocytic naevi and chronically UV-exposed skin were identified. Then cSCC, BCC and melanoma WES and WGS analyses was used to check for the presence of mutated driver genes (as identified in this thesis) in those potential pre-malignant lesions and of chronically UV-exposed skin samples.

5.5.1 Actinic Keratosis

Studies which reported on WES of AKs were ascertained from the published literature; searches were performed up to 24th January 2020. The publications that were identified as containing WES data on AKs comprised of seven samples from Chitsazzadeh et al., 2016 and five samples from Albibas et al., 2018. A larger study of AKs was identified while the AK analysis was being conducted and 30 samples from Thomson et al., 2021 was included, which in total contained 42 AKs that had undergone WES investigations, and which provided WES data on 42 AKs. No studies were identified which included WGS information on AKs.

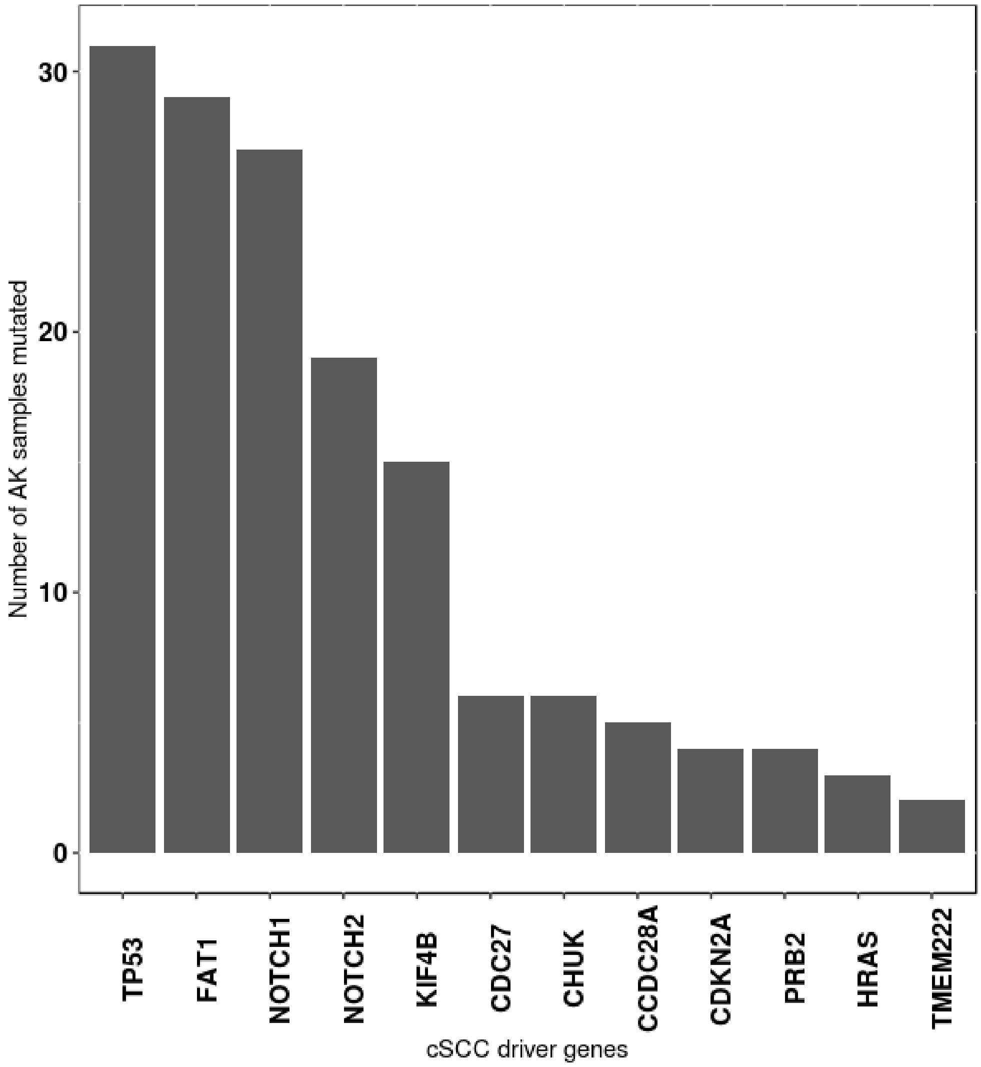


Figure 5-33: The cSCC driver genes which were mutated in actinic keratosis (AK) WES data that was obtained from Chitsazzadeh et al., 2016, Albibas et al., 2018 and Thomson et al., 2021 studies. *The graph shows the number of AK samples which have mutations in the respective cSCC driver genes.*

Due to the limited number of AK samples, driver genes in AKs were not investigated using the four bioinformatics programs (i.e. MutSig2CV, OncodriveCLUST, OncodriveCLUSTL and dNdScv) as had been undertaken for the different types of SCCs and other types of skin cancer earlier in this thesis. Instead, the AK data was examined to determine whether the AKs showed mutations in the driver genes that had been identified in cSCCs (figures 3-10 and 5-25), because of the previous genetic evidence that cSCCs can arise from AKs (Albibas et al., 2018). All the 12 cSCC driver genes were noted to be mutated in this cohort of AKs, with variation in the frequency of the individual driver genes that were mutated in the AKs (figure 5-33). The *TP53* gene was mutated in the greatest number of AK samples and *TMEM222* is mutated in the least amount of AK samples. The top five most mutated driver genes in AKs were *TP53*, *FAT1*, *NOTCH1*, *NOTCH2* and *KIF4B*. The top five most mutated cSCC driver genes which were mutated in AKs were mutated in 15 to 31 samples.

5.5.2 Melanocytic naevi

Although many melanocytic naevi (commonly known as “moles”) are benign and never develop into a melanoma (Tsao et al., 2003), some melanomas arise from pre-existing melanocytic naevi (Pampena et al., 2017). Therefore, known published literature was identified which reported on WES of melanocytic naevi. One publication with WES on melanocytic naevi, which contained 30 samples, was identified (Stark et al., 2018). The WES data on melanocytic naevi was examined to identify whether the driver genes that had been identified in cutaneous melanoma in this thesis were mutated in the melanocytic naevi.

Table 5-2 List of melanoma driver genes identified in melanocytic naevi. *The table shows the number of samples which have mutations in these driver genes.*

Melanoma driver genes	**Number of samples mutated (Stark *et al.*, 2018)**	**Melanoma driver genes**	**Number of samples mutated (Stark *et al.*, 2018)**	**Melanoma driver genes**	**Number of samples mutated (Stark *et al.*, 2018)**
BRAF	16	GRM3	3	KLF12	2
MYH7	16	ITSN1	3	LEPR	2
ACTC1	11	KIAA2022	3	LHCGR	2
HYDIN	10	MYH1	3	LRP2	2
XIRP2	10	MYH2	3	MYBPC1	2
PTPRB	9	MYLK	3	OR52J3	2
PCDH15	8	MYOM3	3	PCDHA4	2
PTPRT	8	NLRP11	3	PDE4DIP	2
TRRAP	7	NLRP4	3	PDE8B	2
EPHA7	6	NLRP5	3	PDZD2	2
APOB	5	PDE7B	3	PHKA1	2
BRWD1	5	PDE9A	3	PLCH1	2
C6	5	SCN10A	3	PPP1R13L	2
DNAH6	5	SCN1A	3	SALL1	2
DSG3	5	SEC23B	3	SETD5	2
LAMA2	5	SH3RF2	3	SI	2
THSD7B	5	SNCAIP	3	SLC15A2	2
CNTNAP2	4	TIGIT	3	SNX31	2
DCC	4	VCAN	3	SPAG17	2
DSG4	4	ADAMTS18	2	STAB2	2
KALRN	4	ADCYAP1R1	2	TMC5	2
KCNQ5	4	BMP5	2	TRHDE	2
MXRA5	4	C1orf168	2	UGT1A3	2
NEBL	4	CAPN6	2	USP29	2
NFASC	4	CD1C	2	ZFX	2
NLRP13	4	CD300E	2	ACSBG1	1
PAK7	4	CDH2	2	ADAM22	1
PCDH18	4	CEACAM6	2	ADAM7	1
PLCE1	4	CHD6	2	ADH1A	1
ZNF536	4	COL17A1	2	ALPK2	1
ALPPL2	3	COL5A2	2	ANKRA2	1
BCLAF1	3	CYP7B1	2	ANO4	1
BMPER	3	DNAH2	2	AP1M1	1
CBL	3	FCRL5	2	ARHGAP21	1
CNTN5	3	FMO3	2	ARMC4	1
CSMD3	3	GPR179	2	ASTN1	1
DNAH3	3	GRIN3A	2	C2CD3	1
ERC2	3	KCNH5	2	C9	1

Melanoma driver genes	**Number of samples mutated (Stark *et al*., 2018)**	**Melanoma driver genes**	**Number of samples mutated (Stark *et al*., 2018)**	**Melanoma driver genes**	**Number of samples mutated (Stark *et al*., 2018)**
CD22	1	MTR	1	TMEM156	1
CDH6	1	MYO9A	1	TP63	1
CDH7	1	MYOCD	1	TSKS	1
CEACAM5	1	N4BP2	1	TTC3	1
CEP63	1	NLRP8	1	TUBA3C	1
CHGB	1	NLRP9	1	UGT2B4	1
COL3A1	1	NRAS	1	WDR76	1
COL7A1	1	NRXN3	1	ZNF365	1
CRB1	1	OR13C8	1	ZNF385D	1
DCAKD	1	OR4D5	1	ZNF667	1
DDX17	1	OR51S1	1	ZNF804A	1
DMBT1	1	OR8D2	1		
DMXL2	1	OTC	1		
EFEMP1	1	PAH	1		
FGD6	1	PCDHA12	1		
FILIP1	1	PCDHA2	1		
GM2A	1	PCDHB7	1		
GML	1	PDE11A	1		
GRID2	1	PLCB4	1		
IL2RA	1	PMFBP1	1		
ITGB3	1	POLN	1		
ITGB6	1	PROL1	1		
ITPR2	1	PTEN	1		
KCNQ3	1	PTPRH	1		
KDSR	1	RAC1	1		
KHDRBS1	1	RHAG	1		
KIAA1109	1	RPRD2	1		
KIF2C	1	RQCD1	1		
KIF5A	1	SELP	1		
KLHL20	1	SEMG2	1		
KRT26	1	SETD2	1		
KRTAP5-10	1	SLC16A9	1		
LGR6	1	SLC46A3	1		
LIPI	1	SLC9A4	1		
MAGI1	1	SRGAP3	1		
MKX	1	SUN5	1		
MME	1	TDRD1	1		
MPP7	1	TEX15	1		

Table 5-2 shows that 201 of the melanoma driver genes identified in chapter 5, section3.2, were mutated in the melanocytic naevi. There were five melanoma driver genes which were mutated in 10 or more samples of melanocytic naevi. The *BRAF* gene was the melanoma driver gene that was mutated in the largest number of melanocytic naevi (i.e., 16 naevi). There were 100 melanoma driver genes which were each mutated only once in the cohort of melanocytic naevi.

5.5.3 Normal skin

Over recent years, some studies that used next generation sequencing to look for genetic mutations in chronically sun-exposed skin have been reported. Therefore, known studies for all relevant publications that reported on next generation sequencing in skin were identified. Three studies which had investigated normal skin in an unbiased way were identified, including Martincorena et al., 2015, Lynch et al., 2017 and Fowler et al., 2021; the study by Albibas et al., 2018 had focussed on p53 immunopositive patches, rather than on entire skin samples. None of the studies had undertaken WES or WGS on normal skin, but the studies by Martincorena et al., 2015 and Fowler et al., 2021 conducted targeted sequencing for 74 genes and Lynch et al., 2017 had conducted targeted sequencing on 121 genes. The skin samples in Martincorena et al., 2015 were from the eyelid skin of four individuals and spanned 0.8 – 4.7mm^2. The skin samples from Lynch et al., 2017 were from the head and neck region of 10 individuals and the skin samples from Fowler et al., 2021 were from different body sites of 35 patients. While it is accepted that WES or WGS data is preferable to targeted sequencing, it was considered that it would be better to use the data from these three normal skin studies rather than excluding all normal skin data from this thesis. As normal skin contains keratinocytes and melanocytes, the data was examined to see whether driver genes from cSCC, BCC and melanoma were mutated in chronically sun-exposed skin.

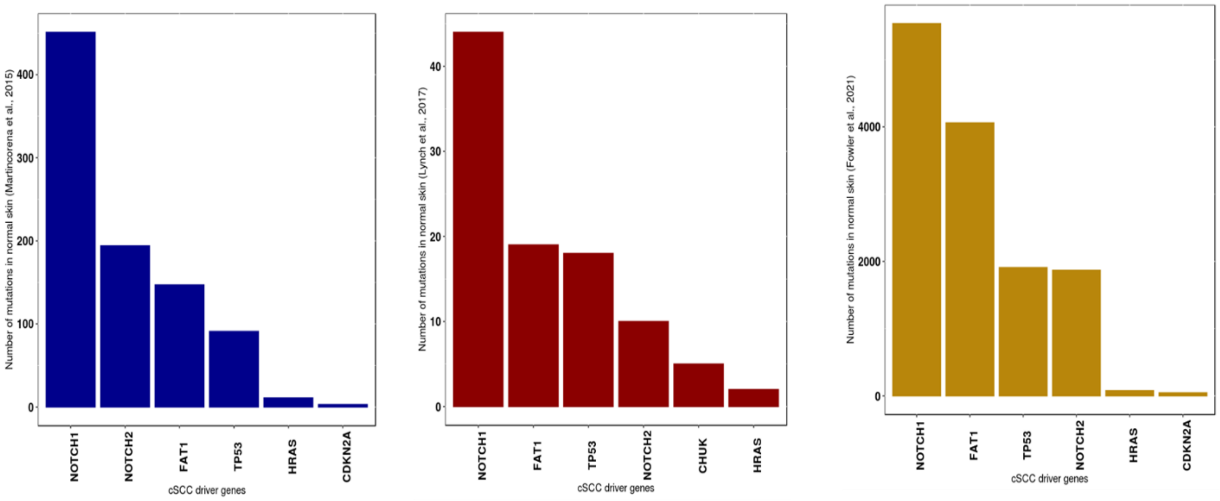


__Figure 5-34: Three graphs showing the cSCC driver genes which were mutated in 0.8 – 4.7mm^2 sections of normal skin from the Martincorena et al., 2015, 16mm^2 sections of normal skin from Lynch et al., 2017 and 2mm^2 sections of normal skin from Fowler et al., 2021 studies__. The graphs show the total number of mutations identified in each of these cSCC driver genes that were identified in normal skin.

The cSCC driver gene that was most frequently mutated across all three cohorts of normal skin was *NOTCH1* (figure 5-34). In addition to *NOTCH1*, the *TP53*, *NOTCH2*, *FAT1* and *HRAS* cSCC driver genes were mutated in normal skin in all three cohorts. *HRAS* was less commonly mutated than *NOTCH1*, *FAT1*, *TP53* and *NOTCH2* across all these cohorts. While the *CDKN2A* gene was mutated infrequently in the Martincorena et al., 2015 and Fowler et al., 2021 studies, *CDKN2A* gene capture failed across all samples in the Lynch et al., 2017 study and so this gene was excluded from the analysis in the Lynch et al., 2017 paper. The *CHUK* cSCC driver gene was only mutated in the Lynch et al., 2017 study, however, this gene was not included in the targeted sequencing conducted in Martincorena et al., 2015 and Fowler et al., 2021.

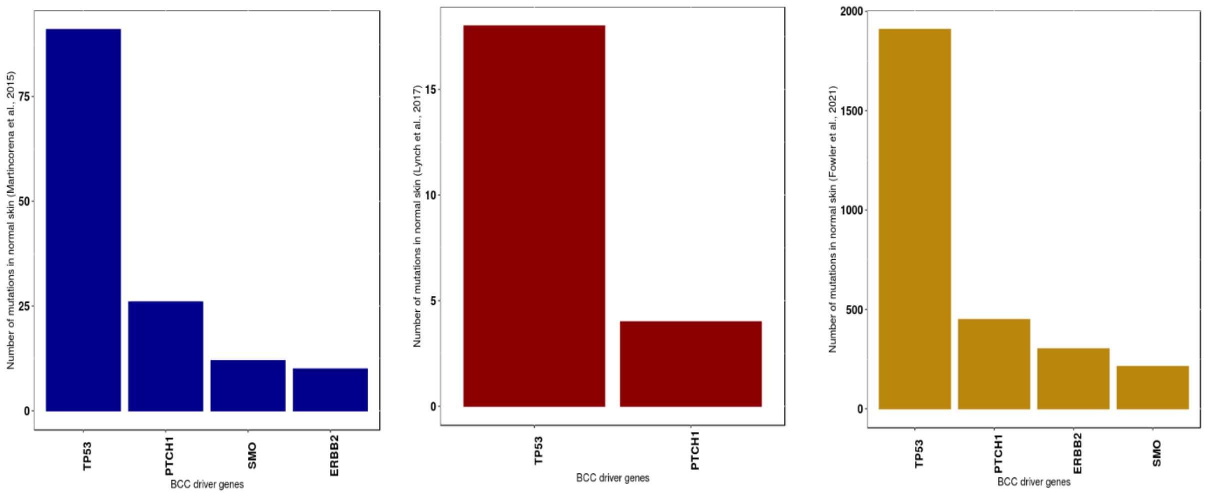


__Figure 5-35: BCC driver genes which were mutated in 0.8 – 4.7mm^2 sections of normal skin from the Martincorena et al., 2015, 16mm^2 sections of normal skin from Lynch et al., 2017 and 2mm^2 sections of normal skin from Fowler et al., 2021 studies__. The graphs show the total number of mutations identified in each of these BCC driver genes that were identified in normal skin.

TP53 was the most commonly mutated BCC driver gene in all three cohorts of normal skin (figure 5-35). In addition to *TP53*, the PTCH1 gene which was a BCC driver gene was also mutated in all three normal skin groups. The *ERBB2* and *SMO* BCC driver genes were mutated less frequently in the Martincorena et al., 2015 and Fowler et al., 2021 samples and were not sequenced in Lynch et al., 2017.

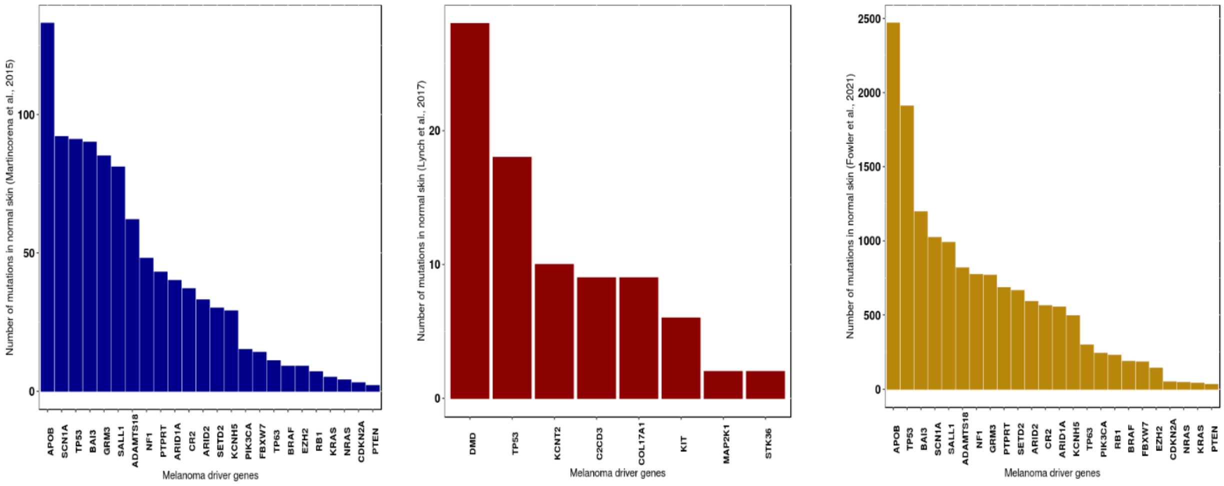


Figure 5-36: The melanoma driver genes which were mutated in 0.8 – 4.7mm² sections of normal skin from the Martincorena et al., 2015, 16mm² sections of normal skin from Lynch et al., 2017 and 2mm² sections of normal skin from Fowler et al., 2021 studies*. The graphs show the total number of mutations identified in each of these melanoma driver genes that were identified in normal skin.*

The next generation targeted sequencing data from the normal skin samples was next examined for mutations in melanoma driver genes. Admittedly, melanocytes constitute only a small fraction of the cells in the epidermis, so it is not possible to know whether the driver gene mutations in the Martincorena et al., 2015, Lynch et al., 2017 and Fowler et al., 2021 studies were present in the melanocytes, but it was thought sensible to look for the presence of these melanoma driver genes, nonetheless. Multiple melanoma driver genes were noted to be mutated in normal skin (figure 5-36). *TP53* which was a driver gene in melanoma, as well as in cSCC and BCC, was mutated in all three normal skin studies. The *APOB* melanoma driver gene was more frequently mutated than *TP53* in the Martincorena et al., 2015 and Fowler et al., 2021 data, whereas in the Lynch et al., 2017 study, the *DMD* gene is more commonly mutated than *TP53*. While *NF1* was mutated fairly frequently in the Martincorena et al., 2015 and Fowler et al., 2021 papers, *BRAF* and *NRAS* mutations were much less common in these cohorts.

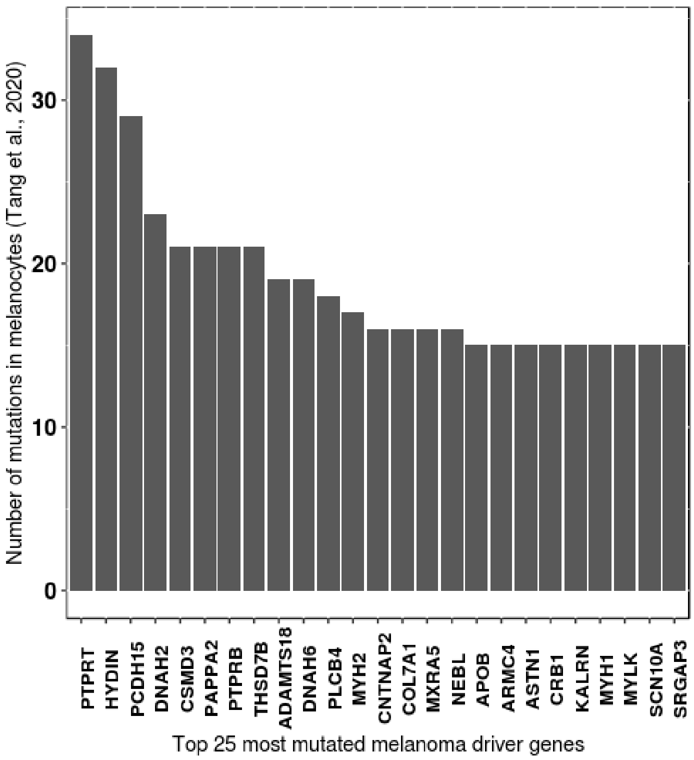


Figure 5-37: The top 25 most frequently mutated melanoma driver genes in all the melanocytes from the Tang et al., 2020 study. *The bars represent the number of mutations from all the melanocytes in the top 25 most mutated melanoma driver genes.*

As stated earlier, the vast majority of cells in the epidermis are keratinocytes, making it difficult to know whether melanoma driver genes that were seen to be mutated in whole epidermis studies were mutated in keratinocytes or melanocytes. Fortunately, Tang et al., 2020 conducted a study where they isolated melanocytes from skin of people aged 63 to 85 years old and then cultured the melanocytes to expand the number of cells from each melanocyte in order to investigate for mutations. Admittedly, the culture of the melanocytes might have allowed the introduction of additional mutations during cell division, although the authors did compare their results with those from cultures of neonatal melanocytes which would not have been exposed to UV *in vivo*, and concluded that the mutations detected in the adult skin were likely to have arisen *in vivo* rather than in vitro (Tang et al., 2020). There were 322 melanoma driver genes which were mutated in the melanocytes from the Tang et al., 2020 study. Figure 5-37 shows the 25 melanoma driver genes with the most mutations. The two most commonly mutated driver genes (*PTPRT* and *HYDIN*) in these melanocytes were also identified as mutated driver genes in melanocytic naevi samples in table 5-2. The *PTPRT* gene was also mutated in normal skin the Martincorena et al., 2015 and Fowler et al., 2021 studies shown in figure 5-36.

5.6 Proposed adjunct to future pipeline for assessment of carcinogenicity of repeated UV exposure during phototherapy for skin disease

It is unclear at this stage whether skin that receives repeated UV exposure over a short-term period, for example during a course of NB-UVB for skin disease or during a “sun-holiday” where people travel to a sunny climate to sunbathe, will contain many (or a few or no) *de novo* mutations because to date no studies have been conducted to investigate for this. However, it is hoped that the identification of driver genes in the three common types of skin cancers and in potentially pre-malignant lesions and normal skin in this thesis may help to delineate whether mutations from repeated UV exposure over a short-term period are likely to be carcinogenic, that is whether they are present in driver genes rather than in other genes, including passenger genes. Therefore, figure 5-38 and the accompanying list of genes in table 5-2 and appendix 7.11.3 and 7.11.4 highlights multiple genes that have been identified as driver genes in skin cancer, and whether they have been previously shown to be mutated in normal skin and potentially pre-malignant lesions, so that these can be used as an assistance for future studies investigating mutations arising from repeated UV exposure during phototherapy and/or as a result of a “sun-holiday”.

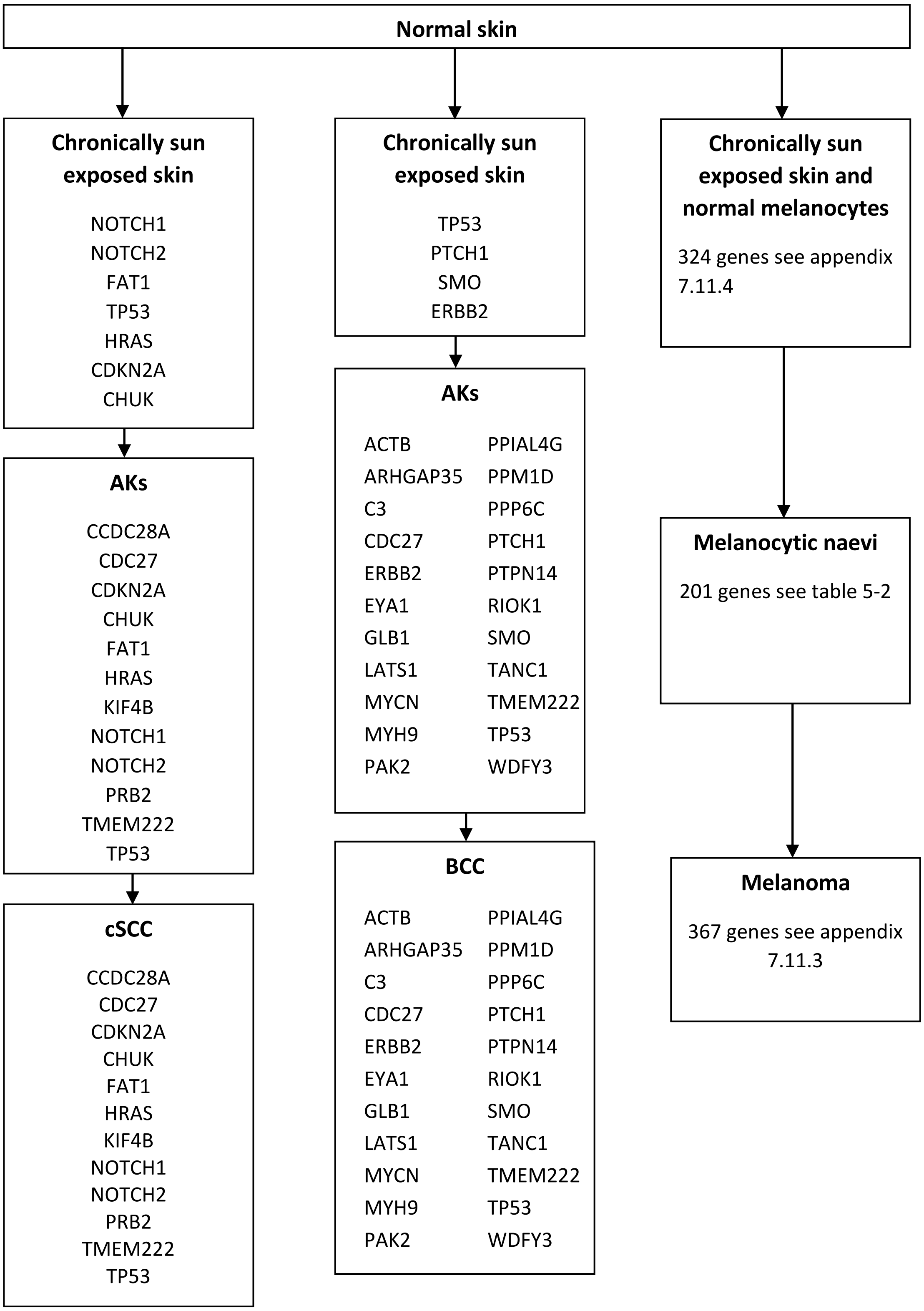


Figure 5-38: Driver genes which have been identified as mutated in sun-exposed skin, precancerous lesions and skin cancers in this thesis. *The driver genes have been listed separately in a flowchart style set of pathways according to whether they are likely to be relevant to development of cSCC, BCC and melanoma.*

5.7 Discussion

The comparison between the three most common types of skin cancer in this chapter highlights that the highest number of skin cancer samples which have undergone WGS, and WES are melanomas. Fewer BCCs and cSCCs have been investigated using WES and WGS. This is probably the main reason why more driver genes were identified in melanoma, especially as cSCCs and BCCs have a higher mutation burden than cutaneous melanoma (Jayaraman et al., 2014, Pickering et al., 2014). As would be expected from the fact that skin is frequently exposed to UV during normal daily living and UV has been reported to be associated with the development of skin cancer in epidemiological studies (D'Orazio et al., 2013), the mutations in BCC and melanoma in this chapter (as well as the mutations in cSCC in chapter 3) were mainly C>T alterations. Related to this, the single base substitution mutation signatures and double base substitution mutation signatures in BCC, melanoma and cSCC suggested that UV was likely to play a considerable role in the development of each of these different types of skin cancer.

However, there was some variation in the mutation signatures in the three types of skin cancer. The SBS7b mutation signature was identified in cSCC and BCC, whereas melanoma and BCC shared the SBS7a mutation signature. While one could argue that this might simply have been due to the limited number of samples of BCC and cSCC that had been studied using WES and/or WGS, the fact that the SBS7a but not the SBS7b mutation signature was seen in melanoma whereas the SBS7b mutation signature was seen in keratinocyte cancers could suggest that the mutational effects of UV differ according to the cell type receiving UV exposure. Indeed, the presence of SBS7a and SBS7b in BCC but not in melanoma nor in cSCC could also suggest that BCC is caused by a combination of UV signatures, but future research will be necessary to determine whether this is the case. All three types of skin cancer shared the DBS1 mutation signature associated with UV which suggests that UV can cause these double base changes in cSCC, BCC and melanoma.

The only driver gene that common to all three types of skin cancer was *TP53*. This is not surprising because *TP53* has been reported to be mutated in many different types of cancer (Olivier et al., 2010). However, only a small number of driver genes were common to cSCC and BCC and/or to cSCC and melanoma and/or to BCC and melanoma. While this might have partially resulted from the limited number of cSCCs and BCCs included in the analysis (as above), this lack of common driver genes might also explain why the three types of skin cancer behave differently in the clinical setting. The risk of metastasis is more frequent in melanoma than in keratinocyte cancer, and is higher in cSCC than in BCC (Nguyen, 2004), which may result from variations in the numbers or types of driver genes that are mutated in these different cancer types. Despite the lower numbers of cSCCs and BCCs that could be included in the comparison with melanoma in this chapter, a number of driver genes that were detected were unique to cSCC and, separately, to BCC. It is unlikely that all of these driver genes in cSCC and BCC were false positives because they had to have a q value <0.1 in MutSig2CV and at least one of the three other bioinformatic programs (dNdScv, OncodriveCLUSTL, OncodriveCLUST) to be classified as a driver gene, thus it seems likely that they were "true" driver genes in these cancers. Consequently, the fact that they did not appear in the list of driver genes in melanoma, which included 1157 tumour samples, suggests that they do not function as driver genes in cutaneous melanoma and that genuine differences exist in the driver genes that lead to the development of keratinocyte cancer and melanoma in skin.

Admittedly, *CDKN2A* gene was seen as a driver gene in cSCC and melanoma, however this is a common cancer gene, reported as a driver gene in other cancer types and is not unique to skin cancer (Bailey et al., 2018). There were certain mutations which were present in this gene in cSCC and melanoma, but future research similar to the analyses in this thesis would need to be conducted to see whether these mutations within *CDKN2A* are unique to skin cancer or are seen in many types of cancer. The *PPP6C* gene was identified as a driver gene in BCC and melanoma, affecting 14% of BCCs and 7% of melanomas, but does not seem to have been reported in the published literature as mutated in many other types of cancer. High expression of *PPP6C* has been associated with poorer survival in glioblastoma multiforme (Leone et al., 2012) whereas expression of PPP6C protein has been noted as a favourable prognostic marker in renal cancer (https://www.proteinatlas.org/ENSG00000119414-PPP6C/pathology).

MUC16 was identified as one of the most frequently mutated genes in melanoma and BCC. When identifying most frequently mutated genes, genes were only excluded if they were not in the Cancer Gene Census. *MUC16* was only identified as frequently mutated gene in melanoma and BCC and was not considered a driver gene in these cancers.

Mucins are characterised by their tandem repeat regions and the number of tandem repeats differ depending on the Mucin gene (Haridas et al., 2014). The number of repeats in the tandem repeat region of *MUC16* can give rise to different isoforms of *MUC16* which can result in functional heterogeneity (Haridas et al., 2014). It is unclear whether (all) *MUC16* mutations identified by sequencing in BCC and melanoma are factually correct, because of a study which investigated the DNA sequence of *MUC2* and *MUC6* and its repetitive nature (Svensson et al., 2018). This latter study identified difficulty in sequencing repetitive regions using short read sequencing and resolved the DNA sequence of these two genes via long read sequencing. Due to long-read sequencing recently being used to sequence cancer genomes, it is highly likely that Illumina short-read sequencing was used in the BCC and melanoma analyses and *MUC16* has large repeat regions, so it is possible that the high mutation frequency of this gene in melanoma and BCC could be due sequencing artefacts (Slatko et al., 2018). This high mutation frequency of *MUC16* could be the reason that it was not identified as a cancer driver gene as it may be identified as frequently mutated but is not mutated higher than expected due to chance using the MutSig2CV algorithm. *MUC16* could also have more synonymous mutations compared to non-synonymous which is why it is not identified as a driver gene using dNdScv algorithm. OncodriveCLUST and OncodriveCLUSTL considers genes as driver genes if they have clustered mutation, the results showed that *MUC16* was a significant driver gene for melanoma and BCC using this algorithm. This suggests that *MUC16* has clustered mutations in BCC and melanoma. This study only considered a gene as a driver gene if it was significant in MutSig2CV and one other program and *MUC16* was not identified as significant in MutSig2CV so was not considered a cancer driver gene for either BCC or melanoma.

The use of short-read sequencing can result in errors to alignment in repetitive regions, therefore this could be the reason for the clustering of mutations and the reason why *MUC16* was considered a driver gene using the OncodriveCLUST and OncodriveCLUSTL programs. The use of long-read sequencing would ensure these repeat regions are sequenced correctly as there would no longer be misalignment of short-reads to this region. Long-read sequencing would also enable the clarification of the mutation frequency of *MUC16*.

MUC16 is however, commonly expressed in ovarian cancers and is associated with disease progression (Aithal et al., 2018, Thomas et al., 2021). There has also been a study which investigated the association of *MUC16* with tumour mutation burden and its prognostic effect on cutaneous melanoma (Wang et al., 2020a). It was suggested that *MUC16* appeared to be a useful predictive marker of tumour mutation burden and patient survival in melanoma.

The BCC samples that were used in this study were from the Bonilla et al., 2016 study which identified seven significantly mutated genes using MutSig2CV: *TP53*, *PTCH1*, *PTPN14*, *MYCN*, *RPL22*, *SMO* and *PPIAL4G* as significantly mutated genes. Six of these genes were considered cancer driver genes in BCC however *RPL22* was not. The Bonilla study does not provide information on the most frequently mutated genes in BCC; therefore this thesis provides further information about genes which are frequently mutated in this cancer type. A gene could be frequently mutated due to its genetic characteristics such as repeat regions and the gene could have many mutations that do not have a large pathogenic effect on the protein structure so is not identified as a driver gene in these cancers but is still considered as frequently mutated. Therefore, further research needs to be conducted to elucidate the pathogenicity of all the frequently mutated genes identified these cancers.

Several the genes mutated in cSCC, BCC and melanoma affected similar signalling pathways, including the RTK-RAS, NOTCH and WNT pathways. While this might suggest some common mechanisms in the development of the three types of skin cancer, it is worth bearing in mind that the use of BRAF inhibitors, which affects the RTK-RAS pathway, has been beneficial for treatment of melanoma but has also been shown to be associated with development of cSCC (Wu et al., 2017). This indicates that one needs to be cautious in avoiding overinterpreting the involvement of similar signalling pathways in the different forms of skin cancer, because it is unclear whether alterations (including minor differences in alterations) in these signalling pathways could have considerably different effects in melanocyte and keratinocytes, and indeed at different stages along the development pathways for melanoma, BCC and cSCC.

As stated above, melanoma was the skin cancer that was most commonly sequenced using WES and WGS, therefore combining the data from all of those studies meant that there was more power to detect driver genes in melanoma in this thesis than in cSCC and BCC. This highlights a need for more cSCC and BCC tumour samples to be sequenced in the future to validate the results of the driver genes identified in keratinocyte cancers in this thesis and to allow more robust investigations for and comparisons of driver genes common to cSCC, BCC and melanoma. As can be seen from the data on normal skin, limited data exists on the mutations that exist in chronically sun-exposed skin and a much greater number of samples of normal skin will need to be investigated in the future to determine the extent of cancer driver gene mutations and their function in contributing to the early stages of skin tumorigenesis.

6. Discussion

This thesis was ultimately involved with identifying driver genes in the three most common types of skin cancer (BCC, cSCC and cutaneous melanoma), and to determine whether mutations in those driver genes have been reported in normal skin and potentially pre-malignant skin lesions, in order to assist future studies that would assess the likely carcinogenicity of phototherapy and short-term repeated exposures to strong summer sunshine (e.g., during a "sun-holiday"). In the skin cancer work and potentially precancerous skin lesions, WES and WGS data were analysed because this was considered the strongest way to take an unbiased scientific approach to identify driver genes. As WES and WGS data was not available for normal skin, it was accepted that the use of targeted next generation sequencing would allow some insight into the presence of driver genes in chronically sun-exposed skin.

However, another important question that was addressed along the way, while learning the relevant bioinformatic skill, was whether cSCC contained similar driver genes to SCCs in other organs. This is because similarities in the driver genes in cSCC and SCCs of other organ types might have the potential to improve our understanding of the biology of cSCC (by extrapolating from what is known about these driver genes in the other SCCs) and the possibility of allowing future treatments for SCCs of other organs to be considered in aggressive cases of cSCC (or in clinical trials on treatment of aggressive / metastatic cSCCs). While chapter 3 was mainly about learning the bioinformatic skills and methodology, it allowed the author to use those skills to examine cSCCs in detail, thus providing an initial platform from which to build the rest of the work in the thesis.

Combining the data and results from chapter 3 with the data in chapter 4 showed the similarities and differences between cSCC, oropharyngeal SCC, oesophageal SCC, lung SCC and cervical SCC. When the top 100 frequently altered genes were compared between the SCCs, 27 genes containing mutations were common between all five types of SCCs. However, the genes which were altered in all five types of SCCs differed in the proportion of samples which had mutations in these genes within each type of SCC. For example, *ABCA13* was mutated in a proportion of each of the five different categories of SCCs, but 61% of cSCC samples had mutations in this gene whereas it was mutated in 6% of oropharyngeal SCCs, 11% of lung SCCs, 5% of oesophageal SCCs and 5% in cervical SCCs. This is just one example of many to suggest that cSCC contains some similarities in its genetic mutations to SCCs of internal organs, yet is also quite different to the other SCCs.

While at first glance, some of the DNA base changes that were identified in the comparative analysis between the five different types of SCCs may look broadly similar, the DNA base changes did also vary somewhat between the different SCCs. This is most probably due to the different carcinogens which cause the different types of SCCs, for example cSCCs had the highest proportion of C to T changes which were likely due to UV exposure (Brash, 2015) but also had C to A changes due to smoking, whereas in lung SCCs, the highest proportion of changes were C to A changes because smoking is the biggest cause of this cancer (Alexandrov et al., 2016). There were differing proportions of C to T changes in each of the internal categories of SCCs but this was conceivably as a result of spontaneous deamination of methylated cytosine residues which varies in tissues due to the differing rates of replicative turnover (Alexandrov et al., 2015, Blokzijl et al., 2016).

The mutation signatures similarly reflected the varying carcinogenic exposures which contribute to the development of each of the individual cancers, including the SCCs in chapter 4 and the three types of skin cancers analysed in chapters 3 and 5. Formerly, the single base substitution signature 7 (SBS7) related to UV as the carcinogen (Alexandrov et al., 2013, Inman et al., 2018), but more recent analysis of signature 7 in melanoma and all of the mutation signatures in a larger number of cancers identified four subtypes of signature 7, namely SBS7a, SBS7b, SBS7c and SBS7d (Hayward et al., 2017, Alexandrov et al., 2020). However, for the different skin cancers, there was variation in which cancers exhibited mutation signatures 7a and 7b in this thesis. UVB and UVA can form cyclobutane pyrimidine dimers (CPDs) (Rochette et al., 2003), therefore the different SBS7 signatures could reflect the mutational patterns caused by differing wavelengths of UV. Yet, this interpretation may be an over-simplification, because a study in mice that has been published as a preprint suggested that UVA may result in another type of single base substitution signature, called SBS51 (Hennessey et al., 2019). There is evidence that cSCC is as a result of long-term sun exposure (Kivisaari and Kahari, 2013), melanoma is a result of intermittent sun exposure (Oliveria et al., 2006) whereas BCC is a mixture of both of these exposures (Kricker et al., 1995) which could also be a reason for these mutation signatures (Savoye et al., 2018). As a result of these exposure differences, it may not only be the UV wavelength but also the frequency of UV exposure, the dose of UV that is received, the ratio of UVB and UVA (which can differ according to the season) (Nishimura et al., 2021) as well as the time interval between exposures (during which differential repair of the CPD and pyrimidine-pyrimidone (6-4) photoproduct (6-4PP) photoproducts might have occurred) (Mitchell, 1988) that influences the type of SBS7 mutation signature, but further research will be necessary to evaluate this. Admittedly, the presence of sequencing artefacts or the algorithm used to produce the mutation signatures could be possible reasons that slightly different UV-related single base substitution mutation signatures were identified in the three different types of skin cancer in the current study.

With relevance to identifying the driver genes in BCC, cSCC and cutaneous melanoma), and determining whether mutations in those driver genes had been reported in normal skin and potentially pre-cancerous skin lesions, the study did highlight quite several driver genes in the overall group of skin cancers. This is useful data for future studies investigating whether UV exposure of normal skin results in the development of mutations in many or all of these types of driver genes in cells within the epidermis, including in the relevant cells from which skin cancer develops (i.e., keratinocytes for BCC and cSCC, melanocytes for melanoma). In addition, this data is likely to be helpful in determining whether several mutations that arise in normal skin following UV exposure are benign or passenger type mutations or whether they are driver genes that increase the risk of developing skin cancer. The current study has demonstrated that mutations in some of the relevant driver genes have been identified to date by employing the results from the skin cancer driver gene data for comparison with targeted next generation sequencing data on normal skin (particularly chronically sun-exposed skin) that exists in the public domain.

However, it is accepted that the project conducted for this aim of the thesis had some limitations. One of these limitations is that the power to detect significant driver genes varied within the three different types of skin cancer due to the varying sample numbers. The highest power to detect driver genes was in cutaneous melanoma due to the larger sample size, but the numbers of cSCCs and BCCs for which data on WES and WGS were available was much more limited. Therefore, a larger sample size in the future could detect more driver genes, which would strengthen the data obtained in the current thesis, as well as allowing one to identify more similarities or differences between these three skin tumour types. Since the WES and WGS data was obtained from a combination of the COSMIC database, GDC portal and literature searchers, and there were varying sequencing platforms used to generate the whole exome and whole genome data, this may have given rise to some variation in base calls in the samples which could also partially have affected the results. The preparation of the samples was also likely to have been different in the various studies (including whether the tumours were put in formalin for fixation and for how long), and the different studies contained different sized samples and contents, including numbers of primary and metastatic tumours, and probably different combinations of less aggressive and more aggressive primary tumours which could have meant that mutations in different genes would have been identified. There were also cSCC samples from patients that had received azathioprine in organ transplant individuals (Inman et al., 2018) and vemurafenib (BRAF inhibitor) (Li et al., 2015) who subsequently developed cSCC, which would have influenced the data in terms of the genes that were mutated in cSCC cohort. There would also have been many other possible unknown influences on the types of mutations and genes mutated in the different types of skin cancer (as well as the different types of SCCs in chapter 4), which may have included other types of medication (Kaae et al., 2010, Ioannidis et al., 2014) (including those that may not yet be known to affect type of mutation and/or DNA repair processes), variation in the age of the patients (Viros et al., 2008, Bauer et al., 2011), their general health, their gut and skin microbiomes (Grice and Segre, 2011, Wong and Yu, 2019) alcohol intake, tobacco smoking, environmental pollutants, etc.

In addition, the data used to compare samples in the current study was whole genome or whole exome data that was in a MAF format, therefore the method used to filter variants may not have been the same across all of the samples which introduces further variation in the data. Related to this, the fact that not all the whole genome and whole exome data identified in the MEDLINE literature search was used because there were data from studies which were not available in a MAF format and cases in which the author was unable to share the data due to ethical reasons (or authors who chose not to share their data), would have introduced additional limitations . While the current study used data that was available in public domains such as COSMIC and GDC portal for analysis, the study also showed that there were protected data in databases such as dbGaP and EGA which required authorisation to access, so in cases where authorisation was not granted, or it was not possible to get authorisation, data would have been missed from the analysis. While the study tried to be as inclusive as possible in order to increase the sample size for the analysis, the downside of variation in the sequencing platforms and the variant calling could also result in an amount of “noise” in the dataset. On a positive note, since all of the data was processed through Oncotator in the current project, the genome was annotated the same way across all samples.

Additionally, in the case of skin cancer, while there were some details available on the public databases (such as the age of the individual), there were no details about the skin type of the individual, which would likely have influenced the data (e.g. number of mutations and possibly types of mutations because there is lots of evidence associating skin cancer with a fairer skin type (Bradford, 2009) and with *MC1R* gene variants (Tagliabue et al., 2015, Tagliabue et al., 2018), which may affect skin cancer development via pigmentary and non-pigmentary mechanisms)(Robinson and Healy, 2002, Robinson et al., 2010). This information would prove useful to compare the mutation burden according to skin type and polymorphisms in genes that affect skin pigmentation (Valverde et al., 1995, Lamason et al., 2005, Sulem et al., 2007, Crawford et al., 2017), but much greater numbers of cSCC and BCCs, and possibly melanomas, would be needed in order to look for associations of various driver genes with these parameters. In addition, it might be possible that some genes might act as driver genes in certain skin types or certain genetic backgrounds but not in others, especially in the case of melanoma because these pigmentation genes are commonly expressed in melanocytes. This might have relevance during the development of skin cancer and/or at later stages when a skin cancer has developed, because in the case of melanoma clinical differences are seen in the colours of melanomas, including the loss of pigment in amelanotic melanoma (Thomas et al., 2014).

The approach used in this thesis concentrated on the genetic mutations in the cancers that were investigated, and the bioinformatics programs identified certain genes as driver genes in skin cancers, but more research needs to be conducted to identify the effect of these mutations at the transcriptional and protein levels in the affected cells and tumour. As mentioned earlier, this study focused on the whole genome and whole exome analysis, however, there could be copy number changes, epigenetic, or transcriptomic alterations in the skin cancers from UV affecting non-exonic areas of the genome. Indeed, the fact that the protein coding region of genome is only about 1% of the genome (Frazer, 2012), stresses the need for further analysis to be conducted into the non-coding regions of the genome and how these regions could affect skin tumorigenesis. Furthermore, in relation to the exonic regions, the analysis also only looked at mutations in genes, but there could have been homozygous loss of alleles at various sites in the genome and exome, with wild type sequence identified in the date being due to contamination by normal tissue, including infiltration of the tumour with immune cells (Lai et al., 2021). This could mean that alterations at some of the driver gene loci were more common than appreciated, but in the absence of analysing the copy number data it is not possible to know whether this was the case.

The examiners of this PhD thesis correctly identified that the mutations in this analysis from the COSMIC database may not have true somatic status as the database included mutations from studies which did not include matched germline DNA from blood or saliva. The cohorts that were affected were lung SCC, oropharyngeal SCC, oesophageal SCC, cervical SCC and melanoma.

Initially, any cancer cohorts which had less than 98% of variants that were considered as confirmed somatic variants were reanalysed and the reasons for this are outlined in chapter 2, section 6. In appendix 7.11.6, the unconfirmed variants were further investigated for each cancer cohort. It identified that *TP53* was the only driver gene that was shared between skin SCC and cervical SCC and all the *TP53* variants in cervical SCC were confirmed somatic variants. Therefore, reanalysis was only conducted for lung SCC, oropharyngeal SCC, oesophageal SCC and melanoma. Further detail of this reanalysis is outlined in chapter 2, section 6.

The results in appendix 7.11.5 show that the MutSig2CV results after only including confirmed somatic variants. This program was used as a gene was only considered a driver gene if it was significant in MutSig2CV and one other driver gene program. In lung SCC, *HRAS* has a q value which has increased in significance by 10-fold in the reanalysis and in melanoma the q value for *CDKN2A* has also become less significant by 10-fold. These results are more reflective of how significantly mutated these genes are within their cancer cohorts as the reanalysis only included confirmed somatic variants. Most importantly, there were also no new novel driver genes identified in the reanalysis suggesting that the original analysis produced equivalent results to the re-analysis including only confirmed somatic variants.

In the COSMIC database, a confirmed somatic mutation is considered a mutation which is present in a tumour sample that is absent in a matched normal sample. There are limitations associated with using the matched normal sample as DNA from blood as this is also considered to contain somatic mutations as these have been found to accumulate at a low rate with increasing age (Welch et al., 2012, Jaiswal et al., 2014, Genovese et al., 2014, Xie et al., 2014, McKerrell et al., 2015). In the Martincorena et al., 2015 study, instead of using a single matched normal sample, they cut normal eyelid skin and compared these normal samples to each other. The use of multiple normal samples provided an extremely high coverage of matched normal sample for each site which enabled the detection of true somatic variants, this also enabled the removal of sequencing artefacts that are produced as a result of errors from alignment of short reads to a repetitive genomic sequence, base call errors and sequencing errors produced as a result of library preparation.

Another limitation in this study, is that the data used are from multiple studies which use different variant callers. The sequencing platforms used would have been different and the samples would have been handled by different people. The TCGA samples were called using the same variant caller whereas the samples from COSMIC and the external databases used different variant callers and may have used different read depth cut off values during analysis of bases. Therefore, the reliability of the data could vary across studies.

6.1 Future work

Regarding the future studies that would assess the likely carcinogenicity of phototherapy, 6mm punch biopsies have been taken by a clinician in our department from the non-psoriatic skin of the buttock and arm of psoriasis patients before and after a course of NB-UVB treatment. Twenty patients were recruited, although some of these did not give skin biopsies after the UV course, and the buttock and arm samples were fixed in PAXgene tissue FIX and then stabilised in PAXgene stabilizer. The samples were cut in half and one half was sent for paraffin embedding and stored at -20°C. The other half of the sample was snap-frozen in liquid nitrogen and then stored at -80°C. Questionnaires about pigmentation and previous sun exposure were completed by the patients and the UV doses for the course of treatment have been recorded. Pre- and post-UV skin samples have been taken from 16 patients and have been sent to the Wellcome Sanger Institute for gene sequencing. Blood was also collected from all patients to distinguish somatic mutations from germline mutations.

Samples from chronically sun-exposed skin that were approximately 1 – 2cm away from skin cancers that were excised have also been collected from 10 patients and handled in the same manner as the skin biopsies from psoriatic patients. The skin samples which have been collected from the 16 patients pre-UV and post UV will be sequenced using nanorate sequencing (NanoSeq) which has the potential to detect somatic mutations in single DNA molecules (Abascal et al., 2021). Nanorate sequencing uses a duplex sequencing protocol where both strands of DNA are sequenced, thus removing the chance of mutations due to sequencing errors being included in individual reads and PCR errors which might be present in one of the two strands (Abascal et al., 2021). Conversely, bottleneck sequencing will not be used because of concerns that it could introduce errors that would be misinterpreted as UV-induced mutations (Hoang et al., 2016). Nanorate sequencing is considered better than whole exome or targeted sequencing as it identifies somatic mutations independently of the requirement for clonality that can be required for whole exome or targeted sequencing of precancerous and cancerous lesions. The NanoSeq libraries are also able to be produced from as little as 1ng of DNA, therefore are capable of sequencing small tissue sections from the 6mm punch skin biopsies.

The results from the skin cancer analysis in the current thesis will hopefully assist researchers in analysing whether any mutations identified in normal skin following a course of NB-UVB are likely to increase the risk of future development of skin cancer. The table of variants and the "proposed adjunct to future pipeline for assessment of carcinogenicity of repeated UV exposure during phototherapy for skin disease" in chapter 5 can be used as a database of potentially pathogenic genes and variants that have been identified in precancerous lesions and skin cancers. Therefore, these genes and variants could be compared with the variants identified in the normal skin of psoriatic patients after UV treatment to contribute to understanding how harmful a course of NB-UVB treatment is. This in turn may allow estimation of how many courses of NB-UVB a patient could safely have over the course of their lifetime without significantly increasing their risk of developing skin cancer in future years.

The clinical impact of the UV study will largely depend on how many or how few mutations are identified in the post-UV skin samples compared to the pre-UV skin samples. It will also tell us whether a single NB-UVB course causes a few or lots of mutations in the skin, something we currently don't know. Modelling this data in relation to the data from the chronically sun-exposed skin samples in the separate group of patients with skin cancer could allow the identification of whether a patient could theoretically safely have only a few or, alternatively, lots of NB-UVB courses over their lifetime. However, this project also has the potential to help us understand why people develop keratinocyte cancers at a later stage of life. The current assumption is that it takes multiple UV exposures over life to allow sufficient numbers of mutations in keratinocytes to form a keratinocyte cancer, but this may be incorrect because another factor such as survival of competing mutated clones in the epidermis may be more relevant once a limited number of mutations arise in a cell (Murai et al., 2018). Moreover, it is hoped that the results of this study will advance the understanding of early cancer development in general.

In addition to the usefulness of the data in this thesis to the NB-UVB study as mentioned above, the data in this thesis in combination with the NB-UVB NanoSeq data is also likely to be helpful to allows researchers to begin to understand clonal evolution in sun-exposed skin. A recent study compared 450 individual matched sun-exposed and non-sun-exposed normal human skin samples (Wei et al., 2021) and looked at the number of clonal mutations in the skin samples. There were hotspots associated with sun exposure in *TP53*, *NOTCH1* and *GRM3*. They then compared the normal skin from patients with cSCC and identified that UV induced mutations were mutations that had previously been reported in cSCC. A comparison of the results from that study and from other clonal mutations studies on sun-exposed skin (Martincorena et al., 2015, Albibas et al., 2017, Lynch et al., 2017, Fowler et al., 2021) with the aforementioned NanoSeq of the NB-UVB exposed skin, and chronically exposed skin adjacent to skin cancers, would allow one to compare mutations that were detected in epidermal clones in those studies against those NanoSeq mutations that have not been identified in any epidermal clones to date. This type of comparison could form a basis upon which researchers can then begin to unpick why certain mutations that likely affect protein function offer no clonal advantage in the epidermis whereas other mutations in the same or different genes can lead to the development of clones and, in a number of cases, to skin cancer.

A similar study to that conducted by Wei et al., 2021 could be carried out in the clinic using the driver genes identified in this thesis. Clinicians could measure the number of clonal mutations in skin before and after a course of NB-UVB, and this could be repeated after each subsequent course of NB-UVB in biopsies sampled in areas adjacent to the original biopsies to provide insight into whether similar or greater numbers of mutated clones arise after the later courses of NB-UVB, thus informing on how safe or harmful the NB-UVB is. If whole genome sequencing were conducted on the skin biopsies in this type of future study, it would enable clinicians to identify the mutation burden and clonal mutations affecting non-coding and coding regions of the genome, including in upstream promoter regions, regulatory regions or epigenetic changes which might increase expression of cancer driver genes.

A similar approach could also be used in the clinic to look at patients with chronically sun-exposed skin by taking a skin biopsy and analysing their current mutation burden. They could also be screened to identify if they have specific pathogenic mutations in cancer driver genes or if the mutation burden is high in the cancer driver genes. However, the mutation burden in different body sites also varies (Fowler et al., 2020) therefore, it would be important that if biopsies are taken from different body sites, the background mutation rate is considered and standardised appropriately. While at the present time, it is unlikely that this type of study would be able to predict that a particular individual is more or less likely to develop skin cancer in the future, undertaking this type of study on large numbers of people who are followed as a cohort to see who develops skin cancer would subsequently allow the identification of biomarkers (e.g. mutation burden, numbers and/or types of driver genes mutated, etc.) that could be used to predict risk of skin cancer. That information might also allow clinicians to target messages about limiting future sun exposure to people who are most at risk of developing skin cancer.

The results presented in this thesis have identified a number of driver genes in skin SCC, BCC and melanoma. The importance of new mutations that may be seen in the post-UV skin samples using nanorate sequencing could be ranked using the driver genes identified in this thesis and could help clarify whether patients develop mutations in driver genes for keratinocyte cancer or melanoma from a single course of UV therapy. However, due to the power of the studies included in the bioinformatic analyses in this thesis, it is possible that all the driver genes for keratinocyte cancer and melanoma may not have been identified in this thesis, therefore this is a limitation in the use of the data from this thesis in future work. In addition, this thesis also only investigated the effect of coding regions of the genome and the mutations identified in the post-UV samples might have mutations in the non-coding regions of the genome, thus this is another limitation of this thesis. However, the mutation burden of the driver genes pre and post-UV could be used to measure the effect of UV in normal skin and if there are lots of mutations in these driver genes, it would raise some concerns that too many courses of NB-UVB over a patient's lifetime would likely lead to skin cancer development. Conversely, if the driver genes are infrequently mutated following a course of NB-UVB, this could suggest that NB-UVB is relatively safe as a treatment. This thesis has also identified individual mutations which have been identified in cancers and the information from the NB-UVB samples will also help identify mutations which are associated with UV. Therefore, this will also enable the stratification of pathogenic mutations which are associated with UV and those which are not.

7. Appendix

7.1 SCC bash script

```
#!/bin/bash


## now uses array to do look up, not relying on paste, outputs "not in vcf" if not found

## also using both coding and noncoding vcfs for look up

## also searching for site matched in column 8 of CosmicMutantExport.tsv/SCC_mutations.txt - to avoid counting twice if site matches multiple columns


#Download and gunzip files from COSMIC b37 v91

#classification.csv

#CosmicMutantExport.tsv

#CosmicCodingMuts.vcf

#CosmicNonCodingVariants.vcf


#combine coding and noncoding vcf file bodies

grep -v "#" CosmicCodingMuts.vcf > Coding

grep -v "#" CosmicNonCodingVariants.vcf > NonCoding

cat Coding NonCoding > all.vcf

rm Coding NonCoding


#get unique list of primary sites in SCC

awk 'BEGIN{FS=",";OFS="\t"}{if($7 == "squamous_cell_carcinoma" && $6 == "carcinoma") print $2}' classification.csv | sort | uniq > SCC_sites.txt


#get mutation list with header- only those from genome wide screens and with genomic location on b37. If histology subtype 1 is ‘squamous cell carcinoma’ and primary histology is ‘carcinoma’ and genome wide screen is ‘y’ and GRCh is ‘37’ then print and make file SCC_mutations.txt.

head -1 CosmicMutantExport.tsv > SCC_mutations.txt

awk 'BEGIN{FS=OFS="\t"}{if($13 == "squamous_cell_carcinoma" && $12 == "carcinoma" && $16 == "y" && $25 == "37") print $0}' CosmicMutantExport.tsv > SCC_mutations.txt
```

```
#list and count mutation types. Print ‘mutation description’ in SCC_ mutations.txt and sort them.

echo "All sites SSC mutation types:"

awk 'BEGIN{FS="\t"}{print $22}' SCC_mutations.txt | sort | uniq -c
```

```
#loop sites, extract mutations, and output sub, ins, and del and save in separate files

cat SCC_sites.txt | while read site; do
```

```
#list and count mutation types per site

echo "${site}:"

awk -v site=$site 'BEGIN{FS=OFS="\t"}{if($8 == site) print $0}' SCC_mutations.txt > ${site}_squamous_cell_carcinoma.txt

awk 'BEGIN{FS="\t"}{print $20}' ${site}_squamous_cell_carcinoma.txt | sort | uniq -c
```

```
#output simple substitutions. If mutation description is ‘substitution’ then print sample name, genomic mutation id (C0SV), mutation cds, mutation AA, mutation description, mutation zygosity, LOH, GrCh, Mutation genome position and Mutation strand, sample name to make site_cosmic_ids_patient.sub. Find all the lines with COSMIC ID in all.vcf sort cosmic ID column 3 making site_vcf_details.sub

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Substitution") print $17}' ${site}_squamous_cell_carcinoma.txt > ${site}_cosmic_ids.sub

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Substitution") print $5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}' ${site}_squamous_cell_carcinoma.txt | sort -k2,2 > ${site}_cosmic_ids_patient.sub

fgrep -w -f ${site}_cosmic_ids.sub all.vcf | sort -k3,3 > ${site}_vcf_details.sub
```

```
#match up vcf details with array. Read in both site_vcf_details.sub and site_cosmic_ids_patient.sub as tab delimited files, read in the vcf file as an array indexed by column 3 (the cosmic ID) in vcf_details.sub, then loop through each line of the patient file and match column 2 (the cosmic ID) in cosmic_ids_patient.sub with the cosmic IDs in the array. When they match output the patient file columns followed by the vcf file columns (all in the same row), else print out “not in vcf”. Then make files all_details.sub.

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}' ${site}_vcf_details.sub ${site}_cosmic_ids_patient.sub > ${site}_all_details.sub
```

```
#extract location - print 6 columns wanted up front. Read in site_all_details.sub as a tab delimited file. Split column 9 (19:52249948-52249948) by the : and save the two part in an array called a. (so a[1]=19 and a[2]=52249948-52249948). The split a[2] further this time by the – character and
```

save in an array called b. (so b[1]=52249948, and b[2]=52249948). Now print these parts as separate column (chromosome, startbasepair, endbasepair) followed by other columns of interest eg, columns 15,16,1,2,3, etc.

```
awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print
a[1],b[1],b[2],$15,$16,$1,$2,$3,$4,$5,$19}' ${site}_all_details.sub > ${site}.sub


echo "Check all sub found:"

grep "not in vcf" ${site}_all_details.sub

wc -l ${site}_*.sub


#output simple deletions


awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Deletion") print $17}' ${site}_squamous_cell_carcinoma.txt >
${site}_cosmic_ids.del

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Deletion") print $5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}'
${site}_squamous_cell_carcinoma.txt | sort -k2,2 > ${site}_cosmic_ids_patient.del

fgrep -w -f ${site}_cosmic_ids.del all.vcf | sort -k3,3 > ${site}_vcf_details.del


#match up vcf details with array

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}'
${site}_vcf_details.del ${site}_cosmic_ids_patient.del > ${site}_all_details.del


#extract location - print 6 columns wanted up front

awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print a[1],b[1],b[2],substr($15,2),"-
",$1,$2,$3,$4,$5,$19}' ${site}_all_details.del > ${site}.del


echo "Check all del found:"

grep "not in vcf" ${site}_all_details.del

wc -l ${site}_*.del


#output simple insertions

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Insertion") print $17}' ${site}_squamous_cell_carcinoma.txt >
${site}_cosmic_ids.ins
```

```
awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Insertion") print $5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}'
${site}_squamous_cell_carcinoma.txt | sort -k2,2 > ${site}_cosmic_ids_patient.ins

fgrep -w -f ${site}_cosmic_ids.ins all.vcf | sort -k3,3 > ${site}_vcf_details.ins


#match up vcf details with array

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}'
${site}_vcf_details.ins ${site}_cosmic_ids_patient.ins > ${site}_all_details.ins


#extract location - print 6 columns wanted up front

awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print a[1],b[1],b[2],"-
",substr($16,2),$1,$2,$3,$4,$5,$19}' ${site}_all_details.ins > ${site}.ins


echo "Check all ins found:"

grep "not in vcf" ${site}_all_details.ins

wc -l ${site}_*.ins

done


##### NOTES #####

# 1. *tsv and *vcf - b37 cosmic v91 (07th April 2020)
```

7.2 Melanoma bash script

```
#!/bin/bash


## 2nd version of this script - previously not outputting all muts vcf details so matching up on
Cosmic id wrong

## now uses array to do look up, not relying on paste, outputs "not in vcf" if not found

## also using both coding and noncoving vcfs for look up

## also searching for site matched in column 8 of CosmicMutantExport.tsv/SCC_mutations.txt - to
avoid counting twice if site matches multiple columns


#Download and gunzip files from COSMIC b37 v88

#classification.csv

#CosmicMutantExport.tsv
```

```
#CosmicCodingMuts.vcf
#CosmicNonCodingVariants.vcf


#combine coding and noncoding vcf file bodies
grep -v "#" CosmicCodingMuts.vcf > Coding
grep -v "#" CosmicNonCodingVariants.vcf > NonCoding
cat Coding NonCoding > all.vcf
rm Coding NonCoding


#get unique list of primary sites in SCC
awk 'BEGIN{FS=",";OFS="\t"}{if($6 == "malignant_melanoma") print $2}' classification.csv | sort | uniq > cutmelanoma_sites.txt


#get mutation list with header- only those from genome wide screens and with genomic location on b37
head -1 CosmicMutantExport.tsv > cutmelanoma_mutations.txt
awk 'BEGIN{FS=OFS="\t"}{if($12 == "malignant_melanoma" && $16 == "y" && $25 == "37") print $0}' CosmicMutantExport.tsv > cutmelanoma_mutations.txt


#list and count mutation types
echo "All sites cutaneous melanoma mutation types:"
awk 'BEGIN{FS="\t"}{print $22}' cutmelanoma_mutations.txt | sort | uniq -c


#loop sites, extract mutations, and output sub, ins, and del and save in separate files
cat cutmelanoma_sites.txt | while read site; do


#list and count mutation types per site
echo "${site}:"
awk -v site=$site 'BEGIN{FS=OFS="\t"}{if($8 == site) print $0}' cutmelanoma_mutations.txt > ${site}_cutaneous_melanoma.txt
awk 'BEGIN{FS="\t"}{print $20}' ${site}_cutaneous_melanoma.txt | sort | uniq -c
```

```
#output simple subsutitutions

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Substitution") print $17}' ${site}_cutaneous_melanoma.txt >
${site}_cosmic_ids.sub

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Substitution") print
$5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}' ${site}_cutaneous_melanoma.txt | sort -k2,2 >
${site}_cosmic_ids_patient.sub

fgrep -w -f ${site}_cosmic_ids.sub all.vcf | sort -k3,3 > ${site}_vcf_details.sub


#match up vcf details with array

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}'
${site}_vcf_details.sub ${site}_cosmic_ids_patient.sub > ${site}_all_details.sub


#extract location - print 6 columns wanted up front

awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print
a[1],b[1],b[2],$15,$16,$1,$2,$3,$4,$5,$19}' ${site}_all_details.sub > ${site}.sub


echo "Check all sub found:"

grep "not in vcf" ${site}_all_details.sub

wc -l ${site}_*.sub


#output simple deletions


awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Deletion") print $17}' ${site}_cutaneous_melanoma.txt >
${site}_cosmic_ids.del

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Deletion") print $5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}'
${site}_cutaneous_melanoma.txt | sort -k2,2 > ${site}_cosmic_ids_patient.del

fgrep -w -f ${site}_cosmic_ids.del all.vcf | sort -k3,3 > ${site}_vcf_details.del


#match up vcf details with array

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}'
${site}_vcf_details.del ${site}_cosmic_ids_patient.del > ${site}_all_details.del


#extract location - print 6 columns wanted up front
```

```
awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print a[1],b[1],b[2],substr($15,2),"-",$1,$2,$3,$4,$5,$19}' ${site}_all_details.del > ${site}.del


echo "Check all del found:"

grep "not in vcf" ${site}_all_details.del

wc -l ${site}_*.del


#output simple insertions

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Insertion") print $17}' ${site}_cutaneous_melanoma.txt > ${site}_cosmic_ids.ins

awk 'BEGIN{FS=OFS="\t"}{if($22 ~ "Insertion") print $5,$17,$20,$21,$22,$23,$24,$25,$26,$27,$5}' ${site}_cutaneous_melanoma.txt | sort -k2,2 > ${site}_cosmic_ids_patient.ins

fgrep -w -f ${site}_cosmic_ids.ins all.vcf | sort -k3,3 > ${site}_vcf_details.ins


#match up vcf details with array

awk 'BEGIN{FS=OFS="\t"} NR==FNR{a[$3]=$0;next}{print $0,a[$2]?a[$2]:"not in vcf"}' ${site}_vcf_details.ins ${site}_cosmic_ids_patient.ins > ${site}_all_details.ins


#extract location - print 6 columns wanted up front

awk 'BEGIN{FS=OFS="\t"}{split($9,a,":");split(a[2],b,"-"); print a[1],b[1],b[2],"-",substr($16,2),$1,$2,$3,$4,$5,$19}' ${site}_all_details.ins > ${site}.ins


echo "Check all ins found:"

grep "not in vcf" ${site}_all_details.ins

wc -l ${site}_*.ins

done


##### NOTES #####

# 1. *tsv and *vcf - b37 cosmic v91 (20th April 2020)
```

7.3 Script for formatting COSMIC and literature search data in Linux for Oncotator compatibility

7.3.1 Skin SCC

COSMIC formatting

```
cut -f1,2,3,4,5,6 *.ins > skininsonc.txt
cut -f1,2,3,4,5,6 *.sub > skinsubonc.txt
cut -f1,2,3,4,5,6 *.del > skindelonc.txt
cat skininsonc.txt | sed -e 's/^/chr/' > skininsoncchr.txt
cat skinsubonc.txt | sed -e 's/^/chr/' > skinsuboncchr.txt
cat skindelonc.txt | sed -e 's/^/chr/' > skindeloncchr.txt
cat skininsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > skininsoncchrxy.txt
cat skinsuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > skinsuboncchrxy.txt
cat skindeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > skindeloncchrxy.txt
wc -l skininsocchr.txt
wc -l skininsoncchrxy.txt
awk '!seen[$0]++' skininsoncchrxy.txt > nodupskininsonc.txt
wc -l nodupskininsonc.txt
awk '!seen[$0]++' skindeloncchrxy.txt > nodupskindelonc.txt
awk '!seen[$0]++' skinsuboncchrxy.txt > nodupskinsubonc.txt
wc -l nodupskinsubonc.txt
wc -l skinsuboncchrxy.txt
head nodupskinsubonc.txt
cat nodupskinsubonc.txt nodupskindelonc.txt nodupskininsonc.txt > mayskininput.txt
head mayskininput.txt
grep -Ev $'^\t|\t\t|\t$' nodupskininsonc.txt > nospaceskininsonc.txt
wc -l nospaceskininsonc.txt
grep -Ev $'^\t|\t\t|\t$' nodupskinsubonc.txt > nospaceskinsubonc.txt
wc -l nospaceskinsubonc.txt
wc -l nodupskinsubonc.txt
grep -Ev $'^\t|\t\t|\t$' nodupskindelonc.txt > nospaceskindelonc.txt
cat nospaceskin*.txt > nospaceskinjuly.txt
wc -l nospaceskinjuly.txt
cat 2headerskininput.txt nospaceskinjuly.txt > nospaceskinfinaljuly1.txt
```

EGA data formatting

```
gunzip refGeneExtent.hg19.bed.gz
```

```
grep ‘exon’ refGeneExtent.hg19.bed > query_exons.bed
```

#The script below shows how the refGeneExtent.hg19.bed.gz file was decompressed, and the second line of the script represents how the exons were extracted from the reference genome file.

```
grep 'SOMATIC' CSCC_0014_M1.strelka.filtered.vcf > CSCC_0014_M1_somatic.vcf
```

#There were 13 Strelka VCF files containing whole genome information for each skin SCC tumour. Each file was prepared individually and first the somatic mutations were extracted from each Strelka VCF file using the command shown.

```
cut -f1,2 CSCC_0014_M1_somatic.vcf > chr_startpos.txt

sed 1,3d chr_startpos.txt > noheader_chr_startpos.txt



cut -f3 CSCC_0014_M1_somatic.vcf > id.txt

sed 1,3d id.txt > noheader_id.txt

sed 's/\./CSCC_0014_M1/g' noheader_id.txt > idtest.txt

cut -f4,5 CSCC_0014_M1_somatic.vcf > ref_alt.txt

sed 1,3d ref_alt.txt > noheader_ref_alt.txt
```

#Then the chromosome number and position were extracted from column one and two of the Strelka VCF file which had been filtered for somatic mutations and a new file ‘pos_somatic.txt’ file was made. The second line of the script below shows that the first three lines of the ‘pos_somatic.txt’ file was removed from the file to remove the column headings and a new file was made ‘noheader_pos.txt’. The tumour identifier (column three), reference and alternate base (column four and five) were extracted from the somatic VCF file for each individual tumour and individual files were made. The first three lines were removed from these files to ensure the column headings were not included in the files. The script summarising this is shown.

```
cut -f2 CSCC_0014_M1_somatic.vcf > pos_somatic.txt

sed 1,3d pos_somatic.txt > noheader_pos.txt
```

#Column two of the somatic VCF file was also extracted on its own to represent the end position of each mutation. The first three lines containing the column headings were also removed as shown in the script.

```
paste noheader_chr_startpos.txt noheader_pos.txt noheader_ref_alt.txt idtest.txt >
0014_M1_input_final.txt
```

#The four files with the chromosome number, start position, end position, reference base, alternate base and the tumour identifier were joined together to produce a file with four columns for each individual skin SCC tumour (0014_M1_input_final.txt) which is shown in the script..

```
cat 0014_M1_input_final.txt |sed -e 's/^/chr/'> 0014_M1_input_final_chr.txt
```

```
sed 's/\./-/g' 0014_M1_input_final_chr.txt > CSCC_0014_M1_input_final_chr.txt
```

#Then the characters ‘chr’ were added to the beginning of each row to represent the chromosome number. Each row which contained a ‘.’ to represent a nucleotide base that was not present was converted to a ‘-‘ so the file would be compatible for Oncotator. A new file was created ‘CSCC_0014_M1_input_final_chr.txt’ file which contained both changes. The commands and files produced are shown as in script

```
cut -f1,2,3 CSCC_0014_M1_input_final_chr_.txt > CSCC_0014_M1.bed
```

#Next the first three columns from the ‘CSCC_0014_M1_input_final_chr.txt’ file, which contained the chromosome number, start position and end position, were extracted from the file to produce a CSCC_0014_M1.bed. The file was converted from a text file into a BED file so it could be used in bedtools (version 2.21.0)(Quinlan and Hall, 2010). The script summarising this is shown.

```
scp /Volumes/“My Passport”/Bed_files/CSCC_0014_M1.bed
username@iridis4_a.soton.ac.uk:/home/username
```

```
scp /Volumes/“My Passport”/Bed_files/query_exons.bed
username@iridis4_a.soton.ac.uk:/home/username
```

#The University of Southampton computer cluster, Iridis was used to access bedtools and the bed files produced for each individual tumour were uploaded to the cluster using the script shown below. The reference file with all the exon coding regions (query_exons.bed) were also uploaded to Iridis.

```
ssh -Y username@iridis4_a.soton.ac.uk
```

```
module load bedtools/2.21.0
```

#The command was used to log into Iridis and then the second command was used to load bedtools.

```
bedtools intersect -a query_exons.bed -b 0014_M1.bed > CSCC_0014_M1_exons.bed
```

#The command used bedtools and the reference file with the exon chromosome co-ordinates (query_exons.bed) to identify all the exonic regions in the whole genome bed file of an individual skin SCC tumour from the Mueller et al., 2019 study. Bedtools then produced a file with all the genome co-ordinates which reside in exons which is shown in the command below as 'CSCCC_0014_M1_exons.bed'.

```
scp username@iridis4_a.soton.ac.uk:/home/username/CSCC_0014_M1_exons.bed
/Users/username/Documents
```

#The file with all the chromosome co-ordinates that reside in exons was then downloaded from the Iridis computer cluster to the local computer using the command shown here.

```
fgrep -w -f CSCC_0014_M1_exons.bed CSCC_0014_M1_input_final_chr.txt >
CSCC_0014_M1_exon_variants.txt
```

```
cat oncotatorheader.txt CSCC_0004_M1_exon_variants.txt >
CSCC_0004_M1_exon_variants_final.txt
```

#The exonic chromosome co-ordinates that were unique to each individual skin SCC tumour file was then extracted from the file which was created previously with all the whole genome data for each individual skin SCC (CSCC_0014_M1_input_final_chr.txt). A new file was produced showing the variants for the coding regions and this is summarised in the first command shown below. The second command showed a file with column headings collated with the 'CSCC_0014_M1_exon_variants.txt' file. This column headings file was to ensure the file was compatible for Oncotator.

```
oncotator -v --db-dir /file location/oncotator_v1_ds_April052016 /file location/
0014_M1_exon_variants_final.txt CSCC_0014_M1_exon_variants_final_output.tsv hg19
```

#The file was then run through Oncotator, shown in the command using default settings, to ensure all the exonic variants were annotated to produce MAF files. This process was completed for each of the 13 Strelka VCF files.

7.3.2 Oropharyngeal SCC

```
cut -f1,2,3,4,5,6 *.ins > upperaerodigestivetractinsonc.txt

cut -f1,2,3,4,5,6 *.sub > upperaerodigestivetractsubonc.txt

cut -f1,2,3,4,5,6 *.del > upperaerodigestivetractdelonc.txt

cat upperaerodigestivetractinsonc.txt | sed -e 's/^/chr/' > upperaerodigestivetractinsoncchr.txt

cat upperaerodigestivetractsubonc.txt | sed -e 's/^/chr/' > upperaerodigestivetractsuboncchr.txt

cat upperaerodigestivetractdelonc.txt | sed -e 's/^/chr/' > upperaerodigestivetractdeloncchr.txt

cat upperaerodigestivetractinsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > upperaerodigestivetractinsoncchrxy.txt

cat upperaerodigestivetractsuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > upperaerodigestivetractsuboncchrxy.txt

cat upperaerodigestivetractdeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > upperaerodigestivetractdeloncchrxy.txt

wc -l upperaerodigestivetractinsocchr.txt

wc -l upperaerodigestivetractinsoncchrxy.txt

awk '!seen[$0]++' upperaerodigestivetractinsoncchrxy.txt > nodupupperaerodigestivetractinsonc.txt

wc -l nodupupperaerodigestivetractinsonc.txt

awk '!seen[$0]++' upperaerodigestivetractdeloncchrxy.txt > nodupupperaerodigestivetractdelonc.txt

awk '!seen[$0]++' upperaerodigestivetractsuboncchrxy.txt > nodupupperaerodigestivetractsubonc.txt

wc -l nodupupperaerodigestivetractsubonc.txt

wc -l upperaerodigestivetractsuboncchrxy.txt

head nodupupperaerodigestivetractsubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupupperaerodigestivetractinsonc.txt > nospaceupperaerodigestivetractinsonc.txt

wc -l nospaceupperaerodigestivetractinsonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupupperaerodigestivetractsubonc.txt > nospaceupperaerodigestivetractsubonc.txt

wc -l nospaceupperaerodigestivetractsubonc.txt

wc -l nodupupperaerodigestivetractsubonc.txt
```

```
grep -Ev $'^\t|\t\t|\t$' nodupupperaerodigestivetractdelonc.txt > nospaceupperaerodigestivetractdelonc.txt

cat nospaceupperaerodigestivetract*.txt > nospaceupperaerodigestivetractmay.txt

wc -l nospaceupperaerodigestivetractmay.txt

cat 2headerupperaerodigestivetractinput.txt nospaceupperaerodigestivetractmay.txt > nospaceupperaerodigestivetractfinal_v91.txt

wc -l nospaceupperaerodigestivetractfinal_v91.txt
```

7.3.3 Oesophageal SCC

```
cut -f1,2,3,4,5,6 *.ins > oesophagusinsonc.txt

cut -f1,2,3,4,5,6 *.sub > oesophagussubonc.txt

cut -f1,2,3,4,5,6 *.del > oesophagusdelonc.txt

cat oesophagusinsonc.txt | sed -e 's/^/chr/' > oesophagusinsoncchr.txt

cat oesophagussubonc.txt | sed -e 's/^/chr/' > oesophagussuboncchr.txt

cat oesophagusdelonc.txt | sed -e 's/^/chr/' > oesophagusdeloncchr.txt

cat oesophagusinsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > oesophagusinsoncchrxy.txt

cat oesophagussuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > oesophagussuboncchrxy.txt

cat oesophagusdeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > oesophagusdeloncchrxy.txt

wc -l oesophagusinsocchr.txt

wc -l oesophagusinsoncchrxy.txt

awk '!seen[$0]++' oesophagusinsoncchrxy.txt > nodupoesophagusinsonc.txt

wc -l nodupoesophagusinsonc.txt

awk '!seen[$0]++' oesophagusdeloncchrxy.txt > nodupoesophagusdelonc.txt

awk '!seen[$0]++' oesophagussuboncchrxy.txt > nodupoesophagussubonc.txt

wc -l nodupoesophagussubonc.txt

wc -l oesophagussuboncchrxy.txt

head nodupoesophagussubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupoesophagusinsonc.txt > nospaceoesophagusinsonc.txt

wc -l nospaceoesophagusinsonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupoesophagussubonc.txt > nospaceoesophagussubonc.txt

wc -l nospaceoesophagussubonc.txt
wc -l nodupoesophagussubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupoesophagusdelonc.txt > nospaceoesophagusdelonc.txt

cat nospaceoesophagus*.txt > nospaceoesophagusmay.txt

wc -l nospaceoesophagusmay.txt

cat 2headeroesophagusinput.txt nospaceoesophagusmay.txt > nospaceoesophagusfinal_v91.txt

wc -l nospaceoesophagusfinal_v91.txt
```

7.3.4 Lung SCC

```
cut -f1,2,3,4,5,6 *.ins > lunginsonc.txt

cut -f1,2,3,4,5,6 *.sub > lungsubonc.txt

cut -f1,2,3,4,5,6 *.del > lungdelonc.txt

cat lunginsonc.txt | sed -e 's/^/chr/' > lunginsoncchr.txt

cat lungsubonc.txt | sed -e 's/^/chr/' > lungsuboncchr.txt

cat lungdelonc.txt | sed -e 's/^/chr/' > lungdeloncchr.txt

cat lunginsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > lunginsoncchrxy.txt

cat lungsuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > lungsuboncchrxy.txt

cat lungdeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > lungdeloncchrxy.txt

wc -l lunginsocchr.txt

wc -l lunginsoncchrxy.txt

awk '!seen[$0]++' lunginsoncchrxy.txt > noduplunginsonc.txt

wc -l noduplunginsonc.txt

awk '!seen[$0]++' lungdeloncchrxy.txt > noduplungdelonc.txt

awk '!seen[$0]++' lungsuboncchrxy.txt > noduplungsubonc.txt

wc -l noduplungsubonc.txt

wc -l lungsuboncchrxy.txt

head noduplungsubonc.txt

grep -Ev $'^\t|\t\t|\t$' noduplunginsonc.txt > nospacelunginsonc.txt

wc -l nospacelunginsonc.txt

grep -Ev $'^\t|\t\t|\t$' noduplungsubonc.txt > nospacelungsubonc.txt

wc -l nospacelungsubonc.txt

wc -l noduplungsubonc.txt
```

```
grep -Ev $'^\t|\t\t|\t$' noduplungdelonc.txt > nospacelungdelonc.txt

cat nospacelung*.txt > nospacelungmay.txt

wc -l nospacelungmay.txt

cat 2headerlunginput.txt nospacelungmay.txt > nospacelungfinal_v91.txt

wc -l nospacelungfinal_v91.txt
```

7.3.5 Cervical SCC

```
cut -f1,2,3,4,5,6 *.ins > cervixinsonc.txt

cut -f1,2,3,4,5,6 *.sub > cervixsubonc.txt

cut -f1,2,3,4,5,6 *.del > cervixdelonc.txt

cat cervixinsonc.txt | sed -e 's/^/chr/' > cervixinsoncchr.txt

cat cervixsubonc.txt | sed -e 's/^/chr/' > cervixsuboncchr.txt

cat cervixdelonc.txt | sed -e 's/^/chr/' > cervixdeloncchr.txt

cat cervixinsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > cervixinsoncchrxy.txt

cat cervixsuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > cervixsuboncchrxy.txt

cat cervixdeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' > cervixdeloncchrxy.txt

wc -l cervixinsocchr.txt

wc -l cervixinsoncchrxy.txt

awk '!seen[$0]++' cervixinsoncchrxy.txt > nodupcervixinsonc.txt

wc -l nodupcervixinsonc.txt

awk '!seen[$0]++' cervixdeloncchrxy.txt > nodupcervixdelonc.txt

awk '!seen[$0]++' cervixsuboncchrxy.txt > nodupcervixsubonc.txt

wc -l nodupcervixsubonc.txt

wc -l cervixsuboncchrxy.txt

head nodupcervixsubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupcervixinsonc.txt > nospacecervixinsonc.txt

wc -l nospacecervixinsonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupcervixsubonc.txt > nospacecervixsubonc.txt

wc -l nospacecervixsubonc.txt

wc -l nodupcervixsubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupcervixdelonc.txt > nospacecervixdelonc.txt
```

```
cat nospacecervix*.txt > nospacecervixmay.txt

wc -l nospacecervixmay.txt

cat 2headercervixinput.txt nospacecervixmay.txt > nospacecervixfinal_v91.txt

wc -l nospacecervixfinal_v91.txt
```

7.3.6 Melanoma

```
cut -f1,2,3,4,5,6 *.ins > melanomaskininsonc.txt

cut -f1,2,3,4,5,6 *.sub > melanomaskinsubonc.txt

cut -f1,2,3,4,5,6 *.del > melanomaskindelonc.txt

cat melanomaskininsonc.txt | sed -e 's/^/chr/' > melanomaskininsoncchr.txt

cat melanomaskinsubonc.txt | sed -e 's/^/chr/' > melanomaskinsuboncchr.txt

cat melanomaskindelonc.txt | sed -e 's/^/chr/' > melanomaskindeloncchr.txt

cat melanomaskininsoncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' >
melanomaskininsoncchrxy.txt

cat melanomaskinsuboncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' >
melanomaskinsuboncchrxy.txt

cat melanomaskindeloncchr.txt | sed -e 's/chr23/chrX/g'|sed -e 's/chr24/chrY/g' >
melanomaskindeloncchrxy.txt

wc -l melanomaskininsocchr.txt

wc -l melanomaskininsoncchrxy.txt

awk '!seen[$0]++' melanomaskininsoncchrxy.txt > nodupmelanomaskininsonc.txt

wc -l nodupmelanomaskininsonc.txt

awk '!seen[$0]++' melanomaskindeloncchrxy.txt > nodupmelanomaskindelonc.txt

awk '!seen[$0]++' melanomaskinsuboncchrxy.txt > nodupmelanomaskinsubonc.txt

wc -l nodupmelanomaskinsubonc.txt

wc -l melanomaskinsuboncchrxy.txt

head nodupmelanomaskinsubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupmelanomaskininsonc.txt > nospacemelanomaskininsonc.txt

wc -l nospacemelanomaskininsonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupmelanomaskinsubonc.txt > nospacemelanomaskinsubonc.txt

wc -l nospacemelanomaskinsubonc.txt

wc -l nodupmelanomaskinsubonc.txt

grep -Ev $'^\t|\t\t|\t$' nodupmelanomaskindelonc.txt > nospacemelanomaskindelonc.txt
```

```
cat nospacemelanomaskin*.txt > nospacemelanomaskinmay.txt

wc -l nospacemelanomaskinmay.txt

cat 2headermelanomaskininput.txt nospacemelanomaskinmay.txt >
nospacemelanomaskinfinal_v91.txt

wc -l nospacemelanomaskinfinal_v91.txt
```

7.4 Script for GDC Portal/ MC3 data and COSMIC data preparation and merging

7.4.1 Oropharyngeal SCC

EXTRACT BARCODES FROM GDC

```
cut -f16 TCGA.HNSC.somaticsniper.3f73f93c-a372-4097-bdff-77ca9760c6d3.DR-10.0.somatic.maf
> gdc_upperaerodigestivetract_barcodes.txt

#Extract the tumour barcode column from GDC cervical cancer MAF file


awk '!seen[$0]++' gdc_upperaerodigestivetract_barcodes.txt >
nodup_gdc_upperaerodigestivetract_barcodes.txt

#Delete all duplicates from the cervical cancer tumour barcode file


wc -l nodup_gdc_upperaerodigestivetract_barcodes.txt

#Count the number of lines in tumour barcode file


sed 1,5d nodup_gdc_upperaerodigestivetract_barcodes.txt >
noheader_gdc_upperaerodigestivetract_barcodes.txt

#delete header of tumour barcode file
```

EXTRACT GDC BARCODES FROM MC3 FILE

```
fgrep -w -f noheader_gdc_upperaerodigestivetract_barcodes.txt mc3.v0.2.8.PUBLIC.maf >
mc3_upperaerodigestivetract_variants.txt

#Extract all the cervical cancer barcodes from the public MAF file


wc -l mc3_upperaerodigestivetract_variants.txt

#Count the number of lines in mc3 cervical cancer file


cut -f5,6,7 mc3_upperaerodigestivetract_variants.txt >
chr_mc3_upperaerodigestivetract_variants.txt

#Cut columns with chromosome number, start position, end position
```

```
cut -f11 mc3_upperaerodigestivetract_variants.txt > ref_mc3_upperaerodigestivetract_variants.txt

#Cut column with reference allele
```

```
cut -f13 mc3_upperaerodigestivetract_variants.txt > alt_mc3_upperaerodigestivetract_variants.txt

#Cut column with alternate allele
```

```
cut -f16 mc3_upperaerodigestivetract_variants.txt > mc3_upperaerodigestivetract_barcodes.txt

#Cut column with tumour barcode
```

```
cat mc3_upperaerodigestivetract_barcodes.txt | sed 's/.\{13\}$//' > 2mc3_upperaerodigestivetract_barcodes.txt

#Delete last 13 characters of each line to shorten each tumour barcode so matches with identifiers in COSMIC
```

```
paste chr_mc3_upperaerodigestivetract_variants.txt ref_mc3_upperaerodigestivetract_variants.txt alt_mc3_upperaerodigestivetract_variants.txt 2mc3_upperaerodigestivetract_barcodes.txt > FINAL_mc3_upperaerodigestivetract_variants.txt

#Paste together chromosome number, start position, end position, reference allele, alternate allele and tumour barcode
```

```
sed '1d' FINAL_mc3_upperaerodigestivetract_variants.txt > FINAL2_mc3_upperaerodigestivetract_variants.txt

#Delete the header of the variant file
```

```
cat FINAL2_mc3_upperaerodigestivetract_variants.txt | sed -e 's/^/chr/' > FINAL3_mc3_upperaerodigestivetract_variants.txt

#Add 'Chr' to start of every line
```

COSMIC AND GDC FILE

```
cut -f6 nospaceupperaerodigestivetractfinal_v91.txt > upperaerodigestivetract_cosmic_ids.txt

#Cut barcode column from COSMIC cervical cancer
```

```
awk '!seen[$0]++' 2mc3_upperaerodigestivetract_barcodes.txt > nodupmc3barcodes.txt

#Delete all duplicates


awk '!seen[$0]++' upperaerodigestivetract_cosmic_ids.txt > nodupupperaerodigestivetract_cosmic_ids.txt

#Delete all duplicates


grep -v -f nodupmc3barcodes.txt nodupupperaerodigestivetract_cosmic_ids.txt > upperaerodigestivetract_cosmic_only_barcodes.txt

#Take lines from a mc3 barcodes and remove them from cosmic ids to make a new file with ids which are not present in mc3 barcodes


wc -l upperaerodigestivetract_cosmic_only_barcodes.txt

#Count the number of lines


cat upper_aerodigestive_tract_squamous_cell_carcinoma.txt | awk '$35 == "cell-line"' | cut -f5,35 | awk '!seen[$0]++' > upperaero_cellline.txt

cat upper_aerodigestive_tract_squamous_cell_carcinoma.txt | awk '$35 == "short-term"'| cut -f5,35 | awk '!seen[$0]++' > upperaero_culture.txt

cat upperaero_cellline.txt upperaero_culture.txt | cut -f1 > upperaero_celllineculture_ids.txt

cut -f5,10 upper_aerodigestive_tract_squamous_cell_carcinoma.txt | grep 'lip'| awk '!seen[$0]++' > upperaero_lip_subtype2.txt

cut -f1 upperaero_lip_subtype2.txt > upperaero_lip_subtype2ids.txt

cat upperaero_lip_subtype2ids.txt upperaero_celllineculture_ids.txt > upperaero_celllineculture_lip_ids.txt

#Tumour barcodes for data which is cell line or from the lip

grep -v -f upperaero_celllineculture_lip_ids.txt upperaerodigestivetract_cosmic_only_barcodes.txt > upperaerodigestivetract_tissue_cosmic_only_barcodes.txt

#find all the tumour barcodes which are only from tissues

wc -l upperaerodigestivetract_tissue_cosmic_only_barcodes.txt

#count the number of lines
```

2run.sh file:

```
#!/bin/bash

rm output.txt

while read id; do
 echo "searching for $id"
 if [ -z "$id" ]
 then
  echo "Empty line in $1"
 else
  grep "\s$id\b" $2 >> output.txt
 fi
done < $1
<END of Script>
```

```
sh 2run.sh upperaerodigestivetract_tissue_cosmic_only_barcodes.txt nospaceupperaerodigestivetractfinal_v91.txt
```

Result: Output.txt

```
#Take all the lines with cosmic only barcodes and match it no the cosmic upperaerodigestivetract file to output all variants of interest
```

MERGE GDC AND COSMIC VARIANTS

```
cat 2headerupperaerodigestivetractinput.txt FINAL3_mc3_upperaerodigestivetract_variants.txt output.txt > final_upperaerodigestivetract_mc3_cosmic.txt
#Add mc3 variants to cosmic variants
```

```
wc -l final_upperaerodigestivetract_mc3_cosmic.txt
#Count the number of lines
```

```
awk '!seen[$0]++' final_upperaerodigestivetract_mc3_cosmic.txt > 2final_upperaerodigestivetract_mc3_cosmic.txt
```

```
#Delete all duplicates
```

```
wc -l final_upperaerodigestivetract_mc3_cosmic.txt
```

```
#Count the number of lines
```

```
wc -l 2final_upperaerodigestivetract_mc3_cosmic.txt
```

```
#Count the number of lines
```

7.4.2 Oesophageal SCC

EXTRACT BARCODES FROM GDC

```
cut -f16 TCGA.ESCA.somaticsniper.56890408-24a5-4b5a-b822-aaf872c057b8.DR-10.0.somatic.maf  > gdc_oesophagus_barcodes.txt
```

```
#Extract the tumour barcode column from GDC cervical cancer MAF file
```

```
awk '!seen[$0]++' gdc_oesophagus_barcodes.txt > nodup_gdc_oesophagus_barcodes.txt
```

```
#Delete all duplicates from the cervical cancer tumour barcode file
```

```
wc -l nodup_gdc_oesophagus_barcodes.txt
```

```
#Count the number of lines in tumour barcode file
```

```
sed 1,5d nodup_gdc_oesophagus_barcodes.txt > noheader_gdc_oesophagus_barcodes.txt
```

```
#delete header of tumour barcode file
```

EXTRACT GDC BARCODES FROM MC3 FILE

```
fgrep -w -f noheader_gdc_oesophagus_barcodes.txt mc3.v0.2.8.PUBLIC.maf > mc3_oesophagus_variants.txt
```

```
#Extract all the cervical cancer barcodes from the public MAF file
```

```
wc -l mc3_oesophagus_variants.txt

#Count the number of lines in mc3 cervical cancer file
```

```
cut -f5,6,7 mc3_oesophagus_variants.txt > chr_mc3_oesophagus_variants.txt

#Cut columns with chromosome number, start position, end position
```

```
cut -f11 mc3_oesophagus_variants.txt > ref_mc3_oesophagus_variants.txt

#Cut column with reference allele
```

```
cut -f13 mc3_oesophagus_variants.txt > alt_mc3_oesophagus_variants.txt

#Cut column with alternate allele
```

```
cut -f16 mc3_oesophagus_variants.txt > mc3_oesophagus_barcodes.txt

#Cut column with tumour barcode
```

```
cat mc3_oesophagus_barcodes.txt | sed 's/.\{13\}$//' > 2mc3_oesophagus_barcodes.txt

#Delete last 13 characters of each line to shorten each tumour barcode so matches with
identifiers in COSMIC
```

```
paste chr_mc3_oesophagus_variants.txt ref_mc3_oesophagus_variants.txt
alt_mc3_oesophagus_variants.txt 2mc3_oesophagus_barcodes.txt >
FINAL_mc3_oesophagus_variants.txt

#Paste together chromosome number, start position, end position, reference allele, alternate
allele and tumour barcode
```

```
sed '1d' FINAL_mc3_oesophagus_variants.txt > FINAL2_mc3_oesophagus_variants.txt

#Delete the header of the variant file
```

```
cat FINAL2_mc3_oesophagus_variants.txt | sed -e 's/^/chr/' >
FINAL3_mc3_oesophagus_variants.txt

#Add 'Chr' to start of every line
```

COSMIC AND GDC FILE

```
cut -f6 nospaceoesophagusfinal_v91.txt > oesophagus_cosmic_ids.txt

#Cut barcode column from COSMIC cervical cancer


awk '!seen[$0]++' 2mc3_oesophagus_barcodes.txt > nodupmc3barcodes.txt

#Delete all duplicates


awk '!seen[$0]++' oesophagus_cosmic_ids.txt > nodupoesophagus_cosmic_ids.txt

#Delete all duplicates


grep -v -f nodupmc3barcodes.txt nodupoesophagus_cosmic_ids.txt >
oesophagus_cosmic_only_barcodes.txt

#Take lines from a mc3 barcodes and remove them from cosmic ids to make a new file with ids
which are not present in mc3 barcodes


wc -l oesophagus_cosmic_only_barcodes.txt

#Count the number of lines


grep -v -f oesophagus_celllineculture_ids.txt oesophagus_cosmic_only_barcodes.txt >
oesophagus_tissue_cosmic_only_barcodes.txt


wc -l oesophagus_tissue_cosmic_only_barcodes.txt
```

2run.sh file:

```
#!/bin/bash


rm output.txt


while read id; do

 echo "searching for $id"

 if [ -z "$id" ]

 then
```

```
  echo "Empty line in $1"
 else
  grep "\s$id\b" $2 >> output.txt
 fi
done < $1
```

<END of Script>

```
sh 2run.sh oesophagus_tissue_cosmic_only_barcodes.txt nospaceoesophagusfinal_v91.txt

#Result: output.txt
#Take all the lines with cosmic only barcodes and match it no the cosmic oesophagus file to output all variants of interest
```

MERGE GDC AND COSMIC VARIANTS

```
cat 2headeroesophagusinput.txt FINAL3_mc3_oesophagus_variants.txt output.txt > final_oesophagus_mc3_cosmic.txt
#Add mc3 variants to cosmic variants

wc -l final_oesophagus_mc3_cosmic.txt
#Count the number of lines

awk '!seen[$0]++' final_oesophagus_mc3_cosmic.txt > 2final_oesophagus_mc3_cosmic.txt
#Delete all duplicates

wc -l final_oesophagus_mc3_cosmic.txt
#Count the number of lines

wc -l 2final_oesophagus_mc3_cosmic.txt
#Count the number of lines
```

7.4.3 Lung SCC

EXTRACT BARCODES FROM GDC

```
cut -f16 TCGA.LUSC.somaticsniper.153233d7-0731-4252-a835-ecd067c5ae71.DR-10.0.somatic.maf > gdc_lung_barcodes.txt

#Extract the tumour barcode column from GDC cervical cancer MAF file
```

```
awk '!seen[$0]++' gdc_lung_barcodes.txt > nodup_gdc_lung_barcodes.txt

#Delete all duplicates from the cervical cancer tumour barcode file
```

```
wc -l nodup_gdc_lung_barcodes.txt

#Count the number of lines in tumour barcode file
```

```
sed 1,5d nodup_gdc_lung_barcodes.txt > noheader_gdc_lung_barcodes.txt

#delete header of tumour barcode file
```

EXTRACT GDC BARCODES FROM MC3 FILE

```
fgrep -w -f noheader_gdc_lung_barcodes.txt mc3.v0.2.8.PUBLIC.maf > mc3_lung_variants.txt

#Extract all the cervical cancer barcodes from the public MAF file
```

```
wc -l mc3_lung_variants.txt

#Count the number of lines in mc3 cervical cancer file
```

```
cut -f5,6,7 mc3_lung_variants.txt > chr_mc3_lung_variants.txt

#Cut columns with chromosome number, start position, end position
```

```
cut -f11 mc3_lung_variants.txt > ref_mc3_lung_variants.txt

#Cut column with reference allele
```

```
cut -f13 mc3_lung_variants.txt > alt_mc3_lung_variants.txt

#Cut column with alternate allele
```

```
cut -f16 mc3_lung_variants.txt > mc3_lung_barcodes.txt

#Cut column with tumour barcode
```

```
cat mc3_lung_barcodes.txt | sed 's/.\{13\}$//' > 2mc3_lung_barcodes.txt

#Delete last 13 characters of each line to shorten each tumour barcode so matches with
identifiers in COSMIC
```

```
paste chr_mc3_lung_variants.txt ref_mc3_lung_variants.txt alt_mc3_lung_variants.txt
2mc3_lung_barcodes.txt > FINAL_mc3_lung_variants.txt

#Paste together chromosome number, start position, end position, reference allele, alternate
allele and tumour barcode
```

```
sed '1d' FINAL_mc3_lung_variants.txt > FINAL2_mc3_lung_variants.txt

#Delete the header of the variant file
```

```
cat FINAL2_mc3_lung_variants.txt | sed -e 's/^/chr/' > FINAL3_mc3_lung_variants.txt

#Add 'Chr' to start of every line
```

COSMIC AND GDC FILE

```
cut -f6 nospacelungfinal_v91.txt > lung_cosmic_ids.txt

#Cut barcode column from COSMIC cervical cancer
```

```
awk '!seen[$0]++' 2mc3_lung_barcodes.txt > nodupmc3barcodes.txt

#Delete all duplicates
```

```
awk '!seen[$0]++' lung_cosmic_ids.txt > noduplung_cosmic_ids.txt

#Delete all duplicates
```

```
grep -v -f nodupmc3barcodes.txt noduplung_cosmic_ids.txt > lung_cosmic_only_barcodes.txt
```

```
#Remove mc3 barcodes from COSMIC ids to make a new file with ids which are unique to COSMIC and not present in mc3 barcodes
```

```
wc -l lung_cosmic_only_barcodes.txt
#Count the number of lines
```

```
grep -v -f lung_celllineculture_ids.txt lung_cosmic_only_barcodes.txt > lung_tissue_cosmic_only_barcodes.txt
```

```
wc -l lung_tissue_cosmic_only_barcodes.txt
```

```
fgrep -w -f lung_tissue_cosmic_only_barcodes.txt nospacelungfinal_v91.txt > cosmic_only_lung_variants.txt
#Take all the lines with cosmic only barcodes and match it no the cosmic lung file to output all variants of interest
```

MERGE GDC AND COSMIC VARIANTS

```
cat 2headerlunginput.txt FINAL3_mc3_lung_variants.txt cosmic_only_lung_variants.txt > final_lung_mc3_cosmic.txt
#Add mc3 variants to cosmic variants
```

```
wc -l final_lung_mc3_cosmic.txt
#Count the number of lines
```

```
awk '!seen[$0]++' final_lung_mc3_cosmic.txt > 2final_lung_mc3_cosmic.txt
#Delete all duplicates
```

```
wc -l final_lung_mc3_cosmic.txt
#Count the number of lines
```

```
wc -l 2final_lung_mc3_cosmic.txt
#Count the number of lines
```

7.4.4 Cervical SCC

EXTRACT BARCODES FROM GDC

```
cut -f16 TCGA.CESC.somaticsniper.45747e1f-7a37-4d82-b722-ae76fe5a0fcf.DR-10.0.somatic.maf > gdc_cervix_barcodes.txt

#Extract the tumour barcode column from GDC cervical cancer MAF file


awk '!seen[$0]++' gdc_cervix_barcodes.txt > nodup_gdc_cervix_barcodes.txt

#Delete all duplicates from the cervical cancer tumour barcode file


wc -l nodup_gdc_cervix_barcodes.txt

#Count the number of lines in tumour barcode file


sed 1,5d nodup_gdc_cervix_barcodes.txt > noheader_gdc_cervix_barcodes.txt

#delete header of tumour barcode file
```

EXTRACT GDC BARCODES FROM MC3 FILE

```
fgrep -w -f noheader_gdc_cervix_barcodes.txt mc3.v0.2.8.PUBLIC.maf > mc3_cervix_variants.txt

#Extract all the cervical cancer barcodes from the public MAF file


wc -l mc3_cervix_variants.txt

#Count the number of lines in mc3 cervical cancer file


cut -f5,6,7 mc3_cervix_variants.txt > chr_mc3_cervix_variants.txt

#Cut columns with chromosome number, start position, end position


cut -f11 mc3_cervix_variants.txt > ref_mc3_cervix_variants.txt

#Cut column with reference allele


cut -f13 mc3_cervix_variants.txt > alt_mc3_cervix_variants.txt

#Cut column with alternate allele
```

```
cut -f16 mc3_cervix_variants.txt > mc3_cervix_barcodes.txt

#Cut column with tumour barcode
```

```
cat mc3_cervix_barcodes.txt | sed 's/.\{13\}$//' > 2mc3_cervix_barcodes.txt

#Delete last 13 characters of each line to shorten each tumour barcode so matches with identifiers in COSMIC
```

```
paste chr_mc3_cervix_variants.txt ref_mc3_cervix_variants.txt alt_mc3_cervix_variants.txt 2mc3_cervix_barcodes.txt > FINAL_mc3_cervix_variants.txt

#Paste together chromosome number, start position, end position, reference allele, alternate allele and tumour barcode
```

```
sed '1d' FINAL_mc3_cervix_variants.txt > FINAL2_mc3_cervix_variants.txt

#Delete the header of the variant file
```

```
cat FINAL2_mc3_cervix_variants.txt | sed -e 's/^/chr/' > FINAL3_mc3_cervix_variants.txt

#Add 'Chr' to start of every line
```

COSMIC AND GDC FILE

```
cut -f6 nospacecervixfinal_v91.txt > cervix_cosmic_ids.txt

#Cut barcode column from COSMIC cervical cancer
```

```
awk '!seen[$0]++' 2mc3_cervix_barcodes.txt > nodupmc3barcodes.txt

#Delete all duplicates
```

```
awk '!seen[$0]++' cervix_cosmic_ids.txt > nodupcervix_cosmic_ids.txt

#Delete all duplicates
```

```
grep -v -f nodupmc3barcodes.txt nodupcervix_cosmic_ids.txt > cervix_cosmic_only_barcodes.txt

#Take lines from a mc3 barcodes and remove them from cosmic ids to make a new file with ids which are not present in mc3 barcodes
```

```
wc -l cervix_cosmic_only_barcodes.txt

#Count the number of lines


cat cervix_squamous_cell_carcinoma.txt | awk '$35 == "cell-line"' | cut -f5,35 | awk '!seen[$0]++'
> cervix_cellline.txt

cat cervix_squamous_cell_carcinoma.txt | awk '$35 == "short-term"'| cut -f5,35 | awk
'!seen[$0]++'> cervix_culture.txt

# For the cervix_squamous_cell_carcinoma.txt file,  column 5 was extracted which was the sample
name and column 35 which included a description of the tumour origin (cell-line). The next part of
the script deleted duplicates to produce a file with all the sample names which have originated
from cell lines. The same was done for samples which have originated from short-term cultures.

cat cervix_cellline.txt cervix_culture.txt | cut -f1 > cervix_celllineculture_ids.txt

# The next part of the script then collates both files with sample names for genomic data which
has originated from cells and not tumour samples to produce a file with this information.


grep -v -f cervix_celllineculture_ids.txt cervix_cosmic_only_barcodes.txt >
cervix_tissue_cosmic_only_barcodes.txt

#Remove cell line and cultured cell samples

wc -l cervix_tissue_cosmic_only_barcodes.txt

#count number of lines

fgrep -w -f cervix_cosmic_only_barcodes.txt nospacecervixfinal_v91.txt >
cosmic_only_cervix_variants.txt

#Take all the lines with cosmic only barcodes and match it no the cosmic cervix file to output all
variants of interest
```

MERGE GDC AND COSMIC VARIANTS

```
cat 2headercervixinput.txt FINAL3_mc3_cervix_variants.txt cosmic_only_cervix_variants.txt >
final_cervix_mc3_cosmic.txt

#Add mc3 variants to cosmic variants


wc -l final_cervix_mc3_cosmic.txt

#Count the number of lines
```

```
awk '!seen[$0]++' final_cervix_mc3_cosmic.txt > 2final_cervix_mc3_cosmic.txt
```

#Delete all duplicates

```
wc -l final_cervix_mc3_cosmic.txt
```

#Count the number of lines

```
wc -l 2final_cervix_mc3_cosmic.txt
```

#Count the number of lines

7.4.5 Melanoma

EXTRACT BARCODES FROM GDC

```
cut -f16 TCGA.SKCM.somaticsniper.b8ef7b54-adb8-4751-93bd-c26349be4252.DR-10.0.somatic.maf  > gdc_melanoma_barcodes.txt
```

#Extract the tumour barcode column from GDC cervical cancer MAF file

```
awk '!seen[$0]++' gdc_melanoma_barcodes.txt > nodup_gdc_melanoma_barcodes.txt
```

#Delete all duplicates from the cervical cancer tumour barcode file

```
wc -l nodup_gdc_melanoma_barcodes.txt
```

#Count the number of lines in tumour barcode file

```
sed 1,5d nodup_gdc_melanoma_barcodes.txt > noheader_gdc_melanoma_barcodes.txt
```

#delete header of tumour barcode file

EXTRACT GDC BARCODES FROM MC3 FILE

```
fgrep -w -f noheader_gdc_melanoma_barcodes.txt mc3.v0.2.8.PUBLIC.maf > mc3_melanoma_variants.txt
```

#Extract all the cervical cancer barcodes from the public MAF file

```
wc -l mc3_melanoma_variants.txt

#Count the number of lines in mc3 cervical cancer file


cut -f5,6,7 mc3_melanoma_variants.txt > chr_mc3_melanoma_variants.txt

#Cut columns with chromosome number, start position, end position


cut -f11 mc3_melanoma_variants.txt > ref_mc3_melanoma_variants.txt

#Cut column with reference allele


cut -f13 mc3_melanoma_variants.txt > alt_mc3_melanoma_variants.txt

#Cut column with alternate allele


cut -f16 mc3_melanoma_variants.txt > mc3_melanoma_barcodes.txt

#Cut column with tumour barcode


cat mc3_melanoma_barcodes.txt | sed 's/.\{13\}$//' > 2mc3_melanoma_barcodes.txt

#Delete last 13 characters of each line to shorten each tumour barcode so matches with identifiers in COSMIC


paste chr_mc3_melanoma_variants.txt ref_mc3_melanoma_variants.txt alt_mc3_melanoma_variants.txt 2mc3_melanoma_barcodes.txt > FINAL_mc3_melanoma_variants.txt

#Paste together chromosome number, start position, end position, reference allele, alternate allele and tumour barcode


sed '1d' FINAL_mc3_melanoma_variants.txt > FINAL2_mc3_melanoma_variants.txt

#Delete the header of the variant file


cat FINAL2_mc3_melanoma_variants.txt | sed -e 's/^/chr/' > FINAL3_mc3_melanoma_variants.txt

#Add 'Chr' to start of every line
```

COSMIC AND GDC FILE

```
cut -f6 nospacemelanomaskinfinal_v91.txt > melanoma_cosmic_ids.txt

#Cut barcode column from COSMIC cervical cancer


awk '!seen[$0]++' 2mc3_melanoma_barcodes.txt > nodupmc3barcodes.txt

#Delete all duplicates


awk '!seen[$0]++' melanoma_cosmic_ids.txt > nodupmelanoma_cosmic_ids.txt

#Delete all duplicates


grep -v -f nodupmc3barcodes.txt nodupmelanoma_cosmic_ids.txt >
melanoma_cosmic_only_barcodes.txt

#Take lines from a mc3 barcodes and remove them from cosmic ids to make a new file with ids
which are not present in mc3 barcodes


wc -l melanoma_cosmic_only_barcodes.txt

#Count the number of lines

cat skin_cutaneous_melanoma.txt | awk '$35 == "cell-line"' | cut -f5,35 | awk '!seen[$0]++' >
melanoma_cellline.txt

cat skin_cutaneous_melanoma.txt | awk '$35 == "short-term"'| cut -f5,35 | awk '!seen[$0]++' >
melanoma_culture.txt

cat melanoma_cellline.txt melanoma_culture.txt > melanoma_cell_culture.txt

cut -f1 melanoma_cell_culture.txt > melanoma_celllineculture_ids.txt

#Identify the cell line and cultured samples from COSMIC melanoma time and then combine
samples into one file and extract melanoma sample identifers.

cut -f5,9 skin_cutaneous_melanoma.txt | grep 'mucosal'| awk '!seen[$0]++' >
melanoma_mucosal_subtype1.txt

cut -f1 melanoma_mucosal_subtype1.txt > melanoma_mucosal_subtype1ids.txt

# In the second part of this script, the column with the sample names for mucosal melanoma
were mined to produce a new file, melanoma_mucosal_subtype1ids.txt.

cut -f5,9 skin_cutaneous_melanoma.txt | grep 'foot'| awk '!seen[$0]++' >
melanoma_foot_subtype1.txt

cut -f1 melanoma_foot_subtype1.txt > melanoma_foot_subtype1ids.txt
```

#The commands were used to remove any melanoma samples which originated from the foot region. Column five of the skin_cutaneous_melanoma.txt file included the sample names and column nine of the file included the site subtype 1 of where the genetic information has originated from. All samples which were described as originating from site subtype 1 ‘foot’ were extracted and placed in a new file ‘melanoma_foot_subtype1.txt’. Then all the sample names that originated from the foot were extracted from the melanoma_foot_subtype1.txt file to produce a new file ‘melanoma_foot_subtype1ids.txt’.

```
cat melanoma_mucosal_subtype1ids.txt melanoma_foot_subtype1ids.txt
melanoma_celllineculture_ids.txt | awk '!seen[$0]++' >
melanoma_celllineculture_mucosal_foot_ids.txt
```

#The next script then collates both files with sample names for genomic data which has originated from cell lines, cultured cells, melanomas from the foot and mucosal melanomas. All duplicate samples were then removed and a new file ‘melanoma_celllineculture_mucosal_foot_ids.txt’ was produced.

```
grep -v -f melanoma_celllineculture_mucosal_foot_ids.txt melanoma_cosmic_only_barcodes.txt >
melanoma_tissue_cosmic_only_barcodes.txt
```

#A file with all the melanoma sample names which were only present in the COSMIC database and were not in GDC portal was produced called ‘melanoma_cosmic_only_barcodes.txt’. The commands below show that all the melanoma sample names which originated from cell lines, cell culture, mucosal melanoma and melanoma from the foot were removed from the melanoma_cosmic_only_barcodes.txt file. A new file was then produced called melanoma_tissue_cosmic_only_barcodes.txt.

```
wc -l melanoma_tissue_cosmic_only_barcodes.txt
```

#count the number of lines in file

```
sh 2run.sh melanoma_tissue_cosmic_only_barcodes.txt nospacemelanomaskinfinal_v91.txt
```

#Result: output.txt

#Take all the lines with cosmic only barcodes and match it no the cosmic melanoma file to output all variants of interest

MERGE GDC AND COSMIC VARIANTS

```
cat 2headermelanomaskininput.txt FINAL3_mc3_melanoma_variants.txt output.txt >
final_melanoma_mc3_cosmic.txt
```

#Add mc3 variants to cosmic variants

```
wc -l final_melanoma_mc3_cosmic.txt
```

```
#Count the number of lines


awk '!seen[$0]++' final_melanoma_mc3_cosmic.txt > 2final_melanoma_mc3_cosmic.txt

#Delete all duplicates


wc -l final_melanoma_mc3_cosmic.txt

#Count the number of lines


wc -l 2final_melanoma_mc3_cosmic.txt

#Count the number of lines
```

7.5 Script for Maftools using R version 3.5.1

7.5.1 List of false positive genes

ABCC9	CSMD2	GUCY1A2	OR2W3	SCN9A	OR2T4
ACAN	CUBN	HERC2	OR4C46	SLC4A10	CNTNAP4
ADAMTS20	DCLK1	HRNR	OR4C6	SLITRK6	PARK2
ADAMTSL3	DMD	HSPA1L	OR4L1	SPTA1	
ADGRL2	DNAH11	IQGAP2	OR5L2	SRCAP	
AMPH	DNAH5	LAMA2	OR5T1	ST6GAL2	
ASH1L	DNAH7	LAMA4	OR6A2	TG	
ASTN2	DOCK4	LRP2	OR6K3	THBS2	
ASXL3	DPP6	LRRIQ1	OTOGL	THSD7B	
ATXN1	DTNA	MAGI2	OTUD7A	TMTC2	
BNC2	DYNC1I1	MGAM	PAIP1	TRIO	
BRINP3	DYSF	MGAT3	PCDH17	TRPA1	
CACHD1	EP400	MUC6	PCDH18	TRPS1	
CALCR	EPHA6	MXRA5	PCLO	TTN	
CASZ1	EPHB1	MYOM2	PCSK5	TXNIP	
CDC27	ESRRG	NALCN	PDZD2	WAC	
CEP170	EXOC2	NEB	PEG3	WDFY3	
CHD7	EYA4	NRXN1	PLEC	ZBTB20	

CNTN1	FBN2	NTRK2	PTCHD4	ZNF292
CNTN5	FLG	OR10A7	PXDN	BAGE2
CNTNAP5	FLT1	OR10G9	RAG1	TPTE
COL14A1	FOXQ1	OR10R2	RANBP6	RYR3
COL25A1	FREM2	OR2L13	RYR2	MUC5B
CPS1	GPC6	OR2M4	SACS	OR2G6
CSMD1	GRIA3	OR2T33	SCN5A	OR4M2

7.5.2 Skin SCC

```
library(maftools)


devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")


#read in file

lamlskin = read.maf(maf = "skin_final_april2021.tsv")

#plot summary

plotmafSummary(maf = lamlskin, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE, titvRaw = FALSE)

OncogenicPathways(maf = lamlskin)


#read in genes to ignore

genestoignore <- c(readLines("false_pos_genes_20190410.txt"))


#Oncoplot

oncoplot(maf = lamlskin, top = 38, genesToIgnore = genestoignore, sampleOrder = lamlskin@variants.per.sample$Tumor_Sample_Barcode)


#Lollipop plot

lollipopPlot(maf = lamlskin, gene = 'CDKN2A', AACol = 'Protein_Change', showMutationRate = TRUE)
lollipopPlot(maf = lamlskin, gene = 'CDKN2A', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'FAT1', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'HRAS', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'NOTCH1', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'NOTCH2', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'TP53', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'CDC27', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlskin, gene = 'TMEM222', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)

oncoplot(maf = lamlskin, genes = c("CDC27","TP53","CDKN2A","FAT1","HRAS","NOTCH1","NOTCH2","CCDC28A", "CHUK", "KIF4B", "PRB2", "TMEM222"), sampleOrder = lamlskin@variants.per.sample$Tumor_Sample_Barcode)



somaticInteractions(maf = lamlskin, genes = c("CDC27","TP53","CDKN2A","FAT1","HRAS","NOTCH1","NOTCH2","CCDC28A", "CHUK", "KIF4B", "PRB2", "TMEM222"), font = 0.6, nShiftSymbols = 2, showSum = FALSE, pvalue = c(0.05, 0.1))


#Oncogenic pathways

OncogenicPathways(maf = lamlskin)

PlotOncogenicPathways(maf = lamlskin, pathways = "RTK-RAS")

PlotOncogenicPathways(maf = lamlskin, pathways = "NOTCH")
PlotOncogenicPathways(maf = lamlskin, pathways = "WNT")
dgi = drugInteractions(maf = lamlskin, fontSize = 0.75)

#Mutation types
laml.titv = titv(maf = lamlskin, plot = FALSE, useSyn = TRUE)
#plot titv summary
plotTiTv(res = laml.titv)

#OncodriveCLUST
laml.sig = oncodrive(maf = lamlskin, AACol = 'Protein_Change', minMut = 5, pvalMethod = 'combined')

plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)

write.table(laml.sig, "skin_oncodrive_300321.txt")
```

7.5.3 Oropharyngeal SCC

```
library(maftools)

devtools::install_github(repo = "PoisonAlien/TCGAmutations")
library(TCGAmutations)
TCGAmutations::tcga_load(study = "SKCM")

#read in file
lamlorophar = read.maf(maf = "OROPHARYNGEAL_SCC_FINAL_v91_gdc_extra_sept.txt")

#plot summary
plotmafSummary(maf = lamlorophar, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE, titvRaw = FALSE)

#read in genes to ignore
genestoignore <- c(readLines("false_pos_genes_20190410.txt"))
```

```
slotNames(lamlorophar)
View (lamlorophar@variants.per.sample)
#oncoplot for top 100 most frequently mutated genes
oncoplot(maf = lamlorophar, top = 126, fontSize = 6, genesToIgnore = genestoignore,
sampleOrder = lamlorophar@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamlorophar, top = 32, fontSize = 8, genesToIgnore = genestoignore, sampleOrder
= lamlorophar@variants.per.sample$Tumor_Sample_Barcode)


if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")


BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)
laml.tnm = trinucleotideMatrix(maf = lamlorophar, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")
plotApobecDiff(tnm = laml.tnm, maf = lamlorophar)
library('NMF')
laml.sign = extractSignatures(mat = laml.tnm, nTry = 6, plotBestFitRes = FALSE)


plotApobecDiff(tnm = laml.tnm, maf = lamlorophar, pVal = 0.2)
plotSignatures(laml.sign, title_size = 1.4)


#Mutation types
laml.titv = titv(maf = lamlorophar, plot = FALSE, useSyn = TRUE)
#plot titv summary
plotTiTv(res = laml.titv)
```

```
laml.sig = oncodrive(maf = lamlorophar, AACol = 'Protein_Change', minMut = 5, pvalMethod =
'combined')

plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)

write.table(laml.sig, "oropharyngeal_oncodrive.txt")



laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")



lollipopPlot(maf = lamlorophar, gene = 'FAT1', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlorophar, gene = 'HRAS', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlorophar, gene = 'TP53', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)

lollipopPlot(maf = lamlorophar, gene = 'NOTCH2', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)
```

7.5.4 Oesophageal SCC

```
library(maftools)

devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")

#read in file

lamloeso = read.maf(maf = "oesophagus_GDC_Cosmicv91_extra_sept2020.txt")
```

```
#plot summary

plotmafSummary(maf = lamloeso, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE, titvRaw = FALSE)


#read in genes to ignore

genestoignore <- c(readLines("false_pos_genes_20190410.txt"))


slotNames(lamloeso)

View (lamloeso@variants.per.sample)

#oncoplot for top 100 most frequently mutated genes

oncoplot(maf = lamloeso, top = 125, fontSize = 6, genesToIgnore = genestoignore, sampleOrder = lamloeso@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamloeso, top = 33, fontSize = 8, genesToIgnore = genestoignore, sampleOrder = lamloeso@variants.per.sample$Tumor_Sample_Barcode)


if (!requireNamespace("BiocManager", quietly = TRUE))

  install.packages("BiocManager")


BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamloeso, prefix = 'chr', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamloeso)

library('NMF')

laml.sign = extractSignatures(mat = laml.tnm, nTry = 6, plotBestFitRes = FALSE)


plotApobecDiff(tnm = laml.tnm, maf = lamloeso, pVal = 0.2)

plotSignatures(laml.sign, title_size = 1.4)


#Mutation types
```

```
laml.titv = titv(maf = lamloeso, plot = FALSE, useSyn = TRUE)

#plot titv summary

plotTiTv(res = laml.titv)


laml.sig = oncodrive(maf = lamloeso, AACol = 'Protein_Change', minMut = 5, pvalMethod =
'combined')


plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)


write.table(laml.sig, "oesophageal_oncodrive.txt")


lollipopPlot(maf = lamloeso, gene = 'FAT1', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)


lollipopPlot(maf = lamloeso, gene = 'NOTCH1', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamloeso, gene = 'NOTCH2', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamloeso, gene = 'TP53', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)
```

7.5.5 Lung SCC

```
library(maftools)


devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")


#read in file

lamllung = read.maf(maf = "LUNG_SCC_FINAL_v91_gdc_extra.txt")
```

```
#plot summary

plotmafSummary(maf = lamllung, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE,
titvRaw = FALSE)


#read in genes to ignore

genestoignore <- c(readLines("false_pos_genes_20190410.txt"))


slotNames(lamllung)

View (lamllung@variants.per.sample)

#oncoplot for top 100 most frequently mutated genes

oncoplot(maf = lamllung, top = 128, fontSize = 6, genesToIgnore = genestoignore, sampleOrder =
lamllung@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamllung, top = 35, fontSize = 8, genesToIgnore = genestoignore, sampleOrder =
lamllung@variants.per.sample$Tumor_Sample_Barcode)


if (!requireNamespace("BiocManager", quietly = TRUE))

  install.packages("BiocManager")


BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamllung, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamllung)

library('NMF')

laml.sign = extractSignatures(mat = laml.tnm, nTry = 6, plotBestFitRes = FALSE)


plotApobecDiff(tnm = laml.tnm, maf = lamllung, pVal = 0.2)

plotSignatures(laml.sign, title_size = 1.4)
```

```
#Mutation types

laml.titv = titv(maf = lamllung, plot = FALSE, useSyn = TRUE)

#plot titv summary

plotTiTv(res = laml.titv)




laml.sig = oncodrive(maf = lamllung, AACol = 'Protein_Change', minMut = 5, pvalMethod = 'poisson')


plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)

write.table(laml.sig, "lung_oncodrive.txt")




lollipopPlot(maf = lamllung, gene = 'FAT1', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamllung, gene = 'HRAS', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamllung, gene = 'TP53', AACol = 'Protein_Change', showMutationRate = TRUE, showDomainLabel = FALSE)
```

7.5.6 Cervical SCC

```
library(maftools)


devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")


#read in file

lamlcervix = read.maf(maf = "CERVIX_SCC_FINAL_v91_gdc_extra.txt")
```

```
#plot summary

plotmafSummary(maf = lamlcervix, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE, titvRaw = FALSE)


#read in genes to ignore

genestoignore <- c(readLines("false_pos_genes_20190410.txt"))


slotNames(lamlcervix)

View (lamlcervix@variants.per.sample)

#oncoplot for top 100 most frequently mutated genes

oncoplot(maf = lamlcervix, top = 122, fontSize = 6, genesToIgnore = genestoignore, sampleOrder = lamlcervix@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamlcervix, top = 36, fontSize = 8, genesToIgnore = genestoignore, sampleOrder = lamlcervix@variants.per.sample$Tumor_Sample_Barcode)


if (!requireNamespace("BiocManager", quietly = TRUE))

  install.packages("BiocManager")


BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamlcervix, prefix = 'chr', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamlcervix)

library('NMF')

laml.sign = extractSignatures(mat = laml.tnm, nTry = 6, plotBestFitRes = FALSE)


plotApobecDiff(tnm = laml.tnm, maf = laml, pVal = 0.2)

plotSignatures(laml.sign, title_size = 1.4)
```

```
#Mutation types

laml.titv = titv(maf = lamlcervix, plot = FALSE, useSyn = TRUE)

#plot titv summary

plotTiTv(res = laml.titv)



laml.sig = oncodrive(maf = lamlcervix, AACol = 'Protein_Change', minMut = 5, pvalMethod =
'combined')


plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)

write.table(laml.sig, "cervix_oncodrive")



lollipopPlot(maf = lamlcervix, gene = 'TP53', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)
```

7.5.7 Basal Cell Carcinoma

```
library(maftools)


devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")


#read in file

lamlbcc = read.maf(maf = "26950094_bcc_Output.tsv")


#plot summary

plotmafSummary(maf = lamlbcc, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE,
titvRaw = FALSE)


#read in genes to ignore

genestoignore <- c(readLines("false_pos_genes_20190410.txt"))
```

```
#oncoplot for top 100 most frequently mutated genes

oncoplot(maf = lamlbcc, top = 130, fontSize = 0.5, genesToIgnore = genestoignore, sampleOrder =
lamlbcc@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamlbcc, top = 33, genesToIgnore = genestoignore, sampleOrder =
lamlbcc@variants.per.sample$Tumor_Sample_Barcode)


#Mutation types

laml.titv = titv(maf = lamlbcc, plot = FALSE, useSyn = TRUE)

#plot titv summary

plotTiTv(res = laml.titv)


laml.sig = oncodrive(maf = lamlbcc, AACol = 'Protein_Change', minMut = 5, pvalMethod =
'combined')


plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE)


write.table(laml.sig, "bcc_oncodrive.txt")




library(maftools)


devtools::install_github(repo = "PoisonAlien/TCGAmutations")

library(TCGAmutations)

TCGAmutations::tcga_load(study = "SKCM")
```

7.5.8 Melanoma

```
#read in file

lamloeso = read.maf(maf = "melanoma_GDC_Cosmicv91Outputoct2020.tsv")

lamlmela = read.maf(maf = "final_melanoma_gdc_cosmic_2021.tsv")

lamlmela = read.maf(maf = "melanoma_final_may2021_new.tsv")

lamlmela = read.maf(maf = "melanoma_GDC_COSMIC_EXTRA_JUN2021_final5.tsv")


oncoplot(maf = lamlmela, genes = c("TP53","CDKN2A","PPP6C"), sampleOrder =
lamlmela@variants.per.sample$Tumor_Sample_Barcode)


lollipopPlot(maf = lamlmela, gene='C3', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)




lollipopPlot(maf = lamlmela, gene = 'CDKN2A', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'ERBB2', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'EYA1', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'PPP6C', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'HRAS', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'TP53', AACol = 'Protein_Change', showMutationRate = TRUE,
showDomainLabel = FALSE)


lollipopPlot(maf = lamlmela, gene = 'NOTCH2', AACol = 'Protein_Change', showMutationRate =
TRUE, showDomainLabel = FALSE)
```

```
oncoplot(maf = lamlmela, genes =
c("CDKN2A","TP53","ERBB2","EYA1","PPP6C","NOTCH2","C3","HRAS"), sampleOrder =
lamlmela@variants.per.sample$Tumor_Sample_Barcode)


oncoplot(maf = lamlmela, pathways = "auto", gene_mar = 8, fontSize = 0.6)




OncogenicPathways(maf = lamlmela)

PlotOncogenicPathways(maf = lamlmela, pathways = "RTK-RAS")

PlotOncogenicPathways(maf = lamlmela, pathways = "NOTCH")

PlotOncogenicPathways(maf = lamlmela, pathways = "WNT")
```

7.6 Script for producing mutation signatures

7.6.1 Single base substitution mutation signatures produced using Maftools in R version 3.5.1

```
BiocManager::install("BSgenome.Hsapiens.UCSC.hg19")

library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamlcervix, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamlcervix)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")


maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")
```

```
install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main =
"cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamlcervix, sig_res = laml.sig)



plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamllung, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamllung)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 5)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")


maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")


install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main =
"cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamllung, sig_res = laml.sig)
```

```
plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamlorophar, prefix = 'chr', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamlorophar)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 4)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")


samplesforsignatures <- laml.sig$contributions

write.csv(samplesforsignatures, file = "oropharyngeal_signature_samples.csv")



maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")


install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main = "cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamlorophar, sig_res = laml.sig)



plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)
```

```
library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamloeso, prefix = 'chr', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamloeso)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")


maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")


install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main = "cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamloeso, sig_res = laml.sig)



plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamlbcc, prefix = 'chr', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamlbcc)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)
```

```
plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")


maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")


install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main =
"cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamlbcc, sig_res = laml.sig)



plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)


library(BSgenome.Hsapiens.UCSC.hg19, quietly = TRUE)

laml.tnm = trinucleotideMatrix(maf = lamlskin, prefix = 'chr', add = TRUE, ref_genome =
"BSgenome.Hsapiens.UCSC.hg19")

plotApobecDiff(tnm = laml.tnm, maf = lamlskin)

library('NMF')

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)


laml.og30.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "legacy")


laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")
```

```
maftools::plotSignatures(nmfRes = laml.sig, title_size = 0.8, sig_db = "SBS")


install.packages('pheatmap')

library('pheatmap')

pheatmap::pheatmap(mat = laml.og30.cosm$cosine_similarities, cluster_rows = FALSE, main = 
"cosine similarity against validated signatures")


laml.se = signatureEnrichment(maf = lamlbcc, sig_res = laml.sig)



plotEnrichmentResults(enrich_res = laml.se, pVal = 0.05)
```

7.6.2 Double base substitution mutation signatures using Sigminer in R version 3.5.1

```
install.packages("sigminer", dependencies = TRUE)

library("sigminer")

library(maftools)

laml <- read_maf(maf='CERVIX_SCC_FINAL_v91_gdc_extra.txt')

head(laml@data)

slotNames(laml)


mt_tally_DBS <- sig_tally(

  laml,

  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",

  useSyn = TRUE,

  mode = "DBS",

  add_trans_bias = TRUE

)



str(mt_tally_DBS$all_matrices, max.level = 1)
```

```
library('NMF')
mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,
            range = 2:5,
            nrun = 2,
            use_random = TRUE,
            cores = 4,
            pConstant = 1e-13,
            verbose = TRUE
)


show_sig_number_survey(mt_est$survey, right_y = NULL)



mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,
           n_sig = 3,
           nrun = 10,
           cores = 4,
           pConstant = 1e-13

)

sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")



show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)

show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)
```

```
show_cosmic_sig_profile(sig_index = c(11, 2, 4), style = "cosmic", sig_db = "DBS")

#>




laml <- read_maf(maf='LUNG_SCC_FINAL_v91_gdc_extra.txt')

head(laml@data)

slotNames(laml)


mt_tally_DBS <- sig_tally(

  laml,

  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",

  useSyn = TRUE,

  mode = "DBS",

  add_trans_bias = TRUE

)



str(mt_tally_DBS$all_matrices, max.level = 1)


library('NMF')

mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,

                      range = 2:5,

                      nrun = 2,

                      use_random = TRUE,

                      cores = 4,

                      pConstant = 1e-13,

                      verbose = TRUE

)
```

```
show_sig_number_survey(mt_est$survey, right_y = NULL)



mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,
             n_sig = 3,
             nrun = 10,
             cores = 4,
             pConstant = 1e-13

)

sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")



show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)

show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)


show_cosmic_sig_profile(sig_index = c(2, 6, 1), style = "cosmic", sig_db = "DBS")
#>



laml <- read_maf(maf='OROPHARYNGEAL_SCC_FINAL_v91_gdc_extra_sept.txt')
head(laml@data)
slotNames(laml)
```

```
mt_tally_DBS <- sig_tally(
  laml,
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  useSyn = TRUE,
  mode = "DBS",
  add_trans_bias = TRUE
)


str(mt_tally_DBS$all_matrices, max.level = 1)

library('NMF')
mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,
                       range = 2:5,
                       nrun = 2,
                       use_random = TRUE,
                       cores = 4,
                       pConstant = 1e-13,
                       verbose = TRUE
)


show_sig_number_survey(mt_est$survey, right_y = NULL)



mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,
                     n_sig = 3,
                     nrun = 10,
                     cores = 4,
                     pConstant = 1e-13
)

sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")



show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)

show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)


show_cosmic_sig_profile(sig_index = c(1, 2, 4), style = "cosmic", sig_db = "DBS")
#>


laml <- read_maf(maf='oesophagus_GDC_Cosmicv91_extra_sept2020.txt')
head(laml@data)
slotNames(laml)

mt_tally_DBS <- sig_tally(
  laml,
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  useSyn = TRUE,
  mode = "DBS",
  add_trans_bias = TRUE
)


str(mt_tally_DBS$all_matrices, max.level = 1)
```

```
library('NMF')

mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,

                    range = 2:5,

                    nrun = 2,

                    use_random = TRUE,

                    cores = 4,

                    pConstant = 1e-13,

                    verbose = TRUE

)



show_sig_number_survey(mt_est$survey, right_y = NULL)




mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,

                  n_sig = 3,

                  nrun = 10,

                  cores = 4,

                  pConstant = 1e-13


)


sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")




show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)


show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)
```

```
show_cosmic_sig_profile(sig_index = c(11, 4, 2), style = "cosmic", sig_db = "DBS")
#>



laml <- read_maf(maf='26950094_bcc_Output.tsv')
head(laml@data)
slotNames(laml)

mt_tally_DBS <- sig_tally(
  laml,
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  useSyn = TRUE,
  mode = "DBS",
  add_trans_bias = TRUE
)



str(mt_tally_DBS$all_matrices, max.level = 1)

library('NMF')
mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,
                       range = 2:5,
                       nrun = 2,
                       use_random = TRUE,
                       cores = 4,
                       pConstant = 1e-13,
                       verbose = TRUE
)
```

```
show_sig_number_survey(mt_est$survey, right_y = NULL)




mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,
                  n_sig = 4,
                  nrun = 10,
                  cores = 4,
                  pConstant = 1e-13

)

sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")




show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)

show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)



show_cosmic_sig_profile(sig_index = c(1, 6), style = "cosmic", sig_db = "DBS")
#>





laml <- read_maf(maf='skin_final_oct2020.tsv')
head(laml@data)
slotNames(laml)
```

```
mt_tally_DBS <- sig_tally(
  laml,
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  useSyn = TRUE,
  mode = "DBS",
  add_trans_bias = TRUE
)


str(mt_tally_DBS$all_matrices, max.level = 1)

library('NMF')
mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,
                    range = 2:5,
                    nrun = 2,
                    use_random = TRUE,
                    cores = 4,
                    pConstant = 1e-13,
                    verbose = TRUE
)


show_sig_number_survey(mt_est$survey, right_y = NULL)



mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,
                  n_sig = 2,
                  nrun = 10,
                  cores = 4,
                  pConstant = 1e-13
)

sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")



show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)

show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)


show_cosmic_sig_profile(sig_index = c(1, 6), style = "cosmic", sig_db = "DBS")
#>



laml <- read_maf(maf='skin_final_april2021.tsv')
head(laml@data)
slotNames(laml)

mt_tally_DBS <- sig_tally(
  laml,
  ref_genome = "BSgenome.Hsapiens.UCSC.hg19",
  useSyn = TRUE,
  mode = "DBS",
  add_trans_bias = TRUE
)


str(mt_tally_DBS$all_matrices, max.level = 1)
```

```
library('NMF')

mt_est <- sig_estimate(mt_tally_DBS$all_matrices$DBS_78,

                range = 2:5,

                nrun = 2,

                use_random = TRUE,

                cores = 4,

                pConstant = 1e-13,

                verbose = TRUE

)



show_sig_number_survey(mt_est$survey, right_y = NULL)




mt_sig<- sig_extract(mt_tally_DBS$all_matrices$DBS_78,

               n_sig = 2,

               nrun = 10,

               cores = 4,

               pConstant = 1e-13


)


sim_v3 <- get_sig_similarity(mt_sig, sig_db = "DBS")




show_sig_profile(mt_sig, mode = "DBS", paint_axis_text = FALSE, x_label_angle = 90)


show_sig_profile(mt_sig, mode = "DBS", style = "cosmic", x_label_angle = 90)
```

show_cosmic_sig_profile(sig_index = c(1, 6), style = "cosmic", sig_db = "DBS")

#>

7.7 MutSig2CV

MutSig2CV installation

download from here:

https://www.google.com/url?q=https%3A%2F%2Fsoftware.broadinstitute.org%2Fcancer%2Fcga%2Fsites%2Fdefault%2Ffiles%2Fdata%2Ftools%2Fmutsig%2FMutSig2CV.tar.gz&sa=D&sntz=1&usg=AFQjCNHGsvWELJJd5q_8O7r75HK1zFregA

Extract the .tar.gz to Downloads (10.1GB directory)

TO INSTALL:

At this time, MutSig is available for 64 bit Linux systems only. MutSig

requires the MATLAB R2013a runtime to be installed. This runtime environment

should be universally compatible with any recent Linux distribution; we have

successfully tested it on on 64 bit CentOS 5, RHEL 6, and Debian 8.2.

Users must download and install the runtime environment from here:

http://www.mathworks.com/supportfiles/MCR_Runtime/R2013a/MCR_R2013a_glnxa64_installer.zip

Installation instructions can be found here:

http://www.mathworks.com/help/compiler/install-the-matlab-runtime.html

matlab runtime installation notes:

$cd Downsload/MCR_2013a

$sudo ./install

$sudo gedit /etc/ld.so.conf.d/randomLibs.conf

Copy this into the randomLibs.conf file that opens in gedit:

```
/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/runtime/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/bin/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/os/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64/native_threads:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64/server:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64
```

Next, set the XAPPLRESDIR environment variable to the following value:

```
/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/X11/app-defaults
```

Save the gedit file.

```
$sudo ldconfig
$export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/runtime/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/bin/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/os/glnxa64:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64/native_threads:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64/server:/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/sys/java/jre/glnxa64/jre/lib/amd64
$export XAPPLRESDIR=/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81/X11/app-defaults
```

Once the runtime is successfully installed, you must add it to your

LD_LIBRARY_PATH. You will likely want to add the following lines to your

.bashrc/.cshrc, so that the MATLAB runtime is always on your path.

For a bash shell:

```
$sudo nano ~/.bashrc
```

paste in these lines into ./bashrc and save file.

```
mcr_root=/usr/local/MATLAB/MATLAB_Compiler_Runtime/v81
```

```
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/bin/glnxa64/
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/sys/java/jre/glnxa64/jre/lib/amd64
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/sys/java/jre/glnxa64/jre/lib/amd64/server
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/sys/java/jre/glnxa64/jre/lib/amd64/native_threads
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/sys/os/glnxa64
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/bin/glnxa64
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/runtime/glnxa64
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$mcr_root/lib
```

log out of ubuntu and log back in.

If all is well, running test/MCR_test from the mutsig2cv root directory should

output the runtime installation path. If you receive the following error:

```
test/MCR_test: error while loading shared libraries: libmwlaunchermain.so:
cannot open shared object file: No such file or directory
```

you have likely not updated your LD_LIBRARY_PATH correctly.

```
###error here:
mjrz@mjrz-Vig800S:~/Downloads/mutsig2cv$ test/MCR_test
bash: test/MCR_test: No such file or directory
####error stop
```

Attempted to run MutSig2CV:

move the following files to the /home/mjrz/Downloads/mutsig2cv directory:

LUSC.coverage.txt

LUSC.maf

gene.coveriates.txt

create a new directory: /home/mjrz/Downloads/mutsig2cv/LUSC - this is where the results will be saved to.

```
mjrz@mjrz-Vig800S:~/Downloads/mutsig2cv$ nohup ./MutSig2CV LUSC.maf ./LUSC/
```

7.8 dNdScv

dndscv

```
cervix <- read.delim("cervixdndscvfinal1sept.txt")

cervix$chr = gsub("chr","",as.vector(cervix$chr))

dndsout = dndscv(cervix)

sel_cv = dndsout$sel_cv

write.csv(sel_cv, "2cervix_selcv.csv")

print(head(sel_cv), digits = 3)

signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]

rownames(signif_genes) = NULL

print(signif_genes)

write.table(signif_genes, "cervix_signifgenes.txt")


library(dndscv)

lung <- read.delim("nospacefinallung1.txt")

lung$chr = gsub("chr","",as.vector(lung$chr))

dndsout = dndscv(lung)

sel_cv = dndsout$sel_cv

write.csv(sel_cv, "2lung_selcv.csv")

print(head(sel_cv), digits = 3)

signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]

rownames(signif_genes) = NULL

print(signif_genes)

write.table(signif_genes, "lung_dndscv2.txt")



oesophagus <- read.delim("nospaceoesofinal1.txt")
```

```
oesophagus$chr = gsub("chr","",as.vector(oesophagus$chr))
dndsout = dndscv(oesophagus)
sel_cv = dndsout$sel_cv
write.csv(sel_cv, "2oeso_selcv.csv")
print(head(sel_cv), digits = 3)
signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]
rownames(signif_genes) = NULL
print(signif_genes)
write.table(signif_genes, "oesophagus_dndscv2.txt")

library("dndscv")
hnscc <- read.delim("nospaceoropharyngealfinal1.txt")
hnscc$chr = gsub("chr","",as.vector(hnscc$chr))
dndsout = dndscv(hnscc)
sel_cv = dndsout$sel_cv
write.csv(sel_cv, "2hnscc_selcv.csv")
print(head(sel_cv), digits = 3)
signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]
rownames(signif_genes) = NULL
print(signif_genes)
write.table(signif_genes, "hnscc_dndscv2.txt")



library("dndscv")
skin <- read.delim("FINAL_skin_dndscv.txt")
skin$chr = gsub("chr","",as.vector(skin$chr))
dndsout = dndscv(skin)
sel_cv = dndsout$sel_cv
write.csv(sel_cv, "2skin_selcv.csv")
print(head(sel_cv), digits = 3)
```

```
signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]
rownames(signif_genes) = NULL
print(signif_genes)
write.table(signif_genes, "skin_dndscv2.txt")


library("dndscv")
skin <- read.delim("skin_final_april2021_dndscv.txt")
skin$chr = gsub("chr","",as.vector(skin$chr))
dndsout = dndscv(skin)
sel_cv = dndsout$sel_cv
write.csv(sel_cv, "2skin_selcv_april2021.csv")
print(head(sel_cv), digits = 3)
signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]
rownames(signif_genes) = NULL
print(signif_genes)
write.table(signif_genes, "skin_dndscv2_april2021.txt")

library("dndscv")
skin <- read.delim("dndscvfinal_header_melanoma.txt")
skin$chr = gsub("chr","",as.vector(skin$chr))
dndsout = dndscv(skin)
sel_cv = dndsout$sel_cv
write.csv(sel_cv, "2skin_selcv_melanoma_may2021.csv")
print(head(sel_cv), digits = 3)
signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]
rownames(signif_genes) = NULL
print(signif_genes)
write.table(signif_genes, "skin_dndscv2_melanoma_may2021.txt")

library("dndscv")
```

```
skin <- read.delim("dndscvfinal_header_bcc.txt")

skin$chr = gsub("chr","",as.vector(skin$chr))

dndsout = dndscv(skin)

sel_cv = dndsout$sel_cv

write.csv(sel_cv, "2skin_selcv_bcc_may2021.csv")

print(head(sel_cv), digits = 3)

signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]

rownames(signif_genes) = NULL

print(signif_genes)

write.table(signif_genes, "skin_dndscv2_bcc_may2021.txt")


library("dndscv")

skin <- read.delim("dndscv_melanoma_jun21_header.txt")

skin$chr = gsub("chr","",as.vector(skin$chr))

dndsout = dndscv(skin)

sel_cv = dndsout$sel_cv

write.csv(sel_cv, "2skin_selcv_melanoma_june2021.csv")

print(head(sel_cv), digits = 3)

signif_genes = sel_cv[sel_cv$qglobal_cv<0.1, c("gene_name","qglobal_cv")]

rownames(signif_genes) = NULL

print(signif_genes)

write.table(signif_genes, "skin_dndscv2_melanoma_jun2021.txt")
```

7.9 Literature Search Terms

Medline OVID was used to conduct the literature search.

7.9.1 Skin search terms

1. carcinoma, squamous cell/ or bowen's disease/

2. limit 1 to yr="2007 -Current"

3. Keratosis, Actinic/

4. limit 3 to yr="2007 -Current"

5. (squamous adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

6. limit 5 to yr="2007 -Current"

7. SCC.mp.

8. limit 7 to yr="2007 -Current"

9. 2 or 4 or 6 or 8

10. skin/ or epidermis/ or hair follicle/

11. limit 10 to yr="2007 -Current"

12. cutaneous.mp.

13. limit 12 to yr="2007 -Current"

14. (skin or epiderm*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

15. limit 14 to yr="2007 -Current"

16. 11 or 13 or 15

17. whole genome sequencing/ or whole exome sequencing/

18. limit 17 to yr="2007 -Current"

19. Genomics/

20. limit 19 to yr="2007 -Current"

21. DNA/

22. limit 21 to yr="2007 -Current"

23. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

24. limit 23 to yr="2007 -Current"

25. 18 or 20 or 22 or 24

26. Mutation/

27. limit 26 to yr="2007 -Current"

28. Selection, Genetic/

29. limit 28 to yr="2007 -Current"

30. Genetic Variation/

31. limit 30 to yr="2007 -Current"

32. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word,

protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

33. limit 32 to yr="2007 -Current"

34. 27 or 29 or 31 or 33

7.9.2 Oropharyngeal SCC search terms

1. carcinoma, squamous cell/

2. limit 1 to yr="2007 -Current"

3. (squamous adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

4. limit 3 to yr="2007 -Current"

5. SCC.mp.

6. limit 5 to yr="2007 -Current"

7. 2 or 4 or 6

8. pharynx/ or hypopharynx/ or nasopharynx/ or oropharynx/

9. limit 8 to yr="2007 -Current"

10. (head and neck).mp.

11. limit 10 to yr="2007 -Current"

12. oral cavit*.mp.

13. limit 12 to yr="2007 -Current"

14. 9 or 11 or 13

15. whole genome sequencing/ or whole exome sequencing/

16. limit 15 to yr="2007 -Current"

17. Genomics/

18. limit 17 to yr="2007 -Current"

19. DNA/

20. limit 19 to yr="2007 -Current"

21. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

22. limit 21 to yr="2007 -Current"

23. 16 or 18 or 20 or 22

24. Mutation/

25. limit 24 to yr="2007 -Current"

26. Selection, Genetic/

27. limit 26 to yr="2007 -Current"

28. Genetic Variation/

29. limit 28 to yr="2007 -Current"

30. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

31. limit 30 to yr="2007 -Current"

32. 25 or 27 or 29 or 31

33. 7 and 14 and 23 and 32

7.9.3 Oesophageal SCC search terms

1. carcinoma, squamous cell/

2. limit 1 to yr="2007 -Current"

3. (squamous adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

4. limit 3 to yr="2007 -Current"

5. SCC.mp.

6. limit 5 to yr="2007 -Current"

7. 2 or 4 or 6

8. Esophagus/

9. limit 8 to yr="2007 -Current"

10. Esophageal.mp.

11. limit 10 to yr="2007 -Current"

12. oesophag*.mp.

13. limit 12 to yr="2007 -Current"

14. 9 or 11 or 13

15. whole genome sequencing/ or whole exome sequencing/

16. limit 15 to yr="2007 -Current"

17. Genomics/

18. limit 17 to yr="2007 -Current"

19. DNA/

20. limit 19 to yr="2007 -Current"

21. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

22. limit 21 to yr="2007 -Current"

23. 16 or 18 or 20 or 22

24. Mutation/

25. limit 24 to yr="2007 -Current"

26. Selection, Genetic/

27. limit 26 to yr="2007 -Current"

28. Genetic Variation/

29. limit 28 to yr="2007 -Current"

30. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

31. limit 30 to yr="2007 -Current"

32. 25 or 27 or 29 or 31

33. 7 and 14 and 23 and 32

7.9.4 Lung SCC search terms

1. carcinoma, squamous cell/

2. limit 1 to yr="2007 -Current"

3. (squamous adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

4. limit 3 to yr="2007 -Current"

5. SCC.mp.

6. limit 5 to yr="2007 -Current"

7. 2 or 4 or 6

8. respiratory system/ or lung/ or bronchi/ or bronchioles/ or pulmonary alveoli/ or respiratory mucosa/ or trachea/

9. limit 8 to yr="2007 -Current"

10. Carcinoma, Non-Small-Cell Lung/

11. limit 10 to yr="2007 -Current"

12. NSCLC.mp.

13. limit 12 to yr="2007 -Current"

14. respiratory.mp.

15. limit 14 to yr="2007 -Current"

16. bronchus.mp.

17. limit 16 to yr="2007 -Current"

18. bronchial.mp.

19. limit 18 to yr="2007 -Current"

20. 9 or 11 or 13 or 15 or 17 or 19

21. whole genome sequencing/ or whole exome sequencing/

22. limit 21 to yr="2007 -Current"

23. Genomics/

24. limit 23 to yr="2007 -Current"

25. DNA/

26. limit 25 to yr="2007 -Current"

27. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

28. limit 27 to yr="2007 -Current"

29. 22 or 24 or 26 or 28

30. Mutation/

31. limit 30 to yr="2007 -Current"

32. Selection, Genetic/

33. limit 32 to yr="2007 -Current"

34. Genetic Variation/

35. limit 34 to yr="2007 -Current"

36. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

37. limit 36 to yr="2007 -Current"

38. 31 or 33 or 35 or 37

39. 7 and 20 and 29 and 38

7.9.5 Cervical SCC search terms

1. carcinoma, squamous cell/

2. limit 1 to yr="2007 -Current"

3. (squamous adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

4. limit 3 to yr="2007 -Current"

5. SCC.mp.

6. limit 5 to yr="2007 -Current"

7. 2 or 4 or 6

8. CERVICAL INTRAEPITHELIAL NEOPLASIA/

9. limit 8 to yr="2007 -Current"

10. cervical.mp.

11. limit 10 to yr="2007 -Current"

12. cervix.mp.

13. limit 12 to yr="2007 -Current"

14. 9 or 11 or 13

15. whole genome sequencing/ or whole exome sequencing/

16. limit 15 to yr="2007 -Current"

17. Genomics/

18. limit 17 to yr="2007 -Current"

19. DNA/

20. limit 19 to yr="2007 -Current"

21. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

22. limit 21 to yr="2007 -Current"

23. 16 or 18 or 20 or 22

24. Mutation/

25. limit 24 to yr="2007 -Current"

26. Selection, Genetic/

27. limit 26 to yr="2007 -Current"

28. Genetic Variation/

29. limit 28 to yr="2007 -Current"

30. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

31. limit 30 to yr="2007 -Current"

32. 25 or 27 or 29 or 31

33. 7 and 14 and 23 and 32

7.9.6 Basal cell carcinoma search terms

1. Carcinoma, Basal Cell/

2. limit 1 to yr="2007 -Current"

3. BCC.mp.

4. limit 3 to yr="2007 -Current"

5. (basal adj3 (cancer* or carcinoma* or tumor* or tumour*)).mp.

6. limit 5 to yr="2007 -Current"

7. rodent ulcer*.mp.

8. limit 7 to yr="2007 -Current"

9. 2 or 4 or 6 or 8

10. skin/ or epidermis/ or hair follicle/

11. limit 10 to yr="2007 -Current"

12. cutaneous.mp.

13. limit 12 to yr="2007 -Current"

14. (skin or epiderm*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

15. limit 14 to yr="2007 -Current"

16. 11 or 13 or 15

17. whole genome sequencing/ or whole exome sequencing/

18. limit 17 to yr="2007 -Current"

19. Genomics/

20. limit 19 to yr="2007 -Current"

21. DNA/

22. limit 21 to yr="2007 -Current"

23. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

24. limit 23 to yr="2007 -Current"

25. 18 or 20 or 22 or 24

26. Mutation/

27. limit 26 to yr="2007 -Current"

28. Selection, Genetic/

29. limit 28 to yr="2007 -Current"

30. Genetic Variation/

31. limit 30 to yr="2007 -Current"

32. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

33. limit 32 to yr="2007 -Current"

34. 27 or 29 or 31 or 33

35. 9 and 16 and 25 and 34

36. basal cell papilloma*.mp.

37. 35 not 36

38. limit 37 to (case reports or practice guideline or published erratum or "review")

39. 37 not 38

7.9.7 Melanoma search terms

1. Melanoma/

2. limit 1 to yr="2007 -Current"

3. melanocytic neoplasia.mp.

4. limit 3 to yr="2007 -Current"

5. 2 or 4

6. skin/ or epidermis/ or hair follicle/

7. limit 6 to yr="2007 -Current"

8. cutaneous.mp.

9. limit 8 to yr="2007 -Current"

10. (skin or epiderm*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

11. limit 10 to yr="2007 -Current"

12. 7 or 9 or 11

13. whole genome sequencing/ or whole exome sequencing/

14. limit 13 to yr="2007 -Current"

15. Genomics/

16. limit 15 to yr="2007 -Current"

17. DNA/

18. limit 17 to yr="2007 -Current"

19. (exome* or genom* or DNA or Deoxyribonucleic or mutation* or genetic*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

20. limit 19 to yr="2007 -Current"

21. 14 or 16 or 18 or 20

22. Mutation/

23. limit 22 to yr="2007 -Current"

24. Selection, Genetic/

25. limit 24 to yr="2007 -Current"

26. Genetic Variation/

27. limit 26 to yr="2007 -Current"

28. (target* sequenc* or mutation* or spectrum* or selection* or gen* driver*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

29. limit 28 to yr="2007 -Current"

30. 23 or 25 or 27 or 29

31. 5 and 12 and 21 and 30

32. limit 31 to (case reports or practice guideline or published erratum or "review")

33. 31 not 32

7.10 Sources of tumour samples

Type of Cancer	**Source (PMIDs)**	**Number of Samples**
Skin SCC	COSMIC	67
Skin SCC	27574101	7
Skin SCC	27906449	14
Skin SCC	30202019	21
Skin SCC	30684551	13
Oropharyngeal SCC	23304554	2
Oropharyngeal SCC	26790612	19
Oropharyngeal SCC	29423084	8
Oropharyngeal SCC	30046007	19
Oropharyngeal SCC	30308005	22
Oropharyngeal SCC	31135957	86
Oropharyngeal SCC	COSMIC	281
Oropharyngeal SCC	GDC/mc3	503
Lung SCC	26943773	37
Lung SCC	30992440	113
Lung SCC	COSMIC	253
Lung SCC	GDC/mc3	480
Oesophageal SCC	COSMIC	635
Oesophageal SCC	GDC/mc3	183
Oesophageal SCC	24670651	88
Oesophageal SCC	26619400	2
Oesophageal SCC	27058444	67
Oesophageal SCC	28365443	66
Oesophageal SCC	28608921	56
Oesophageal SCC	29358502	23
Oesophageal SCC	30012096	9
Oesophageal SCC	30975989	39
Oesophageal SCC	31289612	16
Cervical SCC	24390348	114
Cervical SCC	31624127	54
Cervical SCC	COSMIC	16
Cervical SCC	GDC/mc3	288
BCC	26950094	131
Melanoma	COSMIC	915
Melanoma	29991680	53
Melanoma	29170503	24
Melanoma	28193624	25
Melanoma	27095580	10
Melanoma	26359337	110
Melanoma	25268584	20
Total Skin SCC		122

Total Oropharyngeal SCC		940
Total lung SCC		883
Total oesophageal SCC		1184
Total cervical SCC		472
Total BCC		131
Total melanoma		1157

7.11 Driver genes

7.11.1 Skin SCC driver genes

Driver Gene	**Cancer**
CCDC28A	Skin SCC
CDC27	Skin SCC
CDKN2A	Skin SCC
CHUK	Skin SCC
FAT1	Skin SCC
HRAS	Skin SCC
KIF4B	Skin SCC
NOTCH1	Skin SCC
NOTCH2	Skin SCC
PRB2	Skin SCC
TMEM222	Skin SCC
TP53	Skin SCC

7.11.2 BCC driver genes

Driver Gene	**Cancer**
ACTB	BCC
ARHGAP35	BCC
C3	BCC
CDC27	BCC
ERBB2	BCC
EYA1	BCC
GLB1	BCC
LATS1	BCC

MYCN	BCC
MYH9	BCC
PAK2	BCC
PPIAL4G	BCC
PPM1D	BCC
PPP6C	BCC
PTCH1	BCC
PTPN14	BCC
RIOK1	BCC
SMO	BCC
TANC1	BCC
TMEM222	BCC
TP53	BCC
WDFY3	BCC

7.11.3 Melanoma driver genes

Driver Gene	**Cancer**	**Driver Gene**	**Cancer**	**Driver Gene**	**Cancer**
ABCB1	Melanoma	CD22	Melanoma	EIF3D	Melanoma
ACOT6	Melanoma	CD300E	Melanoma	ENPP2	Melanoma
ACSBG1	Melanoma	CDH2	Melanoma	EPHA7	Melanoma
ACTC1	Melanoma	CDH6	Melanoma	EPRS	Melanoma
ADAM22	Melanoma	CDH7	Melanoma	ERC2	Melanoma
ADAM7	Melanoma	CDH9	Melanoma	ESRRG	Melanoma
ADAMTS18	Melanoma	CDHR5	Melanoma	EZH2	Melanoma
ADCYAP1R1	Melanoma	CDKN2A	Melanoma	FAM107B	Melanoma
ADH1A	Melanoma	CEACAM5	Melanoma	FAM131B	Melanoma
AHI1	Melanoma	CEACAM6	Melanoma	FAM83B	Melanoma
AKR1C4	Melanoma	CEP55	Melanoma	FBXW7	Melanoma
ALDH5A1	Melanoma	CEP63	Melanoma	FCRL5	Melanoma
ALPK2	Melanoma	CHD6	Melanoma	FGD6	Melanoma
ALPPL2	Melanoma	CHGB	Melanoma	FILIP1	Melanoma
AMBN	Melanoma	CNTN5	Melanoma	FMO3	Melanoma
AMBP	Melanoma	CNTNAP2	Melanoma	FOXP1	Melanoma
AMICA1	Melanoma	COL17A1	Melanoma	GLT8D2	Melanoma
ANKRA2	Melanoma	COL3A1	Melanoma	GM2A	Melanoma
ANO4	Melanoma	COL5A2	Melanoma	GML	Melanoma
AP1M1	Melanoma	COL7A1	Melanoma	GNA11	Melanoma
APC	Melanoma	CPN1	Melanoma	GNAI2	Melanoma
APOB	Melanoma	CR2	Melanoma	GPA33	Melanoma
ARHGAP21	Melanoma	CRB1	Melanoma	GPR133	Melanoma
ARHGEF6	Melanoma	CSMD3	Melanoma	GPR179	Melanoma
ARID1A	Melanoma	CTNNB1	Melanoma	GRID2	Melanoma
ARID2	Melanoma	CUBN	Melanoma	GRIN3A	Melanoma
ARMC4	Melanoma	CXCR2	Melanoma	GRM3	Melanoma
ASTN1	Melanoma	CYP3A7	Melanoma	GSDMC	Melanoma

B2M	Melanoma	CYP7B1	Melanoma	GTPBP4	Melanoma
BAI3	Melanoma	DCAKD	Melanoma	GUCY2C	Melanoma
BCLAF1	Melanoma	DCC	Melanoma	GZMA	Melanoma
BMP3	Melanoma	DDX17	Melanoma	HNF4G	Melanoma
BMP5	Melanoma	DDX3X	Melanoma	HYDIN	Melanoma
BMPER	Melanoma	DDX4	Melanoma	IDH1	Melanoma
BRAF	Melanoma	DGKI	Melanoma	IL1R1	Melanoma
BRD7	Melanoma	DHX57	Melanoma	IL2RA	Melanoma
BRWD1	Melanoma	DMBT1	Melanoma	ITGA2	Melanoma
C10orf12	Melanoma	DMD	Melanoma	ITGA5	Melanoma
C16orf71	Melanoma	DMXL2	Melanoma	ITGAD	Melanoma
C1orf168	Melanoma	DNAH2	Melanoma	ITGB3	Melanoma
C1orf210	Melanoma	DNAH3	Melanoma	ITGB6	Melanoma
C2CD3	Melanoma	DNAH6	Melanoma	ITM2A	Melanoma
C6	Melanoma	DNAJC27	Melanoma	ITPR2	Melanoma
C6orf165	Melanoma	DNMT3L	Melanoma	ITSN1	Melanoma
C9	Melanoma	DPYD	Melanoma	KALRN	Melanoma
CAPN6	Melanoma	DSG3	Melanoma	KBTBD8	Melanoma
CBL	Melanoma	DSG4	Melanoma	KCNB2	Melanoma
CCDC11	Melanoma	EBF2	Melanoma	KCNH5	Melanoma
CD1C	Melanoma	EFEMP1	Melanoma	KCNQ3	Melanoma

Driver Gene	**Cancer**	**Driver Gene**	**Cancer**	**Driver Gene**	**Cancer**
KCNQ5	Melanoma	MYOM3	Melanoma	PLCE1	Melanoma
KCNT2	Melanoma	N4BP2	Melanoma	PLCH1	Melanoma
KDR	Melanoma	NEBL	Melanoma	PMFBP1	Melanoma
KDSR	Melanoma	NF1	Melanoma	POLN	Melanoma
KERA	Melanoma	NFASC	Melanoma	PPP1R13L	Melanoma
KHDRBS1	Melanoma	NLRP11	Melanoma	PPP6C	Melanoma
KIAA1109	Melanoma	NLRP13	Melanoma	PRG4	Melanoma
KIAA1199	Melanoma	NLRP4	Melanoma	PRKACA	Melanoma
KIAA2022	Melanoma	NLRP5	Melanoma	PROL1	Melanoma
KIF2B	Melanoma	NLRP8	Melanoma	PRRG3	Melanoma
KIF2C	Melanoma	NLRP9	Melanoma	PSG5	Melanoma
KIF3A	Melanoma	NPHP1	Melanoma	PTEN	Melanoma
KIF5A	Melanoma	NRAS	Melanoma	PTPRB	Melanoma
KIT	Melanoma	NRXN3	Melanoma	PTPRH	Melanoma
KLF12	Melanoma	NSUN6	Melanoma	PTPRO	Melanoma
KLHL20	Melanoma	OR11H1	Melanoma	PTPRT	Melanoma
KLHL4	Melanoma	OR13C8	Melanoma	PZP	Melanoma
KRAS	Melanoma	OR1A1	Melanoma	RAC1	Melanoma
KRT1	Melanoma	OR4C3	Melanoma	RASA2	Melanoma
KRT26	Melanoma	OR4D5	Melanoma	RB1	Melanoma
KRTAP10-8	Melanoma	OR4N4	Melanoma	RGS1	Melanoma

KRTAP5-10	Melanoma	OR51S1	Melanoma	RGS7	Melanoma
LAMA2	Melanoma	OR52J3	Melanoma	RHAG	Melanoma
LAMC2	Melanoma	OR7D2	Melanoma	RPL5	Melanoma
LEPR	Melanoma	OR8D2	Melanoma	RPRD2	Melanoma
LGR6	Melanoma	OR8K5	Melanoma	RQCD1	Melanoma
LHCGR	Melanoma	OSBP	Melanoma	SAG	Melanoma
LIPI	Melanoma	OSMR	Melanoma	SALL1	Melanoma
LRP2	Melanoma	OTC	Melanoma	SCAND3	Melanoma
MAGEC1	Melanoma	PAH	Melanoma	SCN10A	Melanoma
MAGI1	Melanoma	PAK7	Melanoma	SCN1A	Melanoma
MAN1A2	Melanoma	PAPPA2	Melanoma	SCUBE1	Melanoma
MAP2K1	Melanoma	PCDH15	Melanoma	SEC23B	Melanoma
MFN1	Melanoma	PCDH18	Melanoma	SELP	Melanoma
MKX	Melanoma	PCDHA12	Melanoma	SEMG2	Melanoma
MME	Melanoma	PCDHA2	Melanoma	SERPINA10	Melanoma
MPP7	Melanoma	PCDHA4	Melanoma	SETD2	Melanoma
MST4	Melanoma	PCDHB7	Melanoma	SETD5	Melanoma
MTR	Melanoma	PDE11A	Melanoma	SF3B1	Melanoma
MXRA5	Melanoma	PDE1A	Melanoma	SH3RF2	Melanoma
MYBPC1	Melanoma	PDE4DIP	Melanoma	SI	Melanoma
MYH1	Melanoma	PDE7B	Melanoma	SIGLEC7	Melanoma
MYH2	Melanoma	PDE8B	Melanoma	SIGLEC8	Melanoma
MYH7	Melanoma	PDE9A	Melanoma	SIKE1	Melanoma
MYLK	Melanoma	PDZD2	Melanoma	SLC12A5	Melanoma
MYO9A	Melanoma	PEG3	Melanoma	SLC15A2	Melanoma
MYOCD	Melanoma	PHKA1	Melanoma	SLC16A9	Melanoma
MYOF	Melanoma	PIK3CA	Melanoma	SLC25A16	Melanoma
MYOM2	Melanoma	PLCB4	Melanoma	SLC27A5	Melanoma

Driver Gene	**Cancer**	**Driver Gene**	**Cancer**
SLC28A2	Melanoma	USP29	Melanoma
SLC46A3	Melanoma	USP36	Melanoma
SLC9A4	Melanoma	USP9X	Melanoma
SLTM	Melanoma	VCAN	Melanoma
SMARCA1	Melanoma	VNN2	Melanoma
SMC1B	Melanoma	WDR76	Melanoma
SNCAIP	Melanoma	XIRP2	Melanoma
SNX31	Melanoma	ZBTB17	Melanoma
SORBS1	Melanoma	ZFX	Melanoma
SPAG16	Melanoma	ZNF229	Melanoma
SPAG17	Melanoma	ZNF318	Melanoma
SPATA19	Melanoma	ZNF334	Melanoma
SPPL2A	Melanoma	ZNF365	Melanoma
SRGAP3	Melanoma	ZNF385D	Melanoma

STAB2	Melanoma	ZNF484	Melanoma
STAT4	Melanoma	ZNF536	Melanoma
STK36	Melanoma	ZNF585B	Melanoma
SUN3	Melanoma	ZNF667	Melanoma
SUN5	Melanoma	ZNF684	Melanoma
TAOK1	Melanoma	ZNF737	Melanoma
TDRD1	Melanoma	ZNF750	Melanoma
TEX15	Melanoma	ZNF780B	Melanoma
THSD7B	Melanoma	ZNF804A	Melanoma
TIGIT	Melanoma	ZP2	Melanoma
TM4SF5	Melanoma		
TMC5	Melanoma		
TMEM156	Melanoma		
TMPO	Melanoma		
TP53	Melanoma		
TP63	Melanoma		
TPTE	Melanoma		
TPX2	Melanoma		
TREX2	Melanoma		
TRHDE	Melanoma		
TRRAP	Melanoma		
TSHZ2	Melanoma		
TSKS	Melanoma		
TTC3	Melanoma		
TUBA3C	Melanoma		
TUBAL3	Melanoma		
TUFM	Melanoma		
TULP1	Melanoma		
TUSC3	Melanoma		
UBE2J2	Melanoma		
UBR7	Melanoma		
UGT1A10	Melanoma		
UGT1A3	Melanoma		
UGT1A5	Melanoma		
UGT2B4	Melanoma		

7.11.4 Melanoma driver genes in normal melanocytes

Melanoma driver genes	**Number of mutations (Tang et al., 2020)**	**Melanoma driver genes**	**Number of mutations (Tang et al., 2020)**	**Melanoma driver genes**	**Number of mutations (Tang et al., 2020)**
PTPRT	34	DNAH3	11	MAGI1	7

HYDIN	32	KCNH5	11	MYOM2	7
PCDH15	29	KCNT2	11	NLRP11	7
DNAH2	23	PEG3	11	NLRP13	7
CSMD3	21	CDH9	10	PCDHB7	7
PAPPA2	21	COL5A2	10	SAG	7
PTPRB	21	DSG4	10	TTC3	7
THSD7B	21	GPR179	10	ZNF780B	7
ADAMTS18	19	GRID2	10	ALPK2	6
DNAH6	19	ITSN1	10	CBL	6
PLCB4	18	LAMA2	10	CDH2	6
MYH2	17	NLRP8	10	DHX57	6
CNTNAP2	16	PDE1A	10	FAM131B	6
COL7A1	16	PDZD2	10	FAM83B	6
MXRA5	16	PLCE1	10	ITPR2	6
NEBL	16	SCN1A	10	LEPR	6
APOB	15	SLC15A2	10	LHCGR	6
ARMC4	15	TPTE	10	MTR	6
ASTN1	15	ZNF536	10	SH3RF2	6
CRB1	15	ANO4	9	SIGLEC7	6
KALRN	15	BAI3	9	SLC12A5	6
MYH1	15	BMPER	9	SLC9A4	6
MYLK	15	C1orf168	9	SUN5	6
SCN10A	15	CDH7	9	VCAN	6
SRGAP3	15	DMBT1	9	ABCB1	5
MME	14	KIAA2022	9	AKR1C4	5
MYOCD	14	MYBPC1	9	ARHGAP21	5
SPAG17	14	PAH	9	BMP5	5
TP63	14	PCDHA2	9	C9	5
XIRP2	14	PDE7B	9	CD1C	5
ERC2	13	SLC16A9	9	CHD6	5
KCNQ3	13	ADCYAP1R1	8	DDX4	5
PLCH1	13	DGKI	8	EBF2	5
CDH6	12	EPHA7	8	ESRRG	5
CNTN5	12	FCRL5	8	GLT8D2	5
DCC	12	FOXP1	8	GTPBP4	5
DSG3	12	KCNB2	8	IL1R1	5
KDR	12	NLRP5	8	KBTBD8	5
LRP2	12	PAK7	8	KRT1	5
MYOM3	12	PDE4DIP	8	MYH7	5
PTPRO	12	PMFBP1	8	MYO9A	5
RGS7	12	TSHZ2	8	NFASC	5
SI	12	ACTC1	7	NLRP4	5
STAB2	12	CUBN	7	NRXN3	5
STK36	12	DPYD	7	SALL1	5
ZNF365	12	GRIN3A	7	SETD2	5
ADAM7	11	ITGAD	7	SNCAIP	5

C6	11	KLF12	7	TDRD1	5
COL3A1	11	MAGEC1	7	TSKS	5

Melanoma driver genes	Number of mutations (Tang et al., 2020)	Melanoma driver genes	Number of mutations (Tang et al., 2020)	Melanoma driver genes	Number of mutations (Tang et al., 2020)
USP36	5	CYP7B1	3	DCAKD	2
ALDH5A1	4	EIF3D	3	DMD	2
ARID1A	4	FBXW7	3	ENPP2	2
ARID2	4	GPR133	3	EPRS	2
BRAF	4	GSDMC	3	FGD6	2
C2CD3	4	GZMA	3	ITGB3	2
CD22	4	HNF4G	3	KDSR	2
CD300E	4	KCNQ5	3	KIF2C	2
CXCR2	4	KERA	3	KLHL4	2
EZH2	4	KHDRBS1	3	KRAS	2
FILIP1	4	KIAA1199	3	LAMC2	2
FMO3	4	KIF2B	3	MAP2K1	2
GPA33	4	KRT26	3	MKX	2
GRM3	4	MYOF	3	MST4	2
GUCY2C	4	NLRP9	3	NRAS	2
IL2RA	4	OR51S1	3	NSUN6	2
ITGA5	4	OR8D2	3	OR11H1	2
ITGB6	4	OSMR	3	OR52J3	2
ITM2A	4	OTC	3	OR7D2	2
KIAA1109	4	PCDH18	3	OR8K5	2
KIF3A	4	PCDHA12	3	PDE8B	2
KIF5A	4	PCDHA4	3	PSG5	2
LGR6	4	PRG4	3	PTPRH	2
MPP7	4	PRKACA	3	RB1	2
NF1	4	RASA2	3	SLC46A3	2
OR13C8	4	RGS1	3	SPAG16	2
OR4D5	4	RPRD2	3	STAT4	2
OSBP	4	RQCD1	3	TPX2	2
PDE11A	4	SELP	3	TUBA3C	2
PHKA1	4	SERPINA10	3	TUBAL3	2
PROL1	4	SETD5	3	TULP1	2
PRRG3	4	SIGLEC8	3	USP29	2
PZP	4	SNX31	3	ZFX	2
RHAG	4	SPATA19	3	ACOT6	1
SORBS1	4	TRHDE	3	ADH1A	1

TEX15	4	ZBTB17	3	BMP3	1
TIGIT	4	ZNF385D	3	BRD7	1
TMC5	4	ZNF737	3	C10orf12	1
TRRAP	4	ZNF804A	3	C1orf210	1
UGT2B4	4	AHI1	2	CDKN2A	1
ZNF334	4	ANKRA2	2	CEP63	1
ZP2	4	AP1M1	2	CPN1	1
ADAM22	3	BCLAF1	2	CR2	1
ALPPL2	3	C16orf71	2	CYP3A7	1
AMBN	3	C6orf165	2	DDX17	1
ARHGEF6	3	CAPN6	2	DDX3X	1
BRWD1	3	CEACAM6	2	DNAJC27	1
CEACAM5	3	CEP55	2	FAM107B	1
CTNNB1	3	COL17A1	2	GNA11	1

Melanoma driver genes	**Number of mutations (Tang et al., 2020)**
IDH1	1
ITGA2	1
KRTAP10-8	1
MAN1A2	1
MFN1	1
NPHP1	1
PDE9A	1
PIK3CA	1
POLN	1
PPP6C	1
SCUBE1	1
SEC23B	1
SEMG2	1
SF3B1	1
SLC25A16	1
SLC28A2	1
SMC1B	1
TMEM156	1
TMPO	1
TP53	1
TREX2	1
TUFM	1
TUSC3	1
USP9X	1
VNN2	1
ZNF229	1

ZNF585B	1
ZNF750	1

7.11.5 Validation results

Cancer	Confirmed somatic variant	Reported in another cancer sample as somatic	Variant of unknown origin	Total	Confirmed somatic variant	Reported in another cancer sample as somatic	Variant of unknown origin
Melanoma	2118056	11064	32345	2161465	98%	1%	1%
Oesophageal SCC	283262	937	4862	289061	98%	0%	2%
Oropharyngeal SCC	417126	173290	121510	711926	59%	24%	17%
Lung SCC	669837	19230	115408	804475	83%	2%	14%
Cervical SCC	238805	1339	3052	243196	98%	1%	1%
BCC	219052	0	0	219052	100%	0%	0%
cSCC	428317	0	0	428317	100%	0%	0%

These analyses were repeated for using an edited script from appendix 7.1. The bold command below shows the specific line of the script that was altered to ensure only confirmed variants were extracted from the COSMIC database. The script shows that column 32 of the file contained the somatic mutation status.

```
#get mutation list with header- only those from genome wide screens and with genomic location on b37. If histology subtype 1 is ‘squamous cell carcinoma’ and primary histology is ‘carcinoma’ and genome wide screen is ‘y’ and GRCh is ‘37’ then print and make file SCC_mutations.txt.
head -1 CosmicMutantExport.tsv > SCC_mutations.txt
awk 'BEGIN{FS=OFS="\t"}{if($13 == "squamous_cell_carcinoma" && $12 == "carcinoma" && $16 == "y" && $25 == "37" && $32 == “Confirmed somatic variant”) print $0}' CosmicMutantExport.tsv > SCC_mutations.txt
```

The rest of the analysis was repeated as outlined in the methods of the thesis and MutSig2CV was also run on the samples to identify if any of the driver genes which were common to skin SCC were still significant after MutSig2CV analysis and if any novel driver genes were identified. This would ensure that there was a reliable comparison between skin SCC samples and other SCCs. The appendix 7.11.5 shows a table of the driver genes which were significant in skin SCC and the cancers which were reanalysed and a comparison of their original p and q value after MutSig2CV analysis compared to their new p and q values after reanalysis.

Cancer	Gene	Original p value in Mutsig2CV	Original q value in MutSig2CV	p value in MutSig2CV after reanalysis in	q value in MutSig2CV after reanalysis
Lung SCC	TP53	1.00E-16	6.98E-13	1.00E-16	9.43E-13
Lung SCC	CDKN2A	3.33E-16	1.57E-12	1.00E-16	9.43E-13
Lung SCC	FAT1	1.71E-14	4.61E-11	1.99E-14	5.35E-11
Lung SCC	HRAS	3.68E-06	0.002572497	3.24E-07	0.00035896
Oropharyngeal SCC	TP53	1.00E-16	3.77E-13	1.00E-16	3.77E-13
Oropharyngeal SCC	CDKN2A	1.00E-16	3.77E-13	1.00E-16	3.77E-13
Oropharyngeal SCC	FAT1	6.66E-16	1.79E-12	6.66E-16	1.79E-12
Oropharyngeal SCC	HRAS	1.78E-07	0.000145568	1.78E-07	0.000145568
Oropharyngeal SCC	NOTCH2	2.20E-05	0.0111915	2.20E-05	0.0111915
Oesophageal SCC	TP53	1E-16	4.18821E-13	1E-16	4.18821E-13
Oesophageal SCC	CDKN2A	1E-16	4.18821E-13	1E-16	4.18821E-13
Oesophageal SCC	FAT1	4.88498E-15	1.02378E-11	4.88498E-15	1.02378E-11
Oesophageal SCC	NOTCH1	1E-16	4.18821E-13	1E-16	4.18821E-13
Oesophageal SCC	NOTCH2	0.000237149	0.07462077	0.000237149	0.07462077
Melanoma	TP53	1E-16	3.14367E-13	1E-16	3.49017E-13
Melanoma	CDKN2A	1E-16	3.14367E-13	1.22125E-15	2.87939E-12

Melanoma	PPP6C	1E-16	3.14367E-13	1E-16	3.49017E-13

7.11.6 Driver gene somatic variant status

This table shows the shared driver genes between the SCCs and skin cancers. The tables summarise their somatic status within their cohorts.

Cervical SCC	Lung SCC	Oesophageal SCC	Oropharyngeal SCC	Skin SCC
	CDKN2A- all true somatic variants	CDKN2A - 12 variants reported in other cancer as somatic = 12/650 = 1.84% of variants not confirmed	CDKN2A - all true somatic variants	CDKN2A - all true somatic variants
	HRAS - all true somatic variants		HRAS- all true somatic variants	HRAS - all true somatic variants
TP53 - all true somatic variants	TP53 - all true somatic variants	TP53 - all true somatic variants	TP53 - all true somatic variants	TP53 - all true somatic variants
	FAT1 - all true somatic variants	FAT1 - all true somatic variants	FAT1 - all true somatic variants	FAT1 - all true somatic variants
		NOTCH1 - all true somatic variants		NOTCH1 - all true somatic variants all true somatic variants
		NOTCH2 - all true somatic variants	NOTCH2 - all true somatic variants	NOTCH2 - all true somatic variants

Skin SCC	**BCC**	**Melanoma**
TP53 - all true somatic variants	TP53 - all true somatic variants	TP53 - all true somatic variants
CDKN2A - all true somatic variants		CDKN2A - 132 variants reported as somatic in other cancer, 72 variants unknown origin = 204/1378 = 14.8% of variants not confirmed
	PPP6C - all true somatic variants	PPP6C - 32 variants reported as somatic in other cancer, 4 variants unknown origin = 36/347 = 10.4% of variants not confirmed
CDC27 - all true somatic variants	CDC27 - all true somatic variants	
TMEM222 - all true somatic variants	TMEM222 - all true somatic variants	

7.12 Oropharyngeal SCC mutation signature analysis

Cancer type	Sample Name	Number of mutations	Source (PMID)	Details
Oropharyngeal SCC	HNPTS_16	1	COSMIC	Upper aerodigestive tract; Head neck (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	UPHN7B	2	Literature search – 30046007	Oral cavity squamous cell carcinomas
Oropharyngeal SCC	2014037004	2	Literature search – 30308005	Oral cavity squamous cell carcinoma
Oropharyngeal SCC	HN19PT	3	COSMIC	Upper aerodigestive tract; Pharynx; Oropharynx (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	TCGA-CV-A468-01	321	GDC/mc3	Tumour location – Upper aerodigestive tract; Head neck (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	41T	587	COSMIC	Mouth; Gingiva (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	B8T_B8N	870	Literature search – 29423084	right parotid gland
Oropharyngeal SCC	TCGA-D6-6516-01	1442	GDC/mc3	Upper aerodigestive tract; Head neck (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	B2T_B2N	1726	Literature search – 29423084	left ear
Oropharyngeal SCC	TCGA-CV-7568-01	1810	GDC/mc3	Upper aerodigestive tract; Head neck (Carcinoma; Squamous cell carcinoma)
Oropharyngeal SCC	B5T_B5N	4157	Literature search – 29423084	Skin of face

8. References

ABASCAL, F., HARVEY, L. M. R., MITCHELL, E., LAWSON, A. R. J., LENSING, S. V., ELLIS, P., RUSSELL, A. J. C., ALCANTARA, R. E., BAEZ-ORTEGA, A., WANG, Y., KWA, E. J., LEE-SIX, H., CAGAN, A., COORENS, T. H. H., CHAPMAN, M. S., OLAFSSON, S., LEONARD, S., JONES, D., MACHADO, H. E., DAVIES, M., OBRO, N. F., MAHUBANI, K. T., ALLINSON, K., GERSTUNG, M., SAEB-PARSY, K., KENT, D. G., LAURENTI, E., STRATTON, M. R., RAHBARI, R., CAMPBELL, P. J., OSBORNE, R. J. & MARTINCORENA, I. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature,* 593**,** 405-410.

ABRAHAM, J. & MATHEW, S. 2019. Merkel Cells: A Collective Review of Current Concepts. *Int J Appl Basic Med Res,* 9**,** 9-13.

AGANEZOV, S., GOODWIN, S., SHERMAN, R. M., SEDLAZECK, F. J., ARUN, G., BHATIA, S., LEE, I., KIRSCHE, M., WAPPEL, R., KRAMER, M., KOSTROFF, K., SPECTOR, D. L., TIMP, W., MCCOMBIE, W. R. & SCHATZ, M. C. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res,* 30**,** 1258-1273.

AHMED, N. U., UEDA, M., NIKAIDO, O., OSAWA, T. & ICHIHASHI, M. 1999. High levels of 8-hydroxy-2'-deoxyguanosine appear in normal human epidermis after a single dose of ultraviolet radiation. *Br J Dermatol,* 140**,** 226-31.

AITHAL, A., RAUTH, S., KSHIRSAGAR, P., SHAH, A., LAKSHMANAN, I., JUNKER, W. M., JAIN, M., PONNUSAMY, M. P. & BATRA, S. K. 2018. MUC16 as a novel target for cancer therapy. *Expert Opin Ther Targets,* 22**,** 675-686.

ALAM, M. & RATNER, D. 2001. Cutaneous squamous-cell carcinoma. *N Engl J Med,* 344**,** 975-83.

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., WALTER, P. 2002. *Molecular Biology of the Cell,* New York, Garland Science.

ALBIBAS, A. A., ROSE-ZERILLI, M. J. J., LAI, C., PENGELLY, R. J., LOCKETT, G. A., THEAKER, J., ENNIS, S., HOLLOWAY, J. W. & HEALY, E. 2017. Subclonal Evolution of Cancer-Related Gene Mutations in p53 Immunopositive Patches in Human Skin. *J Invest Dermatol*.

ALEXANDROV, L. B., JONES, P. H., WEDGE, D. C., SALE, J. E., CAMPBELL, P. J., NIK-ZAINAL, S. & STRATTON, M. R. 2015. Clock-like mutational processes in human somatic cells. *Nat Genet,* 47**,** 1402-7.

ALEXANDROV, L. B., JU, Y. S., HAASE, K., VAN LOO, P., MARTINCORENA, I., NIK-ZAINAL, S., TOTOKI, Y., FUJIMOTO, A., NAKAGAWA, H., SHIBATA, T., CAMPBELL, P. J., VINEIS, P., PHILLIPS, D. H. & STRATTON, M. R. 2016. Mutational signatures associated with tobacco smoking in human cancer. *Science,* 354**,** 618-622.

ALEXANDROV, L. B., KIM, J., HARADHVALA, N. J., HUANG, M. N., TIAN NG, A. W., WU, Y., BOOT, A., COVINGTON, K. R., GORDENIN, D. A., BERGSTROM, E. N., ISLAM, S. M. A., LOPEZ-BIGAS, N., KLIMCZAK, L. J., MCPHERSON, J. R., MORGANELLA, S., SABARINATHAN, R., WHEELER, D. A., MUSTONEN, V., GROUP, P. M. S. W., GETZ, G., ROZEN, S. G., STRATTON, M. R. & CONSORTIUM, P. 2020. The repertoire of mutational signatures in human cancer. *Nature,* 578**,** 94-101.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A., BORRESEN-DALE, A. L., BOYAULT, S., BURKHARDT, B., BUTLER, A. P., CALDAS, C., DAVIES, H. R., DESMEDT, C., EILS, R., EYFJORD, J. E., FOEKENS, J. A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JAGER, N., JONES, D. T., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S. R., LOPEZ-OTIN, C., MARTIN, S., MUNSHI, N. C., NAKAMURA, H., NORTHCOTT, P. A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J. V., PUENTE, X. S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A. L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T. N., SPAN, P. N., TEAGUE, J. W., TOTOKI, Y., TUTT, A. N., VALDES-MAS, R., VAN BUUREN, M. M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L. R., AUSTRALIAN PANCREATIC CANCER GENOME, I., CONSORTIUM, I. B. C., CONSORTIUM, I. M.-S., PEDBRAIN, I., ZUCMAN-ROSSI, J., FUTREAL, P. A., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S. M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S. M., CAMPBELL, P. J. & STRATTON, M. R. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-21.

AMAKYE, D., JAGANI, Z. & DORSCH, M. 2013. Unraveling the therapeutic potential of the Hedgehog pathway in cancer. *Nat Med,* 19**,** 1410-22.

AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology,* 21**,** 30.

ANZAR, I., SVERCHKOVA, A., STRATFORD, R. & CLANCY, T. 2019. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Medical Genomics,* 12**,** 63.

ARCHIER, E., DEVAUX, S., CASTELA, E., GALLINI, A., AUBIN, F., LE MAITRE, M., ARACTINGI, S., BACHELEZ, H., CRIBIER, B., JOLY, P., JULLIEN, D., MISERY, L., PAUL, C., ORTONNE, J. P. & RICHARD, M. A. 2012. Carcinogenic risks of psoralen UV-A therapy and narrowband UV-B therapy in chronic plaque psoriasis: a systematic literature review. *J Eur Acad Dermatol Venereol,* 26 Suppl 3**,** 22-31.

ARKWRIGHT, P. D., MOTALA, C., SUBRAMANIAN, H., SPERGEL, J., SCHNEIDER, L. C., WOLLENBERG, A. & ATOPIC DERMATITIS WORKING GROUP OF THE ALLERGIC SKIN DISEASES COMMITTEE OF THE, A. 2013. Management of difficult-to-treat atopic dermatitis. *J Allergy Clin Immunol Pract,* 1**,** 142-51.

ARMITAGE, P. & DOLL, R. 1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer,* 8**,** 1-12.

ARMSTRONG, B. K. & KRICKER, A. 2001. The epidemiology of UV induced skin cancer. *J Photochem Photobiol B,* 63**,** 8-18.

ARNEDO-PAC, C., MULARONI, L., MUINOS, F., GONZALEZ-PEREZ, A. & LOPEZ-BIGAS, N. 2019. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics,* 35**,** 5396.

AZEN, E. A., O'CONNELL, P. & KIM, H. S. 1992. PRB2/1 fusion gene: a product of unequal and homologous crossing-over between proline-rich protein (PRP) genes PRB1 and PRB2. *Am J Hum Genet,* 50**,** 842-51.

BAILEY, M. H., TOKHEIM, C., PORTA-PARDO, E., SENGUPTA, S., BERTRAND, D., WEERASINGHE, A., COLAPRICO, A., WENDL, M. C., KIM, J., REARDON, B., KWOK-SHING NG, P., JEONG, K. J., CAO, S., WANG, Z., GAO, J., GAO, Q., WANG, F., LIU, E. M., MULARONI, L., RUBIO-PEREZ, C., NAGARAJAN, N., CORTES-CIRIANO, I., ZHOU, D. C., LIANG, W. W., HESS, J. M., YELLAPANTULA, V. D., TAMBORERO, D., GONZALEZ-PEREZ, A., SUPHAVILAI, C., KO, J. Y., KHURANA, E., PARK, P. J., VAN ALLEN, E. M., LIANG, H., GROUP, M. C. W., CANCER GENOME ATLAS RESEARCH, N., LAWRENCE, M. S., GODZIK, A., LOPEZ-BIGAS, N., STUART, J., WHEELER, D., GETZ, G., CHEN, K., LAZAR, A. J., MILLS, G. B., KARCHIN, R. & DING, L. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell,* 174**,** 1034-1035.

BARKER, W. C., KETCHAM, L. K. & DAYHOFF, M. O. 1978. A comprehensive examination of protein sequences for evidence of internal gene duplication. *J Mol Evol,* 10**,** 265-81.

BARREA, L., SAVANELLI, M. C., DI SOMMA, C., NAPOLITANO, M., MEGNA, M., COLAO, A. & SAVASTANO, S. 2017. Vitamin D and its role in psoriasis: An overview of the dermatologist and nutritionist. *Rev Endocr Metab Disord,* 18**,** 195-205.

BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHAR, J., KRYUKOV, G. V., SONKIN, D., REDDY, A., LIU, M., MURRAY, L., BERGER, M. F., MONAHAN, J. E., MORAIS, P., MELTZER, J., KOREJWA, A., JANE-VALBUENA, J., MAPA, F. A., THIBAULT, J., BRIC-FURLONG, E., RAMAN, P., SHIPWAY, A., ENGELS, I. H., CHENG, J., YU, G. K., YU, J., ASPESI, P., JR., DE SILVA, M., JAGTAP, K., JONES, M. D., WANG, L., HATTON, C., PALESCANDOLO, E., GUPTA, S., MAHAN, S., SOUGNEZ, C., ONOFRIO, R. C., LIEFELD, T., MACCONAILL, L., WINCKLER, W., REICH, M., LI, N., MESIROV, J. P., GABRIEL, S.

B., GETZ, G., ARDLIE, K., CHAN, V., MYER, V. E., WEBER, B. L., PORTER, J., WARMUTH, M., FINAN, P., HARRIS, J. L., MEYERSON, M., GOLUB, T. R., MORRISSEY, M. P., SELLERS, W. R., SCHLEGEL, R. & GARRAWAY, L. A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature,* 483**,** 603-7.

BAUER, J., BUTTNER, P., MURALI, R., OKAMOTO, I., KOLAITIS, N. A., LANDI, M. T., SCOLYER, R. A. & BASTIAN, B. C. 2011. BRAF mutations in cutaneous melanoma are independently associated with age, anatomic site of the primary tumor, and the degree of solar elastosis at the primary tumor site. *Pigment Cell Melanoma Res,* 24**,** 345-51.

BAXTER, L. L. & PAVAN, W. J. 2013. The etiology and molecular genetics of human pigmentation disorders.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., CHEETHAM, R. K., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X. H., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M. J., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X. L., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M. D., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CATENAZZI, M. C. E., CHANG, S., COOLEY, R. N., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FAJARDO, K. V. F., FUREY, W. S., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-59.

BERG, R. J., VAN KRANEN, H. J., REBEL, H. G., DE VRIES, A., VAN VLOTEN, W. A., VAN KREIJL, C. F., VAN DER LEUN, J. C. & DE GRUIJL, F. R. 1996. Early p53 alterations in mouse skin carcinogenesis by UVB radiation: immunohistochemical detection of mutant p53 protein in clusters of preneoplastic epidermal cells. *Proc Natl Acad Sci U S A,* 93**,** 274-8.

BERKING, C., HAUSCHILD, A., KOLBL, O., MAST, G. & GUTZMER, R. 2014. Basal cell carcinoma-treatments for the commonest skin cancer. *Dtsch Arztebl Int,* 111**,** 389-95.

BHUTANI, T. & LIAO, W. 2010. A Practical Approach to Home UVB Phototherapy for the Treatment of Generalized Psoriasis. *Pract Dermatol,* 7**,** 31-35.

BIDKHORI, G., NARIMANI, Z., HOSSEINI ASHTIANI, S., MOEINI, A., NOWZARI-DALINI, A. & MASOUDI-NEJAD, A. 2013. Reconstruction of an integrated genome-scale co-expression network reveals key modules involved in lung adenocarcinoma. *PLoS One,* 8**,** e67552.

BIERIE, B. & MOSES, H. L. 2006. Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nat Rev Cancer,* 6**,** 506-20.

BIN, L., EDWARDS, M. G., HEISER, R., STREIB, J. E., RICHERS, B., HALL, C. F. & LEUNG, D. Y. 2014. Identification of novel gene signatures in patients with atopic dermatitis complicated by eczema herpeticum. *J Allergy Clin Immunol,* 134**,** 848-55.

BLOKZIJL, F., DE LIGT, J., JAGER, M., SASSELLI, V., ROERINK, S., SASAKI, N., HUCH, M., BOYMANS, S., KUIJK, E., PRINS, P., NIJMAN, I. J., MARTINCORENA, I., MOKRY, M., WIEGERINCK, C. L., MIDDENDORP, S., SATO, T., SCHWANK, G., NIEUWENHUIS, E. E., VERSTEGEN, M. M., VAN DER LAAN, L. J., DE JONGE, J., JN, I. J., VRIES, R. G., VAN DE WETERING, M., STRATTON, M. R., CLEVERS, H., CUPPEN, E. & VAN BOXTEL, R. 2016. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature,* 538**,** 260-264.

BLUMTHALER, M. & AMBACH, W. 1988. Solar UVB-albedo of various surfaces. *Photochem Photobiol,* 48**,** 85-8.

BOEHNCKE, W. H. & SCHON, M. P. 2015. Psoriasis. *Lancet,* 386**,** 983-94.

BOEKE, J. D., GARFINKEL, D. J., STYLES, C. A. & FINK, G. R. 1985. Ty elements transpose through an RNA intermediate. *Cell,* 40**,** 491-500.

BONILLA, X., PARMENTIER, L., KING, B., BEZRUKOV, F., KAYA, G., ZOETE, V., SEPLYARSKIY, V. B., SHARPE, H. J., MCKEE, T., LETOURNEAU, A., RIBAUX, P. G., POPADIN, K., BASSET-SEGUIN, N., BEN CHAABENE, R., SANTONI, F. A., ANDRIANOVA, M. A., GUIPPONI, M., GARIERI, M., VERDAN, C., GROSDEMANGE, K., SUMARA, O., EILERS, M., AIFANTIS, I., MICHIELIN, O., DE SAUVAGE, F. J., ANTONARAKIS, S. E. & NIKOLAEV, S. I. 2016. Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. *Nature Genetics,* 48**,** 398-+.

BONNET, D. & DICK, J. E. 1997. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med,* 3**,** 730-7.

BONYADI RAD, E., HAMMERLINDL, H., WELS, C., POPPER, U., RAVINDRAN MENON, D., BREITENEDER, H., KITZWOEGERER, M., HAFNER, C., HERLYN, M., BERGLER, H. & SCHAIDER, H. 2016. Notch4 Signaling Induces a Mesenchymal-Epithelial-like Transition in Melanoma Cells to Suppress Malignant Behaviors. *Cancer Res,* 76**,** 1690-7.

BOWDEN, R., DAVIES, R. W., HEGER, A., PAGNAMENTA, A. T., DE CESARE, M., OIKKONEN, L. E., PARKES, D., FREEMAN, C., DHALLA, F., PATEL, S. Y., POPITSCH, N., IP, C. L. C., ROBERTS, H. E., SALATINO, S., LOCKSTONE, H., LUNTER, G., TAYLOR, J. C., BUCK, D., SIMPSON, M. A. & DONNELLY, P. 2019. Sequencing of human genomes with nanopore technology. *Nature Communications,* 10**,** 1869.

BRADFORD, P. T. 2009. Skin cancer in skin of color. *Dermatol Nurs,* 21**,** 170-7, 206; quiz 178.

BRADFORD, P. T., GOLDSTEIN, A. M., MCMASTER, M. L. & TUCKER, M. A. 2009. Acral lentiginous melanoma: incidence and survival patterns in the United States, 1986-2005. *Arch Dermatol,* 145**,** 427-34.

BRADFORD, P. T., GOLDSTEIN, A. M., TAMURA, D., KHAN, S. G., UEDA, T., BOYLE, J., OH, K. S., IMOTO, K., INUI, H., MORIWAKI, S., EMMERT, S., PIKE, K. M., RAZIUDDIN, A., PLONA, T. M., DIGIOVANNA, J. J., TUCKER, M. A. & KRAEMER, K. H. 2011. Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair. *J Med Genet,* 48**,** 168-76.

BRANTON, D., DEAMER, D. W., MARZIALI, A., BAYLEY, H., BENNER, S. A., BUTLER, T., DI VENTRA, M., GARAJ, S., HIBBS, A., HUANG, X., JOVANOVICH, S. B., KRSTIC, P. S., LINDSAY, S., LING, X. S., MASTRANGELO, C. H., MELLER, A., OLIVER, J. S., PERSHIN, Y. V., RAMSEY, J. M., RIEHN, R., SONI, G. V., TABARD-COSSA, V., WANUNU, M., WIGGIN, M. & SCHLOSS, J. A. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol,* 26**,** 1146-53.

BRASH, D. E. 1988. UV mutagenic photoproducts in Escherichia coli and human cells: a molecular genetics perspective on human skin cancer. *Photochem Photobiol,* 48**,** 59-66.

BRASH, D. E. 2015. UV signature mutations. *Photochem Photobiol,* 91**,** 15-26.

BRASH, D. E., ZIEGLER, A., JONASON, A. S., SIMON, J. A., KUNALA, S. & LEFFELL, D. J. 1996. Sunlight and sunburn in human skin cancer: p53, apoptosis, and tumor promotion. *J Investig Dermatol Symp Proc,* 1**,** 136-42.

BRENNER, M. & HEARING, V. J. 2008. The protective role of melanin against UV damage in human skin. *Photochem Photobiol,* 84**,** 539-49.

BROOKE, R. C., NEWBOLD, S. A., TELFER, N. R. & GRIFFITHS, C. E. 2001. Discordance between facial wrinkling and the presence of basal cell carcinoma. *Arch Dermatol,* 137**,** 751-4.

BRUNET, J. P., TAMAYO, P., GOLUB, T. R. & MESIROV, J. P. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A,* 101**,** 4164-9.

CALBO, J., PAGES, D. & GONZALEZ, J. A. 2005. Empirical studies of cloud effects on UV radiation: A review. *Reviews of Geophysics,* 43.

CALLEN, J. P., BICKERS, D. R. & MOY, R. L. 1997. Actinic keratoses. *J Am Acad Dermatol,* 36**,** 650-3.

CAMMARERI, P., ROSE, A. M., VINCENT, D. F., WANG, J., NAGANO, A., LIBERTINI, S., RIDGWAY, R. A., ATHINEOS, D., COATES, P. J., MCHUGH, A., POURREYRON, C., DAYAL, J. H., LARSSON, J., WEIDLICH, S., SPENDER, L. C., SAPKOTA, G. P., PURDIE, K. J., PROBY, C. M., HARWOOD, C. A., LEIGH, I. M., CLEVERS, H., BARKER, N., KARLSSON, S., PRITCHARD, C., MARAIS, R., CHELALA, C., SOUTH, A. P., SANSOM, O. J. & INMAN, G. J. 2016. Inactivation of TGFbeta receptors in stem cells drives cutaneous squamous cell carcinoma. *Nat Commun,* 7**,** 12493.

CAMPBELL, J. D., YAU, C., BOWLBY, R., LIU, Y. X., BRENNAN, K., FAN, H. H., TAYLOR, A. M., WANG, C., WALTER, V., AKBANI, R., BYERS, L. A., CREIGHTON, C. J., COARFA, C., SHIH, J., CHERNIACK, A. D., GEVAERT, O., PRUNELLO, M., SHEN, H., ANUR, P., CHEN, J. H., CHENG, H., HAYES, D. N., BULLMAN, S., PEDAMALLU, C. S., OJESINA, A. I., SADEGHI, S., MUNGALL, K. L., ROBERTSON, A. G., BENZ, C., SCHULTZ, A., KANCHI, R. S., GAY, C. M., HEGDE, A., DIAO, L. X., WANG, J., MA, W. C., SUMAZIN, P., CHIU, H. S., CHEN, T. W., GUNARATNE, P., DONEHOWER, L., RADER, J. S., ZUNA, R., AL-AHMADIE, H., LAZAR, A. J., FLORES, E. R., TSAI, K. Y., ZHOU, J. H., RUSTGI, A. K., DRILL, E., SHEN, R. L., WONG, C. K., STUART, J. M., LAIRD, P. W., HOADLEY, K. A., WEINSTEIN, J. N., PETO, M., PICKERING, C. R., CHEN, Z., WAES, C. & NETWORK, C. G. A. R. 2018. Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Reports,* 23**,** 194-+.

CAMPBELL, P. J., YACHIDA, S., MUDIE, L. J., STEPHENS, P. J., PLEASANCE, E. D., STEBBINGS, L. A., MORSBERGER, L. A., LATIMER, C., MCLAREN, S., LIN, M. L., MCBRIDE, D. J., VARELA, I., NIK-ZAINAL, S. A., LEROY, C., JIA, M., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., GRIFFIN, C. A., BURTON, J., SWERDLOW, H., QUAIL, M. A., STRATTON, M. R., IACOBUZIO-DONAHUE, C. & FUTREAL, P. A. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature,* 467**,** 1109-13.

CAPACCIONE, K. M. & PINE, S. R. 2013. The Notch signaling pathway as a mediator of tumor survival. *Carcinogenesis,* 34**,** 1420-30.

CARBONE, M. & LEVINE, A. S. 1990. Oncogenes, antioncogenes, and the regulation of cell growth. *Trends Endocrinol Metab,* 1**,** 248-53.

CHADYŠIENE, R., GIRGŽDYS, A. 2010. Ultraviolet radiation albedo of natural surfaces. *Journal of Environmental Engineering and Landscape Management,* 16**,** 83-88.

CHANDRASEKARAN, G., TATRAI, P. & GERGELY, F. 2015. Hitting the brakes: targeting microtubule motors in cancer. *Br J Cancer,* 113**,** 693-8.

CHANG, D. & SHAIN, A. H. 2021. The landscape of driver mutations in cutaneous squamous cell carcinoma. *NPJ Genom Med,* 6**,** 61.

CHEN, X., YANG, M., CHENG, Y., LIU, G. J. & ZHANG, M. 2013. Narrow-band ultraviolet B phototherapy versus broad-band ultraviolet B or psoralen-ultraviolet A photochemotherapy for psoriasis. *Cochrane Database Syst Rev***,** CD009481.

CHIANG, A., JAJU, P. D., BATRA, P., REZAEE, M., EPSTEIN, E. H., JR., TANG, J. Y. & SARIN, K. Y. 2018. Genomic Stability in Syndromic Basal Cell Carcinoma. *J Invest Dermatol,* 138**,** 1044-1051.

CHITSAZZADEH, V., COARFA, C., DRUMMOND, J. A., NGUYEN, T., JOSEPH, A., CHILUKURI, S., CHARPIOT, E., ADELMANN, C. H., CHING, G., NGUYEN, T. N., NICHOLAS, C., THOMAS, V. D., MIGDEN, M., MACFARLANE, D., THOMPSON, E., SHEN, J. J., TAKATA, Y., MCNIECE, K., POLANSKY, M. A., ABBAS, H. A., RAJAPAKSHE, K., GOWER, A., SPIRA, A., COVINGTON, K. R., XIAO, W. M., GUNARATNE, P., PICKERING, C., FREDERICK, M., MYERS, J. N., SHEN, L., YAO, H., SU, X. P., RAPINI, R. P., WHEELER, D. A., HAWK, E. T., FLORES, E. R. & TSAI, K. Y. 2016. Cross-species identification of genomic drivers of squamous cell carcinoma development across preneoplastic intermediates. *Nature Communications,* 7.

CHONG, M. & FONACIER, L. 2016. Treatment of Eczema: Corticosteroids and Beyond. *Clin Rev Allergy Immunol,* 51**,** 249-262.

CHUNG, C. H., GUTHRIE, V. B., MASICA, D. L., TOKHEIM, C., KANG, H., RICHMON, J., AGRAWAL, N., FAKHRY, C., QUON, H., SUBRAMANIAM, R. M., ZUO, Z., SEIWERT, T., CHALMERS, Z. R., FRAMPTON, G. M., ALI, S. M., YELENSKY, R., STEPHENS, P. J., MILLER, V. A., KARCHIN, R. & BISHOP, J. A. 2015. Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing. *Ann Oncol,* 26**,** 1216-23.

CICHOREK, M., WACHULSKA, M., STASIEWICZ, A. & TYMINSKA, A. 2013. Skin melanocytes: biology and development. *Postepy Dermatol Alergol,* 30**,** 30-41.

CLAYTON, K., VALLEJO, A. F., DAVIES, J., SIRVENT, S. & POLAK, M. E. 2017. Langerhans Cells-Programmed by the Epidermis. *Front Immunol,* 8**,** 1676.

CLEAVER, J. E. 2005. Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nat Rev Cancer,* 5**,** 564-73.

COLEBATCH, A. J., FERGUSON, P., NEWELL, F., KAZAKOFF, S. H., WITKOWSKI, T., DOBROVIC, A., JOHANSSON, P. A., SAW, R. P. M., STRETCH, J. R., MCARTHUR, G. A., LONG, G. V., THOMPSON, J. F., PEARSON, J. V., MANN, G. J., HAYWARD, N. K., WADDELL, N., SCOLYER, R. A. & WILMOTT, J. S. 2019. Molecular Genomic Profiling of Melanocytic Nevi. *J Invest Dermatol,* 139**,** 1762-1768.

COLOTTA, F., ALLAVENA, P., SICA, A., GARLANDA, C. & MANTOVANI, A. 2009. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis,* 30**,** 1073-81.

CONSCIENCE, I., JOVENIN, N., COISSARD, C., LORENZATO, M., DURLACH, A., GRANGE, F., BIREMBAUT, P., CLAVEL, C. & BERNARD, P. 2006. P16 is overexpressed in cutaneous carcinomas located on sun-exposed areas. *Eur J Dermatol,* 16**,** 518-22.

COSTELLO, M., PUGH, T. J., FENNELL, T. J., STEWART, C., LICHTENSTEIN, L., MELDRIM, J. C., FOSTEL, J. L., FRIEDRICH, D. C., PERRIN, D., DIONNE, D., KIM, S., GABRIEL, S. B., LANDER, E. S., FISHER, S. & GETZ, G. 2013. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res,* 41**,** e67.

COUTELIER, M., HOLTGREWE, M., JÄGER, M., FLÖTTMAN, R., MENSAH, M. A., SPIELMANN, M., KRAWITZ, P., HORN, D., BEULE, D. & MUNDLOS, S. 2022. Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics,* 30**,** 178-186.

COX, N. H., EEDY, D. J. & MORTON, C. A. 1999. Guidelines for management of Bowen's disease. British Association of Dermatologists. *Br J Dermatol,* 141**,** 633-41.

CRAWFORD, N. G., KELLY, D. E., HANSEN, M. E. B., BELTRAME, M. H., FAN, S., BOWMAN, S. L., JEWETT, E., RANCIARO, A., THOMPSON, S., LO, Y., PFEIFER, S. P., JENSEN, J. D., CAMPBELL, M. C., BEGGS, W., HORMOZDIARI, F., MPOLOKA, S. W., MOKONE, G. G., NYAMBO, T., MESKEL, D. W., BELAY, G., HAUT, J., PROGRAM, N. C. S., ROTHSCHILD, H., ZON, L., ZHOU, Y., KOVACS, M. A., XU, M., ZHANG, T., BISHOP, K., SINCLAIR, J., RIVAS, C., ELLIOT, E., CHOI, J., LI, S. A., HICKS, B., BURGESS, S., ABNET, C., WATKINS-CHOW, D. E., OCEANA, E., SONG, Y. S., ESKIN, E., BROWN, K. M., MARKS, M. S., LOFTUS, S. K., PAVAN, W. J., YEAGER, M., CHANOCK, S. & TISHKOFF, S. A. 2017. Loci associated with skin pigmentation identified in African populations. *Science,* 358.

CRISCIONE, V. D., WEINSTOCK, M. A., NAYLOR, M. F., LUQUE, C., EIDE, M. J., BINGHAM, S. F. & DEPARTMENT OF VETERAN AFFAIRS TOPICAL TRETINOIN CHEMOPREVENTION TRIAL, G. 2009. Actinic keratoses: Natural history and risk of malignant transformation in the Veterans Affairs Topical Tretinoin Chemoprevention Trial. *Cancer,* 115**,** 2523-30.

CROOK, T., TIDY, J. A. & VOUSDEN, K. H. 1991. Degradation of p53 can be targeted by HPV E6 sequences distinct from those required for p53 binding and trans-activation. *Cell,* 67**,** 547-56.

CRUK. 2018. Cancer Research UK. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence [Accessed July 2021].

D'ORAZIO, J., JARRETT, S., AMARO-ORTIZ, A. & SCOTT, T. 2013. UV radiation and the skin. *Int J Mol Sci,* 14**,** 12222-48.

DAS, P. M. & SINGAL, R. 2004. DNA methylation and cancer. *J Clin Oncol,* 22**,** 4632-42.

DEMEHRI, S., TURKOZ, A. & KOPAN, R. 2009. Epidermal Notch1 loss promotes skin tumorigenesis by impacting the stromal microenvironment. *Cancer Cell,* 16**,** 55-66.

DENARDO, D. G., ANDREU, P. & COUSSENS, L. M. 2010. Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity. *Cancer Metastasis Rev,* 29**,** 309-16.

DERHEIMER, F. A., HICKS, J. K., PAULSEN, M. T., CANMAN, C. E. & LJUNGMAN, M. 2009. Psoralen-induced DNA interstrand cross-links block transcription and induce p53 in an ataxia-telangiectasia and rad3-related-dependent manner. *Mol Pharmacol,* 75**,** 599-607.

DICK, J. E. 2003. Breast cancer stem cells revealed. *Proc Natl Acad Sci U S A,* 100**,** 3547-9.

DIGIOVANNA, J. J. & KRAEMER, K. H. 2012. Shining a light on xeroderma pigmentosum. *J Invest Dermatol,* 132**,** 785-96.

DIXON, K. M., NORMAN, A. W., SEQUEIRA, V. B., MOHAN, R., RYBCHYN, M. S., REEVE, V. E., HALLIDAY, G. M. & MASON, R. S. 2011. 1alpha,25(OH)(2)-vitamin D and a nongenomic vitamin D analogue inhibit ultraviolet radiation-induced skin carcinogenesis. *Cancer Prev Res (Phila),* 4**,** 1485-94.

DOORBAR, J. 2006. Molecular biology of human papillomavirus infection and cervical cancer. *Clin Sci (Lond),* 110**,** 525-41.

DOTTO, G. P. & RUSTGI, A. K. 2016. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. *Cancer Cell,* 29**,** 622-637.

DROSTEN, M., DHAWAHIR, A., SUM, E. Y., UROSEVIC, J., LECHUGA, C. G., ESTEBAN, L. M., CASTELLANO, E., GUERRA, C., SANTOS, E. & BARBACID, M. 2010. Genetic analysis of Ras signalling pathways in cell proliferation, migration and survival. *EMBO J,* 29**,** 1091-104.

DUNCAVAGE, E. J., SCHROEDER, M. C., O'LAUGHLIN, M., WILSON, R., MACMILLAN, S., BOHANNON, A., KRUCHOWSKI, S., GARZA, J., DU, F., HUGHES, A. E. O., ROBINSON, J., HUGHES, E., HEATH, S. E., BATY, J. D., NEIDICH, J., CHRISTOPHER, M. J., JACOBY, M. A., UY, G. L., FULTON, R. S., MILLER, C. A., PAYTON, J. E., LINK, D. C., WALTER, M. J., WESTERVELT, P., DIPERSIO, J. F., LEY, T. J. & SPENCER, D. H. 2021. Genome Sequencing as an Alternative to Cytogenetic Analysis in Myeloid Cancers. *New England Journal of Medicine,* 384**,** 924-935.

DUNN, J., POTTER, M., REES, A. & RUNGER, T. M. 2006. Activation of the Fanconi anemia/BRCA pathway and recombination repair in the cellular response to solar ultraviolet light. *Cancer Res,* 66**,** 11140-7.

EKLOF SPINK, K., FRIDMAN, S. G. & WEIS, W. I. 2001. Molecular mechanisms of beta-catenin recognition by adenomatous polyposis coli revealed by the structure of an APC-beta-catenin complex. *EMBO J,* 20**,** 6203-12.

ELLROTT, K., BAILEY, M. H., SAKSENA, G., COVINGTON, K. R., KANDOTH, C., STEWART, C., HESS, J., MA, S., CHIOTTI, K. E., MCLELLAN, M., SOFIA, H. J., HUTTER, C., GETZ, G., WHEELER, D., DING, L., GROUP, M. C. W. & CANCER GENOME ATLAS RESEARCH, N. 2018. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst,* 6**,** 271-281 e7.

ENZO, E., SECONE SECONETTI, A., FORCATO, M., TENEDINI, E., POLITO, M. P., SALA, I., CARULLI, S., CONTIN, R., PEANO, C., TAGLIAFICO, E., BICCIATO, S., BONDANZA, S. & DE LUCA, M. 2021. Single-keratinocyte transcriptomic analyses identify different clonal types and proliferative potential mediated by FOXM1 in human epidermal stem cells. *Nat Commun,* 12**,** 2505.

EPSTEIN, E. H. 2008. Basal cell carcinomas: attack of the hedgehog. *Nat Rev Cancer,* 8**,** 743-54.

FAN, X., MIKOLAENKO, I., ELHASSAN, I., NI, X., WANG, Y., BALL, D., BRAT, D. J., PERRY, A. & EBERHART, C. G. 2004. Notch1 and notch2 have opposite effects on embryonal brain tumor growth. *Cancer Res,* 64**,** 7787-93.

FEARON, E. R. & VOGELSTEIN, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-67.

FINNEY, S. A. 2001. Real-time data collection in Linux: a case study. *Behav Res Methods Instrum Comput,* 33**,** 167-73.

FISCHER, M. 2017. Census and evaluation of p53 target genes. *Oncogene,* 36**,** 3943-3956.

FITZPATRICK, T. B. 1988. The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol,* 124**,** 869-71.

FLINDT-HANSEN, H., MCFADDEN, N., EEG-LARSEN, T. & THUNE, P. 1991. Effect of a new narrow-band UVB lamp on photocarcinogenesis in mice. *Acta Derm Venereol,* 71**,** 245-8.

FORBES, S. A., BINDAL, N., BEARE, D., BAMFORD, S., COLE, C. G., WARD, S., LEUNG, K., KOK, C. Y., JIA, M. M., DE, T. S., SONDKA, Z., STRATTON, M. R. & CAMPBELL, P. J. 2016. COSMIC: comprehensively exploring oncogenomics. *Cancer Research,* 76.

FORTINI, P., PARLANTI, E., SIDORKINA, O. M., LAVAL, J. & DOGLIOTTI, E. 1999. The type of DNA glycosylase determines the base excision repair pathway in mammalian cells. *J Biol Chem,* 274**,** 15230-6.

FOURMENT, M. & GILLINGS, M. R. 2008. A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics,* 9**,** 82.

FOUSTERI, M. & MULLENDERS, L. H. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res,* 18**,** 73-84.

FOWLER, J. C., KING, C., BRYANT, C., HALL, M., SOOD, R., ONG, S. H., EARP, E., FERNANDEZ-ANTORAN, D., KOEPPEL, J., DENTRO, S. C., SHORTHOUSE, D., DURRANI, A., FIFE, K., RYTINA, E., MILNE, D., ROSHAN, A., MAHUBUBANI, K., SAEB-PARSY, K., HALL, B. A., GERSTUNG, M. & JONES, P. H. 2020. Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discov*.

FRAZER, K. A. 2012. Decoding the human genome. *Genome Res,* 22**,** 1599-601.

FREEMAN, S. E. 1988. Variations in excision repair of UVB-induced pyrimidine dimers in DNA of human skin in situ. *J Invest Dermatol,* 90**,** 814-7.

FUKUSUMI, T., GUO, T. W., SAKAI, A., ANDO, M., REN, S., HAFT, S., LIU, C., AMORNPHIMOLTHAM, P., GUTKIND, J. S. & CALIFANO, J. A. 2018. The NOTCH4-HEY1 Pathway Induces Epithelial-Mesenchymal Transition in Head and Neck Squamous Cell Carcinoma. *Clin Cancer Res,* 24**,** 619-633.

GAO, G. F., PARKER, J. S., REYNOLDS, S. M., SILVA, T. C., WANG, L. B., ZHOU, W., AKBANI, R., BAILEY, M., BALU, S., BERMAN, B. P., BROOKS, D., CHEN, H., CHERNIACK, A. D., DEMCHOK, J. A., DING, L., FELAU, I., GAHEEN, S., GERHARD, D. S., HEIMAN, D. I., HERNANDEZ, K. M., HOADLEY, K. A., JAYASINGHE, R., KEMAL, A., KNIJNENBURG, T. A., LAIRD, P. W., MENSAH, M. K. A., MUNGALL, A. J., ROBERTSON, A. G., SHEN, H., TARNUZZER, R., WANG, Z., WYCZALKOWSKI, M., YANG, L., ZENKLUSEN, J. C., ZHANG, Z., GENOMIC DATA ANALYSIS, N., LIANG, H. & NOBLE, M. S. 2019. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst,* 9**,** 24-34 e10.

GAO, Y. B., CHEN, Z. L., LI, J. G., HU, X. D., SHI, X. J., SUN, Z. M., ZHANG, F., ZHAO, Z. R., LI, Z. T., LIU, Z. Y., ZHAO, Y. D., SUN, J., ZHOU, C. C., YAO, R., WANG, S. Y., WANG, P., SUN, N., ZHANG, B. H., DONG, J. S., YU, Y., LUO, M., FENG, X. L., SHI, S. S., ZHOU, F., TAN, F. W., QIU, B., LI, N., SHAO, K., ZHANG, L. J., ZHANG, L. J., XUE, Q., GAO, S. G. & HE, J. 2014. Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet,* 46**,** 1097-102.

GARFINKEL, D. J., BOEKE, J. D. & FINK, G. R. 1985. Ty element transposition: reverse transcriptase and virus-like particles. *Cell,* 42**,** 507-17.

GATENBY, R. A. & GILLIES, R. J. 2008. A microenvironmental model of carcinogenesis. *Nat Rev Cancer,* 8**,** 56-61.

GAUJOUX, R. & SEOIGHE, C. 2010. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics,* 11**,** 367.

GAUTHIER, J., VINCENT, A. T., CHARETTE, S. J. & DEROME, N. 2019. A brief history of bioinformatics. *Brief Bioinform,* 20**,** 1981-1996.

GENOVESE, G., KÄHLER, A. K., HANDSAKER, R. E., LINDBERG, J., ROSE, S. A., BAKHOUM, S. F., CHAMBERT, K., MICK, E., NEALE, B. M., FROMER, M., PURCELL, S. M., SVANTESSON, O., LANDÉN, M., HÖGLUND, M., LEHMANN, S., GABRIEL, S. B., MORAN, J. L., LANDER, E. S., SULLIVAN, P. F., SKLAR, P., GRÖNBERG, H., HULTMAN, C. M. & MCCARROLL, S. A. 2014. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine,* 371**,** 2477-2487.

GIBBS, N. K., TRAYNOR, N. J., MACKIE, R. M., CAMPBELL, I., JOHNSON, B. E. & FERGUSON, J. 1995. The phototumorigenic potential of broad-band (270-350 nm) and narrow-band (311-313 nm) phototherapy sources cannot be predicted by their edematogenic potential in hairless mouse skin. *J Invest Dermatol,* 104**,** 359-63.

GILCHREST, B. A., ELLER, M. S., GELLER, A. C. & YAAR, M. 1999. The pathogenesis of melanoma induced by ultraviolet radiation. *N Engl J Med,* 340**,** 1341-8.

GLASS, D., VINUELA, A., DAVIES, M. N., RAMASAMY, A., PARTS, L., KNOWLES, D., BROWN, A. A., HEDMAN, A. K., SMALL, K. S., BUIL, A., GRUNDBERG, E., NICA, A. C., DI MEGLIO, P., NESTLE, F. O., RYTEN, M., CONSORTIUM, U. K. B. E., MU, T. C., DURBIN, R., MCCARTHY, M. I., DELOUKAS, P., DERMITZAKIS, E. T., WEALE, M. E., BATAILLE, V. & SPECTOR, T. D. 2013. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol,* 14**,** R75.

GLEASON, B. C., CRUM, C. P. & MURPHY, G. F. 2008. Expression patterns of MITF during human cutaneous embryogenesis: evidence for bulge epithelial expression and persistence of dermal melanoblasts. *J Cutan Pathol,* 35**,** 615-22.

GLOGAU, R. G. 2000. The risk of progression to invasive disease. *J Am Acad Dermatol,* 42**,** 23-4.

GOECKERMAN, W. 1925. Treatment of psoriasis. *Northwest Medicine***,** 229-231.

GOENKA, S. D., GORZYNSKI, J. E., SHAFIN, K., FISK, D. G., PESOUT, T., JENSEN, T. D., MONLONG, J., CHANG, P.-C., BAID, G., BERNSTEIN, J. A., CHRISTLE, J. W., DALTON, K. P., GARALDE, D. R., GROVE, M. E., GUILLORY, J., KOLESNIKOV, A., NATTESTAD, M., RUZHNIKOV, M. R. Z., SAMADI, M., SETHIA, A., SPITERI, E., WRIGHT, C. J., XIONG, K., ZHU, T., JAIN, M., SEDLAZECK, F. J., CARROLL, A., PATEN, B. & ASHLEY, E. A. 2022. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nature Biotechnology,* 40**,** 1035-1041.

GORDON-THOMSON, C., GUPTA, R., TONGKAO-ON, W., RYAN, A., HALLIDAY, G. M. & MASON, R. S. 2012. 1alpha,25 dihydroxyvitamin D3 enhances cellular defences against UV-induced oxidative and other forms of DNA damage in skin. *Photochem Photobiol Sci,* 11**,** 1837-47.

GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

GREEN, A. C., WILLIAMS, G. M., LOGAN, V. & STRUTTON, G. M. 2011. Reduced melanoma after regular sunscreen use: randomized trial follow-up. *J Clin Oncol,* 29**,** 257-63.

GREEN, C., FERGUSON, J., LAKSHMIPATHI, T. & JOHNSON, B. E. 1988. 311 nm UVB phototherapy--an effective treatment for psoriasis. *Br J Dermatol,* 119**,** 691-6.

GREENMAN, C., WOOSTER, R., FUTREAL, P. A., STRATTON, M. R. & EASTON, D. F. 2006. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics,* 173**,** 2187-98.

GREINERT, R., VOLKMER, B., HENNING, S., BREITBART, E. W., GREULICH, K. O., CARDOSO, M. C. & RAPP, A. 2012. UVA-induced DNA double-strand breaks result from the repair of clustered oxidative DNA damages. *Nucleic Acids Research,* 40**,** 10263-10273.

GRICE, E. A. & SEGRE, J. A. 2011. The skin microbiome. *Nat Rev Microbiol,* 9**,** 244-53.

GRIFFITH, M., GRIFFITH, O. L., COFFMAN, A. C., WEIBLE, J. V., MCMICHAEL, J. F., SPIES, N. C., KOVAL, J., DAS, I., CALLAWAY, M. B., ELDRED, J. M., MILLER, C. A., SUBRAMANIAN, J., GOVINDAN, R., KUMAR, R. D., BOSE, R., DING, L., WALKER, J. R., LARSON, D. E., DOOLING,

D. J., SMITH, S. M., LEY, T. J., MARDIS, E. R. & WILSON, R. K. 2013. DGIdb: mining the druggable genome. *Nat Methods,* 10**,** 1209-10.

GRIFFITHS, H. R., MISTRY, P., HERBERT, K. E. & LUNEC, J. 1998. Molecular and cellular effects of ultraviolet light-induced genotoxicity. *Crit Rev Clin Lab Sci,* 35**,** 189-237.

GRIVENNIKOV, S. I., GRETEN, F. R. & KARIN, M. 2010. Immunity, inflammation, and cancer. *Cell,* 140**,** 883-99.

GROSSMAN, R. L., HEATH, A. P., FERRETTI, V., VARMUS, H. E., LOWY, D. R., KIBBE, W. A. & STAUDT, L. M. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med,* 375**,** 1109-12.

GUDJONSSON, J. E., JOHNSTON, A., SIGMUNDSDOTTIR, H. & VALDIMARSSON, H. 2004. Immunopathogenic mechanisms in psoriasis. *Clin Exp Immunol,* 135**,** 1-8.

GUENAY-GREUNKE, Y., BOHAN, D. A., TRAUGOTT, M. & WALLINGER, C. 2021. Handling of targeted amplicon sequencing data focusing on index hopping and demultiplexing using a nested metabarcoding approach in ecology. *Scientific Reports,* 11**,** 19510.

GUO, C., LI, X., WANG, R., YU, J., YE, M., MAO, L., ZHANG, S. & ZHENG, S. 2016. Association between Oxidative DNA Damage and Risk of Colorectal Cancer: Sensitive Determination of Urinary 8-Hydroxy-2'-deoxyguanosine by UPLC-MS/MS Analysis. *Sci Rep,* 6**,** 32581.

HAAKE, A., SCOTT, G.A. & HOLBROOK, K.A. 2001. Structure and function of the skin: overview of the epidermis and dermis. *In:* WOODLY, D. T. F., R.K. (ed.) *The biology of the skin.* New York: The Parthoenon Publishing Group Inc.

HALDER, R. M. & BANG, K. M. 1988. Skin cancer in blacks in the United States. *Dermatol Clin,* 6**,** 397-405.

HAMEETMAN, L., COMMANDEUR, S., BAVINCK, J. N., WISGERHOF, H. C., DE GRUIJL, F. R., WILLEMZE, R., MULLENDERS, L., TENSEN, C. P. & VRIELING, H. 2013. Molecular profiling of cutaneous squamous cell carcinomas and actinic keratoses from organ transplant recipients. *BMC Cancer,* 13**,** 58.

HAMID, O., ROBERT, C., DAUD, A., HODI, F. S., HWU, W. J., KEFFORD, R., WOLCHOK, J. D., HERSEY, P., JOSEPH, R. W., WEBER, J. S., DRONCA, R., GANGADHAR, T. C., PATNAIK, A., ZAROUR, H., JOSHUA, A. M., GERGICH, K., ELASSAISS-SCHAAP, J., ALGAZI, A., MATEUS, C., BOASBERG, P., TUMEH, P. C., CHMIELOWSKI, B., EBBINGHAUS, S. W., LI, X. N., KANG, S. P. & RIBAS, A. 2013. Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N Engl J Med,* 369**,** 134-44.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

HARA, J., HIGUCHI, K., OKAMOTO, R., KAWASHIMA, M. & IMOKAWA, G. 2000. High-expression of sphingomyelin deacylase is an important determinant of ceramide deficiency leading to barrier disruption in atopic dermatitis. *J Invest Dermatol,* 115**,** 406-13.

HARIDAS, D., PONNUSAMY, M. P., CHUGH, S., LAKSHMANAN, I., SESHACHARYULU, P. & BATRA, S. K. 2014. MUC16: molecular analysis and its functional implications in benign and malignant conditions. *FASEB J,* 28**,** 4183-99.

HATTORI, Y., NISHIGORI, C., TANAKA, T., UCHIDA, K., NIKAIDO, O., OSAWA, T., HIAI, H., IMAMURA, S. & TOYOKUNI, S. 1996. 8-hydroxy-2'-deoxyguanosine is increased in epidermal cells of hairless mice after chronic ultraviolet B exposure. *J Invest Dermatol,* 107**,** 733-7.

HAYASHI, Y., SUEMITSU, E., KAJIMOTO, K., SATO, Y., AKHTER, A., SAKURAI, Y., HATAKEYAMA, H., HYODO, M., KAJI, N., BABA, Y. & HARASHIMA, H. 2014. Hepatic Monoacylglycerol O-acyltransferase 1 as a Promising Therapeutic Target for Steatosis, Obesity, and Type 2 Diabetes. *Mol Ther Nucleic Acids,* 3**,** e154.

HAYWARD, N. K., WILMOTT, J. S., WADDELL, N., JOHANSSON, P. A., FIELD, M. A., NONES, K., PATCH, A. M., KAKAVAND, H., ALEXANDROV, L. B., BURKE, H., JAKROT, V., KAZAKOFF, S., HOLMES, O., LEONARD, C., SABARINATHAN, R., MULARONI, L., WOOD, S., XU, Q., WADDELL, N., TEMBE, V., PUPO, G. M., DE PAOLI-ISEPPI, R., VILAIN, R. E., SHANG, P., LAU,

L. M. S., DAGG, R. A., SCHRAMM, S. J., PRITCHARD, A., DUTTON-REGESTER, K., NEWELL, F., FITZGERALD, A., SHANG, C. A., GRIMMOND, S. M., PICKETT, H. A., YANG, J. Y., STRETCH, J. R., BEHREN, A., KEFFORD, R. F., HERSEY, P., LONG, G. V., CEBON, J., SHACKLETON, M., SPILLANE, A. J., SAW, R. P. M., LOPEZ-BIGAS, N., PEARSON, J. V., THOMPSON, J. F., SCOLYER, R. A. & MANN, G. J. 2017. Whole-genome landscapes of major melanoma subtypes. *Nature,* 545**,** 175-180.

HEALY, E., FLANNAGAN, N., RAY, A., TODD, C., JACKSON, I. J., MATTHEWS, J. N., BIRCH-MACHIN, M. A. & REES, J. L. 2000. Melanocortin-1-receptor gene and sun sensitivity in individuals without red hair. *Lancet,* 355**,** 1072-3.

HEARN, R. M., KERR, A. C., RAHIM, K. F., FERGUSON, J. & DAWE, R. S. 2008. Incidence of skin cancers in 3867 patients treated with narrow-band ultraviolet B phototherapy. *Br J Dermatol,* 159**,** 931-5.

HENNESSEY, R. C., BOWMAN, R. L., TALLMAN, D. A., WEISS, T. J., CRAWFORD, E. R., MURPHY, B. M., WEBB, A., ZHANG, S., LA PERLE, K. M. D., BURD, C. J., LEVINE, R. L., SHAIN, A. H. & BURD, C. E. 2019. UVA and UVB elicit distinct mutational signatures in melanoma. *bioRxiv***,** 778449.

HIRSCH, T., ROTHOEFT, T., TEIG, N., BAUER, J. W., PELLEGRINI, G., DE ROSA, L., SCAGLIONE, D., REICHELT, J., KLAUSEGGER, A., KNEISZ, D., ROMANO, O., SECONE SECONETTI, A., CONTIN, R., ENZO, E., JURMAN, I., CARULLI, S., JACOBSEN, F., LUECKE, T., LEHNHARDT, M., FISCHER, M., KUECKELHAUS, M., QUAGLINO, D., MORGANTE, M., BICCIATO, S., BONDANZA, S. & DE LUCA, M. 2017. Regeneration of the entire human epidermis using transgenic stem cells. *Nature,* 551**,** 327-332.

HOANG, M. L., KINDE, I., TOMASETTI, C., MCMAHON, K. W., ROSENQUIST, T. A., GROLLMAN, A. P., KINZLER, K. W., VOGELSTEIN, B. & PAPADOPOULOS, N. 2016. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A,* 113**,** 9846-51.

HOBBS, G. A., DER, C. J. & ROSSMAN, K. L. 2016. RAS isoforms and mutations in cancer at a glance. *J Cell Sci,* 129**,** 1287-92.

HODIS, E., WATSON, I. R., KRYUKOV, G. V., AROLD, S. T., IMIELINSKI, M., THEURILLAT, J. P., NICKERSON, E., AUCLAIR, D., LI, L. R., PLACE, C., DICARA, D., RAMOS, A. H., LAWRENCE, M. S., CIBULSKIS, K., SIVACHENKO, A., VOET, D., SAKSENA, G., STRANSKY, N., ONOFRIO, R. C., WINCKLER, W., ARDLIE, K., WAGLE, N., WARGO, J., CHONG, K., MORTON, D. L., STEMKE-HALE, K., CHEN, G., NOBLE, M., MEYERSON, M., LADBURY, J. E., DAVIES, M. A., GERSHENWALD, J. E., WAGNER, S. N., HOON, D. S. B., SCHADENDORF, D., LANDER, E. S., GABRIEL, S. B., GETZ, G., GARRAWAY, L. A. & CHIN, L. 2012. A Landscape of Driver Mutations in Melanoma. *Cell,* 150**,** 251-263.

HORN, S., FIGL, A., RACHAKONDA, P. S., FISCHER, C., SUCKER, A., GAST, A., KADEL, S., MOLL, I., NAGORE, E., HEMMINKI, K., SCHADENDORF, D. & KUMAR, R. 2013. TERT promoter mutations in familial and sporadic melanoma. *Science,* 339**,** 959-61.

HSU, Y. C., LI, L. & FUCHS, E. 2014. Emerging interactions between skin stem cells and their niches. *Nat Med,* 20**,** 847-56.

HUANG, F. W., HODIS, E., XU, M. J., KRYUKOV, G. V., CHIN, L. & GARRAWAY, L. A. 2013. Highly recurrent TERT promoter mutations in human melanoma. *Science,* 339**,** 957-9.

IANNACONE, M. R., GHEIT, T., WATERBOER, T., GIULIANO, A. R., MESSINA, J. L., FENSKE, N. A., CHERPELIS, B. S., SONDAK, V. K., ROETZHEIM, R. G., MICHAEL, K. M., TOMMASINO, M., PAWLITA, M. & ROLLISON, D. E. 2012. Case-control study of cutaneous human papillomaviruses in squamous cell carcinoma of the skin. *Cancer Epidemiol Biomarkers Prev,* 21**,** 1303-13.

IBBOTSON, S. H., BILSLAND, D., COX, N. H., DAWE, R. S., DIFFEY, B., EDWARDS, C., FARR, P. M., FERGUSON, J., HART, G., HAWK, J., LLOYD, J., MARTIN, C., MOSELEY, H., MCKENNA, K., RHODES, L. E., TAYLOR, D. K. & BRITISH ASSOCIATION OF, D. 2004. An update and guidance on narrowband ultraviolet B phototherapy: a British Photodermatology Group Workshop Report. *Br J Dermatol,* 151**,** 283-97.

IMOKAWA, G. 2001. Lipid abnormalities in atopic dermatitis. *J Am Acad Dermatol,* 45**,** S29-32.

INMAN, G. J., WANG, J., NAGANO, A., ALEXANDROV, L. B., PURDIE, K. J., TAYLOR, R. G., SHERWOOD, V., THOMSON, J., HOGAN, S., SPENDER, L. C., SOUTH, A. P., STRATTON, M., CHELALA, C., HARWOOD, C. A., PROBY, C. M. & LEIGH, I. M. 2018. The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. *Nat Commun,* 9**,** 3667.

IOANNIDIS, J. P., ZHOU, Y., CHANG, C. Q., SCHULLY, S. D., KHOURY, M. J. & FREEDMAN, A. N. 2014. Potential increased risk of cancer from commonly used medications: an umbrella review of meta-analyses. *Ann Oncol,* 25**,** 16-23.

ITO, S. & WAKAMATSU, K. 2003. Quantitative analysis of eumelanin and pheomelanin in humans, mice, and other animals: a comparative review. *Pigment Cell Res,* 16**,** 523-31.

JACKSON, S. P. & BARTEK, J. 2009. The DNA-damage response in human biology and disease. *Nature,* 461**,** 1071-1078.

JAISWAL, S., FONTANILLAS, P., FLANNICK, J., MANNING, A., GRAUMAN, P. V., MAR, B. G., LINDSLEY, R. C., MERMEL, C. H., BURTT, N., CHAVEZ, A., HIGGINS, J. M., MOLTCHANOV, V., KUO, F. C., KLUK, M. J., HENDERSON, B., KINNUNEN, L., KOISTINEN, H. A., LADENVALL, C., GETZ, G., CORREA, A., BANAHAN, B. F., GABRIEL, S., KATHIRESAN, S., STRINGHAM, H. M., MCCARTHY, M. I., BOEHNKE, M., TUOMILEHTO, J., HAIMAN, C., GROOP, L., ATZMON, G., WILSON, J. G., NEUBERG, D., ALTSHULER, D. & EBERT, B. L. 2014. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine,* 371**,** 2488-2498.

JAVERI, A., HUANG, X. X., BERNERD, F., MASON, R. S. & HALLIDAY, G. M. 2008. Human 8-oxoguanine-DNA glycosylase 1 protein and gene are expressed more abundantly in the superficial than basal layer of human epidermis. *DNA Repair (Amst),* 7**,** 1542-50.

JAYARAMAN, S. S., RAYHAN, D. J., HAZANY, S. & KOLODNEY, M. S. 2014. Mutational Landscape of Basal Cell Carcinomas by Whole-Exome Sequencing. *Journal of Investigative Dermatology,* 134**,** 213-220.

JENKINS, T. 2004. The First Language - A Case for Python? *Innovation in Teaching and Learning in Information and Computer Sciences,* 3**,** 1-9.

JOLLY, C. & VAN LOO, P. 2018. Timing somatic events in the evolution of cancer. *Genome Biol,* 19**,** 95.

JONASON, A. S., KUNALA, S., PRICE, G. J., RESTIFO, R. J., SPINELLI, H. M., PERSING, J. A., LEFFELL, D. J., TARONE, R. E. & BRASH, D. E. 1996. Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc Natl Acad Sci U S A,* 93**,** 14025-9.

KAAE, J., BOYD, H. A., HANSEN, A. V., WULF, H. C., WOHLFAHRT, J. & MELBYE, M. 2010. Photosensitizing medication use and risk of skin cancer. *Cancer Epidemiol Biomarkers Prev,* 19**,** 2942-9.

KADALAYIL, L., RAFIQ, S., ROSE-ZERILLI, M. J., PENGELLY, R. J., PARKER, H., OSCIER, D., STREFFORD, J. C., TAPPER, W. J., GIBSON, J., ENNIS, S. & COLLINS, A. 2015. Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform,* 16**,** 380-92.

KALADY, M. F., WHITE, R. R., JOHNSON, J. L., TYLER, D. S. & SEIGLER, H. F. 2003. Thin melanomas: predictive lethal characteristics from a 30-year clinical experience. *Ann Surg,* 238**,** 528-35; discussion 535-7.

KAN, Z., JAISWAL, B. S., STINSON, J., JANAKIRAMAN, V., BHATT, D., STERN, H. M., YUE, P., HAVERTY, P. M., BOURGON, R., ZHENG, J., MOORHEAD, M., CHAUDHURI, S., TOMSHO, L. P., PETERS, B. A., PUJARA, K., CORDES, S., DAVIS, D. P., CARLTON, V. E., YUAN, W., LI, L., WANG, W., EIGENBROT, C., KAMINKER, J. S., EBERHARD, D. A., WARING, P., SCHUSTER, S. C., MODRUSAN, Z., ZHANG, Z., STOKOE, D., DE SAUVAGE, F. J., FAHAM, M. & SESHAGIRI,

S. 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature,* 466**,** 869-73.

KANJILAL, S., STROM, S. S., CLAYMAN, G. L., WEBER, R. S., EL-NAGGAR, A. K., KAPUR, V., CUMMINGS, K. K., HILL, L. A., SPITZ, M. R., KRIPKE, M. L. & ET AL. 1995. p53 mutations in nonmelanoma skin cancer of the head and neck: molecular evidence for field cancerization. *Cancer Res,* 55**,** 3604-9.

KARAOZ, U., MURALI, T. M., LETOVSKY, S., ZHENG, Y., DING, C., CANTOR, C. R. & KASIF, S. 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A,* 101**,** 2888-93.

KAZEMI-SEFAT, G. E., KERAMATIPOUR, M., TALEBI, S., KAVOUSI, K., SAJED, R., KAZEMI-SEFAT, N. A. & MOUSAVIZADEH, K. 2021. The importance of CDC27 in cancer: molecular pathology and clinical aspects. *Cancer Cell Int,* 21**,** 160.

KENFIELD, S. A., WEI, E. K., STAMPFER, M. J., ROSNER, B. A. & COLDITZ, G. A. 2008. Comparison of aspects of smoking among the four histological types of lung cancer. *Tob Control,* 17**,** 198-204.

KIM, R., EMI, M. & TANABE, K. 2007. Cancer immunoediting from immune surveillance to immune escape. *Immunology,* 121**,** 1-14.

KIM, S. H., HO, J. N., JIN, H., LEE, S. C., LEE, S. E., HONG, S. K., LEE, J. W., LEE, E. S. & BYUN, S. S. 2016. Upregulated expression of BCL2, MCM7, and CCNE1 indicate cisplatin-resistance in the set of two human bladder cancer cell lines: T24 cisplatin sensitive and T24R2 cisplatin resistant bladder cancer cell lines. *Investig Clin Urol,* 57**,** 63-72.

KIRKE, S. M., LOWDER, S., LLOYD, J. J., DIFFEY, B. L., MATTHEWS, J. N. & FARR, P. M. 2007. A randomized comparison of selective broadband UVB and narrowband UVB in the treatment of psoriasis. *J Invest Dermatol,* 127**,** 1641-6.

KIVISAARI, A. & KAHARI, V. M. 2013. Squamous cell carcinoma of the skin: Emerging need for novel biomarkers. *World J Clin Oncol,* 4**,** 85-90.

KNUDSON, A. G., JR. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

KOBAYASHI, K., HISAMATSU, K., SUZUI, N., HARA, A., TOMITA, H. & MIYAZAKI, T. 2018. A Review of HPV-Related Head and Neck Cancer. *J Clin Med,* 7.

KRICKER, A., ARMSTRONG, B. K., ENGLISH, D. R. & HEENAN, P. J. 1995. Does intermittent sun exposure cause basal cell carcinoma? a case-control study in Western Australia. *Int J Cancer,* 60**,** 489-94.

KROKAN, H. E. & BJORAS, M. 2013. Base excision repair. *Cold Spring Harb Perspect Biol,* 5**,** a012583.

KRUEGER, J. G., WOLFE, J. T., NABEYA, R. T., VALLAT, V. P., GILLEAUDEAU, P., HEFTLER, N. S., AUSTIN, L. M. & GOTTLIEB, A. B. 1995. Successful ultraviolet B treatment of psoriasis is accompanied by a reversal of keratinocyte pathology and by selective depletion of intraepidermal T cells. *J Exp Med,* 182**,** 2057-68.

KUMAR, L., SKIDMORE, A. K. & KNOWLES, E. 1997. Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science,* 11**,** 475-497.

KUNISADA, M., KUMIMOTO, H., ISHIZAKI, K., SAKUMI, K., NAKABEPPU, Y. & NISHIGORI, C. 2007. Narrow-band UVB induces more carcinogenic skin tumors than broad-band UVB through the formation of cyclobutane pyrimidine dimer. *J Invest Dermatol,* 127**,** 2865-71.

KUNISADA, M., SAKUMI, K., TOMINAGA, Y., BUDIYANTO, A., UEDA, M., ICHIHASHI, M., NAKABEPPU, Y. & NISHIGORI, C. 2005. 8-Oxoguanine formation induced by chronic UVB exposure makes Ogg1 knockout mice susceptible to skin carcinogenesis. *Cancer Res,* 65**,** 6006-10.

LAI, C., COLTART, G., SHAPANIS, A., HEALY, C., ALABDULKAREEM, A., SELVENDRAN, S., THEAKER, J., SOMMERLAD, M., ROSE-ZERILLI, M., AL-SHAMKHANI, A. & HEALY, E. 2021.

CD8+CD103+ tissue-resident memory T cells convey reduced protective immunity in cutaneous squamous cell carcinoma. *J Immunother Cancer,* 9.

LAMASON, R. L., MOHIDEEN, M. A., MEST, J. R., WONG, A. C., NORTON, H. L., AROS, M. C., JURYNEC, M. J., MAO, X., HUMPHREVILLE, V. R., HUMBERT, J. E., SINHA, S., MOORE, J. L., JAGADEESWARAN, P., ZHAO, W., NING, G., MAKALOWSKA, I., MCKEIGUE, P. M., O'DONNELL, D., KITTLES, R., PARRA, E. J., MANGINI, N. J., GRUNWALD, D. J., SHRIVER, M. D., CANFIELD, V. A. & CHENG, K. C. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science,* 310**,** 1782-6.

LAPPALAINEN, I., ALMEIDA-KING, J., KUMANDURI, V., SENF, A., SPALDING, J. D., UR-REHMAN, S., SAUNDERS, G., KANDASAMY, J., CACCAMO, M., LEINONEN, R., VAUGHAN, B., LAURENT, T., ROWLAND, F., MARIN-GARCIA, P., BARKER, J., JOKINEN, P., TORRES, A. C., DE ARGILA, J. R., LLOBET, O. M., MEDINA, I., PUY, M. S., ALBERICH, M., DE LA TORRE, S., NAVARRO, A., PASCHALL, J. & FLICEK, P. 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet,* 47**,** 692-5.

LARSEN, P. A., HEILMAN, A. M. & YODER, A. D. 2014. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics,* 15**,** 720.

LAWRENCE, M. S., STOJANOV, P., MERMEL, C. H., ROBINSON, J. T., GARRAWAY, L. A., GOLUB, T. R., MEYERSON, M., GABRIEL, S. B., LANDER, E. S. & GETZ, G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature,* 505**,** 495-501.

LAWRENCE, M. S., STOJANOV, P., POLAK, P., KRYUKOV, G. V., CIBULSKIS, K., SIVACHENKO, A., CARTER, S. L., STEWART, C., MERMEL, C. H., ROBERTS, S. A., KIEZUN, A., HAMMERMAN, P. S., MCKENNA, A., DRIER, Y., ZOU, L., RAMOS, A. H., PUGH, T. J., STRANSKY, N., HELMAN, E., KIM, J., SOUGNEZ, C., AMBROGIO, L., NICKERSON, E., SHEFLER, E., CORTES, M. L., AUCLAIR, D., SAKSENA, G., VOET, D., NOBLE, M., DICARA, D., LIN, P., LICHTENSTEIN, L., HEIMAN, D. I., FENNELL, T., IMIELINSKI, M., HERNANDEZ, B., HODIS, E., BACA, S., DULAK, A. M., LOHR, J., LANDAU, D. A., WU, C. J., MELENDEZ-ZAJGLA, J., HIDALGO-MIRANDA, A., KOREN, A., MCCARROLL, S. A., MORA, J., CROMPTON, B., ONOFRIO, R., PARKIN, M., WINCKLER, W., ARDLIE, K., GABRIEL, S. B., ROBERTS, C. W. M., BIEGEL, J. A., STEGMAIER, K., BASS, A. J., GARRAWAY, L. A., MEYERSON, M., GOLUB, T. R., GORDENIN, D. A., SUNYAEV, S., LANDER, E. S. & GETZ, G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-218.

LEACH, D. R., KRUMMEL, M. F. & ALLISON, J. P. 1996. Enhancement of antitumor immunity by CTLA-4 blockade. *Science,* 271**,** 1734-6.

LEE, E. Y. H. P. & MULLER, W. J. 2010. Oncogenes and Tumor Suppressor Genes. *Cold Spring Harbor Perspectives in Biology,* 2.

LEE, H. J., KIM, J. S., HA, S. J., ROH, K. Y., SEO, E. J., PARK, W. S., LEE, J. Y., PARK, K. S. & KIM, J. W. 2000. p53 gene mutations in Bowen's disease in Koreans: clustering in exon 5 and multiple mutations. *Cancer Lett,* 158**,** 27-33.

LEEMANS, C. R., BRAAKHUIS, B. J. & BRAKENHOFF, R. H. 2011. The molecular biology of head and neck cancer. *Nat Rev Cancer,* 11**,** 9-22.

LEONE, P. E., GONZALEZ, M. B., ELOSUA, C., GOMEZ-MORETA, J. A., LUMBRERAS, E., ROBLEDO, C., SANTOS-BRIZ, A., VALERO, J. M., DE LA GUARDIA, R. D., GUTIERREZ, N. C., HERNANDEZ, J. M. & GARCIA, J. L. 2012. Integration of global spectral karyotyping, CGH arrays, and expression arrays reveals important genes in the pathogenesis of glioblastoma multiforme. *Ann Surg Oncol,* 19**,** 2367-79.

LEVINE, A. J. 1997. p53, the cellular gatekeeper for growth and division. *Cell,* 88**,** 323-31.

LI, V. S., NG, S. S., BOERSEMA, P. J., LOW, T. Y., KARTHAUS, W. R., GERLACH, J. P., MOHAMMED, S., HECK, A. J., MAURICE, M. M., MAHMOUDI, T. & CLEVERS, H. 2012. Wnt signaling through inhibition of beta-catenin degradation in an intact Axin1 complex. *Cell,* 149**,** 1245-56.

LI, X., UPADHYAY, A. K., BULLOCK, A. J., DICOLANDREA, T., XU, J., BINDER, R. L., ROBINSON, M. K., FINLAY, D. R., MILLS, K. J., BASCOM, C. C., KELLING, C. K., ISFORT, R. J., HAYCOCK, J. W., MACNEIL, S. & SMALLWOOD, R. H. 2013. Skin stem cell hypotheses and long term clone survival--explored using agent-based modelling. *Sci Rep,* 3**,** 1904.

LI, Y. & TOLLEFSBOL, T. O. 2011. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol,* 791**,** 11-21.

LI, Y. Y., HANNA, G. J., LAGA, A. C., HADDAD, R. I., LORCH, J. H. & HAMMERMAN, P. S. 2015. Genomic Analysis of Metastatic Cutaneous Squamous Cell Carcinoma. *Clinical Cancer Research,* 21**,** 1447-1456.

LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GOTZSCHE, P. C., IOANNIDIS, J. P. A., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J. & MOHER, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *Bmj-British Medical Journal,* 339.

LIN, J. Y. & FISHER, D. E. 2007. Melanocyte biology and skin pigmentation. *Nature,* 445**,** 843-50.

LIN, T. L., WU, C. Y., CHANG, Y. T., JUAN, C. K., CHEN, C. C., YU, S. H. & CHEN, Y. J. 2019. Risk of skin cancer in psoriasis patients receiving long-term narrowband ultraviolet phototherapy: Results from a Taiwanese population-based cohort study. *Photodermatol Photoimmunol Photomed,* 35**,** 164-171.

LOBRY, C., OH, P. & AIFANTIS, I. 2011. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. *J Exp Med,* 208**,** 1931-5.

LOGSDON, G. A., VOLLGER, M. R. & EICHLER, E. E. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics,* 21**,** 597-614.

LONG, G. V., MENZIES, A. M., NAGRIAL, A. M., HAYDU, L. E., HAMILTON, A. L., MANN, G. J., HUGHES, T. M., THOMPSON, J. F., SCOLYER, R. A. & KEFFORD, R. F. 2011. Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma. *J Clin Oncol,* 29**,** 1239-46.

LU, H., GIORDANO, F. & NING, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics,* 14**,** 265-279.

LUDMIR, E. B., STEPHENS, S. J., PALTA, M., WILLETT, C. G. & CZITO, B. G. 2015. Human papillomavirus tumor infection in esophageal squamous cell carcinoma. *J Gastrointest Oncol,* 6**,** 287-95.

LYNCH, M. D., LYNCH, C. N. S., CRAYTHORNE, E., LIAKATH-ALI, K., MALLIPEDDI, R., BARKER, J. N. & WATT, F. M. 2017. Spatial constraints govern competition of mutant clones in human epidermis. *Nature Communications,* 8.

MADAN, V., LEAR, J. T. & SZEIMIES, R. M. 2010. Non-melanoma skin cancer. *Lancet,* 375**,** 673-85.

MADRONICH, S., MCKENZIE, R. L., BJORN, L. O. & CALDWELL, M. M. 1998. Changes in biologically active ultraviolet radiation reaching the Earth's surface. *J Photochem Photobiol B,* 46**,** 5-19.

MAILMAN, M. D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., ZIYABARI, L., LEE, M., SHAO, Y., WANG, Z. Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J. & SHERRY, S. T. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet,* 39**,** 1181-6.

MAN, I., CROMBIE, I. K., DAWE, R. S., IBBOTSON, S. H. & FERGUSON, J. 2005. The photocarcinogenic risk of narrowband UVB (TL-01) phototherapy: early follow-up data. *Br J Dermatol,* 152**,** 755-7.

MANTERE, T., KERSTEN, S. & HOISCHEN, A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics,* 10.

MARKS, R., RENNIE, G. & SELWOOD, T. S. 1988. Malignant transformation of solar keratoses to squamous cell carcinoma. *Lancet,* 1**,** 795-7.

MARTINCORENA, I., RAINE, K. M., GERSTUNG, M., DAWSON, K. J., HAASE, K., VAN LOO, P., DAVIES, H., STRATTON, M. R. & CAMPBELL, P. J. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 171**,** 1029-1041 e21.

MARTINCORENA, I., ROSHAN, A., GERSTUNG, M., ELLIS, P., VAN LOO, P., MCLAREN, S., WEDGE, D. C., FULLAM, A., ALEXANDROV, L. B., TUBIO, J. M., STEBBINGS, L., MENZIES, A., WIDAA, S., STRATTON, M. R., JONES, P. H. & CAMPBELL, P. J. 2015. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science,* 348**,** 880-6.

MARTINEZ-JIMENEZ, F., MUINOS, F., SENTIS, I., DEU-PONS, J., REYES-SALAZAR, I., ARNEDO-PAC, C., MULARONI, L., PICH, O., BONET, J., KRANAS, H., GONZALEZ-PEREZ, A. & LOPEZ-BIGAS, N. 2020. A compendium of mutational cancer driver genes. *Nat Rev Cancer,* 20**,** 555-572.

MATSUI, T. & AMAGAI, M. 2015. Dissecting the formation, structure and barrier function of the stratum corneum. *Int Immunol,* 27**,** 269-80.

MAYAKONDA, A., LIN, D. C., ASSENOV, Y., PLASS, C. & KOEFFLER, H. P. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res,* 28**,** 1747-1756.

MCKERRELL, T., PARK, N., MORENO, T., GROVE, C. S., PONSTINGL, H., STEPHENS, J., CRAWLEY, C., CRAIG, J., SCOTT, M. A., HODKINSON, C., BAXTER, J., RAD, R., FORSYTH, D. R., QUAIL, M. A., ZEGGINI, E., OUWEHAND, W., VARELA, I. & VASSILIOU, G. S. 2015. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis.

MEERAN, S. M., PUNATHIL, T. & KATIYAR, S. K. 2008. IL-12 deficiency exacerbates inflammatory responses in UV-irradiated skin and skin tumors. *J Invest Dermatol,* 128**,** 2716-27.

MEIER, B., VOLKOVA, N. V., HONG, Y., SCHOFIELD, P., CAMPBELL, P. J., GERSTUNG, M. & GARTNER, A. 2018. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Res,* 28**,** 666-675.

MEIER, F., WILL, S., ELLWANGER, U., SCHLAGENHAUFF, B., SCHITTEK, B., RASSNER, G. & GARBE, C. 2002. Metastatic pathways and time courses in the orderly progression of cutaneous melanoma. *Br J Dermatol,* 147**,** 62-70.

MENTER, A., KORMAN, N. J., ELMETS, C. A., FELDMAN, S. R., GELFAND, J. M., GORDON, K. B., GOTTLIEB, A., KOO, J. Y., LEBWOHL, M., LIM, H. W., VAN VOORHEES, A. S., BEUTNER, K. R. & BHUSHAN, R. 2010. Guidelines of care for the management of psoriasis and psoriatic arthritis: Section 5. Guidelines of care for the treatment of psoriasis with phototherapy and photochemotherapy. *J Am Acad Dermatol,* 62**,** 114-35.

MEYERSON, M., GABRIEL, S. & GETZ, G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics,* 11**,** 685-696.

MICALI, G., LACARRUBBA, F., BONGU, A. & WEST D. 2001. The Skin barrier. *In:* WOODLY, D. T. F., R.K. (ed.) *The biology of the skin.* New York: The Parthoenon Publishing Group Inc. .

MITCHELL, D. L. 1988. The relative cytotoxicity of (6-4) photoproducts and cyclobutane dimers in mammalian cells. *Photochem Photobiol,* 48**,** 51-7.

MOLL, I., LANE, A. T., FRANKE, W. W. & MOLL, R. 1990. Intraepidermal formation of Merkel cells in xenografts of human fetal skin. *J Invest Dermatol,* 94**,** 359-64.

MOLL, U. M. & PETRENKO, O. 2003. The MDM2-p53 interaction. *Molecular Cancer Research,* 1**,** 1001-1008.

MONTEITH, J. L. & UNSWORTH, M. H. 1990. *Principles of environmental physics,* London ; New York

New York, E. Arnold ;

Distributed in the USA by Routledge, Chapman and Hall.

MOURET, S., BAUDOUIN, C., CHARVERON, M., FAVIER, A., CADET, J. & DOUKI, T. 2006. Cyclobutane pyrimidine dimers are predominant DNA lesions in whole human skin exposed to UVA radiation. *Proc Natl Acad Sci U S A,* 103**,** 13765-70.

MUELLER, S. A., GAUTHIER, M. A., ASHFORD, B., GUPTA, R., GAYEVSKIY, V., CH'NG, S., PALME, C. E., SHANNON, K., CLARK, J. R., RANSON, M. & COWLEY, M. J. 2019. Mutational Patterns in Metastatic Cutaneous Squamous Cell Carcinoma. *J Invest Dermatol,* 139**,** 1449-1458 e1.

MULARONI, L., SABARINATHAN, R., DEU-PONS, J., GONZALEZ-PEREZ, A. & LOPEZ-BIGAS, N. 2016. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol,* 17**,** 128.

MUNOZ-COUSELO, E., GARCIA, J. S., PEREZ-GARCIA, J. M., CEBRIAN, V. O. & CASTAN, J. C. 2015. Recent advances in the treatment of melanoma with BRAF and MEK inhibitors. *Ann Transl Med,* 3**,** 207.

MURAI, K., SKRUPSKELYTE, G., PIEDRAFITA, G., HALL, M., KOSTIOU, V., ONG, S. H., NAGY, T., CAGAN, A., GOULDING, D., KLEIN, A. M., HALL, B. A. & JONES, P. H. 2018. Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. *Cell Stem Cell,* 23**,** 687-699 e8.

MYERS, E., KHERADMAND, S. & MILLER, R. 2021. An Update on Narrowband Ultraviolet B Therapy for the Treatment of Skin Diseases.

NAKAGAWA, H. & FUJITA, M. 2018. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci,* 109**,** 513-522.

NAKAGAWA, H., WARDELL, C. P., FURUTA, M., TANIGUCHI, H. & FUJIMOTO, A. 2015. Cancer whole-genome sequencing: present and future. *Oncogene,* 34**,** 5943-50.

NARAYAN, R. 2009. Biomedical Materials. *Springer Science & Business Media***,** 376.

NARAYANAN, D. L., SALADI, R. N. & FOX, J. L. 2010. Ultraviolet radiation and skin cancer. *Int J Dermatol,* 49**,** 978-86.

NASTI, T. H. & TIMARES, L. 2015. MC1R, eumelanin and pheomelanin: their role in determining the susceptibility to skin cancer. *Photochem Photobiol,* 91**,** 188-200.

NELSON, M. A., EINSPAHR, J. G., ALBERTS, D. S., BALFOUR, C. A., WYMER, J. A., WELCH, K. L., SALASCHE, S. J., BANGERT, J. L., GROGAN, T. M. & BOZZO, P. O. 1994. Analysis of the p53 gene in human precancerous actinic keratosis lesions and squamous cell cancers. *Cancer Lett,* 85**,** 23-9.

NEWELL, F., WILMOTT, J. S., JOHANSSON, P. A., NONES, K., ADDALA, V., MUKHOPADHYAY, P., BROIT, N., AMATO, C. M., VAN GULICK, R., KAZAKOFF, S. H., PATCH, A. M., KOUFARIOTIS, L. T., LAKIS, V., LEONARD, C., WOOD, S., HOLMES, O., XU, Q., LEWIS, K., MEDINA, T., GONZALEZ, R., SAW, R. P. M., SPILLANE, A. J., STRETCH, J. R., RAWSON, R. V., FERGUSON, P. M., DODDS, T. J., THOMPSON, J. F., LONG, G. V., LEVESQUE, M. P., ROBINSON, W. A., PEARSON, J. V., MANN, G. J., SCOLYER, R. A., WADDELL, N. & HAYWARD, N. K. 2020. Whole-genome sequencing of acral melanoma reveals genomic complexity and diversity. *Nat Commun,* 11**,** 5259.

NGUYEN, T. H. 2004. Mechanisms of metastasis. *Clin Dermatol,* 22**,** 209-16.

NICE 2016. Sunlight exposure: risks and benefits. *NICE guideline,* NG34**,** 6-7.

NICOLAS, M., WOLFER, A., RAJ, K., KUMMER, J. A., MILL, P., VAN NOORT, M., HUI, C. C., CLEVERS, H., DOTTO, G. P. & RADTKE, F. 2003. Notch1 functions as a tumor suppressor in mouse skin. *Nat Genet,* 33**,** 416-21.

NIJSTEN, T. E. & STERN, R. S. 2003. The increased risk of skin cancer is persistent after discontinuation of psoralen+ultraviolet A: a cohort study. *J Invest Dermatol,* 121**,** 252-8.

NISHIMURA, K., IKEHATA, H., DOUKI, T., CADET, J., SUGIURA, S. & MORI, T. 2021. Seasonal Differences in the UVA/UVB Ratio of Natural Sunlight Influence the Efficiency of the Photoisomerization of (6-4) Photoproducts into their Dewar Valence Isomers. *Photochem Photobiol,* 97**,** 582-588.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

NURK, S., KOREN, S., RHIE, A., RAUTIAINEN, M., BZIKADZE, A. V., MIKHEENKO, A., VOLLGER, M. R., ALTEMOSE, N., URALSKY, L., GERSHMAN, A., AGANEZOV, S., HOYT, S. J., DIEKHANS, M., LOGSDON, G. A., ALONGE, M., ANTONARAKIS, S. E., BORCHERS, M., BOUFFARD, G. G., BROOKS, S. Y., CALDAS, G. V., CHEN, N.-C., CHENG, H., CHIN, C.-S., CHOW, W., DE LIMA, L. G., DISHUCK, P. C., DURBIN, R., DVORKINA, T., FIDDES, I. T., FORMENTI, G., FULTON, R. S., FUNGTAMMASAN, A., GARRISON, E., GRADY, P. G. S., GRAVES-LINDSAY, T. A., HALL, I. M., HANSEN, N. F., HARTLEY, G. A., HAUKNESS, M., HOWE, K., HUNKAPILLER, M. W., JAIN, C., JAIN, M., JARVIS, E. D., KERPEDJIEV, P., KIRSCHE, M., KOLMOGOROV, M., KORLACH, J., KREMITZKI, M., LI, H., MADURO, V. V., MARSCHALL, T., MCCARTNEY, A. M., MCDANIEL, J., MILLER, D. E., MULLIKIN, J. C., MYERS, E. W., OLSON, N. D., PATEN, B., PELUSO, P., PEVZNER, P. A., PORUBSKY, D., POTAPOVA, T., ROGAEV, E. I., ROSENFELD, J. A., SALZBERG, S. L., SCHNEIDER, V. A., SEDLAZECK, F. J., SHAFIN, K., SHEW, C. J., SHUMATE, A., SIMS, Y., SMIT, A. F. A., SOTO, D. C., SOVIĆ, I., STORER, J. M., STREETS, A., SULLIVAN, B. A., THIBAUD-NISSEN, F., TORRANCE, J., WAGNER, J., WALENZ, B. P., WENGER, A., WOOD, J. M. D., XIAO, C., YAN, S. M., YOUNG, A. C., ZARATE, S., SURTI, U., MCCOY, R. C., DENNIS, M. Y., ALEXANDROV, I. A., GERTON, J. L., O’NEILL, R. J., TIMP, W., ZOOK, J. M., SCHATZ, M. C., EICHLER, E. E., MIGA, K. H. & PHILLIPPY, A. M. 2022. The complete sequence of a human genome. *Science,* 376**,** 44-53.

O'DONNELL, J. S., TENG, M. W. L. & SMYTH, M. J. 2019. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nat Rev Clin Oncol,* 16**,** 151-167.

OLIVERIA, S. A., SARAIYA, M., GELLER, A. C., HENEGHAN, M. K. & JORGENSEN, C. 2006. Sun exposure and risk of melanoma. *Arch Dis Child,* 91**,** 131-8.

OLIVIER, M., HOLLSTEIN, M. & HAINAUT, P. 2010. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology,* 2.

OSHIMORI, N., ORISTIAN, D. & FUCHS, E. 2015. TGF-beta promotes heterogeneity and drug resistance in squamous cell carcinoma. *Cell,* 160**,** 963-976.

PAGET, S. 1989. The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev,* 8**,** 98-101.

PAMPENA, R., KYRGIDIS, A., LALLAS, A., MOSCARELLA, E., ARGENZIANO, G. & LONGO, C. 2017. A meta-analysis of nevus-associated melanoma: Prevalence and practical implications. *J Am Acad Dermatol,* 77**,** 938-945 e4.

PAN, B., REN, L., ONUCHIC, V., GUAN, M., KUSKO, R., BRUINSMA, S., TRIGG, L., SCHERER, A., NING, B., ZHANG, C., GLIDEWELL-KENNEY, C., XIAO, C., DONALDSON, E., SEDLAZECK, F. J., SCHROTH, G., YAVAS, G., GRUNENWALD, H., CHEN, H., MEINHOLZ, H., MEEHAN, J., WANG, J., YANG, J., FOOX, J., SHANG, J., MICLAUS, K., DONG, L., SHI, L., MOHIYUDDIN, M., PIROOZNIA, M., GONG, P., GOLSHANI, R., WOLFINGER, R., LABABIDI, S., SAHRAEIAN, S. M. E., SHERRY, S., HAN, T., CHEN, T., SHI, T., HOU, W., GE, W., ZOU, W., GUO, W., BAO, W., XIAO, W., FAN, X., GONDO, Y., YU, Y., ZHAO, Y., SU, Z., LIU, Z., TONG, W., XIAO, W., ZOOK, J. M., ZHENG, Y. & HONG, H. 2022. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biology,* 23**,** 2.

PARK, H. Y., KOSMADAKI, M., YAAR, M. & GILCHREST, B. A. 2009. Cellular mechanisms regulating human melanogenesis. *Cell Mol Life Sci,* 66**,** 1493-506.

PARRISH, J. A., JAENICKE, K. F. & ANDERSON, R. R. 1982. Erythema and melanogenesis action spectra of normal human skin. *Photochem Photobiol,* 36**,** 187-91.

PATHIRANA, D., ORMEROD, A. D., SAIAG, P., SMITH, C., SPULS, P. I., NAST, A., BARKER, J., BOS, J. D., BURMESTER, G. R., CHIMENTI, S., DUBERTRET, L., EBERLEIN, B., ERDMANN, R., FERGUSON, J., GIROLOMONI, G., GISONDI, P., GIUNTA, A., GRIFFITHS, C., HONIGSMANN, H., HUSSAIN, M., JOBLING, R., KARVONEN, S. L., KEMENY, L., KOPP, I., LEONARDI, C., MACCARONE, M., MENTER, A., MROWIETZ, U., NALDI, L., NIJSTEN, T., ORTONNE, J. P., ORZECHOWSKI, H. D., RANTANEN, T., REICH, K., REYTAN, N., RICHARDS, H., THIO, H. B., VAN DE KERKHOF, P. & RZANY, B. 2009. European S3-guidelines on the systemic treatment of psoriasis vulgaris. *J Eur Acad Dermatol Venereol,* 23 Suppl 2**,** 1-70.

PATRICK, M. H. 1977. Studies on thymine-derived UV photoproducts in DNA--I. Formation and biological role of pyrimidine adducts in DNA. *Photochem Photobiol,* 25**,** 357-72.

PAWAR, S. A., SARKAR, T. R., BALAMURUGAN, K., SHARAN, S., WANG, J., ZHANG, Y., DOWDY, S. F., HUANG, A. M. & STERNECK, E. 2010. C/EBP{delta} targets cyclin D1 for proteasome-mediated degradation via induction of CDC27/APC3 expression. *Proc Natl Acad Sci U S A,* 107**,** 9210-5.

PENTA, D., SOMASHEKAR, B. S. & MEERAN, S. M. 2018. Epigenetics of skin cancer: Interventions by selected bioactive phytochemicals. *Photodermatol Photoimmunol Photomed,* 34**,** 42-49.

PERIS, K., FARGNOLI, M. C., GARBE, C., KAUFMANN, R., BASTHOLT, L., SEGUIN, N. B., BATAILLE, V., MARMOL, V. D., DUMMER, R., HARWOOD, C. A., HAUSCHILD, A., HOLLER, C., HAEDERSDAL, M., MALVEHY, J., MIDDLETON, M. R., MORTON, C. A., NAGORE, E., STRATIGOS, A. J., SZEIMIES, R. M., TAGLIAFERRI, L., TRAKATELLI, M., ZALAUDEK, I., EGGERMONT, A., GROB, J. J., EUROPEAN DERMATOLOGY FORUM, T. E. A. O. D.-O., THE EUROPEAN ORGANIZATION FOR, R. & TREATMENT OF, C. 2019. Diagnosis and treatment of basal cell carcinoma: European consensus-based interdisciplinary guidelines. *Eur J Cancer,* 118**,** 10-34.

PETIT, A., RAGU, C., SOLER, G., OTTOLENGHI, C., SCHLUTH, C., RADFORD-WEISS, I., SCHNEIDER-MAUNOURY, S., CALLEBAUT, I., DASTUGUE, N., DRABKIN, H. A., BERNARD, O. A., ROMANA, S. & PENARD-LACRONIQUE, V. 2012. Functional analysis of the NUP98-CCDC28A fusion protein. *Haematologica,* 97**,** 379-87.

PICKERING, C. R., ZHOU, J. H., LEE, J. J., DRUMMOND, J. A., PENG, S. A., SAADE, R. E., TSAI, K. Y., CURRY, J. L., TETZLAFF, M. T., LAI, S. Y., YU, J., MUZNY, D. M., DODDAPANENI, H., SHINBROT, E., COVINGTON, K. R., ZHANG, J. H., SETH, S., CAULIN, C., CLAYMAN, G. L., EL-NAGGAR, A. K., GIBBS, R. A., WEBER, R. S., MYERS, J. N., WHEELER, D. A. & FREDERICK, M. J. 2014. Mutational Landscape of Aggressive Cutaneous Squamous Cell Carcinoma. *Clinical Cancer Research,* 20**,** 6582-6592.

PLEASANCE, E. D., CHEETHAM, R. K., STEPHENS, P. J., MCBRIDE, D. J., HUMPHRAY, S. J., GREENMAN, C. D., VARELA, I., LIN, M. L., ORDONEZ, G. R., BIGNELL, G. R., YE, K., ALIPAZ, J., BAUER, M. J., BEARE, D., BUTLER, A., CARTER, R. J., CHEN, L., COX, A. J., EDKINS, S., KOKKO-GONZALES, P. I., GORMLEY, N. A., GROCOCK, R. J., HAUDENSCHILD, C. D., HIMS, M. M., JAMES, T., JIA, M., KINGSBURY, Z., LEROY, C., MARSHALL, J., MENZIES, A., MUDIE, L. J., NING, Z., ROYCE, T., SCHULZ-TRIEGLAFF, O. B., SPIRIDOU, A., STEBBINGS, L. A., SZAJKOWSKI, L., TEAGUE, J., WILLIAMSON, D., CHIN, L., ROSS, M. T., CAMPBELL, P. J., BENTLEY, D. R., FUTREAL, P. A. & STRATTON, M. R. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature,* 463**,** 191-6.

PLEASANCE, E. D., STEPHENS, P. J., O'MEARA, S., MCBRIDE, D. J., MEYNERT, A., JONES, D., LIN, M. L., BEARE, D., LAU, K. W., GREENMAN, C., VARELA, I., NIK-ZAINAL, S., DAVIES, H. R., ORDONEZ, G. R., MUDIE, L. J., LATIMER, C., EDKINS, S., STEBBINGS, L., CHEN, L., JIA, M., LEROY, C., MARSHALL, J., MENZIES, A., BUTLER, A., TEAGUE, J. W., MANGION, J., SUN, Y. A., MCLAUGHLIN, S. F., PECKHAM, H. E., TSUNG, E. F., COSTA, G. L., LEE, C. C., MINNA, J. D., GAZDAR, A., BIRNEY, E., RHODES, M. D., MCKERNAN, K. J., STRATTON, M. R., FUTREAL, P. A. & CAMPBELL, P. J. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature,* 463**,** 184-90.

POLLA, D. L., FARAZI FARD, M. A., TABATABAEI, Z., HABIBZADEH, P., LEVCHENKO, O. A., NIKUEI, P., MAKRYTHANASIS, P., HUSSAIN, M., VON HARDENBERG, S., ZEINALI, S., FALLAH, M. S., SCHUURS-HOEIJMAKERS, J. H. M., SHAHZAD, M., FATIMA, F., FATIMA, N., KAAT, L. D., BRUGGENWIRTH, H. T., FLEMING, L. R., CONDIE, J., PLOSKI, R., POLLAK, A., PILCH, J., DEMINA, N. A., CHUKHROVA, A. L., SERGEEVA, V. S., VENSELAAR, H., MASRI, A. T., HAMAMY, H., SANTONI, F. A., LINDA, K., AHMED, Z. M., NADIF KASRI, N., DE BROUWER, A. P. M., BERGMANN, A. K., HETHEY, S., YAVARIAN, M., ANSAR, M., RIAZUDDIN, S., RIAZUDDIN, S., SILAWI, M., RUGGERI, G., PIROZZI, F., EFTEKHAR, E., TAGHIPOUR SHESHDEH, A., BAHRAMJAHAN, S., MIRZAA, G. M., LAVROV, A. V., ANTONARAKIS, S. E., FAGHIHI, M. A. & VAN BOKHOVEN, H. 2021. Biallelic variants in TMEM222 cause a new autosomal recessive neurodevelopmental disorder. *Genet Med,* 23**,** 1246-1254.

POURREYRON, C., COX, G., MAO, X., VOLZ, A., BAKSH, N., WONG, T., FASSIHI, H., ARITA, K., O'TOOLE, E. A., OCAMPO-CANDIANI, J., CHEN, M., HART, I. R., BRUCKNER-TUDERMAN, L., SALAS-ALANIS, J. C., MCGRATH, J. A., LEIGH, I. M. & SOUTH, A. P. 2007. Patients with recessive dystrophic epidermolysis bullosa develop squamous-cell carcinoma regardless of type VII collagen expression. *J Invest Dermatol,* 127**,** 2438-44.

PURDIE, K. J., HARWOOD, C. A., GULATI, A., CHAPLIN, T., LAMBERT, S. R., CERIO, R., KELLY, G. P., CAZIER, J. B., YOUNG, B. D., LEIGH, I. M. & PROBY, C. M. 2009. Single nucleotide polymorphism array analysis defines a specific genetic fingerprint for well-differentiated cutaneous SCCs. *J Invest Dermatol,* 129**,** 1562-8.

QIAN, B. Z. & POLLARD, J. W. 2010. Macrophage diversity enhances tumor progression and metastasis. *Cell,* 141**,** 39-51.

QIU, L., TAN, X., LIN, J., LIU, R. Y., CHEN, S., GENG, R., WU, J. & HUANG, W. 2017. CDC27 Induces Metastasis and Invasion in Colorectal Cancer via the Promotion of Epithelial-To-Mesenchymal Transition. *J Cancer,* 8**,** 2626-2635.

QIU, L., WU, J., PAN, C., TAN, X., LIN, J., LIU, R., CHEN, S., GENG, R. & HUANG, W. 2016. Downregulation of CDC27 inhibits the proliferation of colorectal cancer cells via the accumulation of p21Cip1/Waf1. *Cell Death Dis,* 7**,** e2074.

QUAEDVLIEG, P. J., TIRSI, E., THISSEN, M. R. & KREKELS, G. A. 2006. Actinic keratosis: how to differentiate the good from the bad ones? *Eur J Dermatol,* 16**,** 335-9.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *Bmc Genomics,* 13.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RAJKUMAR, T., GOPAL, G., SELVALUXMI, G. & RAJALEKSHMY, K. R. 2005. CDC27 protein is involved in radiation response in squamous cell cervix carcinoma. *Indian J Biochem Biophys,* 42**,** 271-8.

RAMOS, A. H., LICHTENSTEIN, L., GUPTA, M., LAWRENCE, M. S., PUGH, T. J., SAKSENA, G., MEYERSON, M. & GETZ, G. 2015. Oncotator: cancer variant annotation tool. *Hum Mutat,* 36**,** E2423-9.

RAONE, B., PATRIZI, A., GURIOLI, C., GAZZOLA, A. & RAVAIOLI, G. M. 2018. Cutaneous carcinogenic risk evaluation in 375 patients treated with narrowband-UVB phototherapy: A 15-year experience from our Institute. *Photodermatol Photoimmunol Photomed,* 34**,** 302-306.

RASTOGI, R. P., RICHA, KUMAR, A., TYAGI, M. B. & SINHA, R. P. 2010. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *J Nucleic Acids,* 2010**,** 592980.

REBEL, H., KRAM, N., WESTERMAN, A., BANUS, S., VAN KRANEN, H. J. & DE GRUIJL, F. R. 2005. Relationship between UV-induced mutant p53 patches and skin tumours, analysed by mutation spectra and by induction kinetics in various DNA-repair-deficient mice. *Carcinogenesis,* 26**,** 2123-30.

REES, J. L. 2004. The genetics of sun sensitivity in humans. *Am J Hum Genet,* 75**,** 739-51.

REES, J. L. & HARDING, R. M. 2012. Understanding the evolution of human pigmentation: recent contributions from population genetics. *J Invest Dermatol,* 132**,** 846-53.

REHMAN, I., QUINN, A. G., HEALY, E. & REES, J. L. 1994. High frequency of loss of heterozygosity in actinic keratoses, a usually benign disease. *Lancet,* 344**,** 788-9.

REHMAN, I., TAKATA, M., WU, Y. Y. & REES, J. L. 1996. Genetic change in actinic keratoses. *Oncogene,* 12**,** 2483-90.

REN, L., MATSUDA, T., DENG, B., KIYOTANI, K., KATO, T., PARK, J.-H., SEIWERT, T. Y., VOKES, E. E., AGRAWAL, N. & NAKAMURA, Y. 2017. Similarity and difference in tumor-infiltrating lymphocytes in original tumor tissues and those of in vitro expanded populations in head and neck cancer. *Oncotarget; Vol 9, No 3*.

REN, Z. P., AHMADIAN, A., PONTEN, F., NISTER, M., BERG, C., LUNDEBERG, J., UHLEN, M. & PONTEN, J. 1997. Benign clonal keratinocyte patches with p53 mutations show no genetic link to synchronous squamous cell precancer or cancer in human skin. *Am J Pathol,* 150**,** 1791-803.

RENDON, A. & SCHAKEL, K. 2019. Psoriasis Pathogenesis and Treatment. *Int J Mol Sci,* 20.

REPANA, D., NULSEN, J., DRESSLER, L., BORTOLOMEAZZI, M., VENKATA, S. K., TOURNA, A., YAKOVLEVA, A., PALMIERI, T. & CICCARELLI, F. D. 2019. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol,* 20**,** 1.

RHEINBAY, E., NIELSEN, M. M., ABASCAL, F., WALA, J. A., SHAPIRA, O., TIAO, G., HORNSHOJ, H., HESS, J. M., JUUL, R. I., LIN, Z., FEUERBACH, L., SABARINATHAN, R., MADSEN, T., KIM, J., MULARONI, L., SHUAI, S., LANZOS, A., HERRMANN, C., MARUVKA, Y. E., SHEN, C., AMIN, S. B., BANDOPADHAYAY, P., BERTL, J., BOROEVICH, K. A., BUSANOVICH, J., CARLEVARO-FITA, J., CHAKRAVARTY, D., CHAN, C. W. Y., CRAFT, D., DHINGRA, P., DIAMANTI, K., FONSECA, N. A., GONZALEZ-PEREZ, A., GUO, Q., HAMILTON, M. P., HARADHVALA, N. J., HONG, C., ISAEV, K., JOHNSON, T. A., JUUL, M., KAHLES, A., KAHRAMAN, A., KIM, Y., KOMOROWSKI, J., KUMAR, K., KUMAR, S., LEE, D., LEHMANN, K. V., LI, Y., LIU, E. M., LOCHOVSKY, L., PARK, K., PICH, O., ROBERTS, N. D., SAKSENA, G., SCHUMACHER, S. E., SIDIROPOULOS, N., SIEVERLING, L., SINNOTT-ARMSTRONG, N., STEWART, C., TAMBORERO, D., TUBIO, J. M. C., UMER, H. M., UUSKULA-REIMAND, L., WADELIUS, C., WADI, L., YAO, X., ZHANG, C. Z., ZHANG, J., HABER, J. E., HOBOLTH, A., IMIELINSKI, M., KELLIS, M., LAWRENCE, M. S., VON MERING, C., NAKAGAWA, H., RAPHAEL, B. J., RUBIN, M. A., SANDER, C., STEIN, L. D., STUART, J. M., TSUNODA, T., WHEELER, D. A., JOHNSON, R., REIMAND, J., GERSTEIN, M., KHURANA, E., CAMPBELL, P. J., LOPEZ-BIGAS, N., DRIVERS, P., FUNCTIONAL INTERPRETATION WORKING, G., GROUP, P. S. V. W., WEISCHENFELDT, J., BEROUKHIM, R., MARTINCORENA, I., PEDERSEN, J. S., GETZ, G. & CONSORTIUM, P. 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature,* 578**,** 102-111.

RIBAS, A. 2012. Tumor immunotherapy directed at PD-1. *N Engl J Med,* 366**,** 2517-9.

ROADMAP EPIGENOMICS, C., KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J., AMIN, V., WHITAKER, J. W., SCHULTZ, M. D., WARD, L. D., SARKAR, A., QUON, G., SANDSTROM, R. S., EATON, M. L., WU, Y. C., PFENNING, A. R., WANG, X., CLAUSSNITZER, M., LIU, Y., COARFA, C., HARRIS, R. A., SHORESH, N., EPSTEIN, C. B., GJONESKA, E., LEUNG, D., XIE, W., HAWKINS, R. D., LISTER, R., HONG, C., GASCARD, P., MUNGALL, A. J., MOORE, R., CHUAH, E., TAM, A., CANFIELD, T. K., HANSEN, R. S., KAUL, R., SABO, P. J., BANSAL, M. S., CARLES, A., DIXON, J. R., FARH, K. H., FEIZI, S., KARLIC, R., KIM, A. R., KULKARNI, A., LI, D., LOWDON, R., ELLIOTT, G., MERCER, T. R., NEPH, S. J., ONUCHIC, V., POLAK, P., RAJAGOPAL, N., RAY, P., SALLARI, R. C., SIEBENTHALL, K. T., SINNOTT-ARMSTRONG, N. A., STEVENS, M., THURMAN, R. E., WU, J., ZHANG, B., ZHOU, X., BEAUDET, A. E., BOYER, L. A., DE JAGER, P. L., FARNHAM, P. J., FISHER, S. J., HAUSSLER, D., JONES, S. J., LI, W., MARRA, M. A., MCMANUS, M. T., SUNYAEV, S., THOMSON, J. A., TLSTY, T. D., TSAI, L. H., WANG, W., WATERLAND, R. A., ZHANG, M. Q., CHADWICK, L. H., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A., MILOSAVLJEVIC, A., REN, B., STAMATOYANNOPOULOS, J. A., WANG, T. & KELLIS, M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-30.

ROBERTS, S. A., LAWRENCE, M. S., KLIMCZAK, L. J., GRIMM, S. A., FARGO, D., STOJANOV, P., KIEZUN, A., KRYUKOV, G. V., CARTER, S. L., SAKSENA, G., HARRIS, S., SHAH, R. R., RESNICK, M. A., GETZ, G. & GORDENIN, D. A. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet,* 45**,** 970-6.

ROBINSON, S., DIXON, S., AUGUST, S., DIFFEY, B., WAKAMATSU, K., ITO, S., FRIEDMANN, P. S. & HEALY, E. 2010. Protection against UVR involves MC1R-mediated non-pigmentary and pigmentary mechanisms in vivo. *J Invest Dermatol,* 130**,** 1904-13.

ROBINSON, S. J. & HEALY, E. 2002. Human melanocortin 1 receptor (MC1R) gene variants alter melanoma cell growth and adhesion to extracellular matrix. *Oncogene,* 21**,** 8037-46.

ROBLES-ESPINOZA, C. D., ROBERTS, N. D., CHEN, S., LEACY, F. P., ALEXANDROV, L. B., PORNPUTTAPONG, N., HALABAN, R., KRAUTHAMMER, M., CUI, R., TIMOTHY BISHOP, D. & ADAMS, D. J. 2016. Germline MC1R status influences somatic mutation burden in melanoma. *Nat Commun,* 7**,** 12064.

ROCHETTE, P. J., THERRIEN, J. P., DROUIN, R., PERDIZ, D., BASTIEN, N., DROBETSKY, E. A. & SAGE, E. 2003. UVA-induced cyclobutane pyrimidine dimers form predominantly at thymine-thymine dipyrimidines and correlate with the mutation spectrum in rodent cells. *Nucleic Acids Res,* 31**,** 2786-94.

RODENBECK, D. L., SILVERBERG, J. I. & SILVERBERG, N. B. 2016. Phototherapy for atopic dermatitis. *Clin Dermatol,* 34**,** 607-13.

RODRIGUEZ-PAREDES, M., BORMANN, F., RADDATZ, G., GUTEKUNST, J., LUCENA-PORCEL, C., KOHLER, F., WURZER, E., SCHMIDT, K., GALLINAT, S., WENCK, H., ROWERT-HUBER, J., DENISOVA, E., FEUERBACH, L., PARK, J., BRORS, B., HERPEL, E., NINDL, I., HOFMANN, T. G., WINNEFELD, M. & LYKO, F. 2018. Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nat Commun,* 9**,** 577.

ROMAGOSA, C., SIMONETTI, S., LOPEZ-VICENTE, L., MAZO, A., LLEONART, M. E., CASTELLVI, J. & RAMON Y CAJAL, S. 2011. p16(Ink4a) overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. *Oncogene,* 30**,** 2087-97.

RYBCHYN, M. S., DE SILVA, W. G. M., SEQUEIRA, V. B., MCCARTHY, B. Y., DILLEY, A. V., DIXON, K. M., HALLIDAY, G. M. & MASON, R. S. 2018. Enhanced Repair of UV-Induced DNA Damage by 1,25-Dihydroxyvitamin D3 in Skin Is Linked to Pathways that Control Cellular Energy. *J Invest Dermatol,* 138**,** 1146-1156.

SAGE, E. 1993. Distribution and repair of photolesions in DNA: genetic consequences and the role of sequence context. *Photochem Photobiol,* 57**,** 163-74.

SAKAMOTO, Y., SEREEWATTANAWOOT, S. & SUZUKI, A. 2020. A new era of long-read sequencing for cancer genomics. *Journal of Human Genetics,* 65**,** 3-10.

SAKAMOTO, Y., ZAHA, S., SUZUKI, Y., SEKI, M. & SUZUKI, A. 2021. Application of long-read sequencing to the detection of structural variants in human cancer genomes. *Computational and Structural Biotechnology Journal,* 19**,** 4207-4216.

SALGADO, R., TOLL, A., ALAMEDA, F., BARO, T., MARTIN-EZQUERRA, G., SANMARTIN, O., MARTORELL-CALATAYUD, A., SALIDO, M., ALMENAR, S., SOLE, F., PUJOL, R. M. & ESPINET, B. 2010. CKS1B amplification is a frequent event in cutaneous squamous cell carcinoma with aggressive clinical behaviour. *Genes Chromosomes Cancer,* 49**,** 1054-61.

SANDRU, A., VOINEA, S., PANAITESCU, E. & BLIDARU, A. 2014. Survival rates of patients with metastatic malignant melanoma. *J Med Life,* 7**,** 572-6.

SAVOYE, I., OLSEN, C. M., WHITEMAN, D. C., BIJON, A., WALD, L., DARTOIS, L., CLAVEL-CHAPELON, F., BOUTRON-RUAULT, M. C. & KVASKOFF, M. 2018. Patterns of Ultraviolet Radiation Exposure and Skin Cancer Risk: the E3N-SunExp Study. *J Epidemiol,* 28**,** 27-33.

SCHADENDORF, D., VAN AKKOOI, A. C. J., BERKING, C., GRIEWANK, K. G., GUTZMER, R., HAUSCHILD, A., STANG, A., ROESCH, A. & UGUREL, S. 2018. Melanoma. *Lancet,* 392**,** 971-984.

SCHATZ, M. C., DELCHER, A. L. & SALZBERG, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res,* 20**,** 1165-73.

SCHUSTER-BOCKLER, B. & LEHNER, B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature,* 488**,** 504-7.

SCHWAEDERLE, M., ELKIN, S. K., TOMSON, B. N., CARTER, J. L. & KURZROCK, R. 2015. Squamousness: Next-generation sequencing reveals shared molecular features across squamous tumor types. *Cell Cycle,* 14**,** 2355-2361.

SCHWARZ, T. 2005. Mechanisms of UV-induced immunosuppression. *Keio J Med,* 54**,** 165-71.

SCHWARZ, T. 2008. 25 years of UV-induced immunosuppression mediated by T cells-from disregarded T suppressor cells to highly respected regulatory T cells. *Photochem Photobiol,* 84**,** 10-8.

SEKULIC, A., KIM, S. Y., HOSTETTER, G., SAVAGE, S., EINSPAHR, J. G., PRASAD, A., SAGERMAN, P., CURIEL-LEWANDROWSKI, C., KROUSE, R., BOWDEN, G. T., WARNEKE, J., ALBERTS, D. S., PITTELKOW, M. R., DICAUDO, D., NICKOLOFF, B. J., TRENT, J. M. & BITTNER, M. 2010. Loss of inositol polyphosphate 5-phosphatase is an early event in development of cutaneous squamous cell carcinoma. *Cancer Prev Res (Phila),* 3**,** 1277-83.

SERRANO, M. 1997. The tumor suppressor protein p16INK4a. *Exp Cell Res,* 237**,** 7-13.

SHAIN, A. H., GARRIDO, M., BOTTON, T., TALEVICH, E., YEH, I., SANBORN, J. Z., CHUNG, J., WANG, N. J., KAKAVAND, H., MANN, G. J., THOMPSON, J. F., WIESNER, T., ROY, R., OLSHEN, A. B., GAGNON, A., GRAY, J. W., HUH, N., HUR, J. S., BUSAM, K. J., SCOLYER, R. A., CHO, R. J., MURALI, R. & BASTIAN, B. C. 2015a. Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway. *Nat Genet,* 47**,** 1194-9.

SHAIN, A. H., YEH, I., KOVALYSHYN, I., SRIHARAN, A., TALEVICH, E., GAGNON, A., DUMMER, R., NORTH, J., PINCUS, L., RUBEN, B., RICKABY, W., D'ARRIGO, C., ROBSON, A. & BASTIAN, B. C. 2015b. The Genetic Evolution of Melanoma from Precursor Lesions. *N Engl J Med,* 373**,** 1926-36.

SHAND, M. A.-O., SOTO, J., LICHTENSTEIN, L., BENJAMIN, D., FARJOUN, Y. A.-O., BRODY, Y., MARUVKA, Y., BLAINEY, P. A.-O. & BANKS, E. 2020. A validated lineage-derived somatic truth data set enables benchmarking in cancer genome analysis.

SHAPANIS, A., LAI, C., SMITH, S., COLTART, G., SOMMERLAD, M., SCHOFIELD, J., PARKINSON, E., SKIPP, P. & HEALY, E. 2020a. Identification of proteins associated with development of metastasis from cutaneous squamous cell carcinomas (cSCCs) via proteomic analysis of primary cSCCs. *Br J Dermatol*.

SHAPANIS, A., LAI, C., SOMMERLAD, M., PARKINSON, E., HEALY, E. & SKIPP, P. 2020b. Proteomic Profiling of Archived Tissue of Primary Melanoma Identifies Proteins Associated with Metastasis. *Int J Mol Sci,* 21.

SIMPSON, C. L., PATEL, D. M. & GREEN, K. J. 2011. Deconstructing the skin: cytoarchitectural determinants of epidermal morphogenesis. *Nat Rev Mol Cell Biol,* 12**,** 565-80.

SINHA, R. P. & HADER, D. P. 2002. UV-induced DNA damage and repair: a review. *Photochem Photobiol Sci,* 1**,** 225-36.

SLATKO, B. E., GARDNER, A. F. & AUSUBEL, F. M. 2018. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol,* 122**,** e59.

SOTTORIVA, A., BARNES, C. P. & GRAHAM, T. A. 2017. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochim Biophys Acta,* 1867**,** 95-100.

SOTTORIVA, A., KANG, H., MA, Z., GRAHAM, T. A., SALOMON, M. P., ZHAO, J., MARJORAM, P., SIEGMUND, K., PRESS, M. F., SHIBATA, D. & CURTIS, C. 2015. A Big Bang model of human colorectal tumor growth. *Nat Genet,* 47**,** 209-16.

SOUTH, A. P., PURDIE, K. J., WATT, S. A., HALDENBY, S., DEN BREEMS, N. Y., DIMON, M., ARRON, S. T., KLUK, M. J., ASTER, J. C., MCHUGH, A., XUE, D. J., DAYAL, J. H. S., ROBINSON, K. S., RIZVI, S. M. H., PROBY, C. M., HARWOOD, C. A. & LEIGH, I. M. 2014. NOTCH1 Mutations Occur Early during Cutaneous Squamous Cell Carcinogenesis. *Journal of Investigative Dermatology,* 134**,** 2630-2638.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet,* 41**,** 393-5.

STARK, M. S., TAN, J. M., TOM, L., JAGIRDAR, K., LAMBIE, D., SCHAIDER, H., SOYER, H. P. & STURM, R. A. 2018. Whole-Exome Sequencing of Acquired Nevi Identifies Mechanisms for Development and Maintenance of Benign Neoplasms. *J Invest Dermatol,* 138**,** 1636-1644.

STEPHEN, J. K., DIVINE, G., CHEN, K. M., CHITALE, D., HAVARD, S. & WORSHAM, M. J. 2013. Significance of p16 in Site-specific HPV Positive and HPV Negative Head and Neck Squamous Cell Carcinoma. *Cancer Clin Oncol,* 2**,** 51-61.

STERN, R. S. & STUDY, P. F. U. 2001. The risk of melanoma in association with long-term exposure to PUVA. *J Am Acad Dermatol,* 44**,** 755-61.

STRACHAN, T., READ, A. 2011. *Human Molecular Genetics 4*, Garland Science/ Taylor & Francis Group.

STRATIGOS, A., GARBE, C., LEBBE, C., MALVEHY, J., DEL MARMOL, V., PEHAMBERGER, H., PERIS, K., BECKER, J. C., ZALAUDEK, I., SAIAG, P., MIDDLETON, M. R., BASTHOLT, L., TESTORI, A., GROB, J. J., EUROPEAN DERMATOLOGY, F., EUROPEAN ASSOCIATION OF, D.-O., EUROPEAN ORGANIZATION FOR, R. & TREATMENT OF, C. 2015. Diagnosis and treatment of invasive squamous cell carcinoma of the skin: European consensus-based interdisciplinary guideline. *Eur J Cancer,* 51**,** 1989-2007.

STRATTON, M. R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science,* 331**,** 1553-8.

STRATTON, M. R., CAMPBELL, P. J. & FUTREAL, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-24.

SU, F., VIROS, A., MILAGRE, C., TRUNZER, K., BOLLAG, G., SPLEISS, O., REIS-FILHO, J. S., KONG, X., KOYA, R. C., FLAHERTY, K. T., CHAPMAN, P. B., KIM, M. J., HAYWARD, R., MARTIN, M., YANG, H., WANG, Q., HILTON, H., HANG, J. S., NOE, J., LAMBROS, M., GEYER, F., DHOMEN, N., NICULESCU-DUVAZ, I., ZAMBON, A., NICULESCU-DUVAZ, D., PREECE, N., ROBERT, L., OTTE, N. J., MOK, S., KEE, D., MA, Y., ZHANG, C., HABETS, G., BURTON, E. A., WONG, B., NGUYEN, H., KOCKX, M., ANDRIES, L., LESTINI, B., NOLOP, K. B., LEE, R. J., JOE, A. K., TROY, J. L., GONZALEZ, R., HUTSON, T. E., PUZANOV, I., CHMIELOWSKI, B., SPRINGER, C. J., MCARTHUR, G. A., SOSMAN, J. A., LO, R. S., RIBAS, A. & MARAIS, R. 2012. RAS mutations in cutaneous squamous-cell carcinomas in patients treated with BRAF inhibitors. *N Engl J Med,* 366**,** 207-15.

SULEM, P., GUDBJARTSSON, D. F., STACEY, S. N., HELGASON, A., RAFNAR, T., MAGNUSSON, K. P., MANOLESCU, A., KARASON, A., PALSSON, A., THORLEIFSSON, G., JAKOBSDOTTIR, M., STEINBERG, S., PALSSON, S., JONASSON, F., SIGURGEIRSSON, B., THORISDOTTIR, K., RAGNARSSON, R., BENEDIKTSDOTTIR, K. R., ABEN, K. K., KIEMENEY, L. A., OLAFSSON, J. H., GULCHER, J., KONG, A., THORSTEINSDOTTIR, U. & STEFANSSON, K. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet,* 39**,** 1443-52.

SVENSSON, F., LANG, T., JOHANSSON, M. E. V. & HANSSON, G. C. 2018. The central exons of the human MUC2 and MUC6 mucins are highly repetitive and variable in sequence between individuals. *Sci Rep,* 8**,** 17503.

TABATA, H., NAGANO, T., RAY, A. J., FLANAGAN, N., BIRCH-MACHIN, M. A. & REES, J. L. 1999. Low frequency of genetic change in p53 immunopositive clones in human epidermis. *J Invest Dermatol,* 113**,** 972-6.

TAGLIABUE, E., FARGNOLI, M. C., GANDINI, S., MAISONNEUVE, P., LIU, F., KAYSER, M., NIJSTEN, T., HAN, J., KUMAR, R., GRUIS, N. A., FERRUCCI, L., BRANICKI, W., DWYER, T., BLIZZARD, L., HELSING, P., AUTIER, P., GARCIA-BORRON, J. C., KANETSKY, P. A., LANDI, M. T., LITTLE, J., NEWTON-BISHOP, J., SERA, F., RAIMONDI, S. & GROUP, M. S. S. 2015. MC1R gene variants and non-melanoma skin cancer: a pooled-analysis from the M-SKIP project. *Br J Cancer,* 113**,** 354-63.

TAGLIABUE, E., GANDINI, S., BELLOCCO, R., MAISONNEUVE, P., NEWTON-BISHOP, J., POLSKY, D., LAZOVICH, D., KANETSKY, P. A., GHIORZO, P., GRUIS, N. A., LANDI, M. T., MENIN, C., FARGNOLI, M. C., GARCIA-BORRON, J. C., HAN, J., LITTLE, J., SERA, F. & RAIMONDI, S. 2018. MC1R variants as melanoma risk factors independent of at-risk phenotypic characteristics: a pooled analysis from the M-SKIP project. *Cancer Manag Res,* 10**,** 1143-1154.

TALVINEN, K., KARRA, H., PITKANEN, R., AHONEN, I., NYKANEN, M., LINTUNEN, M., SODERSTROM, M., KUOPIO, T. & KRONQVIST, P. 2013. Low cdc27 and high securin expression predict short survival for breast cancer patients. *APMIS,* 121**,** 945-53.

TAMBORERO, D., GONZALEZ-PEREZ, A. & LOPEZ-BIGAS, N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics,* 29**,** 2238-44.

TANG, J., FEWINGS, E., CHANG, D., ZENG, H., LIU, S., JORAPUR, A., BELOTE, R. L., MCNEAL, A. S., TAN, T. M., YEH, I., ARRON, S. T., JUDSON-TORRES, R. L., BASTIAN, B. C. & SHAIN, A. H. 2020. The genomic landscapes of individual melanocytes from human skin. *Nature,* 586**,** 600-605.

TANNEBERGER, K., PFISTER, A. S., KRIZ, V., BRYJA, V., SCHAMBONY, A. & BEHRENS, J. 2011. Structural and functional characterization of the Wnt inhibitor APC membrane recruitment 1 (Amer1). *J Biol Chem,* 286**,** 19204-14.

TAYLOR, S. J. & SHALLOWAY, D. 1996. Cell cycle-dependent activation of Ras. *Curr Biol,* 6**,** 1621-7.

THELU, J., ROSSIO, P. & FAVIER, B. 2002. Notch signalling is linked to epidermal cell differentiation level in basal cell carcinoma, psoriasis and wound healing. *BMC Dermatol,* 2**,** 7.

THOMAS, D., SAGAR, S., LIU, X., LEE, H.-R., GRUNKEMEYER, J. A., GRANDGENETT, P. M., CAFFREY, T., O'CONNELL, K. A., SWANSON, B., MARCOS-SILVA, L., STEENTOFT, C., WANDALL, H. H., MAURER, H. C., PENG, X. L., YEH, J. J., QIU, F., YU, F., MADIYALAKAN, R., OLIVE, K. P., MANDEL, U., CLAUSEN, H., HOLLINGSWORTH, M. A. & RADHAKRISHNAN, P. 2021. Isoforms of MUC16 activate oncogenic signaling through EGF receptors to enhance the progression of pancreatic cancer. *Molecular Therapy,* 29**,** 1557-1571.

THOMAS, N. E., KRICKER, A., WAXWEILER, W. T., DILLON, P. M., BUSMAN, K. J., FROM, L., GROBEN, P. A., ARMSTRONG, B. K., ANTON-CULVER, H., GRUBER, S. B., MARRETT, L. D., GALLAGHER, R. P., ZANETTI, R., ROSSO, S., DWYER, T., VENN, A., KANETSKY, P. A., ORLOW, I., PAINE, S., OLLILA, D. W., REINER, A. S., LUO, L., HAO, H., FRANK, J. S., BEGG, C. B., BERWICK, M., GENES, E. & MELANOMA STUDY, G. 2014. Comparison of clinicopathologic features and survival of histopathologically amelanotic and pigmented melanomas: a population-based study. *JAMA Dermatol,* 150**,** 1306-314.

THOMSEN, S. F. 2014. Atopic dermatitis: natural history, diagnosis, and treatment. *ISRN Allergy,* 2014**,** 354250.

THOMSON, J., BEWICKE-COPLEY, F., ANENE, C. A., GULATI, A., NAGANO, A., PURDIE, K., INMAN, G. J., PROBY, C. M., LEIGH, I. M., HARWOOD, C. A. & WANG, J. 2021. The Genomic Landscape of Actinic Keratosis. *J Invest Dermatol,* 141**,** 1664-1674 e7.

THORNTON, B. R., NG, T. M., MATYSKIELA, M. E., CARROLL, C. W., MORGAN, D. O. & TOCZYSKI, D. P. 2006. An architectural map of the anaphase-promoting complex. *Genes Dev,* 20**,** 449-60.

THUL, P. J., AKESSON, L., WIKING, M., MAHDESSIAN, D., GELADAKI, A., AIT BLAL, H., ALM, T., ASPLUND, A., BJORK, L., BRECKELS, L. M., BACKSTROM, A., DANIELSSON, F., FAGERBERG, L., FALL, J., GATTO, L., GNANN, C., HOBER, S., HJELMARE, M., JOHANSSON, F., LEE, S., LINDSKOG, C., MULDER, J., MULVEY, C. M., NILSSON, P., OKSVOLD, P., ROCKBERG, J., SCHUTTEN, R., SCHWENK, J. M., SIVERTSSON, A., SJOSTEDT, E., SKOGS, M., STADLER, C., SULLIVAN, D. P., TEGEL, H., WINSNES, C., ZHANG, C., ZWAHLEN, M., MARDINOGLU, A., PONTEN, F., VON FEILITZEN, K., LILLEY, K. S., UHLEN, M. & LUNDBERG, E. 2017. A subcellular map of the human proteome. *Science,* 356.

TOKHEIM, C. J., PAPADOPOULOS, N., KINZLER, K. W., VOGELSTEIN, B. & KARCHIN, R. 2016. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A,* 113**,** 14330-14335.

TORRE, L. A., SIEGEL, R. L., WARD, E. M. & JEMAL, A. 2016. Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol Biomarkers Prev,* 25**,** 16-27.

TSAO, H., BEVONA, C., GOGGINS, W. & QUINN, T. 2003. The transformation rate of moles (melanocytic nevi) into cutaneous melanoma: a population-based estimate. *Arch Dermatol,* 139**,** 282-8.

URANO, Y., ASANO, T., YOSHIMOTO, K., IWAHANA, H., KUBO, Y., KATO, S., SASAKI, S., TAKEUCHI, N., UCHIDA, N., NAKANISHI, H. & ET AL. 1995. Frequent p53 accumulation in the chronically sun-exposed epidermis and clonal expansion of p53 mutant cells in the epidermis adjacent to basal cell carcinoma. *J Invest Dermatol,* 104**,** 928-32.

VALAVANIDIS, A., VLACHOGIANNI, T. & FIOTAKIS, C. 2009. 8-hydroxy-2' -deoxyguanosine (8-OHdG): A critical biomarker of oxidative stress and carcinogenesis. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev,* 27**,** 120-39.

VALVERDE, P., HEALY, E., JACKSON, I., REES, J. L. & THODY, A. J. 1995. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet,* 11**,** 328-30.

VAN WEELDEN, H., DE LA FAILLE, H. B., YOUNG, E. & VAN DER LEUN, J. C. 1988. A new development in UVB phototherapy of psoriasis. *Br J Dermatol,* 119**,** 11-9.

VENABLES, Z. C., NIJSTEN, T., WONG, K. F., AUTIER, P., BROGGIO, J., DEAS, A., HARWOOD, C. A., HOLLESTEIN, L. M., LANGAN, S. M., MORGAN, E., PROBY, C. M., RASHBASS, J. & LEIGH, I. M. 2019. Epidemiology of basal and cutaneous squamous cell carcinoma in the U.K. 2013-15: a cohort study. *Br J Dermatol,* 181**,** 474-482.

VENESS, M. J., MORGAN, G. J., PALME, C. E. & GEBSKI, V. 2005. Surgery and adjuvant radiotherapy in patients with cutaneous head and neck squamous cell carcinoma metastatic to lymph nodes: combined treatment should be considered best practice. *Laryngoscope,* 115**,** 870-5.

VIROS, A., FRIDLYAND, J., BAUER, J., LASITHIOTAKIS, K., GARBE, C., PINKEL, D. & BASTIAN, B. C. 2008. Improving melanoma classification by integrating genetic and morphologic features. *PLoS Med,* 5**,** e120.

VOGELSTEIN, B., PAPADOPOULOS, N., VELCULESCU, V. E., ZHOU, S., DIAZ, L. A., JR. & KINZLER, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546-58.

VOUSDEN, K. H. & LU, X. 2002. Live or let die: the cell's response to p53. *Nat Rev Cancer,* 2**,** 594-604.

WACKER, M. & HOLICK, M. F. 2013. Sunlight and Vitamin D: A global perspective for health. *Dermatoendocrinol,* 5**,** 51-108.

WANG, K., WU, X., WANG, J. & HUANG, J. 2013. Cancer stem cell theory: therapeutic implications for nanomedicine. *Int J Nanomedicine,* 8**,** 899-908.

WANG, S., TAO, Z., WU, T. & LIU, X. S. 2021. Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis. *Bioinformatics,* 37**,** 1590-1592.

WANG, X., YU, X., KRAUTHAMMER, M., HUGO, W., DUAN, C., KANETSKY, P. A., TEER, J. K., THOMPSON, Z. J., KALOS, D., TSAI, K. Y., SMALLEY, K. S. M., SONDAK, V. K., CHEN, Y. A. & CONEJO-GARCIA, J. R. 2020a. The Association of MUC16 Mutation with Tumor Mutation Burden and Its Prognostic Implications in Cutaneous Melanoma. *Cancer Epidemiol Biomarkers Prev,* 29**,** 1792-1799.

WANG, Y., MASHOCK, M., TONG, Z., MU, X., CHEN, H., ZHOU, X., ZHANG, H., ZHAO, G., LIU, B. & LI, X. 2020b. Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. *Front Oncol,* 10**,** 447.

WANG, Y., ZHOU, X., WEINSTEIN, E., MARYLES, B., ZHANG, Y., MOORE, J., GAO, D., ATENCIO, D. P., ROSENSTEIN, B. S., LEBWOHL, M., CHEN, H. D., XIAO, T. & WEI, H. 2008. p53 gene mutations in SKH-1 mouse tumors differentially induced by UVB and combined subcarcinogenic benzo[a]pyrene and UVA. *Photochem Photobiol,* 84**,** 444-9.

WEI, L. A.-O., CHRISTENSEN, S. A.-O., FITZGERALD, M. E., GRAHAM, J., HUTSON, N. A.-O., ZHANG, C., HUANG, Z., HU, Q. A.-O., ZHAN, F., XIE, J., ZHANG, J., LIU, S., REMENYIK, E., GELLEN, E. A.-O., COLEGIO, O. R., BAX, M., XU, J., LIN, H., HUSS, W. A.-O., FOSTER, B. A.-O. & PARAGH, G. A.-O. 2021. Ultradeep sequencing differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. LID - 10.1126/sciadv.abd7703 [doi] LID - eabd7703.

WEINSTEIN, I. B. 2002. Cancer: Addiction to oncogenes - The Achilles heal of cancer. *Science,* 297**,** 63-64.

WEISS, B., BOLLAG, G. & SHANNON, K. 1999. Hyperactive Ras as a therapeutic target in neurofibromatosis type 1. *Am J Med Genet,* 89**,** 14-22.

WELCH, JOHN S., LEY, TIMOTHY J., LINK, DANIEL C., MILLER, CHRISTOPHER A., LARSON, DAVID E., KOBOLDT, DANIEL C., WARTMAN, LUKAS D., LAMPRECHT, TAMARA L., LIU, F., XIA, J., KANDOTH, C., FULTON, ROBERT S., MCLELLAN, MICHAEL D., DOOLING, DAVID J., WALLIS, JOHN W., CHEN, K., HARRIS, CHRISTOPHER C., SCHMIDT, HEATHER K., KALICKI-VEIZER, JOELLE M., LU, C., ZHANG, Q., LIN, L., O'LAUGHLIN, MICHELLE D., MCMICHAEL, JOSHUA F., DELEHAUNTY, KIM D., FULTON, LUCINDA A., MAGRINI, VINCENT J., MCGRATH, SEAN D., DEMETER, RYAN T., VICKERY, TAMMI L., HUNDAL, J., COOK, LISA L., SWIFT, GARY W., REED, JERRY P., ALLDREDGE, PATRICIA A., WYLIE, TODD N., WALKER, JASON R., WATSON, MARK A., HEATH, SHARON E., SHANNON, WILLIAM D., VARGHESE, N., NAGARAJAN, R., PAYTON, JACQUELINE E., BATY, JACK D., KULKARNI, S., KLCO, JEFFERY M., TOMASSON, MICHAEL H., WESTERVELT, P., WALTER, MATTHEW J., GRAUBERT, TIMOTHY A., DIPERSIO, JOHN F., DING, L., MARDIS, ELAINE R. & WILSON, RICHARD K. 2012. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell,* 150**,** 264-278.

WENGER, A. M., PELUSO, P., ROWELL, W. J., CHANG, P.-C., HALL, R. J., CONCEPCION, G. T., EBLER, J., FUNGTAMMASAN, A., KOLESNIKOV, A., OLSON, N. D., TÖPFER, A., ALONGE, M., MAHMOUD, M., QIAN, Y., CHIN, C.-S., PHILLIPPY, A. M., SCHATZ, M. C., MYERS, G., DEPRISTO, M. A., RUAN, J., MARSCHALL, T., SEDLAZECK, F. J., ZOOK, J. M., LI, H., KOREN, S., CARROLL, A., RANK, D. R. & HUNKAPILLER, M. W. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology,* 37**,** 1155-1162.

WIKONKAL, N. M. & BRASH, D. E. 1999. Ultraviolet radiation induced signature mutations in photocarcinogenesis. *J Investig Dermatol Symp Proc,* 4**,** 6-10.

WILLIAMS, H. C. & STRACHAN, D. P. 1998. The natural history of childhood eczema: observations from the British 1958 birth cohort study. *Br J Dermatol,* 139**,** 834-9.

WLASCHEK, M., TANTCHEVA-POOR, I., NADERI, L., MA, W., SCHNEIDER, L. A., RAZI-WOLF, Z., SCHULLER, J. & SCHARFFETTER-KOCHANEK, K. 2001. Solar UV irradiation and dermal photoaging. *J Photochem Photobiol B,* 63**,** 41-51.

WONG, C. C., MARTINCORENA, I., RUST, A. G., RASHID, M., ALIFRANGIS, C., ALEXANDROV, L. B., TIFFEN, J. C., KOBER, C., CHRONIC MYELOID DISORDERS WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C., GREEN, A. R., MASSIE, C. E., NANGALIA, J., LEMPIDAKI, S., DOHNER, H., DOHNER, K., BRAY, S. J., MCDERMOTT, U., PAPAEMMANUIL, E., CAMPBELL, P. J. & ADAMS, D. J. 2014. Inactivating CUX1 mutations promote tumorigenesis. *Nat Genet,* 46**,** 33-8.

WONG, S. H. & YU, J. 2019. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat Rev Gastroenterol Hepatol,* 16**,** 690-704.

WU, J. H., COHEN, D. N., RADY, P. L. & TYRING, S. K. 2017. BRAF inhibitor-associated cutaneous squamous cell carcinoma: new mechanistic insight, emerging evidence for viral involvement and perspectives on clinical management. *Br J Dermatol,* 177**,** 914-923.

WULF, H. C., HANSEN, A. B. & BECH-THOMSEN, N. 1994. Differences in narrow-band ultraviolet B and broad-spectrum ultraviolet photocarcinogenesis in lightly pigmented hairless mice. *Photodermatol Photoimmunol Photomed,* 10**,** 192-7.

XIE, M., LU, C., WANG, J., MCLELLAN, M. A.-O., JOHNSON, K. J., WENDL, M. C., MCMICHAEL, J. F., SCHMIDT, H. K., YELLAPANTULA, V., MILLER, C. A.-O., OZENBERGER, B. A.-O., WELCH, J. S., LINK, D. C., WALTER, M. J., MARDIS, E. R., DIPERSIO, J. F., CHEN, F., WILSON, R. K., LEY, T. J. & DING, L. 2014. Age-related mutations associated with clonal hematopoietic expansion and malignancies.

YACHIDA, S., JONES, S., BOZIC, I., ANTAL, T., LEARY, R., FU, B., KAMIYAMA, M., HRUBAN, R. H., ESHLEMAN, J. R., NOWAK, M. A., VELCULESCU, V. E., KINZLER, K. W., VOGELSTEIN, B. & IACOBUZIO-DONAHUE, C. A. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467**,** 1114-7.

YAEGER, R. & CORCORAN, R. B. 2019. Targeting Alterations in the RAF-MEK Pathway. *Cancer Discov,* 9**,** 329-341.

YANG, Z., RO, S. & RANNALA, B. 2003. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics,* 165**,** 695-705.

YATES, L. R. & CAMPBELL, P. J. 2012. Evolution of the cancer genome. *Nat Rev Genet,* 13**,** 795-806.

YILMAZ, A. S., OZER, H. G., GILLESPIE, J. L., ALLAIN, D. C., BERNHARDT, M. N., FURLAN, K. C., CASTRO, L. T., PETERS, S. B., NAGARAJAN, P., KANG, S. Y., IWENOFU, O. H., OLENCKI, T., TEKNOS, T. N. & TOLAND, A. E. 2017. Differential mutation frequencies in metastatic cutaneous squamous cell carcinomas versus primary tumors. *Cancer,* 123**,** 1184-1193.

YIZHAK, K., AGUET, F., KIM, J., HESS, J. M., KUBLER, K., GRIMSBY, J., FRAZER, R., ZHANG, H., HARADHVALA, N. J., ROSEBROCK, D., LIVITZ, D., LI, X., ARICH-LANDKOF, E., SHORESH, N., STEWART, C., SEGRE, A. V., BRANTON, P. A., POLAK, P., ARDLIE, K. G. & GETZ, G. 2019. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science,* 364.

YOGIANTI, F., KUNISADA, M., ONO, R., SAKUMI, K., NAKABEPPU, Y. & NISHIGORI, C. 2012. Skin tumours induced by narrowband UVB have higher frequency of p53 mutations than tumours induced by broadband UVB independent of Ogg1 genotype. *Mutagenesis,* 27**,** 637-43.

YOU, Y. H., LEE, D. H., YOON, J. H., NAKAJIMA, S., YASUI, A. & PFEIFER, G. P. 2001. Cyclobutane pyrimidine dimers are responsible for the vast majority of mutations induced by UVB irradiation in mammalian cells. *J Biol Chem,* 276**,** 44688-94.

YUAN, X., WU, H., XU, H., XIONG, H., CHU, Q., YU, S., WU, G. S. & WU, K. 2015. Notch signaling: an emerging therapeutic target for cancer treatment. *Cancer Lett,* 369**,** 20-7.

ZHANG, P. & WU, M. X. 2018. A clinical review of phototherapy for psoriasis. *Lasers Med Sci,* 33**,** 173-180.

ZHU, C., ZHAO, J., BIBIKOVA, M., LEVERSON, J. D., BOSSY-WETZEL, E., FAN, J. B., ABRAHAM, R. T. & JIANG, W. 2005. Functional analysis of human microtubule-based motor proteins, the kinesins and dyneins, in mitosis/cytokinesis using RNA interference. *Mol Biol Cell,* 16**,** 3187-99.

ZIEGLER, A., JONASON, A. S., LEFFELL, D. J., SIMON, J. A., SHARMA, H. W., KIMMELMAN, J., REMINGTON, L., JACKS, T. & BRASH, D. E. 1994. Sunburn and p53 in the onset of skin cancer. *Nature,* 372**,** 773-6.

ZUMSTEG, Z. S., COOK-WIENS, G., YOSHIDA, E., SHIAO, S. L., LEE, N. Y., MITA, A., JEON, C., GOODMAN, M. T. & HO, A. S. 2016. Incidence of Oropharyngeal Cancer Among Elderly Patients in the United States. *JAMA Oncol,* 2**,** 1617-1623.