Identification of germline variants that predispose to familial melanoma

Aravind Sankar

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

St Edmund's College 2020

To Appa, Amma and Aditya: without whom this would not be possible.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 60,000 words and has less than 150 figures, exclusive of tables, footnotes, bibliography, and appendices.

Aravind Sankar
2020

Acknowledgements

There is an old adage which goes “It takes a village to raise a child”. I believe it also takes a village to obtain a doctoral degree. I would like to take this opportunity to thank the members of my village who helped me get to where I am today.

First and foremost, I would like to thank my supervisor Dave Adams for his endless support, guidance, wisdom and kindness. I wouldn’t know where to begin thanking you for all the things you have done for me so I would just like to say - thank you for everything.

I would also like to thank my other supervisor, Vivek Iyer, who mentored, trained and tutored me with the patience of a saint. It is due to him that I can confidently call myself a bioinformatician today.

I am extremely grateful to Professor Tim Bishop from the University of Leeds for taking the time to e-mail, phone and make trips to Cambridge to teach me statistics.

I would like to acknowledge the Wellcome Sanger Institute and the MELGEN network for funding my PhD and for providing me with a platform to collaborate, interact and learn from current and future world-leaders in research.

I would like to thank every member of the GenoMEL consortium that provided samples, ideas, inputs and suggestions over the last 4 years. This project would not have been possible without them.

To Sofia - Thank you for starting and finishing your PhD with me, for the dinners, the conference trips, for always lending an ear and for being my voice of reason.

To Gemma - Thank you for being my friend, my driver, my flat mate and my confidante. Every day got a little bit easier with you around.

To Nicky - Thank you for the tea trips, the joint birthday celebrations and for reminding me to always look on the bright side of life.

To Marco - Thank you for taking the effort to make me feel at home in Cambridge on my first day and for continuing to do that ever since.

To Daniela - Thank you for always being there even when you are in Mexico.

To Katharina - Thank you for your positivity, for encouraging me to run and for being a social butterfly.

To Annie - Thank you for taking the time to remind me that I cannot take my time with my thesis!

To everyone else in team113 - Thank you for making me feel excited to come to work everyday. I will see you at lunch at 12.

To Prakaash and Mahathi - Thank you for sticking with me.

Finally, to my family. Appa, Amma, Paati and Aditya - For encouraging me to fly while being my safety net, for being living examples of showing that hard work pays off and for your unconditional love and support - Thank you.

Abstract

Melanoma is an extremely aggressive malignancy with a poor prognosis in advanced disease. While GWAS and exome analysis have helped to identify loci linked to the development of the disease, these studies have explained predisposition to melanoma in only a fraction of cases. Thus, the majority of the genetic factors that contribute to the pathogenesis of melanoma are yet to be defined. This project aims at identifying novel genes and pathways involved in the development of familial melanoma, and also identify loci which predispose individuals to disease development.

308 individuals from 133 different families previously diagnosed with melanoma were sequenced through a mixture of exome or whole genome sequencing. Multiple workflows were established to analyse the dataset for novel driver mutations. A novel approach of combining association and linkage analysis was established for the variants in the coding region to identify genes with high burden of mutations where the variants segregated with the disease within the pedigrees. The role of non-coding variants and structural variants in melanoma onset was also investigated through additional workflows in the whole-genome sequenced individuals.

Non-synonymous mutations were found in *CDKN2A, BRCA1, POT1* and *BAP1*. Disruptive variants were also observed in novel genes such as *EXO5, TP53AIP* and *AMER1*. An increased burden on variants in transcription factor binding motifs were observed in genes including *SYK* and *SRC*. A large deletion upstream of *CDKN2A* was identified. Genes including *ATR* and *FAT1* were identified to have a higher burden of disruptive variants that segregated with the disease within the cases through the novel combined association-linkage analysis.

Disruptive germline variants that could play a role in familial melanoma development were identified in multiple genes through a combination of several approaches.

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

CGC	Cancer Gene Census
ExAC	Exome Aggregation Consortium
GC	Count of individuals for each genotype
GC NFE	Count of Non-Finnish European individuals for each genotype
gnomAD	Genome Aggregation Database
GQ	Genotype Quality
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
gVCF	Genomic Variant Calling Format
HGNC	HUGO Gene Nomenclature Committee
OR	Odds Ratio
PRS	Polygenic Risk Scores
TF	Transcription Factors
TFBM	Transcription Factor Binding Motifs
VCF	Variant Calling Format
VEP	Variant Effect Predictor

Chapter 1

Background

1.1 Melanoma - A statistical overview

Melanoma is a highly aggressive type of skin cancer with a poor prognosis in the advanced stage. It originates from melanocytes, a type of skin cell which are responsible for the production of pigments called melanin. The incidence rates of people suffering from malignant melanoma has increased by 134% since the early 90s and has gone up by 45% since 2006 in the United Kingdom[1]. Skin cancer is the most commonly diagnosed type of cancer in the United States of America with melanoma comprising of less than 1% of total skin cancer cases but the vast majority of skin cancer related deaths[2]. Around 91,270 people are expected to be diagnosed with melanoma in 2018 with 9,320 estimated deaths in the United States of America[2]. Comparatively, melanoma consisted of 10% of all skin cancer diagnoses in 2016 but was responsible for 63% of skin-cancer related deaths in the United Kingdom in 2016[1]. The overall incidence of melanoma in the UK is expected to increase by 7% by 2035[1]. Malignant melanoma mortality rates in the UK have increased by 156% since the 1970's[1]. Melanoma has the second highest increase in mortality (after liver cancer) amongst all cancers for men in the last ten years and the fifth highest increase in mortality for women. The 5 year survival status for melanoma patients is quite high: 99% of all cases in the United States of America[2] and 100% of all cases in the UK were expected to survive for local melanomas[3]. However, this dropped to less than 20% for both for distant metastatic melanomas[2, 3]. 1 in 10 melanoma cases diagnosed were estimated to be metastatic in the UK at diagnosis[1].

1.2 Melanoma through the ages

Melanoma (derived from the Greek words *melas* meaning “dark” and *-oma* referring to abnormal growth or tumours - such as carcinoma or lymphoma) was first coined by Dr Robert Carswell in 1838 as part of his seminal work “Illustrations of the Elementary Forms of Disease”. The earliest chronicles of melanoma as a disease came from Hippocrates in the 5th century B.C. and Rufus of Eupheses in the 1st century A.D. [4]. Metastases in the skeletons of nine mummies from the Pre-Colombian Incas of Peru dated back to 4th century B.C. is commonly cited as the first chronological physical evidence of melanoma[4], a claim that has however been recently disputed. A "cancerous fungous excrescence" was surgically resected by John Hunter, which was eventually confirmed to be a metastatic melanoma tumour. Metastatic melanoma was first described by the French physician Rene Laennec in 1806[5]. He observed metastatic melanoma tumours while studying granulomas in lungs and defined these black masses as melanoses.

The first known description of familial melanoma was provided by Dr. William Norris in 1820[6]. While studying a patient, he observed several salient characteristics: the tumour originated from a mole, the father of the patient(who eventually died of his ailments) also died due to a similar disease and had tumours originating from moles and the son of the patient had similar distribution of moles across his body. These observations led him to conclude that the “disease is hereditary” which was eventually characterised and defined as Familial atypical multiple mole-melanoma syndrome (FAMM) in 1978[7].

The impact of ultraviolet radiation from sunlight on melanoma onset was made in 1956 by Henry Lancaster, who discerned that the intensity of sunlight had a direct influence on the risk of melanoma development, with particular relevance to Caucasian populations[8]. He went on to eventually link features of the skin with playing a principal part in melanoma development.

Different metrics for measuring melanoma progression were designed by Wallace Clark and Alexander Breslow in the late 1960’s. Wallace Clark devised a framework called as Clark’s levels which compares melanoma progression to the level of invasion of the tumour into the skin (Figure 1.1)[9]. A Level I tumour would be restricted to the upper layer of the epidermis while a Level V tumour would be completely invasive and would already be in the subcutaneous tissue. Alexander Breslow, around the same time, determined that melanoma progression is also linked to the size of the tumour as opposed to just the level of invasion, with specific emphasis on the thickness[10]. This was a measurement of the distance between the uppermost layer of the skin to the deepest point of tumour penetration in millimeters and was later defined as Breslow thickness. These metrics of classification are explained in detail

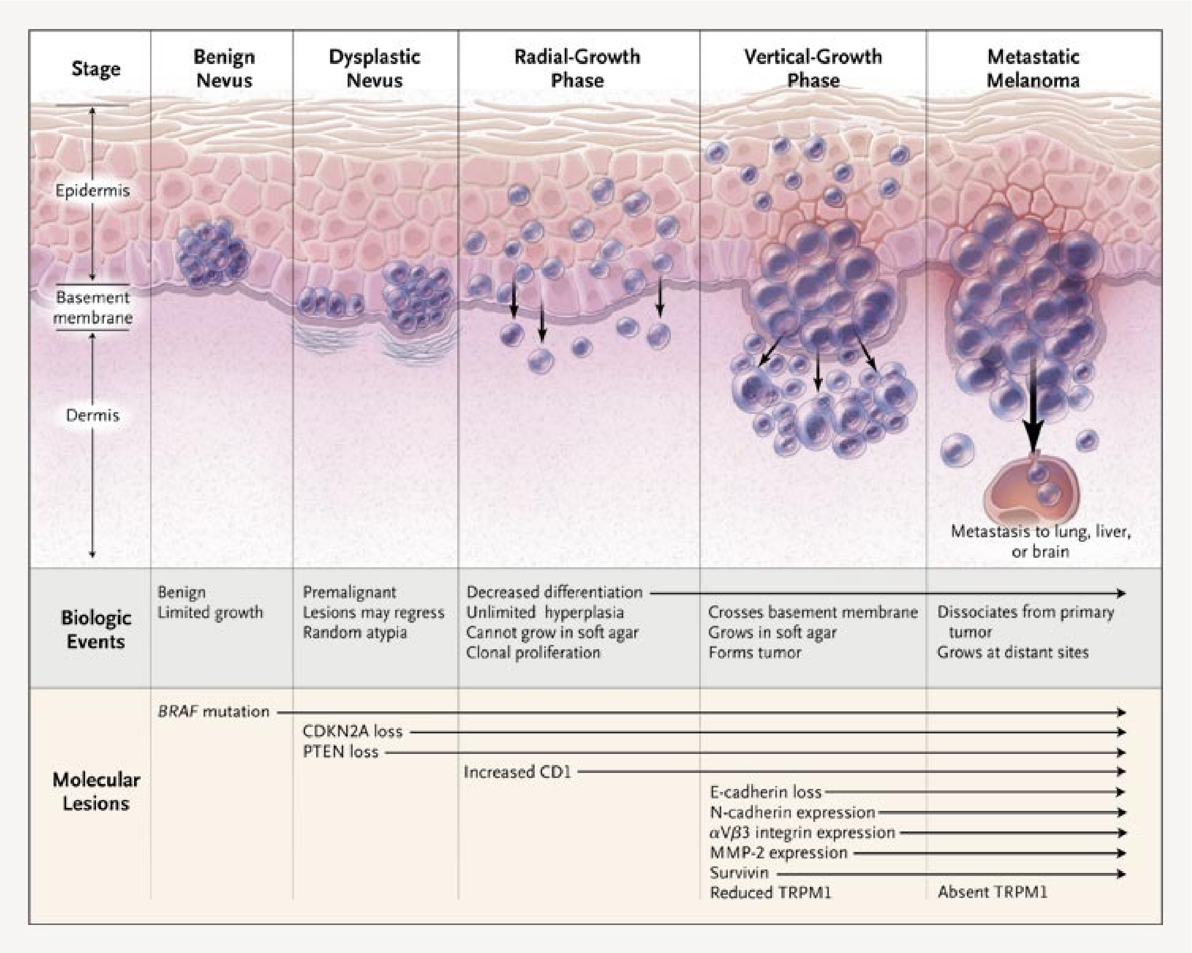


Figure 1.1: The five stages of melanoma progression as described in Clark's model. Reproduced with permission from [11].

in Section 1.4.2.1.

Over time, several insights into the progression of cancer development led to the discovery that genetic mutations in genes related to cancer development, primarily classified into oncogenes and tumour suppressors, play a vital role in tumour formation. One of the earliest group of genes that were determined to play a role in the context of melanoma were the RAS family of genes. *NRAS* was discovered as a novel human transforming gene in 1983[12] with *NRAS* mutations in melanoma cell lines being ascertained soon after[13]. Activating *KRAS* and *HRAS* mutations in melanoma were also eventually discovered. It is estimated that roughly 15-20% of all melanomas have mutations in *NRAS[14]*with 15% of all cancers carrying a RAS mutation[15]. This period also saw the discovery of the RAF family of oncogenes, namely *ARAF*[16], *BRAF[17]* and *CRAF*[18]. These genes play a significant role in the RAS/Raf/MEK/ ERK pathway (Section 1.3.3.1). *BRAF* in particular is responsible for the activation of *MEK1* and *MEK2*. Multiple large scale genetic and genomic screens have since determined *BRAF* to be the most mutated driver gene in melanoma with roughly 50% of all melanomas carrying a *BRAF* mutation[19]. A majority of these mutations, almost 90%, are a specific substitution at the 600th amino acid where a valine is replaced with a glutamic acid (*V600E*)[19]. It is to be noted that *BRAF* and *NRAS* mutations are usually mutually exclusive and, in a study by the Cancer Genome Atlas network, were identified as being responsible for almost 80% of all melanoma tumours[20]. They also show phenotypic difference as $BRAF^{V600E}$ mutations occur in younger patients having non-chronically sun exposed melanoma[21] while *NRAS* mutations occur in older patients with much higher sun-exposure[14]. The role of *BRAF* and *NRAS* mutations in the development of melanoma is discussed in Section 1.3.3. The relationship between skin characteristics and melanoma development, as predicted by Lancaster[8], were finally established through the discovery of Melanocortin receptor 1 (*MC1R*), a gene that was initially determined as a marker for hair-colour and paleness of skin[22]. Experiments confirmed the presence of several low penetrant-high frequency variants in *MC1R* which not only determined these traits but also, response to tanning/sun exposure[23] and risk of melanoma development[24]. The function of *MC1R* in melanin production is discussed in Section 1.3.1.

The early 1990's also saw the discovery of the most prominent familial melanoma gene, *Cyclin-dependent kinase inhibitor 2A* (Section 1.3.4.2), through linkage analysis[25, 26]. This has since then been determined as the single biggest driver gene in familial melanoma with roughly 40% of all familial melanoma genes carrying a *CDKN2A* driver mutation[27]. The growth of next-generation sequencing in the early 2000's and the availability of whole-exome and whole genome sequencing has helped discover multiple low-frequency driver genes in melanoma. *NF1*, *CDK4*, *BAP1*, *POT1* and *TERT* have all been established as being responsible for a significant number of melanoma cases. These genes are discussed in Sections 1.3.3 and 1.3.4. However, despite these efforts, roughly 20 percent of all melanomas and 50% of all familial melanoma cases have unknown and undiscovered genetic driver mutations. Several ongoing projects across the world are aiming to address this conundrum including this project.

1.3 The biology of melanoma

1.3.1 An introduction to melanocytes

The skin primarily comprises of three layers: The epidermis, the dermis and the subcutaneous tissue (hypodermis). Melanocytes are specific types of skin cells that exist between the inner most layer of the epidermis (also called the basal layer) and the dermis. There are roughly 1,200 melanocytes per square millimeter of the epidermis in the average human[28]. They are neural crest-derived cells responsible for the production of melanin; a pigment that plays a key role in the protection of keratinocytes in the skin against ultra-violet radiation. Melanin are also responsible for determining the phenotype of the skin colour of the organism[29]. Almost all melanocytes are derived from the neural crest with the exception of retinal pigment epithelium (RPE), which are derived from the neuroepithelium[30]. Although they are primarily present in skin, melanocytes have also been identified in the eye, cochlea, heart, brain and the adipose tissue[31]. Melanin produced by melanocytes can be classified into three types: eumelanin, pheomelanin and neuromelanin.

- Eumelanin is a dark brown-black heterogenous polymer that act as a protective layer to the skin and help by absorbing hazardous solar radiation, particularly UV radiation[32]. Eumelanin is primarily responsible for the beneficial effects of melanin in terms of protection from chronic sun exposure. Eumelanin also functions as a free radical scavenger and superoxide dismutase that reduce reactive oxygen species (ROS)[33, 34].

- Pheomelanin is a reddish-brown pigment that is composed of sulfur-containing benzothiazine and benzothiazole derivatives[35]. It is responsible for the occurrence of red hair and freckles in the general populace. They are unstable in the presence of light and do not offer much protection against radiation; contrarily they may be responsible for the onset of carcinogenesis[32]. This is reflected in the rate of skin cancer in different ethnic populations as the rate of skin cancer development in populations with lighter skin was more than 70 times the rate of skin cancer development in darker populations[32].

- Neuromelanin is a black-brown pigment found within the brain that is a mixture of both eumelanin and pheomelanin[36]. Not much is known about its function.

The primary components of the epidermis which cells called keratinocytes. Production of melanin from melanocytes is instigated when keratinocytes in the epidermis are exposed to sun-light. When keratinocytes are exposed to ultra violet radiation, they produce several products including α-melanocyte stimulating hormone (α-MSH). αMSH in turn binds to a specific receptor called the melanocortin 1 receptor (*MC1R*) which is expressed in melanocytes[37]. The binding of αMSH to *MC1R* initiates the production of melanin, particularly eumelanin. Once this process has been started, mechanisms of melanin synthesis and survival are initiated within the melanocytes. Melanin is then shipped to cellular organelles within the melanocytes called melanosomes[38]. Once the melanosomes mature, they are transferred through the melanocytic dendrites of the melanocyte to the keratinocytes in the epidermis[39]. A single melanocyte can produce and transfer melanin for up to 40 keratinocytes through the melanocytic dendrites[40].

Mutations in the melanin production process are an integral part of several diseases ranging from pigmental disorders like Neurofibromatosis type 1[41] and Occulocutaneous Albinism[**?**] to skin cancers like melanoma[42].

1.3.2 The progression of melanocytes to metastatic melanoma

Melanoma, as mentioned in Section 1.1, is an aggressive form of skin cancer that develops from uncontrolled proliferation of melanocytes and is referred to as cutaneous melanoma when it occurs in the upper layers of the skin. Cutaneous melanoma can be classified into two major categories based on sun-exposure, chronically sun damaged (CSD) melanoma and non-chronically sun damaged (non-CSD) melanoma. CSD melanoma are associated with increased UV mutations and are observed in parts of the body exposed to the sun while non-CSD are linked to highly penetrant genetic mutations. While there is evidence to suggest that melanoma progression and development predominantly arises de novo[43, 44], a significant percentage of melanoma cases are due to tumours formed from malignant melanocytes associated with pre-existing nevi. The presence of a high number of common and atypical nevi, also known as an atypical mole syndrome, is highly predictive of increased melanoma risk[45]. Previously GWAS studies focussing on nevi counts as a risk phenotype for melanoma have helped identify key driver genes in melanoma[46]. The genes identified from this study are discussed in Section 1.3.4.1.The development of melanoma from melanocytes occurs through several stages, shown in Figure 1.2, and was described in detail by Shain et al[47]. These stages are summarized here.

i) Melanocytic nevus

A melanocytic nevus is a benign neoplasm consisting of melanocytes that appear as a raised, pigmented irregularity. While the majority of such nevi have a low probability of progression to malignancy, increased nevus counts are associated with increased melanoma risk[48]. Melanomas arising from pre-existing nevi are associated with non-CSD melanomas, particularly with superficial spreading melanoma, which are observed with a much lower frequency in CSD melanomas[49]. A specific mutation in the *BRAF* gene called the $BRAF^{V600E}$ variant, described in Section 1.3.3.2, has been identified as a triggering event for nevi formation[50]. While skin related phenotypes such as tanning ability and pigmentation are risk factors for melanoma development, none of the variants in affected genes for these phenotypes play a

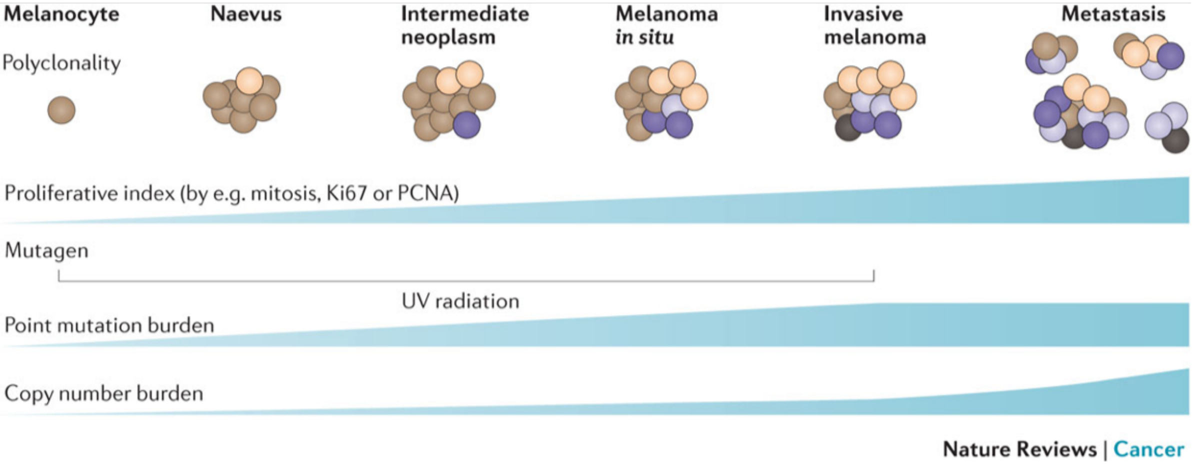


Figure 1.2: Biological characteristics for the progression of melanoma from melanocytes to metastasis. Reproduced with permission from [47].

role in melanocytic nevi formation[51]. However, lighter skin implies weaker protection of the skin against UV radiation which results in higher UV related mutagenesis. This is seen as an implicating factor in nevi development and is therefore linked to increased nevi counts[52–54]. Thus, *BRAF*$^{\mathrm{V600E}}$ variants are seen as the only driver events for melanocytic nevi formation.

ii) Dysplastic nevus

A dysplastic nevus is a transitional category of melanocytic neoplasm with histopathological features that are between a benign nevus and a malignant melanoma[55]. Due to the large variance in the characteristics of such nevi, there is a significant variation in the classification and diagnosis of these neoplasms. They are clinically defined as having a minimum diameter of 5 mm and possessing two of the following characteristics: variable pigmentation, asymmetry and/or irregular or indistinct borders[56]. This type of nevi is commonly observed in familial melanoma pedigrees. While benign melanocytic nevi only carry *BRAF*$^{\mathrm{V600E}}$ mutations, dysplastic nevi carry variants in other pathways associated with melanoma development such as MAPK signalling, telomerase regulation (*TERT* promoter variants) and cell cycle regulation (*CDKN2A* variants)[57]. This indicates that melanocytic nevi may carry pathogenic variants but the proliferation is controlled by regulatory pathways such as the MAPK pathway. Additional mutations that disrupt the MAPK pathway could result in increased proliferation and growth of melanocytic nevi to dysplastic nevi. However, some dysplastic nevi also have a higher proportion of *NRAS* and *BRAF* mutations which are not V600E, implying that they may occur independently without developing from a melanocytic nevus[57].

iii) Melanoma *in situ*

The first point of malignancy is the formation of a melanoma in-situ by the proliferation of melanocytes with asymmetrical growth contained within the epithelium without basal invasion[58]. Melanoma *in situ* is considered a precursor to invasive melanoma. Depending on the type of growth of melanocytes, two distinct types of melanoma are formed : Superficial spreading melanoma and lentigo maligna melanoma, discussed in Sections 1.4.1.1 and 1.4.1.3 respectively.

Superficial spreading melanoma is a result of a pagetoid growth pattern where the growth spreads both upwards and downwards through the different layers of the epidermis. They are commonly observed in melanomas arising from a pre-existing nevus which may be benign or dysplastic. These types of melanoma also have a higher proportion of $BRAF^{V600E}$ mutations[21].

Lentigo maligna melanoma on the other hand is associated with lentiginous growth and is seen in melanomas linked to chronic sun damage. It is also associated with a low proportion of *BRAF* V600E mutations[21] Such a growth is generally not linked to pre-existing nevi and surfaces independently. Acral melanomas, discussed in Section 1.4.1.4, also exhibit lentiginous melanoma.

iv) Invasive melanoma

When melanocytes from a melanoma *in situ* invade secondary layers of the skin and spread into the dermis from the epithelium, they are termed as invasive melanoma. For melanocytes to reach this stage, they must carry multiple driver mutations over its progression, including variants disrupting the MAPK pathway (described in Section 1.3.3.1) and *TERT* promoter mutations (described in Section 1.3.4.4). However, in addition to such mutations, invasive melanomas have a high fraction of p16INK4a inactivation[57]. This is a tumour suppressor protein generated from *CDKN2A* (discussed in Section 1.3.4.2) which is responsible for cell cycle regulation. With prior evidence towards the loss of *INK4A* developing highly penetrant invasive melanoma, it is increasingly evident that *INK4A* plays a critical role in preventing the transition of melanoma *in situ* to invasive melanoma. However, some invasive melanomas also preserve functional *INK4A*, implying that the G1/S checkpoint is disrupted through other mutations. One such alternative mechanism for the promotion of invasive melanoma are variants in *ARID1A* and *ARID2*, which are important components of the SWI/SNF chromatin-remodeling complex[57]. These genes also function as tumour suppressors in melanoma but function by maintaining genomic stability[59]. Loss of function of *ARID1A* or *ARID2* leads to increased chromosomal aberrations which promotes the proliferation of melanocytes. Additionally, *PTEN* and *TP53* mutations have also been reported in a subset of invasive melanomas but these mutations are in much lower frequencies in primary melanomas, indicating that such mutations are formed in the later stages of tumour progression[57].

v) Metastatic melanoma

A melanoma is said to become metastatic when tumour cells have dispersed into other organs and tissues beyond the site of origin and is the final stage of melanoma progression. Metastatic melanoma is usually observed at the regional lymph nodes and progressively spreads to distal sites. However, circulating tumour cells have previously been observed in melanoma cases without any metastases and even after the patients were considered to be disease free, indicating that the metastatic process does not occur sequentially but rather concurrently[60]. The presence of initial metastases at regional lymph nodes could therefore just be an early hallmark of metastatic progression, which has been seen in other cancers. Additionally, a small percentage of individuals without a recognizable primary tumour have been identified to have metastases. Such types of melanoma occurrences are known as melanomas of unknown primary (MUP)[61]. Several hypotheses exist for the presence of such melanomas. The primary theory of origin is the regression theory proposed by Smith and Stehlin in 1965 which suggests that the disappearance of the primary melanoma is due to spontaneous regression of the tumour post-metastasis[62]. Other cases are attributed to misdiagnosis of the primary tumour or unreported treatment/excision of a suspect lesion[61]. Similar to primary melanomas, MUPs have a high burden of UV radiation-induced mutations[63]. This implies that sun exposure plays a role in the genesis of such tumours.

1.3.3 The landscape of somatic variation in melanoma

A significant proportion of melanocytic tumours are created due to somatic mutations in different parts of the MAPK(Ras-Raf-MEK-ERK) pathway. As previously mentioned in Section 1.2, mutations in *BRAF* and *NRAS* have been observed in 80% of all melanoma tumours. However, *BRAF* and *NRAS* mutations are not solely responsible for the formation of melanocytic nevi; on the contrary, they have observed to be present commonly in benign nevi too. The Cancer Genome Atlas (TCGA) is a project started in 2005 to identify and catalogue the different genetic and genomic variations in different types of cancers. Genomic alterations in cutaneous melanoma were reported by TCGA in 2015; cutaneous melanoma was suggested to be classified into one of 4 subtypes: *BRAF* mutant melanoma, *RAS* mutant melanoma, *NF1* mutant melanoma and triple-wild-type (Triple WT) melanoma[20]. While such a classification is practical from a therapeutic and palliative perspective, there are also other mutations such as those in the triple WT classification. In such cases, the cause for disease onset is not very clear. This section explores the role of the MAPK pathway and the different subtypes of cutaneous melanoma as defined by the TCGA.

1.3.3.1 The MAPK (Ras-Raf-MEK-ERK) Pathway

The MAPK (Ras-Raf-MEK-ERK) cascade reaction is an important signalling pathway, shown in Figure 1.3, that plays a vital role in cancer development[64].

The RAS/MAPK pathway has a critical role in normal development through regulation of cell growth, differentiation, and senescence[64]. A detailed description of the role of the MAPK pathway in cellular proliferation was provided by Zhang and Liu in 2002[65].

The signaling pathway starts with the interaction between a ligand and a epidermal growth factor receptor (EGFR), specifically human epidermal growth factor receptor 2 (HER2). Once active, HER2 links to a protein complex consisting of SOS and GRB2. SOS is a prominent guanine nucleotide exchange factor while GRB2 is an adaptor protein that plays a role in intracellular signalling. Once Sos-Grb2 has been activated, it in turn interacts with proteins from the RAS family, a family of small GTPases, and activates RAS. GTPases toggle between inactive and active conformations which acts as a switch for the signalling chains. This is done through the binding of guanosine phosphates. When guanosine diphosphate (GDP) is bound to the GTPase, they are in their inactive state. This is switched to an active state through based on their binding to the guanine nucleotides GDP or GTP. GTPases are in the "OFF" state when bound to GDP and are activated by guanine nucleotide exchange factors (GEFs), such as SOS. These change the confirmation of the GTPase, release GDP and bind GTP, which changes the structure of the protein and activates other targets downstream of the signalling process. A similar process occurs with the RAS family (*NRAS,KRAS,HRAS*). RAS in its original state is loaded with GDP. When it interacts with the Sos-Grb2 complex, it is shifted from its inactive GDP state to an active GTP state through the action of SOS as a GEF. Active RAS in negatively regulated through neurofibromin 1 (*NF1*), a GTPase activating protein which accelerates the rate of hydrolysis of GTP to GDP, thereby inactivating RAS. The active RAS protein then continues the MAPK pathway by binding to the Ras-binding domain of a family of proteins called the Rapidly Accelerated Fibrosarcoma(RAF) proteins, particularly *BRAF*. This activates *BRAF* which phosphorylates and activates two mitogen activated protein kinase kinases, MEK1 and MEK2 respectively. The activated MEK proteins phosphorylate and activate their targets, a pair of extra-cellular signal related kinases (ERK) called ERK1 and ERK2. ERK1 and ERK2 are responsible as regulators of several key cellular processes including cell proliferation, survival and metastasis. Improper regulation of this key pathway through disruption of its components plays a critical role in the oncogenesis of multiple types of cancer, including cutaneous melanoma[64]. The majority of sporadic melanoma cases are caused due to disruptions in key proteins of this pathway, particularly in *BRAF* and *NRAS*. The role of *BRAF* and *NRAS* in melanoma development are discussed in

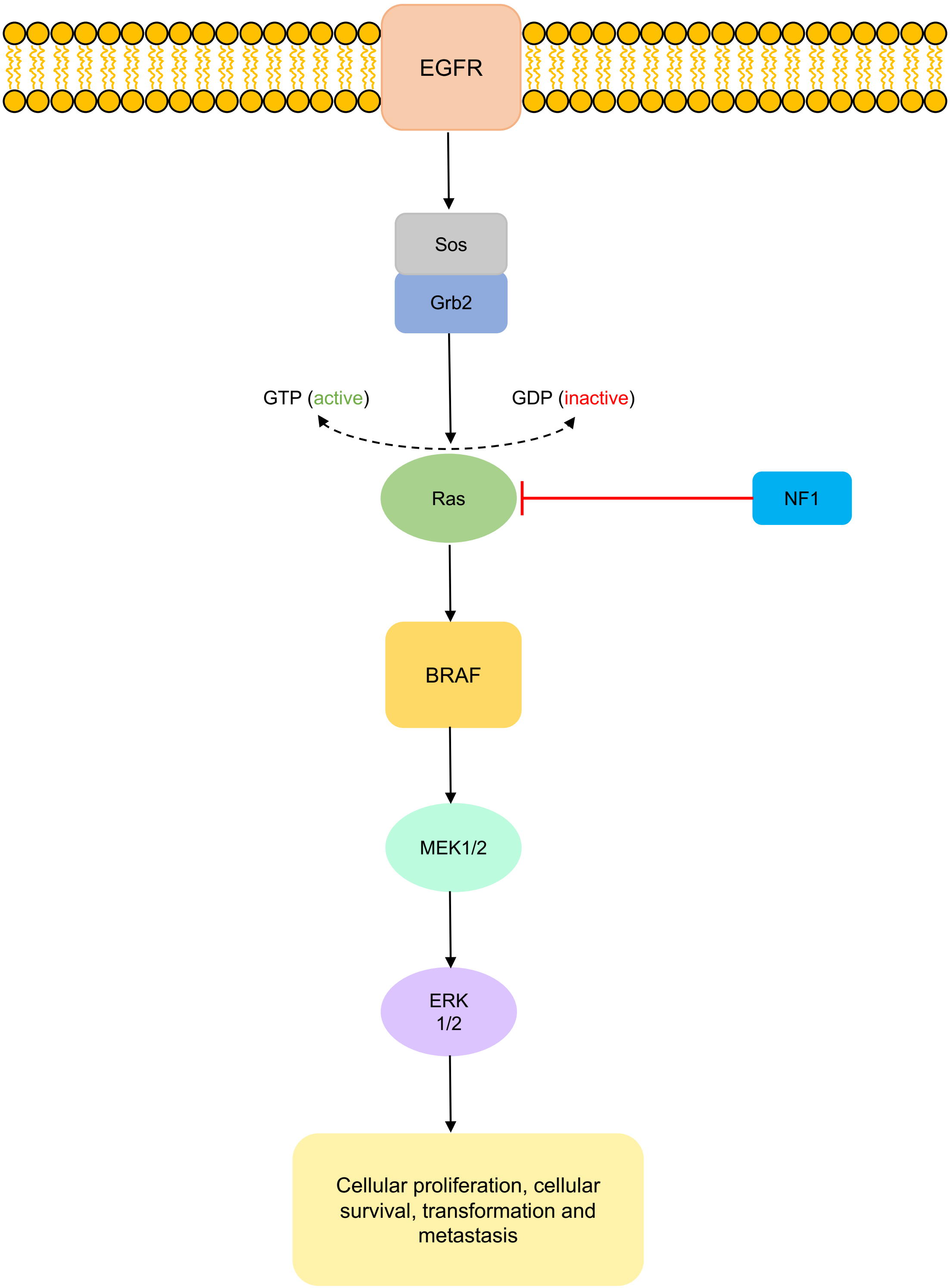

EGFR
Sos
Grb2
GTP (active)
GDP (inactive)
Ras
NF1
BRAF
MEK1/2
ERK
1/2
Cellular proliferation, cellular survival, transformation and metastasis

Figure 1.3: The MAPK (Ras-Raf-MEK-ERK) Pathway.

Sections 1.3.3.2 and 1.3.3.3 respectively. The importance of the MAPK pathway in cancer development has resulted in the development of multiple approaches for cancer therapy that target the key components of the pathway[66–70].

1.3.3.2 *BRAF* mutant melanoma

BRAF is a gene that encodes for a signal transduction protein kinase. It is part of the RAF family of genes and functions as an activator of the the MAPK (Ras-Raf-MEK-ERK) pathway. *BRAF* is phosphorylated and activated by RAS which in turn binds to and phosphorylates MEK1 and MEK2. Mutated *BRAF* is responsible for constitutive phosphorylation of MEK1 which continuously activates ERK and results in uncontrolled cellular growth. Activating *BRAF* mutations have been found in variety of human cancers such as cholangiocarcinoma (22%)[71], papillary thyroid cancer (69%)[72], colorectal cancer (12%)[73] and borderline ovarian cancer (28-48%)[74, 75]. However, cutaneous melanoma has the highest proportion of *BRAF* mutations with roughly 50% of cutaneous melanoma tumours estimated to have disruptive *BRAF* mutations. The vast majority of these mutations (over 90%) occur at the amino acid generated by the 600th codon which is a valine. V600E comprises of 90% of all mutations at codon 600 while V600K, V600R and V600D are other less frequent changes that occur at this position, although these frequencies change in different populations[19]. The V600E mutation results in constitutive activation of *BR*AF, leading to uncontrolled cellular growth[15]. This also prevents negative regulation and cellular senescence [76] and contributes to eventual metastasis[77]. *BRAF* mutations in melanoma occur more commonly in relatively younger patients with tumours in non-sun exposed areas. *BRAF* mutations rarely observed much in rarer types of melanoma such as mucosal and acral melanoma. *BRAF* mutation frequencies also differ between primary and metastatic melanomas with a range of 36 to 45% in primary melanomas and 42-55% in metastatic melanomas[78]. $BRAF^{V600E}$ mutations typically occur in younger patients(<55 years), occur at non-sun exposed areas of the body such as the trunk and the extremities, have moderate overall mutational burden and exist mutually exclusive with *NRAS* mutations[21, 79]. In the study conducted by TCGA, 75% of *BRAF* mutant patients were also identified as having *TERT* promoter mutations. The importance and relevance of both sporadic and germline *TERT* mutations, especially promoter mutations, in melanoma development is discussed in detail in Section 1.3.4.4. 91% of *BRAF* mutant samples harboured a UV-signature[20]. Multiple drugs functioning as kinase inhibitors have been designed specifically to target *BRAF* mutations such as vemurafenib, dabrafenib, and trametinib[80]. These drugs however have not proven to be completely effective; further treatment strategies are required for higher effectiveness.

1.3.3.3 *NRAS* mutant melanoma

A series of experiments on transforming retroviruses rats led to the identification of the Rat Sarcoma (RAS) family of genes. They were initially established as oncogenic viruses which cause formation of sarcomas in infected animals and had the potential to transform cells in culture. The Harvey murine sarcoma virus was identified in 1964 [81] with the discovery of the Kirsten murine sarcoma virus in following soon after in 1967[82]. Eventually, the origins of these viruses were traced back to the oncogenes that were responsible for them and these genes were named *HRAS* and *KRAS* respectively. In 1982, a third human *RAS* gene was identified in neuroblastoma-derived DNA and was therefore named as *NRAS*[83]. The three *RAS* family members, *NRAS*, *HRAS* and *KRAS*, have since been identified as being frequently mutated in human cancers, 20% of all tumours harbor activating mutations in one of their *RAS* genes. Mutations in *NRAS*, *KRAS*, or *HRAS* are known to be present in 20% (sometimes as high as 30%), 2%, and 1% respectively of all melanomas tested, making *NRAS* mutant melanoma the second highest subtype of cutaneous melanoma after *BRAF* mutant melanoma[84].

The most commonly reported hotspot mutation in *NRAS* which is observed in 80% of all *NRAS* mutations is a single nucleotide change that affects codon 61, changing a glutamine to a leucine (Q61L)[85]. Other common *NRAS* mutations include Q61R, Q61K, Q61H and G12R, G12D and G12A. *HRAS* mutations also largely affect codon 61 while *KRAS* mutations affected codon 12[86]. Mutations in position 61 are linked to reduced activity of the GTPase protein. This results in the RAS protein being stuck in its activated state and constantly activates Raf proteins. In the study by the Cancer Genome Atlas Network, 93.5% of RAS mutant melanomas were observed to exhibit UV signature[20]. This study also found copy number amplifications in *NRAS* that co-occurred along with *NRAS* mutations within the tumours. 72% of RAS subtype melanoma tumours were observed to have *TERT* promoter mutations. The typical patient with RAS mutant melanoma also tends to be older (>55 years of age) with higher chronic sun exposure than a patient with a *BRAF* mutant melanoma[79, 84]. *BRAF* mutations are more prevalent in benign nevi that are commonly present across the body with 80% of nevi carrying *BRAF* mutations compared to 14% carrying *NRAS* mutations[87]. Congenital nevi, however, have a much larger proportion of *NRAS* mutations; roughly 80% of congenital nevi have *NRAS* mutations[88]. *BRAF* mutations are rarely observed in congenital nevi. Previous attempts at targeted therapies for *NRAS* mutant melanomas by directly interacting with RAS have not been successful. A class of drugs were designed around farnesyltransferase inhibitors, aimed at preventing RAS modification[86]. However, these therapies exhibited severe off-target effects and were therefore not used. Phase 3 trials using binimetinib and dacarbazine were also applied on advanced *NRAS* mutant patients, binimetinib showing a better response rate but with no difference in overall survival[84].

1.3.3.4 *NF1* mutant melanoma

Neurofibromin 1 (*NF1*) is a gene that was first discovered in the context of studying Neurofibromatosis type I, a genetic disorder that results in the formation of red spots on the eye (Lisch nodules) and in benign skin tumours (neurofibromas). *NF1* was first identified as being in the long arm of chromosome 17 through several concurrent studies in 1989. A precise mapping of *NF1* to a specific genetic locus of 17q11.2 soon followed in 1994. *NF1* encodes a large protein with multiple functional domains, including the GTPase activating protein related domain. This well known domain is responsible for the negative regulating of RAS, which is performed by switching the active state of RAS (Ras-GTP) to its inactive state (Ras-GDP) through hydrolysis. *NF1* functions as a tumour suppressor which is why loss of heterozygosity is also commonly observed in *NF1* mutant tumours through somatic mutations. Unlike *BRAF* and *NRAS*, *NF1* patients do not have a specific phenotypic characterisation, with wide variability being commonly observed. Large-scale next-generation sequencing studies have identified *NF1* as a commonly mutated driver gene in melanoma; 12 to 18% of all melanomas have *NF1* mutations. Although *NF1* mutations sometimes occur alongside *BRAF*/RAS mutations, they are more common in patients with no *BRAF* or *NRAS* mutations. While *NF1* mutations co-occur with hotspot RAS mutations, they do not co-occur with hotspot *BRAF* mutations. *NF1* mutations are also observed at a higher frequency in patients with desmoplastic melanoma, with 45-93% of patients having *NF1* mutations. *NF1* has been found to have somatic mutations in other types of cancers in addition to cutaneous melanoma. 40% of malignant peripheral nerve sheath tumours have *NF1* mutations, with lower frequencies of NF1 mutations also being observed in acute lymphoblastic leukaemia, glioblastoma, lung adenocarcinoma and pancreatic carcinoma. The study by the Cancer Genome Atlas Network focussing on the different subtypes of melanoma found that the average patient with *NF1* mutations was older(>55 years of age), 93% of whom exhibited UV signature and 83.3% of whom carried *TERT* promoter mutations. There was no significant observations with copy-number alterations[20]. While there have been no therapies that target *NF1* directly in melanoma due to its recent discovery, there are other inhibitors targeting the Ras-Raf-MEK-ERK pathway which is regulated by *NF1*. Melanomas with wild-type *BRAF*/*NRAS* are highly sensitive to the MEK inhibitor trametinib, with *NF1* protein expression also being sensitive to MEK inhibition, potentially implicating trametinib as a useful therapy for *NF1* patients[89]. It is also predicted that *NF1* mutations play a role in *BRAF* targeted therapy by increasing resistance to *BRAF* inhibitors. *BRAF* mutant murine tumours that are also mutated in *NF1* are identified as being resistant to *BRAF* inhibitors. These tumours are however sensitive to combinatorial inhibition of MAPK/ERK and mTOR pathways[90]. Increased understanding of the role of *NF1* in cutaneous melanoma onset offers a potential novel target for new therapies which aim at inhibiting the Ras-Raf-MEK-ERK pathway.

1.3.3.5 Triple-Wild type melanoma

Cutaneous melanoma patients that do not have hotspot or any mutations in the tumour in either *BRAF*, *NRAS* or *NF1* are collectively called as triple-wild type melanoma patients[20]. Information on triple-wildtype melanoma is relatively sparse due to low frequency of cases and non-specific mechanism of melanoma development. The study by the Cancer Genome Atlas Network[20] identified several driver mutations in genes that weren't *BRAF*, *RAS* or *NF1* within the triple-wildtype subtype. As compared to the three main subtypes, triple-wildtype tumours also exhibited a much lower rate of UV signature with only 30% of tumours displaying the signature. Similarly, only 6.7% of triple wild type patients carried *TERT* promoter mutations. There was a significant increase in copy-number alterations and structural changes as compared to the other subtypes. Melanoma onset and progression in triple-wildtype patients may be driven by one of several other driver genes; therapeutic treatment of this subtype is therefore not as straightforward.

1.3.4 Germline familial melanoma genes and their clinical impact

1.3.4.1 The role of GWAS in melanoma research

The most common approach used in the investigation of large datasets of genomic data is to use a genome wide association study. These studies are used to help identify association of specific genetic SNPs or loci with a disease by comparing the genetic mutational burden across thousands of affected and unaffected individuals. While GWAS has been used as a tool quite significantly in a lot of other disorders such as type 2 diabetes[91], its use in melanoma research has been limited until recently. An initial GWAS performed in 2009 identified 3 markers involved in pigmentation and nevi formation including *TYR, ASIP* and *MC1R[92]. MTAP* and *PLA2G6* were also identified in this study and verified through follow up meta-analyses of multiple melanoma related GWAS studies in 2015[51] and 2020 [93]. *MTAP* is a gene that is frequently disrupted in cancers due to its proximity to *CDKN2A [94]* while compound mutations in *PLA2G6* lead to early-onset Parkinsons [95], a disease closely associated with melanoma[96]. The discovery of *PLA2G6* as a melanoma marker through GWAS showed that neurological diseases are linked to melanoma risk. Another GWAS study focussing on loci related to tanning response to sun exposure helped identify another 14 novel loci related to this risk phenotype[97]. A GWAS study focussing on hair colour in Europeans determined 124 loci associated with hair colour, another low risk phenotype for melanoma[98]. Such GWAS studies have therefore expanded our knowledge of both low risk and high risk genetic markers in sporadic and familial melanoma. The number of risk markers for melanoma also increased from 3 in 2009 to 68 SNPS in 54 distinct loci in 2020 including genes in DNA damage repair pathways, telomerase pathways and pigmentation pathways, all of which have subsequently yielded novel candidate genes. Improvements in GWAS studies and inclusion of loci from additional risk phenotypes such as naevi count, hair colour and sun exposure will also improve the accuracy of polygenic risk score estimations in the future. Thus, while GWAS analyses might not directly translate into explaining the genetic origins of a disease, they play a significant role in providing insight into the mechanisms involved in disease onset.

1.3.4.2 Cyclin-dependent kinase inhibitor 2A *(CDKN2A)*

A major proportion of familial melanoma cases have a cause attributed to germline mutations in Cyclin-dependent kinase inhibitor 2A (*CDKN2A*). It was linked to melanoma for the first time through linkage analysis in 1994[25, 26]. It is estimated that it is responsible for up to 40% of familial melanoma cases with percentages varying from 20% to 57% in different parts of the world[27]. The *CDKN2A* gene lies on chromosome 9, and encodes two separate tumour suppressor protein products: p16INK4A and p14ARF (ARF = alternative reading frame). These two proteins arise from alternate splicing of *CDKN2A*, as shown in Figure 1.4. While both protein products share exons 2 and 3, they do not have a common amino acid sequence due to being encoded in an alternate reading frame.

The p16INK4a protein regulates the cell cycle by inhibiting the activity of *CDK4* and *CDK6*, two cyclin dependent kinases which are responsible for phosphorylation of the Retinoblastoma protein (RB)[100]. By controlling this activity, *CDKN2A* prevents the phosphorylation of RB and arrests the cell cycle at the G1-S phase. Mutant p16INK4A leads to early phosphorylation of RB which leads to improper progression into the S-phase of the cell cycle[101]. By contrast, p14ARF inhibits the binding of *HDM2* to the p53 tumour suppressor, thereby controlling the negative regulation of p53[102]. Both p53 and p16INK4A are known to play vital roles in cellular damage response and cellular senescence, both of which are integral pathways in cancer onset and progression. Germline in both p16INK4A([103]) and p14ARF([104]) have been shown to lead to melanoma; such mutations significantly increase the risk of melanoma development with a penetrance of up to 80% in the 8th decade of life. However, it is to be noted that *CDKN2A* mutations are extremely rare in population based melanoma cases. The

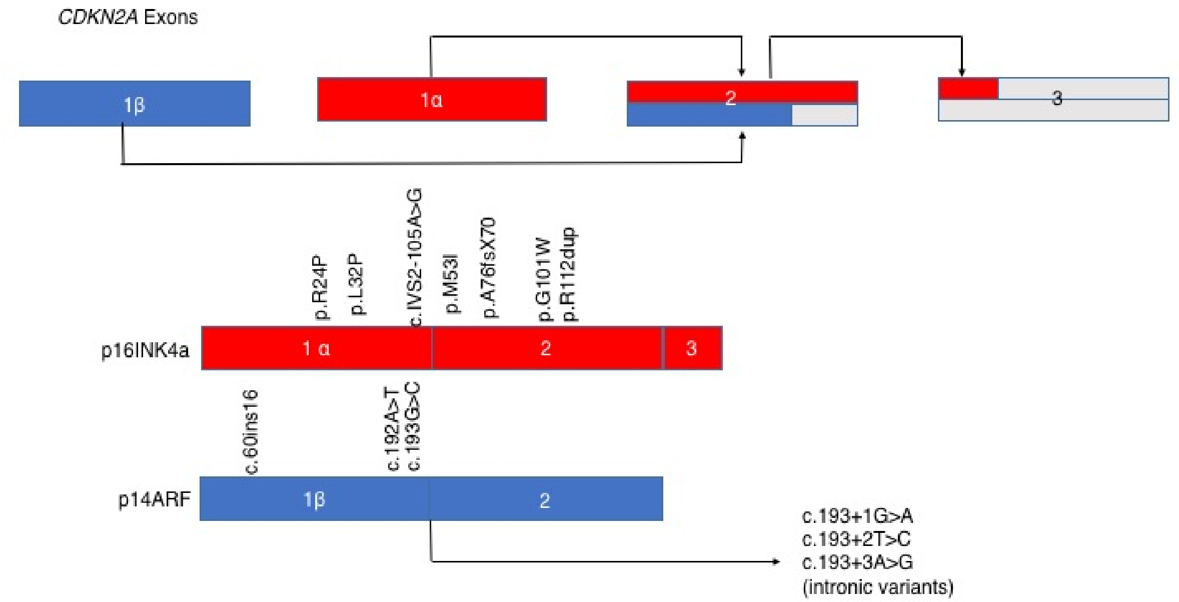


Figure 1.4: Creation of p16INK4A and p14ARF through alternate splicing of *CDKN2A* with the locations of founder mutations in melanoma in each protein also given. Adapted from [99].

collation of multiple studies comparing germline *CDKN2A* mutation carriers with non-carriers has also determined that *CDKN2A* mutation carriers have a lower median age of melanoma diagnosis[27, 105–108].

An increased incidence of pancreatic cancer has been observed in patients from familial melanoma pedigrees germline mutations in *CDKN2A* across several studies*[103, 109–113]*. In addition to giving rise to melanoma, somatic *CDKN2A* variants and disruption have also been observed in a several other types of cancer including oral pediatric lymphoblastic leukaemia, oral squamous cell carcinoma, head and neck squamous cell carcinoma, colon cancer and bladder cancer, indicating that *CDKN2A* has a role to play in general tumour formation and cancer development[114–122]. Germline *CDKN2A* mutations also resulted in increased risk for pancreatic, lung, head and neck cancers[123].

1.3.4.3 Cyclin dependent Kinase 4 *(CDK4)*

Once *CDKN2A* was discovered as a familial melanoma gene, efforts were focussed on discovering other potential familial genes by determining which other genes interact with it. This led to the discovery of the next familial melanoma gene, cyclin dependent Kinase 4 (*CDK4*). *CDK4* plays a key role in controlling cell cycle progression as it is responsible for the phosphorylation of RB. Mutations in *CDK4* which inhibit the activity of p16INK4A - thereby leading to early phosphorylation of RB - have also shown to be causative of melanoma. Due to the similar mechanisms involved in *CDK4* and *CDKN2A* mutations, they result in identical phenotypes, namely, early onset CMM, multiple primary melanomas and distinct nevi[124]. This complicates the process of distinguishing between *CDKN2A* and *CDK4* mutations as the cause of melanoma; families with melanoma that test negative for *CDKN2A* mutations should still be tested for mutations in *CDK4*. All *CDK4* driver mutations identified to date affect the 24th codon[124]. This supports the idea that the arginine amino acid generated by this codon normally is essential for the binding of the p16 tumour suppressor to *CDK4* which in turn prevents the phosphorylation of RB1. Alteration of this amino acid would therefore prevent *CDK4* inactivation. However, while *CDK4* mutations are necessary to understand the role of the p16 pathway in melanoma development, they are extremely rare.

1.3.4.4 Telomere maintenance pathway

Telomeres are terminal DNA structures at the ends of chromosomes responsible for genomic stability and integrity[125]. In humans, telomeres comprise of 9-15 kb double stranded repeats of a "TTAGGG" sequence which ends with a single stranded overhang called as the G-tail or G-overhang[126]. Cells that lack telomere length maintenance mechanisms progressively lose telomeric sequence with every round of cell division. This continues until the length of the protective telomere ends become critically short at which point they stop protecting the cells from DNA damage repair, leading to replicative senescence [126]. In this way, telomere length helps provide a mechanism for controlling the replicative lifespan of cells. Telomere replication, regulation and maintenance are primarily controlled through two protein complexes : the telomerase complex and the shelterin complex[127]. This section discusses the functions of these complexes, the roles of specific genes within these complexes in familial melanoma onset and previously identified germline mutations that implicate telomere dysregulation as a mechanism for melanoma development.

i) Telomerase and telomerase reverse transcriptase *(TERT)*

The maintenance of telomere length in germline cells and stem cells is controlled through the telomerase complex. Telomerase consists of two core subunits: telomerase reverse transcriptase (*TERT*) and telomerase RNA component (*TERC*)[127]. These components are responsible for telomerase extension through the addition of multiple "TTAGGG" repeats at the ends of the chromosomes using the G-overhang as the substrate[126]. *TERC* generates the RNA component of telomerase while the addition of telomeric repeats is controlled by the highly regulated and conserved *TERT*. Additional accessory proteins including Dyskerin, GAR1, NHP2 and NOP10 are recruited to promote TERC accumulation[128]. TERC provides the template for DNA synthesis and is 451 bp long in *homo sapiens*. It also binds to TERT, which acts as the active site for catalysis and binds to the telomeric DNA[129]. The structure of the telomerase complex is shown in Figure 1.6.

Telomere length as a risk phenotype for melanoma was first reported in 2007 where increased telomere lengths were observed in circulating white blood cells which were associated with higher naevi count, another risk phenotype for melanoma[130]. In 2013, a novel promoter mutation 57 bp upstream of the translation start site which created an ETS binding motif was observed to be segregating with the disease in a German pedigree[131]. This was also shown to double the transcription rate of *TERT*. Members of the pedigree with this mutation had also developed multiple other types of cancer including ovarian cancer, renal cell carcinoma and bladder cancer, implying that *TERT* promoter mutations might play a role in general cancer development as opposed to being specific to melanoma. However, follow-up studies showed that germline *TERT* promoter mutations were sparse in the context of familial melanoma[132]. Additional recurrent somatic mutations affecting the *TERT* promoter at positions 124bp and 146 bp upstream of the translation start site have since been identified in multiple melanoma cell lines though whole genome sequencing [133]. These mutations, along with the germline promoter mutation, are shown in Figure 1.5. Somatic *TERT* promoter mutations have also been identified in an array of other cancers including bladder cancer[134], glioblastoma[135], thyroid cancer[136], mesothelioma[137], hepatocellular carcinoma[138] and squamous cell carcinoma[139]. The confirmation of telomere length as a risk phenotype for melanoma after initial studies focussing on precursor phenotypes indicate that this is a cogent approach in the discovery of novel cancer genes and pathways.

ii) Protection of telomeres protein 1 *(POT1)* and the shelterin complex

The shelterin complex are a set of six proteins that protect telomeres from degradation. These proteins regulate the interactions of the telomeres with the telomerase complex, help protect telomeres from the DNA damage repair pathway and maintain genomic integrity[141]. The constituent components of the shelterin complex include the following six core proteins : adrenocortical dysplasia protein homolog *(ACD)*, protection of telomeres 1 *(POT1), TERF2*-interacting protein 1 *(TERF2IP,* also called *Rap1),* telomeric repeat-binding factors 1 and 2 *(TERF1* and *TERF2)* and *TERF1*-interacting protein 2 *(TINF2)* [127].

TERF1 and *TERF2* are responsible for the production of double-stranded DNA binding

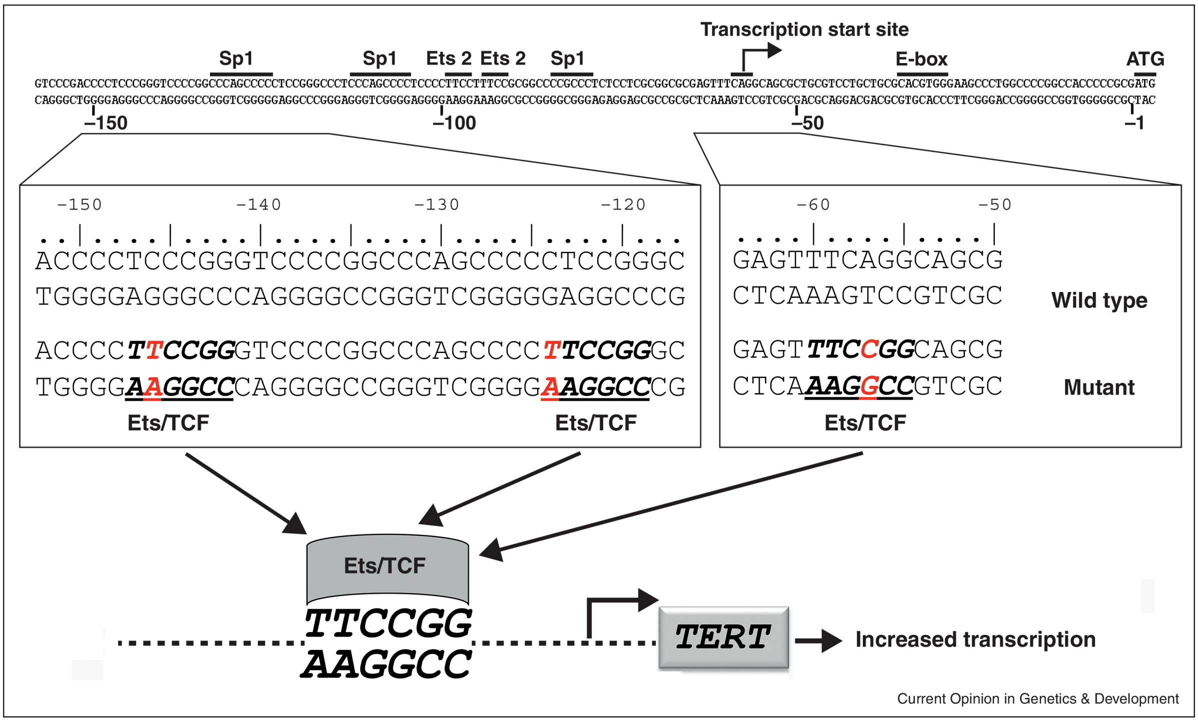

Transcription start site
Sp1
Sp1
Ets 2
Ets 2
Sp1
E-box
ATG
GTCCCGACCCCTCCCGGGTCCCCGGCCCAGCCCCCTCCGGGCCCTCCCAGCCCCTCCCCTTCCTTTCCGCGGCCCCGCCCTCTCCTCGCGGCGCGAGTTTCAGGCAGCGCTGCGTCCTGCTGCGCACGTGGGAAGCCCTGGCCCCGGCCACCCCCGCGATG
CAGGGCTGGGGAGGGCCCAGGGGCCGGGTCGGGGGAGGCCCGGGAGGGTCGGGGAGGGGAAGGAAAGGCGCCGGGGCGGGAGAGGAGCGCCGCGCTCAAAGTCCGTCGCGACGCAGGACGACGCGTGCACCCTTCGGGACCGGGGCCGGTGGGGGCGCTAC
–150
–100
–50
–1
-150
-140
-130
-120
ACCCCTCCCGGGTCCCCGGCCCAGCCCCCTCCGGGC
TGGGGAGGGCCCAGGGGCCGGGTCGGGGGAGGCCCG
ACCCCTTCCGGGTCCCCGGCCCAGCCCCTTCCGGGC
TGGGGAAGGCCCAGGGGCCGGGTCGGGGAAGGCCCG
Ets/TCF
Ets/TCF
-60
-50
GAGTTTCAGGCAGCG
CTCAAAGTCCGTCGC
Wild type
GAGTTTCCGGCAGCG
CTCAAAGGCCGTCGC
Mutant
Ets/TCF
Ets/TCF
TTCCGG
AAGGCC
TERT
Increased transcription
Current Opinion in Genetics & Development

Figure 1.5: Previously identified *TERT* promoter mutations in sporadic and familial melanoma. Reused with permission from [140].

proteins that identify and bind to the telomeric repeats. *POT1* is the most conserved component of the shelterin complex which creates a protein that binds to single stranded telomeric DNA; *ACD* helps in the recruitment of POT1 to the telomere by binding to POT1 and creating a sub-complex[127, 141]. *TERF2IP* does not directly interact with telomeric repeats and instead interacts with *TERF2*[142]. The shelterin complex is bound together as a single entity through *TINF2*. *TINF2* constructs a protein which binds to TERF1, TERF2 and the ACD/POT1 complex[127, 141]. The shelterin complex helps in the generation of a protective structure at the end of the telomere called as telomeric loop or T-loop. The presence of this loop provides a distinct protective cap to telomeres which distinguishes telomeres from double-strand breaks and protects them from the DNA damage repair pathway[126]. The arrangement of the different proteins within the shelterin complex and their interactions with each other resulting in the formation of the T-loop are shown in Figure 1.6.

Large-scale exome sequencing of members belonging to 105 pedigrees from Australia, The UK and The Netherlands was performed for identification of novel germline variants[143]. Three missense mutations (Y89C, Q94E, R273L) affecting the OB domain of *POT1* and a splice set variant in *POT1* were identified from this dataset. These mutations resulted in longer telomeres, which predisposed the individuals with the mutation to developing cuta-

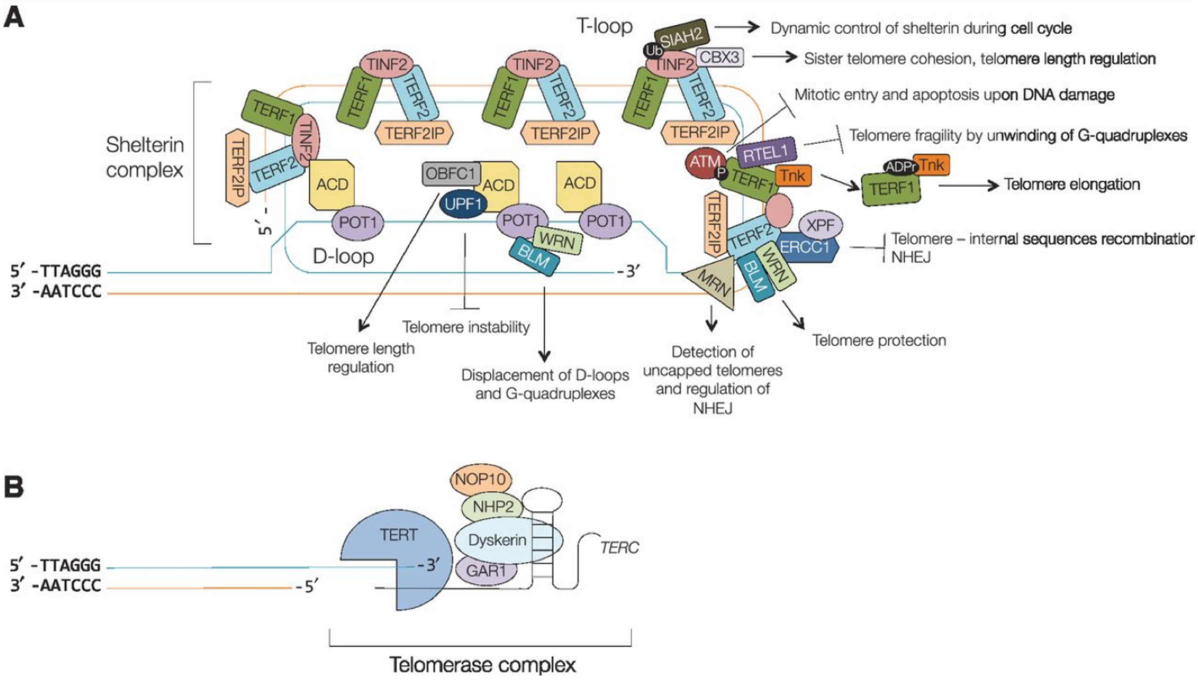

A
T-loop
SIAH2
Ub
TINF2
CBX3
Dynamic control of shelterin during cell cycle
Sister telomere cohesion, telomere length regulation
Mitotic entry and apoptosis upon DNA damage
Telomere fragility by unwinding of G-quadruplexes
Telomere elongation
Telomere – internal sequences recombinatior
NHEJ
Shelterin complex
TERF1
TINF2
TERF2
TERF2IP
ACD
POT1
OBFC1
UPF1
WRN
BLM
ATM
P
RTEL1
Tnk
ADPr
XPF
ERCC1
MRN
5′ -TTAGGG
3′ -AATCCC
5′-
-3′
D-loop
Telomere length regulation
Telomere instability
Displacement of D-loops and G-quadruplexes
Detection of uncapped telomeres and regulation of NHEJ
Telomere protection
B
NOP10
NHP2
Dyskerin
GAR1
TERT
TERC
-3′
-5′
Telomerase complex

Figure 1.6: a) The structure of the shelterin complex showing the interactions between the different components (including the association of *ACD* with *POT1* and *TERF2IP* with *TERF2*), alongside their function in telomerase maintenance and T-loop formation.b) Structure of the telomerase complex showing the interaction between *TERC*, *TERT* and the other proteins necessary for the addition of telomeric repeats. Reproduced with permission from [127].

neous melanoma[143]. Concurrently, a founder mutation and additional rare variants in *POT1* were also identified from another whole-exome sequencing study[144]. With the discovery of inactivating mutations in *POT1*, the remaining components of the shelterin complex were also investigated for their possible role in the melanoma progression. Additional exome and whole genome sequencing of 510 affected families resulted in the discovery of 6 families with mutations in *ACD* and 4 families with mutations in *TERF2IP*[145]. *ACD* binds with *POT1* and interacts with the telomerase complex. It was observed that when the binding domains of *ACD* and *POT1* are mutated, they fail to form a functioning shelterin complex, which leads to increased telomere length due to an active telomerase complex. The mutations in *TERF2IP* were found to disrupt its binding capacity with *TERF2* which in turn affects the repair of the double strand break at the telomere.

Thus, the presence of mutations in genes of the shelterin complex which are associated with familial melanoma strengthen the relevance of telomere dysregulation as a mechanism in melanoma development.

1.3.4.5 BRCA1-associated protein-1 *(BAP1)*

A heterodimer of breast cancer 1(*BRCA1*) and *BRCA1*-associated RING domain (*BARD1*), which has E3 ubiquitin ligase activity, controls the DNA damage repair pathway. *BRCA1*-associated protein-1 (*BAP1*) acts as a deubiquitination enzyme and helps in deubiquitinating *BARD1* and regulating the E3 ligase activity of this complex*[146]*. Inhibition of *BAP1* leads to impaired DNA damage repair process and results in S-phase retardation. *BAP1* also interacts with the Yin Yang1 (YY1) transcription factor to control the transcription of genes involved in cellular proliferation*[146]*.

Inactivating somatic mutations were first identified in *BAP1* in metastasizing uveal melanoma tumours[147]. Additional studies have associated germline *BAP1* mutations with malignant mesothelioma and with distinct morphological neoplasms related to melanocytic tumours leading to cutaneous melanoma and renal cell carcinomas[148–150]. While missense mutations do occur, they were comparatively rare compared to nonsense germline *BAP1* mutations. An extensive study of all known nonsense and missense germline *BAP1* mutations identified 104 unique nonsense variants and 36 unique missense variants from 181 families[151]. In 2013, 15% of *BAP1* mutation carriers developed cutaneous melanoma*[146]* but the larger study in 2018 identified a 24% occurrence of cutaneous melanoma in probands and a 12% occurrence of cutaneous melanoma in non-proband variant carriers[151]. This implies that *BAP1* is a medium-penetrance risk gene for cutaneous melanoma. A list of all truncating germline variants in *BAP1* are shown in Figure 1.7.

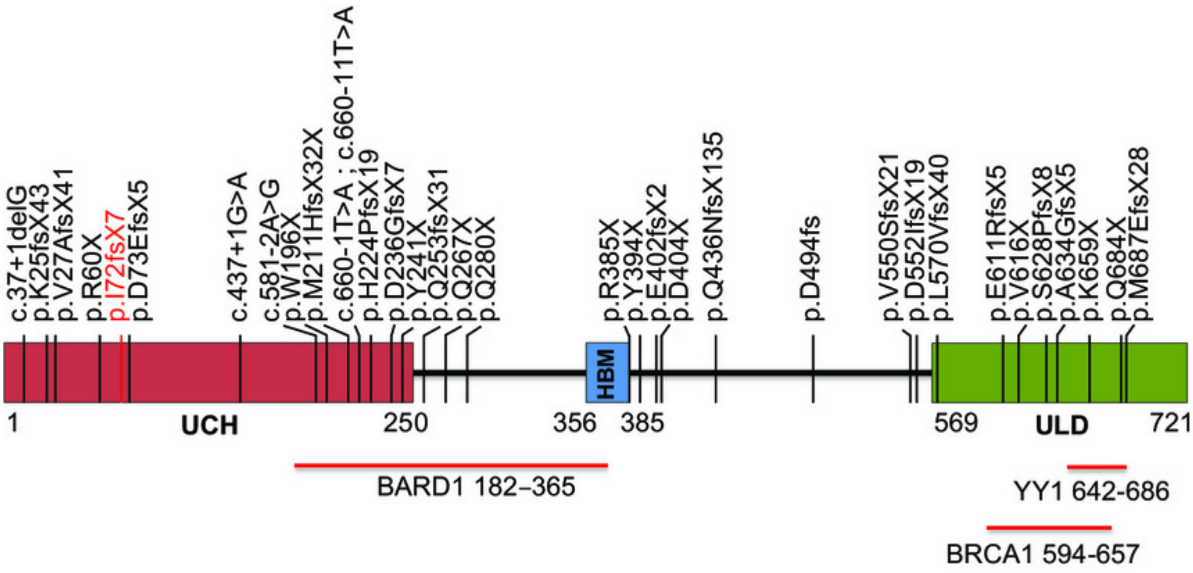


Figure 1.7: Germline truncating *BAP1* mutations along with the protein domains of *BAP1*. The binding sites of *BARD1*, *BRCA1* and *YY1* are shown in red. Reused with permission from [99].

A large scale population study was performed in 2017 by O'Shea et al. to identify the frequency of germline *BAP1* mutations in sporadic melanoma[152]. 1,977 melanoma cases and 754 controls were sequenced for this study with only 30 mutations identified in *BAP1*, shown in Figure 1.8. Only 2 of these were truncating mutations, indicating that germline *BAP1* mutations are extremely rare in sporadic melanoma and strengthening the claim that *BAP1* is a medium-penetrance risk gene for cutaneous melanoma.

1.4 Classification of melanoma

1.4.1 Melanoma subtypes

Wallace Clark not only defined a staging system for the progression of a tumour from a benign nevus malignancy, but had also previously established three specific malignant melanoma subtypes based on the type of tumour : superficial spreading melanoma (SSM), nodular melanoma (NM) and lentigo maligna melanoma (LM)[9]. A fourth subtype of melanoma called acral lentiginous melanoma (ALM) was later identified[154]. Additional, rare melanoma subtypes were identified following this. The characteristics of these subtypes were described by Anand Rotte and Madhuri Bhandaru[155]; these characteristics are summarised here:

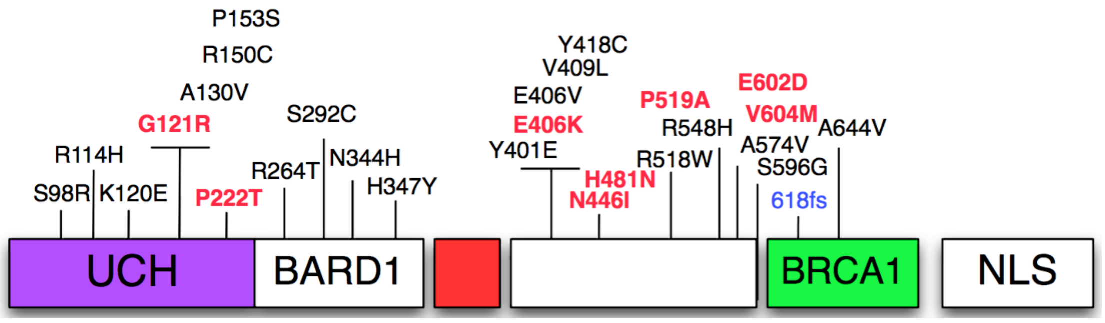


Figure 1.8: List of germline mutations in *BAP1* observed in sporadic melanoma. The nonsense mutations are highlighted in red. Reproduced with permission from [152].

1.4.1.1 Superficial spreading melanoma

This is the most common type of melanoma, accounting for roughly 60% of all melanoma cases. Patients with SSM are usually less than 60 years old (with a median age of 55) and have a high rate of *BRAF*^{V600E} mutations. SSM tumours are not caused due to chronic sun-damage and are usually located in non-sun exposed areas of the body. The tumour appears as a flat, discoloured region which eventually enlarges radially. When it invades the dermal region of the skin, it forms an elevated lump in the region.

1.4.1.2 Nodular melanoma

This is the second most common subtype of melanoma and is observed in 15-30% of all melanoma cases. It is usually observed in the sun-exposed areas (head and neck) of older patients (>60 years of age) with high sun-exposure. NM is the quickest growing subtype in terms of tumour depth and is linked to poor prognosis due to late detection. NM tumours are rigid, symmetric and do not show a lot of colour variation. They also do not tend to change colour on growth, which sometimes contributes to its late detection. On occasions, these tumours ulcerate and potentially start bleeding.

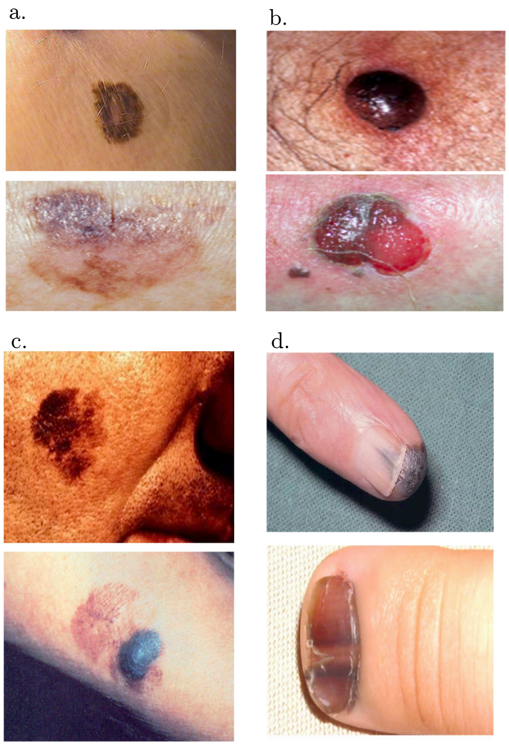


Figure 1.9: The different types of tumours observed in common melanoma subtypes is depicted here. These subtypes are a) Superficial spreading melanoma b) Nodular melanoma c) Lentigo maligna melanoma d) Acral lentiginous melanoma. Images obtained from [153]. Images downloaded and reused under the under Creative Commons Attribution (CC BY) license (http://creativecommons.org/ licenses/by/4.0/).

1.4.1.3 Lentigo maligna melanoma

This subtype refers to the development of melanomas in sun-damaged skin caused by chronic sun exposure. As a result, these are commonly found on parts of the body with high sun-exposure and in older people (>60 years of age). LM accounts for 4-10% of all melanoma cases. The tumour initially presents as a black coloured region of discolouration that is flat and has rhomboidal structures. tumour growth is lentiginous in the epidermal region.

1.4.1.4 Acral lentiginous melanoma

This subtype is responsible for 2-3% of all melanoma cases but is the most common type of melanoma in non-Caucasian populations., ranging from 9% in Hispanic Whites to 36% in people of African descent. LM, similar to ALM, is also diagnosed over the age of 60 on average. It is observed in the extremities of the limbs such as the palms, soles or under the nails. The development of ALM in such irregular sites leads to a later diagnosis as compared to the other subtypes and therefore has a worse survival rate as well. These lesions have specific ridge like patterns and occasionally exhibit symmetric arrangement of globules under dermoscopic observation.

1.4.1.5 Rare melanoma subtypes

In additional to these primary subtypes, there are several extremely rare subtypes of melanoma that are seen in around 1% of all melanoma cases. Some of these subtypes are:

- Desmoplastic melanoma: Rare subtype of melanoma more frequently seen in men. Presents as scar like nodules and lacks prominent clinical features, preventing early detection.

- Nevoid melanoma: Rare subtype of nodular melanoma. tumour presents as a common or Spitz nevus.

- Verrucous melanoma: Rare subtype of cutaneous melanoma often mistaken for seborrhoeic keratosis . More frequent in women. tumour presents as a wart covered lesion and usually develop on the extremities.

- Mucosal melanoma: Comprise of less than 1% of all melanoma cases; present in mucosal surfaces of the body such as nasal passages, sinuses, vagina, bowel, urethra and anus.

- Giant congenital melanocytic nevus: Considered as a precursor for malignant melanoma, consist of melanocytic lesions present at birth that grow in size over time and may eventually progress to melanoma.

Such diversity in melanoma characterisation indicate that cancer progression and development through melanocytes can take one of several paths depending on several factors including gender, age, ethnicity and sun exposure. Such a complex process demands a clear and well-defined system of stratification to categorize patients into distinct groups which would enable better disease treatment and improve survival.

1.4.2 Cancer staging systems

1.4.2.1 A history of staging systems for cancer

Initial diagnoses of melanoma up to the 1950's were made by identifying distinct attributes on the skin, at which point they were often at an advanced stage with a poor prognosis. To improve patient survival and treatments, different staging systems have been developed over time to characterise and stratify patients based on the nature of their tumour and cancer progression. These staging systems have also improved along with the knowledge of the disease and have become more accurate over time.

The first major observation related to prognosis and tumour types was made in 1953 by Allen and Spitz where they had reported that melanoma tumours with higher depth had worse survival[156]. The first attempt at developing a stage system for melanoma based on the primary melanomas was done by Petersen in 1962 where patients were classified based on what stage of the dermis had been invaded by the tumour[157]. Further studies confirmed that there were three distinct groups of patients based on primary tumour, regional lymph metastasis and distant metastasis status of the patients. This led to another three stage system proposed in 1964 by Mcneer and Dasgupta where patients were classified into one of three categories: primary tumour with no metastasis, metastasis confined to regional lymph nodes and multiple, distant metastases[158].

An additional staging system based on tumour depth was established by Wallace Clark in 1969 which led to the Clark levels staging of melanoma invasion[9], which was also independently verified in 1970[159]. Based on this system, cutaneous melanoma was subdivided into five categories:

1. Level I : tumours are restricted to upper membrane of the epidermis, also called as in-situ melanoma.
2. Level II : tumours have extended from basement membrane to the papillary dermis but have not extended into the reticular dermis.
3. Level III : tumours have bridged the interface between the papillary and reticular dermis.
4. Level IV : tumours have completely extended into the reticular dermis.
5. Level V: tumours have invaded the subcutaneous tissue.

As mentioned in Section 1.2, Alexander Breslow came up with Breslow's thickness as a measurement parameter for tumour thickness at the same time[10]. This was found to be more effective as a prognostic measure and has since been incorporated into almost all future major melanoma staging systems. The late 1970's saw melanoma staging systems based on the primary tumours, lymph nodes and metastasis status which incorporated both the Clarks levels and Breslow thickness. These systems were created by the American Joint Committee on Cancer and the Union for International Cancer Control; these were eventually merged into a single

TNM staging system. This system has been regularly updated with multiple editions across the years with several subcategories and subgroups included based on analysis of thousands of patients to better classify different patients and is the current standard.

1.5 Genetic testing and therapies for familial melanoma

Unlike somatic mutations, germline variations affect and are present in all cells of the body. By testing for specific predominant germline mutations, we can estimate a risk for predisposition to familial melanoma. This process is referred to as genetic testing or gene panel testing. The risk of predisposition is higher if there are multiple family members affected with familial melanoma or if an individual in the family has had multiple primary melanomas as this suggests an underlying germline genetic cause over a sporadic mutation. The primary genes that were tested for familial melanoma were *CDKN2A*(p16) and *CDK4*. Recent improvements to multi-gene panel testing include the addition of several other genes of relevance to cancer development. The current NHS cancer gene panel includes 156 genes; the familial m elanoma gene panel amongst these genes include the aforementioned *CDKN2A* and *CDK4* but now also include other familial melanoma driver genes such as *BAP1, BRCA2, POLE, POT1* and *TERT*. Genetic testing is also free in the NHS if there are more than 2 related individuals with melanoma, if a patient has multiple primary melanoms or if melanoma and pancreatic cancer exist in the same pedigree. The addition of knowledge from sequencing studies such as this project regarding the variants in other medium to low penetrance genes in the development of familial melanoma will further refine the quality of genetic testing for familial melanoma in the future. The quality and access to genetic testing varies significantly from place to place; additional education is required for dermatologists and oncologists to refer patients to a clinical geneticist and genetic testing when required.

The primary method of treatment for stage 1 and stage 2 melanomas is surgical excision, where the melanoma is removed along with a small area of the skin around it. Patients are continued to be monitored for a few years to ensure that the melanoma does not return. Stage 3 and 4 melanomas are usually treated with a combination of immunotherapy and targeted gene therapy. Targeted gene therapy involves the use of drugs including dabrafenib, vemurafenib and trametinib in the treatment of specific types of melanoma. Particularly, dabrafenib and vemurafenib are used for the treatment of cases with *BRAFV600E* mutations in sporadic cases[80], while trametinib, which is a MEK inhibitor is used for *NRAS* and *NF1* affected melanoma cases[89].

Immunotherapy involves the use of drugs that target specific components of the immune complex such as cytotoxic T lymphocyte associated antigen 4 (CTLA4*)* and programmed death ligand 1 (PDL1). Cancer cells produce antigens which are detected by the immune system and help identify them. Components of the immune system called as cytotoxic T lymphocytes (CTLs) in the lymph node target these cancer cells based on the antigens and kill them. However, additional inhibitory signals are sometimes created by the dendritic cells that act as a checkpoint and are aimed at preventing the immune system from attacking the body. These inhibitory signals are detected by CTLA4 which are present on CTLs and this results in the cytotoxic reaction of the CTLs being deactivated, resulting in the growth of the cancer cells. Drugs such as ipilimumab can bind and block the function of CTLA4, thus allowing CTLs to function as normal and kill cancer cells[160]. Another similar mechanism is the role of T cells in targeting cancer cells. The surface of T cells contain a protein called programmed cell death 1 (PD1). In order to prevent the T cells from targeting normal cells, a protective protein called PDL1 binds to PD1 to inhibit T cells. However, some cancer cells can also produce PDL1 which prevents the T cells from killing these cancer cells. Nivolumab is a drug that works on this mechanism and obstructs PDL1 from binding to PD1, thereby activating T cells[161].

The presence of germline variants usually has no bearing on the treatment of the disease as they have access to either targeted gene therapy or immunotherapy depending on the presence or absence of *BRAF* mutations. However, patients with germline mutations do tend to have better survival compared to sporadic melanoma cases due to the less aggressive nature of the tumours and increased surveillance of the disease.

1.6 Proposed approaches of sequence analysis undertaken for the identification of novel genes

A variety of approaches ranging from linkage analysis and positional cloning to whole-exome sequencing have helped identify several high penetrance familial melanoma genes including *CDKN2A*, *CDK4, BAP1*, *POT1* and *TERT*. However, germline variants in these genes are jointly responsible for up to 40% of all known familial melanoma cases, leaving the remaining with an unexplained genetic cause. Such cases can be explained with one of three alternatives:

- They have a mutation in a yet-to-be-discovered high penetrance gene.
- They have a mutation in a high penetrance gene but in a region that is still unexplored such as transcription factor binding sites or the 5’ UTR.

- They have a high burden of mutations in several low penetrance genes which cumulatively lead to melanoma development.

- They have an increased risk of cancer due to a combination of genetic and epigenetic factors.

With respect to the pathways involved in melanoma development, most of the genes identified so far are involved in cell cycle regulation (*CDKN2A*, *CDK4* and *BAP1*) or genomic stability through telomere maintenance(*TERT*, *POT1*). Disruption of these processes can lead to multiple types of cancer, as is evident from the different disorders these driver genes cause. The set of families with an unknown cause indicates that there could be either be other potential driver genes within these pathways or genes in uncharted novel pathways that play a role in the onset of familial melanoma. Additionally, while epigentic factors might play a role in oncogenesis, the high number of melanomas in each pedigree indicates towards a genetic mutation because the cause of disease.

This PhD project was directed at exploring these possibilities and shed some light on answering these questions. The project comprises of sequencing and analysing a large cohort of familial melanoma pedigrees comprising of both whole genome and exome sequences. This approach was chosen for the following advantages:

- The falling costs of sequencing (whole genome sequencing in particular) has allowed for the sequencing of hundreds of samples, increasing the statistical power of the project in the ability to observe low-frequency mutations.

- The availability of whole genome sequences enables the exploration of the non-coding region of the genome for structural variants, regulatory region variants and promoter mutations, all of which are relatively unexplored in the context of familial melanoma.

- The presence of publicly available datasets such as ExAC and gnomAD which provide a highly accurate estimate of allele frequencies in normal populations which is vital for the identification of low-frequency, high-penetrant mutations.

- Sequencing the whole genome/exome allows us to neutrally examine the entire human gene set for novel driver genes and variants as opposed to focussing on a specific subset of candidate genes which could be potentially biased.

The chapters in this thesis have been organized into different components based on the type of analysis performed on the dataset, consisting of 4 distinct cohorts. This is shown in Figure 1.10.

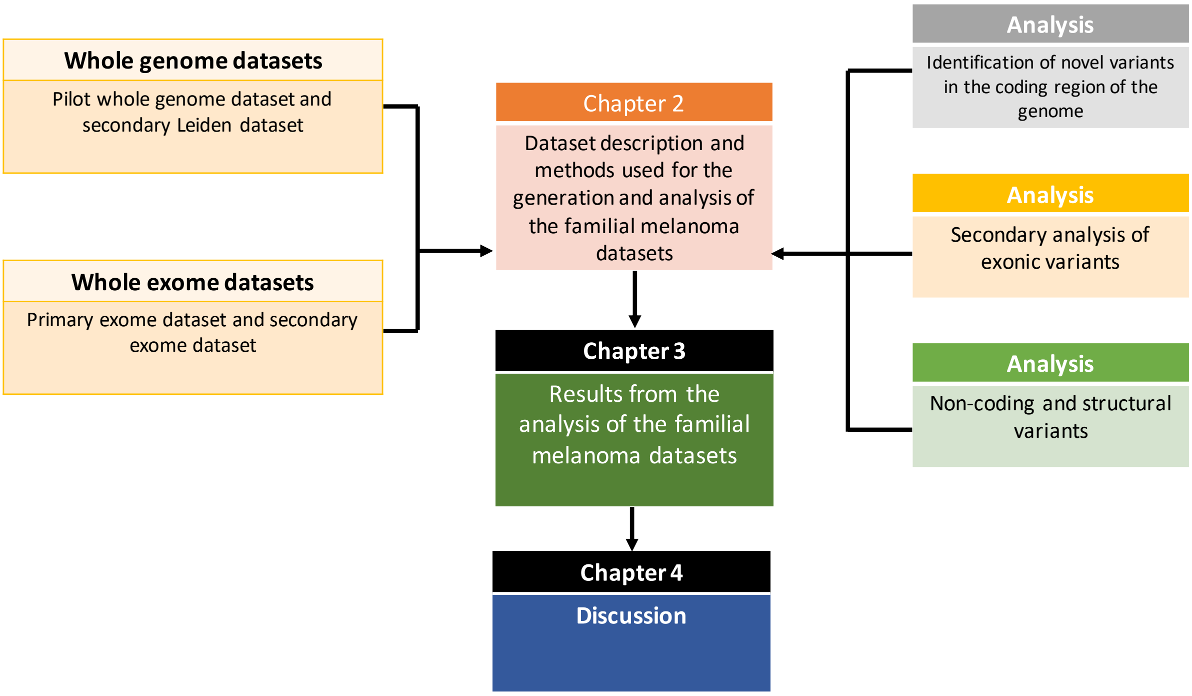

Whole genome datasets
Pilot whole genome dataset and secondary Leiden dataset
Whole exome datasets
Primary exome dataset and secondary exome dataset
Chapter 2
Dataset description and methods used for the generation and analysis of the familial melanoma datasets
Chapter 3
Results from the analysis of the familial melanoma datasets
Chapter 4
Discussion
Analysis
Identification of novel variants in the coding region of the genome
Analysis
Secondary analysis of exonic variants
Analysis
Non-coding and structural variants

Figure 1.10: Outline of PhD project.

Two of these cohorts consist of whole genome sequences while the other two are whole-exome datasets and the samples were obtained from 8 different locations/institutions across the world. The composition and criteria for sample selection in each of these datasets are explained in Chapter 2. Chapter 2 also contains information on the background and methods designed for the different analyses performed on the dataset. This includes an association analysis and a joint association-linkage analysis of coding region variants aimed at identifying novel high penetrance melanoma susceptibility genes. This is followed by the exploration of the variants in the dataset in known melanoma predisposition genes and additional secondary analysis of exonic variants that complement the association and linkage analysis. The final approach that is discussed comprises the search for novel mutations in the non-coding region, particularly with structural variants and mutations in known transcription factor binding sites. The complete workflows for all of these different processes are described in detail in Chapter 2. The results from these different analysis are included altogether in Chapter 3. Finally, in Chapter 4, the relevance of the results from each of the preceding chapters as well as the future directions of this project are discussed.

1.7 Overarching aims of the project

The primary goal of this project is to determine novel variants that predispose individuals carrying these variants to the development of familial melanoma. This goal incorporates several key distinct aims which are individually listed here.

- To obtain the samples of familial melanoma patients from multiple locations/sources and to analyse these samples - through exome or whole genome sequencing.
- To incorporate all the individual datasets sequenced through different methods into a single, consistent dataset.
- To perform variant calling uniformly across the dataset and to annotate each mutation with their predicted consequences on protein function.
- To perform preliminary analyses on the dataset to eliminate potential pre-existing biases related to an increased burden of common risk factors and population stratification.
- To identify rare, deleterious variants in data from cases and controls by filtering on several criteria.
- To utilise a rare variant association analysis for the identification of genes with a higher mutation burden in cases compared to controls.
- To design and execute a joint approach combining association analysis and linkage analysis that employs both variant data from the sequencing and the relatedness data from the pedigrees can be utilised in determining novel candidates for familial melanoma development.
- To establish methods that can determine variants related to cancer development which cannot be identified through a rare-variant association and linkage analysis.
- To determine which of these variants have high segregation within our cases and to account for the presence of potential phenocopies within the pedigrees.
- To identify variants in known melanoma predisposition genes by annotating their clinical significance using ClinVar and to explore potentially pathogenic variants associated with cancer.
- To establish the location of transcription factor binding motifs across the genome.

- To ascertain rare non-coding variants that lie within transcription factor binding motifs.
- To determine a suitable control dataset and to identify genes with increased burden of non-coding variants within transcription factor binding motifs in cases compared to controls and to discern rare variants within the non-coding region of the genome.
- To establish a workflow for the identification of structural variants within the cases.
- To identify novel structural variants disrupting known cancer genes.

Chapter 2

Dataset description and methods used for the generation and analysis of the familial melanoma datasets

2.1 Introduction

This chapter introduces the process of selection for the sequencing dataset used in the project. The pedigrees were initially chosen and sequenced as part of four distinct datasets, two exome and two whole genome. The first exome dataset, referred to as the primary exome dataset in this chapter, was sequenced and partly analysed before I started working on the project. The remaining three datasets were chosen and sequenced after I started the project, for which I collaborated with members of our melanoma consortium called GenoMEL, described in Section 2.2.1. The description of these pedigrees as four distinct datasets is purely for the distinction of the varied choices of pedigree selection, sequencing methodology, technology used and the institutions they were sequenced in. Eventually, these cohorts were merged into a single dataset and analysed as one large dataset for the rest of the project, with the exception of only whole genome sequences being considered for the noncoding and structural variant analysis. This chapter includes a description of assembly and sequencing of all four datasets and the eventual process used to merge them into a single dataset and perform variant calling on them. Additional methods on the filtering of the variants are also described in detail. Following this, all the methods for analysing the dataset including an association analysis on the coding region variants, a joint association and linkage analysis, secondary exonic analyses, and studies on variants disrupting transcription factor binding motifs and large structural alterations are also described in this chapter. The results from all of these approaches are presented in the following chapter.

2.2 Dataset description and assembly

2.2.1 An introduction to GenoMEL

GenoMEL is a melanoma genetics consortium comprising researchers and investigators from 24 institutions across the world focussing on the identification of genes that increase the risk of both familial and sporadic melanoma. It is the largest colletion of familial melanoma data in the world and was started in the early 1990s by Professor Julia Newton Bishop from the University of Leeds, with the number of collaborators constantly increasing every year since then. The consortium also investigates the interaction of these genetic factors with environmental factors and the relevance of the inheritance of these genes to familial melanoma risk. The consortium emphasises open knowledge sharing and transfer of key knowledge related to melanoma genetics research. This project was funded as part of the MELGEN Early Training Network under GenoMEL with data being provided by 8 of the 24 GenoMEL institutions studying familial melanoma.

2.2.2 Cohort description

A total of 308 patients diagnosed with familial melanoma from 133 different pedigrees from across the world were sequenced as a part of this project, making it the largest dataset of its kind to date. These patients were selected to be *CDKN2A* and *CDK4* negative to increase the chances of finding a novel familial melanoma predisposition driver gene. An example of the type of pedigree chosen to be sequenced is provided in Figure 2.1. The pedigree in this figure comprises of 16 members, 6 of whom were diagnosed with melanoma.

The pedigrees that were sequenced as part of this dataset were identified by collaborators in 9 different institutions across the world, shown in Table 2.1.

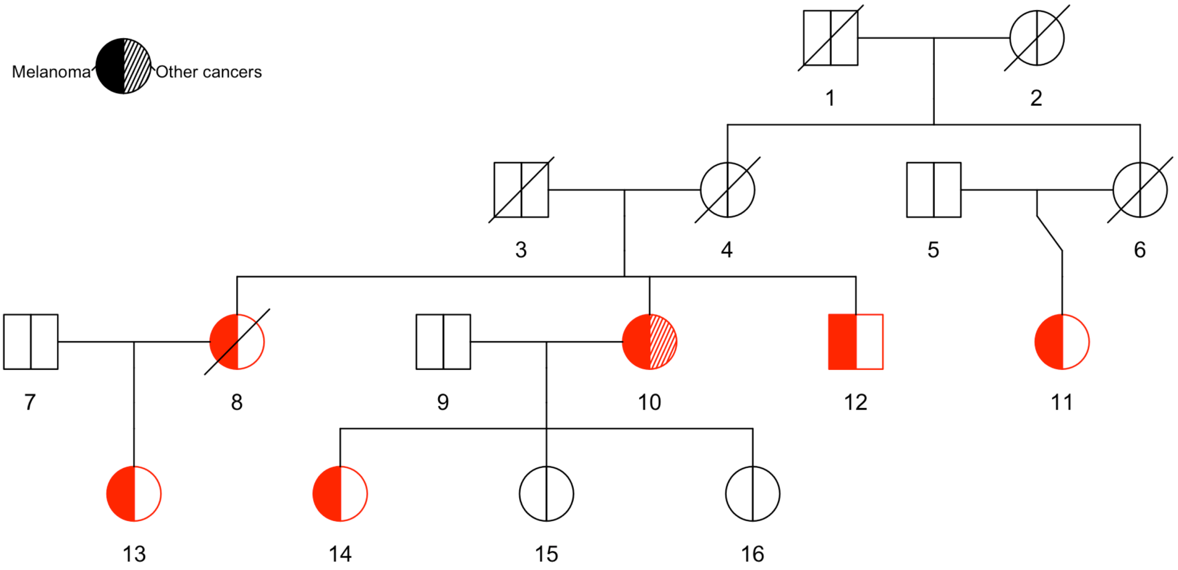


Figure 2.1: An example of a pedigree sequenced as part of the study. Circles indicate female individuals while squares indicate male individuals. A diagonal line across the symbol indicates that the individual is deceased. The members of the pedigree marked in red were the patients sequenced from this pedigree. In this case, there were 6 affected members in the pedigree, all of whom were sequenced.

Institution	Location	Lead Principal Investigator
University of Pennsylvania	Pennsylvania, United States of America	Dr. Peter A. Kanetsky
University of Sydney	Sydney, Australia	Professor Graham Mann
The QIMR Berghofer Medical Research Institute	Brisbane, Australia	Professor Nicholas Hayward
Leiden University Medical Center	Leiden, The Netherlands	Dr. Remco van Doorn and Dr. Nelleke Gruis
University of Leeds	Leeds, United Kingdom	Professor Tim Bishop and Professor Julia Newton-Bishop
Karolinska Institutet	Stockholm, Sweden	Dr. Veronica Höiom
Rigshospitalet	Copenhagen, Denmark	Dr. Karin Wadt
Institut d'Investigacions Biomediques August Pi I Sunyer	Barcelona, Spain	Dr. Susana Puig
Kings College	London, United Kingdom	Dr. Veronique Bataille

Table 2.1: The different collaborative institutions and their corresponding lead investigators who helped provide samples for this project and their respective locations.

Additional criteria based on the number of primary melanomas, age of onset and number of affected members in each pedigree were also used to finalize the list of patients chosen for sequencing. DNA from these patients was collected by our collaborators at institutions mentioned in Table 2.1 and sent to the Wellcome Sanger Institute for sequencing. The information regarding the sequencing of each dataset, including sequencing platforms, baits and read lengths for the whole genome and exome sequences are provided in Sections 2.2.3 and 2.2.4 respectively. The 308 patients were sequenced using a mixture of exome and whole genome sequencing and were initially sequenced as four individual datasets:

1. Pilot whole genome dataset - Consisting of 123 whole genome sequences from 32 pedigrees.

2. Secondary Leiden whole genome dataset - Consisting of 28 whole genome sequences from 6 pedigrees.

3. Primary exome dataset - Consisting of 80 exome sequences from 67 pedigrees.

4. Secondary exome dataset - Consisting of 77 exome sequences from 28 pedigrees.

The distribution of samples and families in each dataset based on the origin of the samples are provided in Table 2.2.

The origin of cases with multiple primary melanomas and early age of onset (<40 years of age) are provided in Table 2.3. Across all datasets, 29.06% of patients were detected to have multiple primary melanomas while 27.92% of patients had an early age of onset.

The average number of people affected and sequenced from each pedigree across all the datasets is given here:

- Pilot whole genome dataset = 6 affected and 4 sequenced.

- Secondary Leiden whole genome dataset = 5 affected and 5 sequenced.

- Primary exome dataset = 4 affected and 2 sequenced.

- Secondary exome dataset = 4 affected and 3 sequenced.

This shows that the whole genome datasets have a higher average number of people affected and sequenced compared to the whole exome datasets. This is also due to the stricter criteria of selection for patients and families imposed on these datasets. In the compiled overall dataset, 5 people were affected on average in every family and 2 were sequenced across all the datasets.

Location	Number of families (number of samples)	Sequence type	Dataset
Pennsylvania	4(15)	Whole genome	Pilot whole genome dataset
Sydney	9(56)	Whole genome	Pilot whole genome dataset
Brisbane	6(27)	Whole genome	Pilot whole genome dataset
Leiden	2(6)	Whole genome	Pilot whole genome dataset
Leeds	1(2)	Whole genome	Pilot whole genome dataset
Stockholm	1(2)	Whole genome	Pilot whole genome dataset
Denmark	4(10)	Whole genome	Pilot whole genome dataset
London (KCL)	5 (5)	Whole genome	Pilot whole genome dataset
Leiden	6(28)	Whole genome	Secondary Leiden whole genome dataset
Leeds	65(75)	Exome	Primary exome dataset
Leiden	2(5)	Exome	Primary exome dataset
Barcelona	10(20)	Exome	Secondary exome dataset
Pennsylvania	1(3)	Exome	Secondary exome dataset
Brisbane	4(15)	Exome	Secondary exome dataset
Stockholm	2(6)	Exome	Secondary exome dataset
Sydney	7(28)	Exome	Secondary exome dataset
London (KCL)	3(3)	Exome	Secondary exome dataset
Leiden	1(2)	Exome	Secondary exome dataset

Table 2.2: Distribution of samples by location, type of sequence and dataset.

Dataset	Proportion of cases with multiple primary melanomas (MPM)	Proportion of early onset cases	Number of cases for which the information was unavailable
Pilot whole genome dataset	36/116	30/116	7 for MPM, 7 for early onset cases
Secondary Leiden whole genome dataset	5/24	11/24	4 for MPM, 4 for early onset cases
Primary exome dataset	3/9	2/9	77 for MPM, 77 for early onset cases
Secondary exome dataset	15/54	12/48	23 for MPM, 29 for early onset

Table 2.3: Distribution of patients with multiple primary melanomas and early age of onset in each dataset.

2.2.3 Whole genome sequences - Sample selection and sequencing

2.2.3.1 Pilot whole genome dataset

In order to determine mutations causative of familial melanoma, the genomes of 123 individuals from 32 families were sequenced. These samples were obtained from our collaborators in Sydney, Pennsylvania, Stockholm, Leiden, Denmark, Leeds and Brisbane who are part of the GenoMEL consortium (https://genomel.org). The dataset also included 8 samples obtained from King's College, London who are not a part of the GenoMEL consortium. These 8 samples were selected due to the presence of MPMs and/or early age of onset as opposed to multiple members of the family affected. The distribution of the 123 individuals across these locations is given in Table 2.2. Informed consent was obtained by each institution (Sydney: HREC/13/CIPHS/71, Pennsylvania: 14.03.0033 Protocol MCC 17751, Sweden: 03-471, 03-713, Leeds: 99/3/045, Leiden: Protocol No. P00.117-gk2, Copenhagen: Protokol af 7. juni 2012, version 3, Brisbane: P452 (H0204-013), London: 07/HO802/84). The distribution of the families and samples are given below in Table 2.2. The criteria for selecting these particular samples varied depending on their origin of the families. These criteria are elucidated in Table 2.4.

Genomic DNA (500 ng) was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. Adapter-ligated libraries were am-

Families	Criteria
All families	Presence of multiple primary melanomas and an early age of onset (<40 years) with 2 or more DNAs available to sequence per family.
European and American families	5 or more cases with 2 or more DNAs available to sequence per family.
Australian families	6 or more cases with 2 or more DNAs available to sequence per family.

Table 2.4: Criteria for selection of whole genome samples in the pilot dataset.

plified by 6 cycles of PCR and subjected to DNA sequencing using the HiSeqX platform (Illumina) according to manufacturer's instructions. Read lengths of 150bp were obtained for this dataset

2.2.3.2 Secondary Leiden whole genome dataset

An additional 29 samples from 6 pedigrees were obtained from our collaborators at the Leiden University Medical Center. The selection criteria for the samples were the same as the ones previously mentioned in Section 2.2.3.1, i.e., 5 or more cases in each pedigree with at least 2 or more DNA samples available to sequence. Ethical approval for sequencing was obtained (Leiden: Protocol No. P00.117-gk2). Genomic DNA was sheared and amplified in the same manner as Section 2.2.3.1. However, this was done using the HiSeq2500 instead of the HiSeqX platform. Read lengths of 100 bp were obtained for this dataset. One sample was removed for low average coverage (<9X across the genome), resulting in a total of 28 samples from 6 pedigrees.

2.2.4 Exome sequences - Sample selection and sequencing

2.2.4.1 Primary exome dataset

The primary criteria for selection of the families for sequencing within this dataset were less stringent compared to the whole genome datasets. These conditions are given in Table 2.5.

The pedigrees for this study were recruited by collaborators from the University of Leeds and the Leiden University Medical Centre. Cases from these pedigrees were confirmed to be negative for *CDKN2A* and *CDK4* mutations. Informed consent was obtained under the Multicentre Research Ethics Committee (UK): 99/3/045 for the Leeds cases and Protocol P00.117-gk2/WK/ib for Leiden cases. Genomic DNA was extracted from blood using standard methods

Families	Criteria
All families	Presence of multiple primary melanomas in multiple members and/or an early age of onset (<40 years)
European families	3 or more cases with 2 or more DNAs available to sequence per family.

Table 2.5: Criteria for selection of exome samples in the primary exome dataset.

from our respective collaborators at these institutions, shown in Table 2.1.

5 µg of genomic DNA were sent for sequencing at the Wellcome Sanger Institute and exonic regions were captured with the Agilent SureSelect Target Enrichment System. Paired-end reads of 75 base pairs (bp) were generated on the HiSeq 2000 platform. A subset of samples from the Leeds cohort were sequenced at the Beijing Genomics Institute (BGI), using the Illumina HiSeq2000 platform, which generated 90 bp paired-end reads.

This dataset was previously a part of a larger dataset used by Dr Carla Daniela Robles Espinoza at the Wellcome Sanger Institute and it led to the discovery of *POT1* as a familial melanoma driver gene[143]. Two of the families from this dataset were also sequenced in the whole genome dataset. However, both the exome and whole genome sequences of these families were included in the analysis to confirm that any variant seen within these families was present in both versions.

2.2.4.2 Secondary exome dataset

The final subset of 77 samples from 28 pedigrees were obtained from collaborators in Barcelona, Sydney, Pennsylvania, Stockholm, Leiden, Leeds and Brisbane who are part of the GenoMEL consortium. Similar to the pilot whole genome dataset, samples with MPMs and/or early age of onset (<40) were again obtained from King's College, London who are not a part of the GenoMEL consortium. The distributions of the 77 samples across these locations, number of samples and number of families are given in Table 2.2. Informed consent was obtained by each institution in the same way as mentioned in Section 2.2.3.1. The criteria for the selection of these pedigrees was again less stringent compared to the whole genome samples. These criteria are shown in Table 2.6. Genomic DNA (500 ng) was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. These samples were then sequenced at the Wellcome Sanger Institute using the Illumina HiSeq 2000 platform with read lengths of 75 bp.

Families	Criteria
All families	Presence of multiple primary melanomas and an early age of onset (<40 years) with 2 or more DNAs available to sequence per family.
European and American families	4 or more cases with 2 or more DNAs available to sequence per family.
Australian families	5 or more cases with 2 or more DNAs available to sequence per family.

Table 2.6: Criteria for selection of exome samples in the secondary exome dataset.

2.3 Alignment of DNA sequence data and variant calling

Samples from the different datasets were sequenced at different times. They were however eventually aligned and processed into a single, larger dataset. The sequences were aligned to the latest version of the reference build of the human genome which was the GRCh38 build of the human genome. A Burrows-Wheeler aligner (BWA-MEM)[162] was used by the core sequencing facility to align the sequences. This was followed by the estimation of sequencing depth in every sample across all positions. To ensure good coverage across all samples, a coverage threshold of 15X was established. A single sample which was whole genome sequenced and belonged to the secondary Leiden whole genome dataset was removed from the dataset due to low mean sequencing depth, i.e., less than 15 reads. Variant calling was then performed using GATK Haplotype Caller[163] which employs a 'joint calling' approach. An intermediate file called a Genomic VCF (gVCF) file is create for each sample which contains the genotype information for that sample across all positions. The caller then jointly calls genotypes across all samples from each gvcf file to create a single multisample file. This ensures that the genotype calls are available for all samples at all positions, regardless of whether a specific sample contains a variant at a given loci or not. This output is termed as a Variant Calling Format file (VCF). The multisample VCF file containing all the variants across all the samples was then annotated with the predicted consequences for each variant using a tool called Variant Effect Predictor[164] which was established by Ensembl. This resulted in the first complete set of variants which comprised of all coding and non-coding mutations, prior to further processing or filtering. The final step in the initial variant calling process was to then filter for variants that were predicted to disrupt or alter the protein produced by different genes. The different consequences that were retained for this step and their predicted impact on the gene are shown in Table 2.7. The second complete set of variants was obtained as the output of this step.

Variant Effect Predictor Consequence	**Description**
Protein altering variant	A variant that affects the protein through a change in the codons.
Missense variant	A non-synonymous variant that changes an amino acid without affecting protein length.
Inframe deletion	An inframe non-synonymous variant which removes bases from the coding sequence.
Inframe insertion	An inframe non-synonymous variant which incorporates bases into the coding sequence.
Transcript amplification	Amplification of a region containing a transcript.
Start lost	A non-synonymous variant that alters a base in the canonical start codon.
Stop lost	A non-synonymous variant that alters a base in a stop codon resulting in longer transcripts.
Frameshift variant	A non-synonymous variant that disrupts the reading frame of the protein through the addition or removal of multiple adjacent bases.
Stop gained	A non-synonymous variant which disrupts a codon in such a way as to introduce a stop codon which results in shorter transcripts.
Splice donor variant	A splice variant that changes the 2 base region at the 5’ end of an intron
Splice acceptor variant	A splice variant that changes the 2 base region at the 3’ end of an intron
Transcript ablation	Deletion of a region containing a transcript.

Table 2.7: The list of predicted protein altering consequences and their impact on the protein, as obtained from the Ensembl variation website[165].

Additional filtering, annotation and processing of variants were then performed on these two sets of variants which are explored in detail in the following chapters. A summary of the workflow up to this point is given in Figure 2.2.

2.4 Exploration of population stratification bias within the dataset

Inherent differences in the ancestry of different population subgroups, especially with respect to the frequency of specific variants and alleles may lead to a bias in population based studies. As this project eventually involved association studies and the comparison of familial melanoma cases to unaffected controls, there was a possibility that an association could be identified with a particular loci due to such a bias within the population and not due to the influence of the loci on the disease status. Such a scenario is defined as population stratification. In order to determine the possible presence of population stratification within the dataset, it was deemed necessary to establish and confirm the reported population subgroups of the cases as accurate.

Genotype and variant information for different population subgroups were obtained from the 1000 genomes project and used as the control dataset for this assessment[166]. While there were several different nationalities reported within the 1000 genomes project, the subgroups were largely classified into: European, Indian, South American, Chinese and African. These variants were filtered to have an allele frequency between 0.2 and 0.8 as variants with extreme frequencies could also potentially bias the accurate estimation of the population subgroup for each sample. These variants were also filtered to ensure that they were in linkage disequilibrium with each other. Once this subgroup of variants, determined to be variable across the population, were identified from the controls, the same variants were also extracted from the complete dataset of familial melanoma cases.

A principal component analysis (PCA) was performed for all samples on this subset of variants using PLINK v1.9. To reliably differentiate between the population subgroups, the first three principal components/eigen vectors were needed (named PC1, PC2 and PC3 respectively). Figure 2.3 highlights the distribution of population clusters as a snapshot of a 3D plot.

The three principal components are represented with the three dimensions of the figure. Each nationality was assigned its own colour for the plot with larger population subgroups sharing similar shades. The familial melanoma cases were all marked in black and are denoted

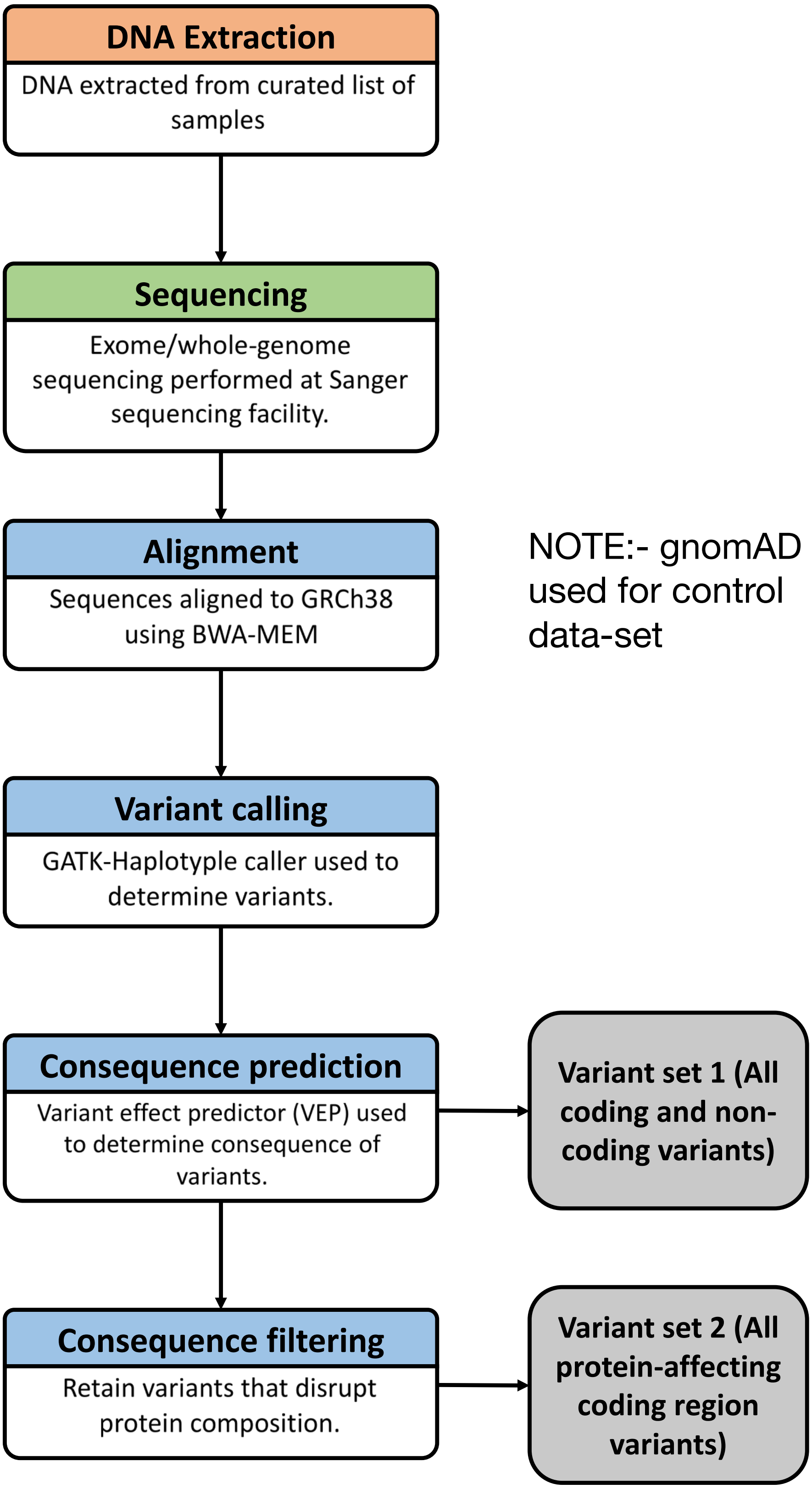

DNA Extraction
DNA extracted from curated list of samples
Sequencing
Exome/whole-genome sequencing performed at Sanger sequencing facility.
Alignment
Sequences aligned to GRCh38 using BWA-MEM
NOTE:- gnomAD used for control data-set
Variant calling
GATK-Haplotype caller used to determine variants.
Consequence prediction
Variant effect predictor (VEP) used to determine consequence of variants.
Variant set 1 (All coding and non-coding variants)
Consequence filtering
Retain variants that disrupt protein composition.
Variant set 2 (All protein-affecting coding region variants)

Figure 2.2: Workflow describing the steps involved in the generation of candidate variant sets in the search for novel driver genes and variants involved in melanoma susceptibility.

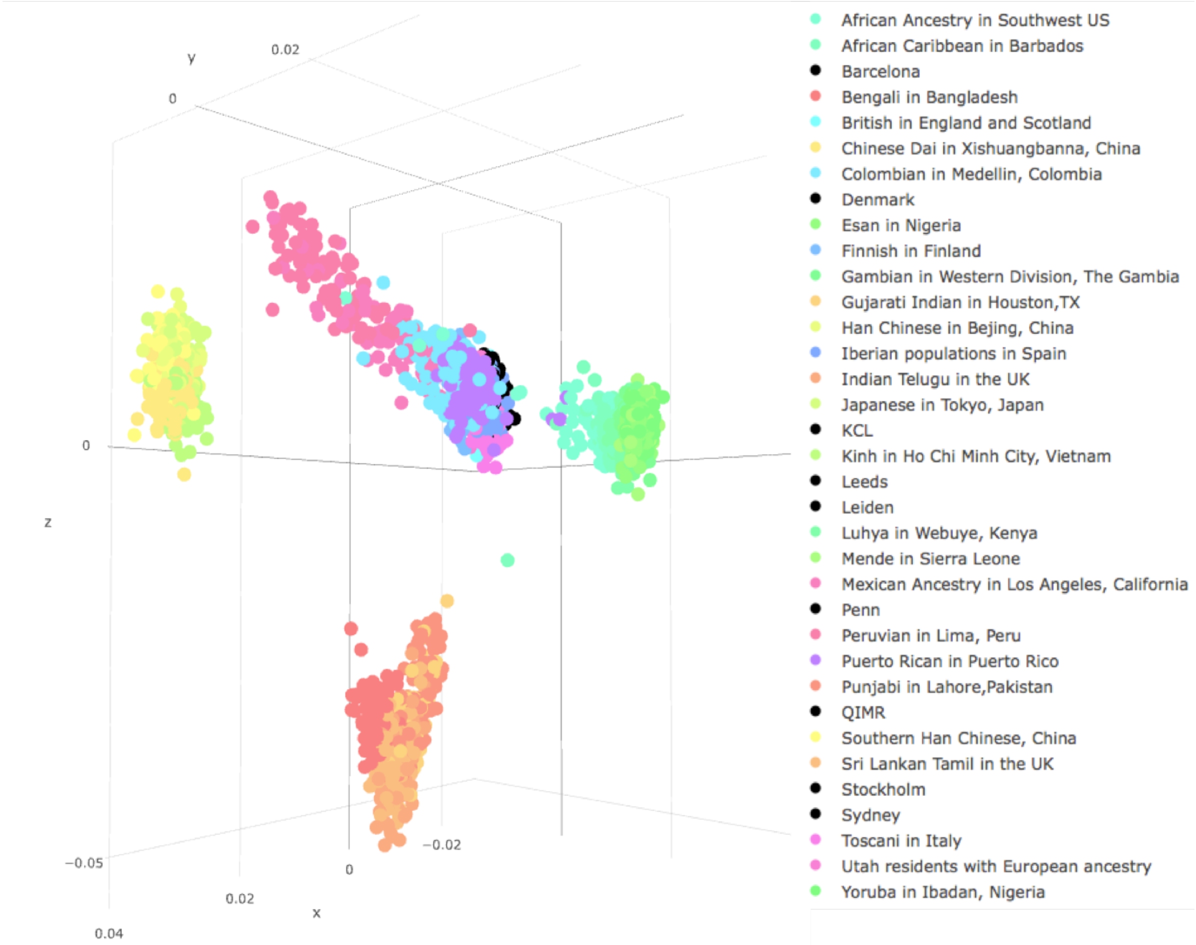


Figure 2.3: Principal component analysis to verify ethnicity.

using a small circle. It is evident that all the individuals from the Indian subcontinent, coloured red to orange in the figure, cluster together. Similarly, there are two distinct green clusters of the Chinese and the African subpopulations as well. The larger cluster at the top is a mixture of the North American, the central American and the European population subgroups. The central and south American populations are however distinctly clustered on the x dimension of the 3D plot which leaves all the Europeans clustered together. The familial melanoma cases are also clustered together with the European population. Therefore, the different population subgroup clusters are observed as being independent of each other when all three principal components are considered together.

Importantly, the cases (marked in black) cluster along with the European subgroup in all three subplots. The familial melanoma cases were all reported to be of European origin when they were sampled and sequenced; this is confirmed through the above analysis. This also implies that there is no population stratification within the dataset and that association analysis can be performed between the familial melanoma cases and a set of unaffected, European origin controls.

2.5 Estimation of polygenic risk scores

2.5.1 Introduction

The origin of a disease with genetic roots may be traced either to the presence of a few single, highly penetrant rare alleles or due to a high burden of common, low risk alleles. The estimation of polygenic risk scores is a measure used to determine if the presence of the disease in a particular individual is more likely due to the former scenario or the latter. Polygenic risk scores are metrics which are a numerical measure of the impact of a combination of genetic variants and their associated weights on a given trait. GWAS studies have helped in determining significant variants with a high association to the phenotype of interest. However, polygenic risk scores are considered to be a better approach when the trait is predicted to be affected or determined by a combination of a large number of variants with lesser impact on the trait than a singular high impact variant. These variants may or may not be statistically significant, and as such may not be identified in a standard GWA study. Such traits which are determined by not a single variant but by a combination of multiple variants in different genes are called as polygenic traits.

Each variant that affects a polygenic trait is assigned a weighted score, which is proportional to the association of the variant with the trait. Higher weighted scores correspond to a larger association with the trait of interest with negative scores corresponding to negative association. Each allele at these variant positions are also given a coefficient corresponding to the impact of the allele on the association at the variant position. If an allele has a higher coefficient, it implies that the presence of the allele increases the impact of the variant on the trait. The risk score of each variant is determined as the product of the coefficient corresponding to the alleles and the weighted score of each variant. This would vary based on the genotypes present in any given sample at the loci. The polygenic risk score of a sample is finally estimated the sum of the risk scores of all the variants affecting the trait[167].

Common examples of polygenic traits include skin colour, eye colour and hair colour. In addition to these normal phenotypes, several disorders are also thought to follow the polygenic model including type 2 diabetes and coronary heart disease[168, 169]. In recent years, polygenic risk scores have also been used in determining patient risk for schizophrenia, bipolar disorders and for certain types of cancer including breast cancer and prostate cancer[170–172].

In order to reliably determine the burden of common, low risk alleles within the familial melanoma cases, a secondary set of sporadic cases and unaffected controls was required. The aim of this approach was to compare the risk scores of the familial melanoma cases and compare these to the risk scores of sporadic cases and unaffected controls to observe if the familial melanoma cases have a higher risk score in general. If they did, it would imply that the presence of melanoma within these pedigrees could possibly be explained by the higher burden of common mutations. If there was no discernible distinction between the different groups, this would implicitly point to the presence of rare, highly penetrant variants within the familial melanoma pedigrees.

2.5.2 Methods

A previously conducted genome wide association study (GWAS) on melanoma and a follow-up meta-analysis study[51] established 20 loci, shown in Table 2.8 as possible SNPs which increase predisposition to melanoma. This GWAS analysis included a estimation of risk bestowed by each of the listed variants.

The individual risk scores of these variants, as estimated by Law et al[51] is used in the estimated of the polygenic risk score of each sample. Each polygenic risk score is determined by combining the risk scores of all 20 SNPs based on the genotypes of the alleles present at those positions within the sample. The polygenic risk scores for all 123 familial melanoma cases within the pilot dataset were estimated using this approach. These scores were calculated using PLINK v1.941 and also verified manually. The secondary dataset required for the

rs_ID	Gene	Chromosome	Position	Reference allele	Alternate allele	Beta value
rs12410869	Intergenic	1	150883677	G	T	-0.130
rs1858550	Intergenic	1	226420403	C	A	-0.143
rs6750047	*RMDN2*	2	38049406	A	G	0.088
rs7582362	*FLACC1*	2	201311571	A	G	0.113
rs380286	*CLPTM1L*	5	1320132	G	A	0.152
rs250417	*SLC45A2*	5	33952273	G	C	-0.891
rs6914598	*CDKAL1*	6	21163688	T	C	0.108
rs1636744	Intergenic	7	16944656	C	T	0.105
rs7852450	*MTAP*	9	21825076	T	C	-0.212
rs10739221	Intergenic	9	106298549	T	C	0.120
rs2995264	*STN1*	10	103909085	G	A	0.144
rs498136	Intergenic	11	69552350	A	C	0.116
rs1393350	*TYR*	11	89277878	G	A	0.198
rs73008229	*ATM*	11	108316962	G	A	-0.188
rs4778138	*OCA2*	15	28090674	A	G	-0.178
rs12596638	*FTO*	16	54081917	G	A	0.143
rs75570604	*FANCA*	16	89780269	G	C	0.600
rs6088372	*RALY*	20	33998942	C	T	0.267
rs408825	*MX2*	21	41371569	C	T	-0.141
rs2092180	*PLA2G6*	22	38175556	A	G	-0.116

Table 2.8: Single Nucleotide Polymorphisms chosen for the polygenic risk score analysis.

comparison of these scores were obtained through the help of collaborators in the University of Leeds, namely Prof. Timothy Bishop and Dr Mark Iles. The genotypes of 1800 sporadic melanoma cases, 148 familial melanoma cases and 489 controls at these 20 loci were provided for this purpose. The polygenic risk scores for these samples were also estimated using the same method as used for the familial melanoma cases in the pilot dataset. The risk scores and their implications for the dataset are discussed in Section 3.2.

2.6 The determination of novel variants through association analysis

2.6.1 Selection of a control dataset

The cases chosen to be analysed for this project comprised familial melanoma patients. During the design of the project, it was decided that there would be no sequencing of unaffected family members from the same pedigrees. The rationale behind this is that familial melanoma is a result of multiple factors (both genetic and environmental) which determines the onset of disease. An individual with a rare allele with high penetrance would have an increased risk of developing melanoma but might not actually develop melanoma. For example, *CDKN2A* is the single most important familial melanoma locus, with 45% of familial melanoma cases being attributed to germline mutations in *CDKN2A*[173]. However, the penetrance of *CDKN2A* mutations varies between 0.3 to 0.67, depending on the age of the carriers, indicating that even the most common familial melanoma gene is not completely penetrant[174]. In such a situation, sequencing unaffected members of the family as matched controls and filtering the common variants could result in the potential loss of the high-risk allele that is causative of the disease within the family.

To compensate for the absence of family-matched controls, a neutral, population matched control set of exome/whole genome sequence data was required. The Exome Aggregation Consortium (ExAC)[175] led by the Broad Institute was originally chosen for this purpose. ExAC was established as a consortium of exome sequencing projects involving unrelated individuals from several population genetics studies across the world. It contains aggregated variant level statistics on 60,706 individuals with population frequencies on each variant and allele provided. The distribution of samples across different population subgroups in ExAC is shown in Table 2.9. As the cases consisted entirely of patients of non-Finnish European ethnicity, the same ethnic group was chosen from ExAC to be used as the control data set. This comprised data from 33,370 individuals.

Population	Description	Genomes	Exomes	Total
AFR	African/African American	1,888	3,315	5,203
AMR	Latino	2,254	3,535	5,789
EAS	East Asian	2,016	2,311	4,327
FIN	Finnish	2,084	1,223	3,307
NFE	Non-Finnish European	18,740	14,630	33,370
SAS	South Asian	6,387	1,869	8,256
OTH	Other (population not assigned)	275	179	454
	Total	33,644	27,062	60,706

Table 2.9: Distribution of samples across different population groups in ExAC.

During the course of the project, a larger data set of controls called the Genome Aggregation Database (gnomAD)[175], also curated by the Broad Institute, became publicly available. gnomAD is a global collaboration of collated summary data from several large-scale sequencing projects. The total number of samples in the gnomAD data set 138,632 individuals compared to 60,706 in ExAC. As the gnomAD data set includes both exome and whole genome sequencing data, it was identified as the more suitable control and was chosen instead of ExAC. The distribution of samples across different population subgroups in gnomAD is given in Table 2.10.

Population	Description	Genomes	Exomes	Total
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632

Table 2.10: Distribution of samples across different population groups in gnomAD.

The non-Finnish European population subgroup was again chosen from gnomAD comprising 7,509 whole genomes. The sequences were all originally aligned to Genome Reference Consortium Human Build 37 (GRCh37). However, as the cases were sequenced and aligned to Genome Reference Consortium Human Build 38 (GRCh38), the gnomAD VCF file consisting of the aggregated variant information across all samples was lifted over from GRCh37 to GRCh38 using CrossMap v0.2.5[176]. The parameters used for this procedure are given in Table 2.11.

The aggregated coverage data for the samples across the genome was also downloaded from the gnomAD repository. The gnomAD dataset included information on 240,779,968 variants across all chromosomes and samples. Data from gnomAD v2.0.2 were used for this purpose.

Note: gnomAD does not include any variants on the Y chromosome . As a result, variants

Parameters	Description
python CrossMap.py	Execution of python script for CrossMap,
vcf	Indicating that the format of the input and the required format for the output is a VCF file.
hg38ToHg19.over.chain.gz	Location of chain files; chain file describe genome-wide pairwise alignments of positions between assemblies. In this case, the chain file is a mapping of alignments between GRCh38 and GRCh37.
input.vcf.gz	Location of input vcf file to be lifted over.
hs37d5.fa	Reference fasta file for the target output genome build, in this case, GRCh37.
output_hg19.vcf	Name of output VCF file to be generated.

Table 2.11: Parameters used for running CrossMap to lift the aligned gnomAD sequences from GRCh37 to GRCh38.

from the Y chromosome were excluded from this analysis.

2.6.2 Initial filtering of variants

2.6.2.1 Control variants from gnomAD

The lifted over variants from the gnomAD VCF file were annotated using the Variant Effect Predictor (VEP) tool from Ensembl. This was performed in order to annotate information including the affected gene and the consequences of the variant on protein function. These consequences include intergenic variants, intronic variants, nonsense mutations and loss of function mutations. Following this, the variants were filtered to retain non-synonymous, nonsense and loss of function mutations. The parameters used for this performing this function are given in Table 2.12.

The list of the consequences chosen to be retained for this purpose was the same as the ones previously mentioned in Table2.7. Genomic locations that were annotated with multiple alternate alleles (multiallelic variants) were split into multiple entries, each entry carrying information on one alternate allele. As gnomAD includes variant data on several population subgroups, as shown in Table 2.10, only the variants present in the non-Finnish European population subgroup were retained.

Parameter	Definition
-i	Input VCF file.
-o	Output VCF file.
-filter "Consequence in missense_variant,inframe_deletion, inframe_insertion, transcript_amplification,stop_lost, frameshift_variant,stop_gained, splice_acceptor_variant, start_lost,protein_altering_variant, splice_donor_variant,transcript_ablation"	List of consequences to be matched. If the annotated consequence is not present in this list, the variant is filtered.
--force_overwrite	Overwrites the output VCF if it already exists to create a new VCF file.
--only_matched	Only writes variants where the annotated consequence exactly matches the consequences provided in the filter step.

Table 2.12: List of parameters used for filtering variants based on their predicted consequence on protein function using VEP.

2.6.2.2 Case variants

Variants from the cases were previously annotated and processed, as described in Section 2.3. These variants were chosen as the initial dataset. Multiallelic variants were split as described in Section 2.6.2.1. There were a total of 131,840 variants in the cases.

During the generation of a VCF file for the sequences, each sample is annotated with a Genotype Quality (GQ) for every position. The GQ is a measure of confidence in the genotype assigned to the sample represented through a Phred-scaled score, which is derived from the probability of error. For instance, a Phred score of 30 indicates that there is a 1 in 1000 or 0.1% rate of error. A higher GQ indicates a higher confidence in the genotype of a sample for a given variant. The median GQ scores were calculated across all cases at every locus. The distribution of median GQ scores across the variants is shown in Figure 2.4. The first quartile(Q1) of the GQ distribution was at 90 while the third quartile (Q3) was at 99. Individual variants with median GQ < 30 were removed from the cases.

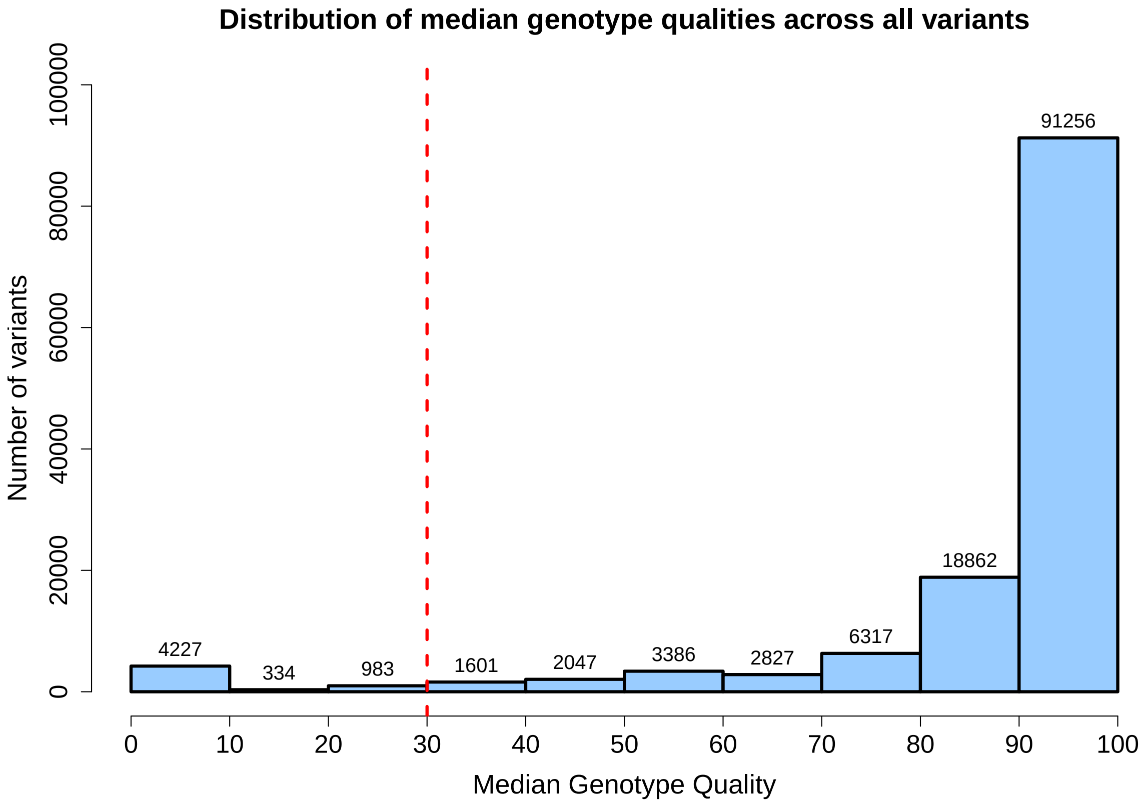


Figure 2.4: Distribution of median genotype qualities across all samples and positions. The median GQ score is denoted on the x-axis while the number of variants with that GQ score are given on the y-axis. The threshold for filtering variants based on the GQ scores was set at 30, this is denoted with the red line. 5,544 variants had median GQ score less than 30 and were removed from the set of variants, resulting in 126,296 variants in the cases.

2.6.3 Joint processing of case and control variants

Following the initial processing, a joint set of variants was established by merging the case and control variant sets together. This set was progressively filtered through several procedures for quality control which are shown in Figure 2.5 and are described here.

2.6.3.1 Annotation and filtering based on frequency of variants in gnomAD

In order to improve the power of the study in detecting important, rare mutations and to reduce the background variant burden, variants were filtered based on a population allele frequency. Prior to this, to focus on low-frequency variants with high penetrance, an sequence artefact filter was used for the cases: If a variant was contained in more than 8 families after the allele frequency filter, it was likely that it is a sequencing artifact as opposed to being involved in the development of the disease . Such variants were therefore removed from the cases. 88,994 variants remained in the cases after this step.

Variants in the cases were annotated with the allele frequencies of the variants from gnomAD. If the case variant was present in gnomAD, the frequency of mutation from the gnomAD dataset was annotated to the variant. If the case variant was not present in gnomAD, it was annotated as being absent. 32,987 of the case variants were not present in the gnomAD dataset while 56,007 variants were present in the gnomAD dataset. The distribution of population allele frequencies for these 56,007 case variants is shown in Figure 2.6. Figure 2.6 A shows the distribution for all allele frequencies while Figure 2.6 B shows the distribution for a subset of variants with allele frequencies less than 0.01. A similar distribution is shown for the control variants from gnomAD in Figure 2.7. A cut-off of one in a thousand (10^{-3}) was set as the threshold for allele frequency to select for rare mutations. Variants with a gnomAD allele frequency > .001 were removed from both the cases and the controls. Variants from the cases which were absent in gnomAD were retained.

Variants from the cases were also filtered based on the number of affected families. The distribution across all variants of the number of families is shown in Figure 2.8. The first quartile(Q1) for the family counts was 1 family while the third quartile(Q3) was 16. The median number of affected families for a variant was 2.

2.6.3.2 Annotation and filtering based on coverage of samples in cases and controls

Despite the removal of samples with low median coverage across the exome/genome in Section 2.3, variant loci with low coverage across most cases and controls were still encountered. Such variants were removed as they would yield unreliable variant calls. Variants from the

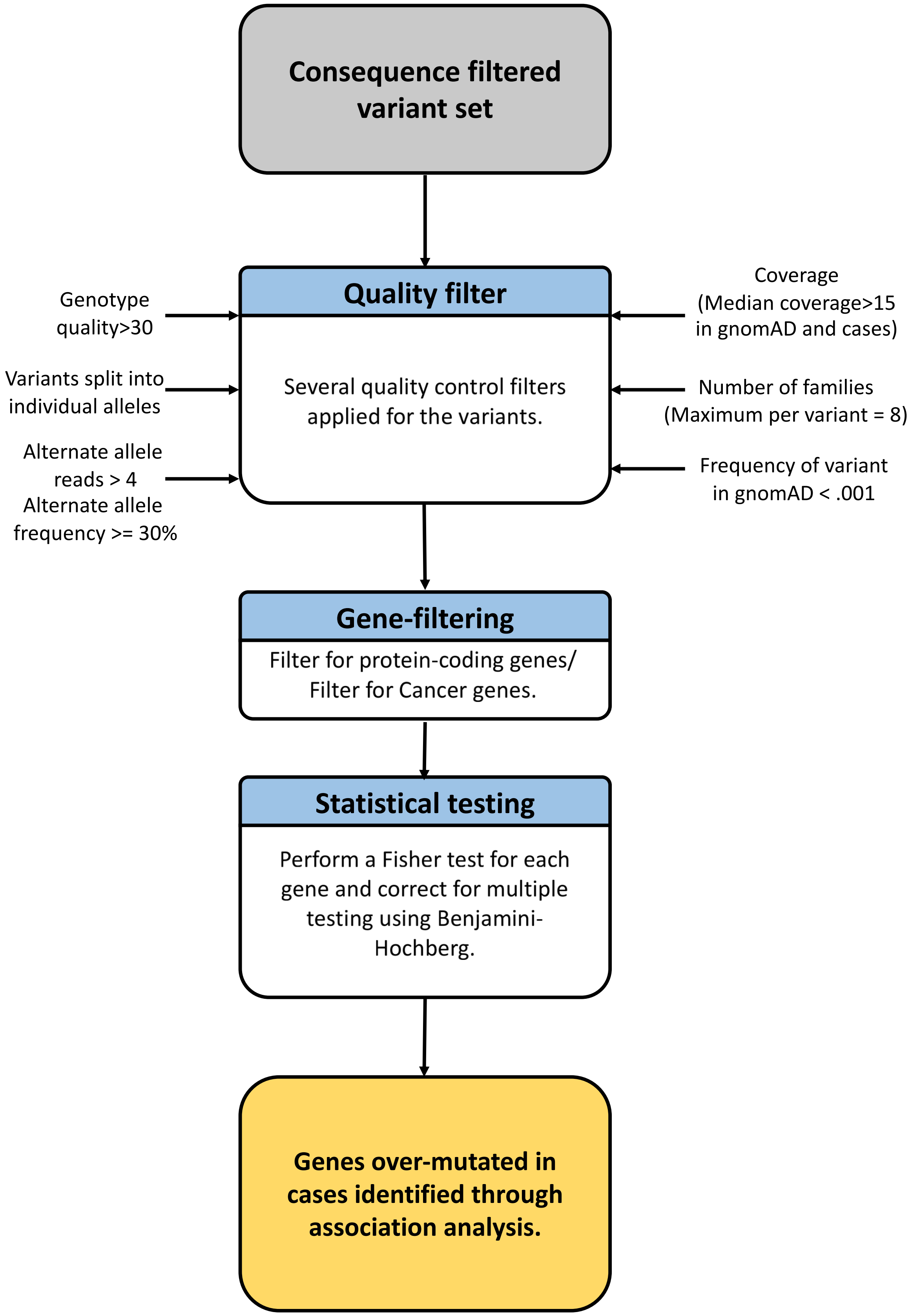

Consequence filtered variant set
Quality filter
Several quality control filters applied for the variants.
Genotype quality>30
Variants split into individual alleles
Alternate allele reads > 4
Alternate allele frequency >= 30%
Coverage (Median coverage>15 in gnomAD and cases)
Number of families (Maximum per variant = 8)
Frequency of variant in gnomAD < .001
Gene-filtering
Filter for protein-coding genes/ Filter for Cancer genes.
Statistical testing
Perform a Fisher test for each gene and correct for multiple testing using Benjamini-Hochberg.
Genes over-mutated in cases identified through association analysis.

Figure 2.5: Overview of the steps involved in determination of genes with an increase burden of mutations in cases through an association analysis.

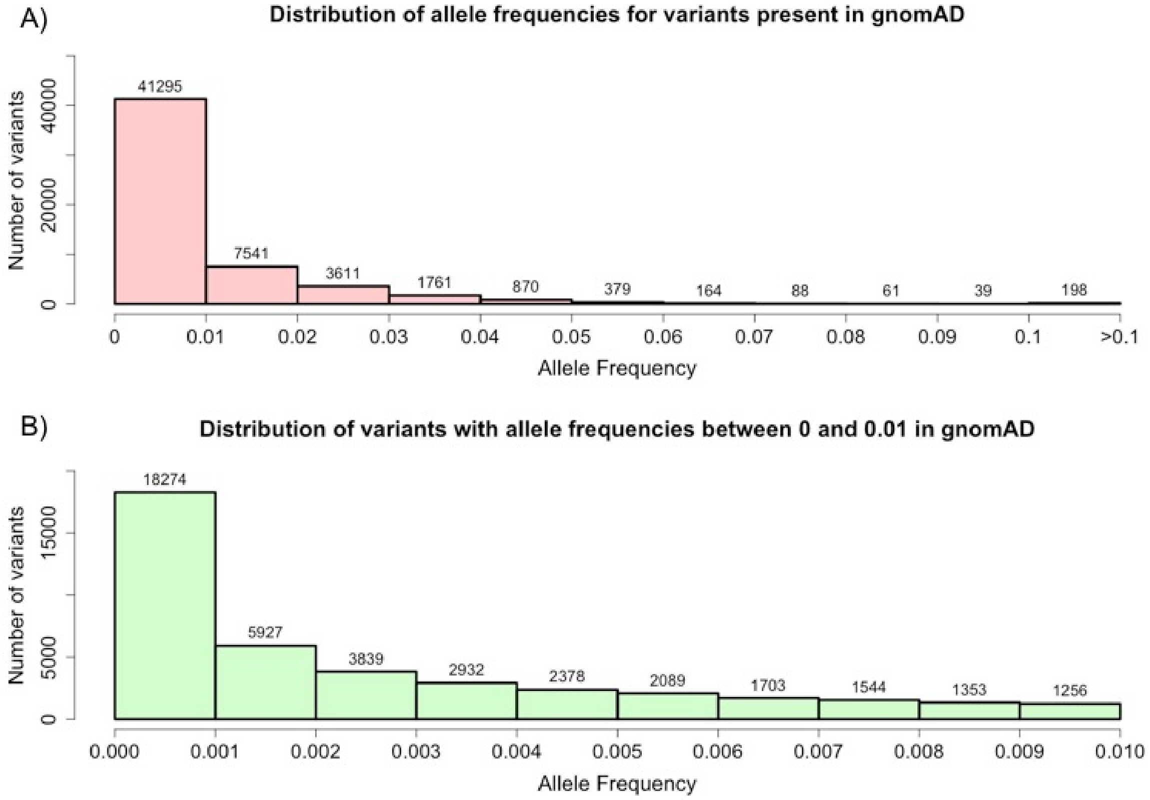


Figure 2.6: Distribution of allele frequencies for variants in cases A) This shows the distribution of allele frequencies for the 56,007 variants from the cases which were also mutated in gnomAD. 74% of the variants from the cases were present in less than 1% of the population in gnomAD. B) As the vast majority of variants had a low allele frequencies in gnomAD (<.01), the distribution of allele frequencies between 0 and 0.01 were plotted in this Figure. The distribution is again skewed as roughly 44% of the 41,295 variants had an allele frequency less than .001 (less than 1 in a 1000 people carried the variant). This was the chosen cut-off for the allele frequency.

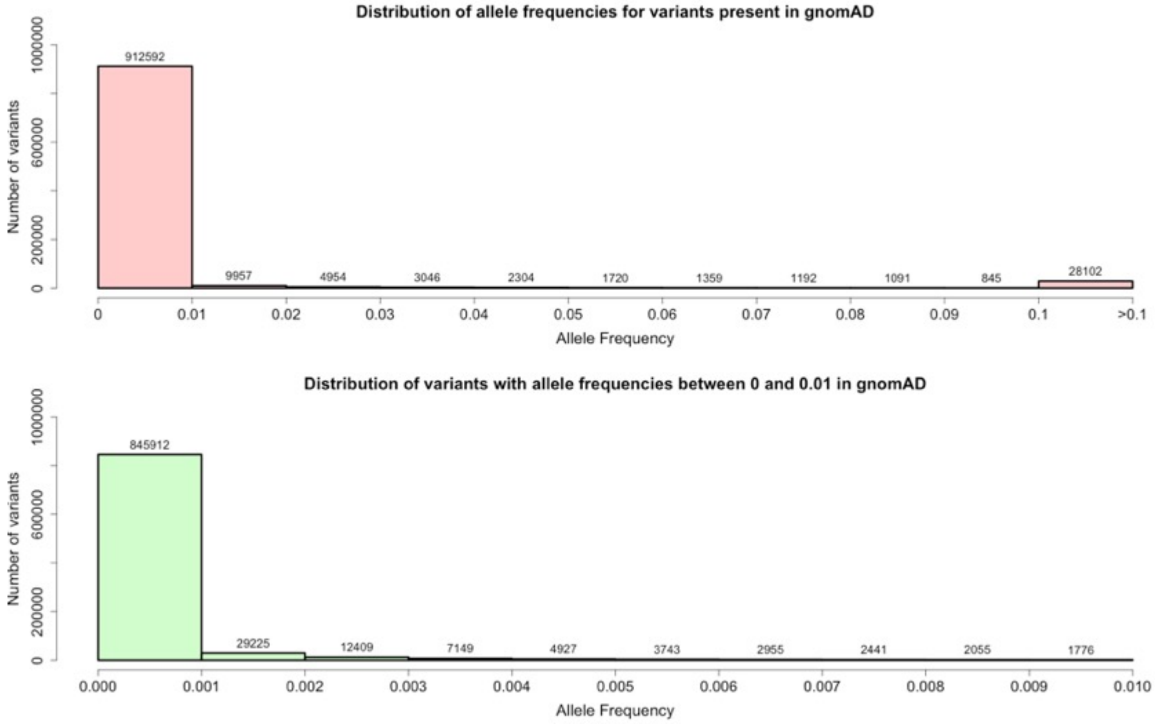


Figure 2.7: Distribution of allele frequencies for variants in controls.A) This shows the distribution of allele frequencies for the 967,162 variants from the gnomAD dataset. 94% of the variants occur in less than 1% of the population in gnomAD. B) This figure shows the distribution of allele frequencies between 0 and 0.01 for variants from gnomAD. 93% of variants occur at a very low frequency of less than 1 in a thousand.

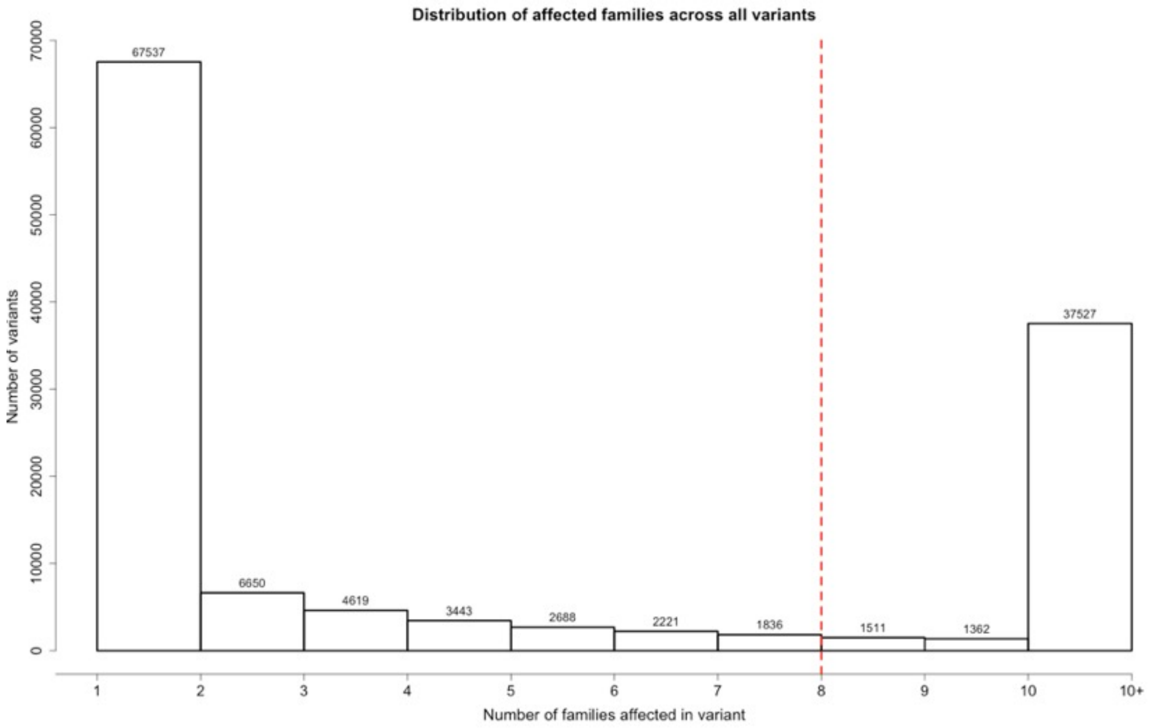


Figure 2.8: Distribution of affected family counts for all variants in cases. More than 50% of the variants in the cases were present in less than 2 families. An artefact threshold of 8 families was used, represented as the red line in the figure.

gnomAD dataset were first annotated with the coverages, which were obtained as a downloadable file. Individual coverages for the samples in gnomAD were not available; summary information across all the samples was provided for each variant. The following summary statistics were available (The number before the x indicates the number of reads covering each variant position): i) Mean coverage across all samples ii) Median coverage across all samples iii) Fraction of samples with a coverage $\geq$ 1x, 5x, 10x, 20x, 30x, 50x and 100x. A similar representation of coverages was established for the variants in the cases to enable a valid comparison with the controls. The read depth of case samples at all variant positions were determined by using the samtools depth command from Samtools v1.9[177]. In this manner, metrics identical to the gnomAD coverages were established for the cases. The coverages from the cases and the controls were annotated to the joint set of variants. Variants with a median coverage of at least 15 in both the cases and the gnomAD dataset were retained for further analysis, i.e., at least half of our cases had a coverage of at least 15x, and at least half our our controls had a coverage of at least 15x for the variant locus to our filters. The complete workflow of this process is shown in Figure 2.9.

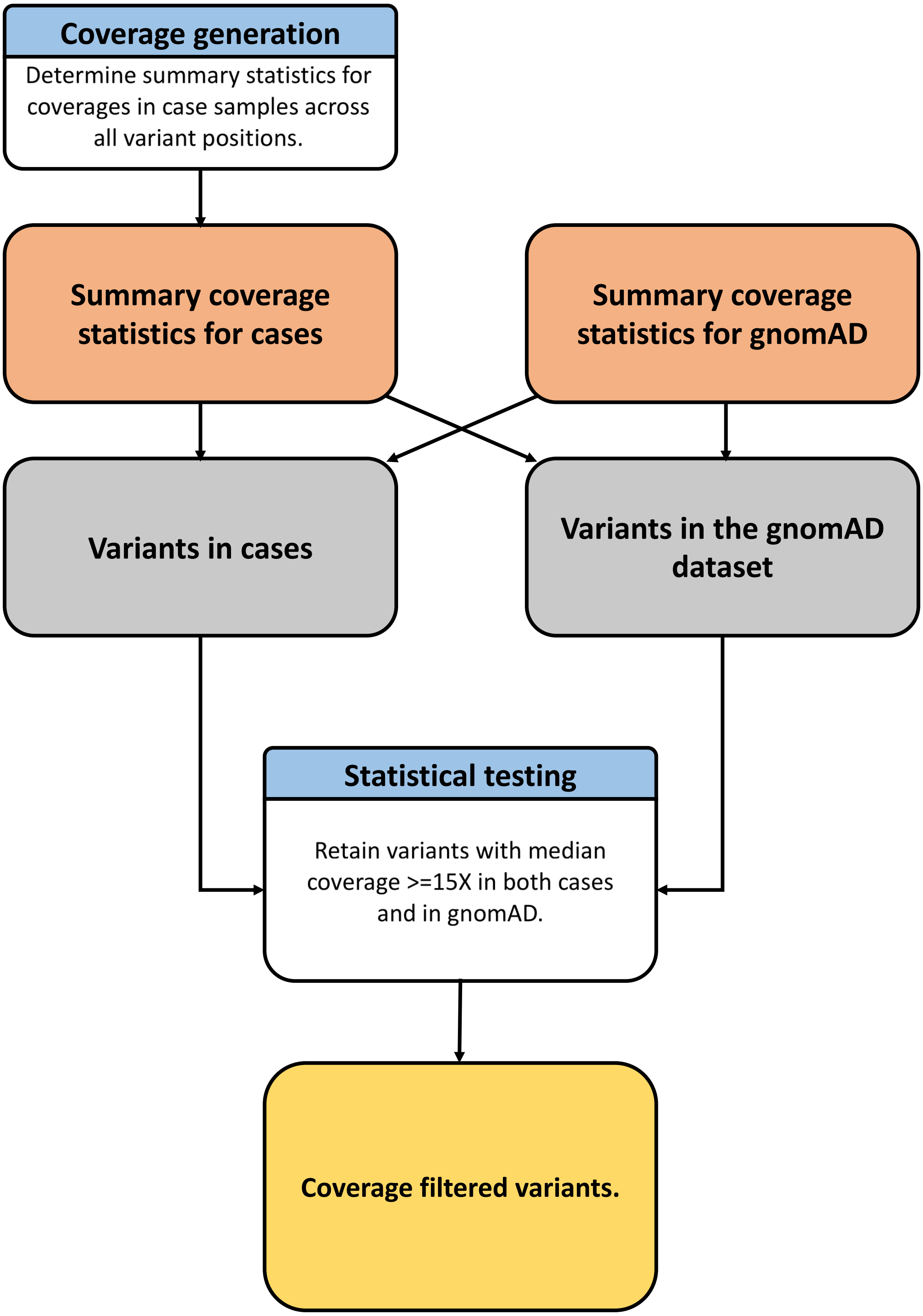

Coverage generation
Determine summary statistics for coverages in case samples across all variant positions.
Summary coverage statistics for cases
Summary coverage statistics for gnomAD
Variants in cases
Variants in the gnomAD dataset
Statistical testing
Retain variants with median coverage >=15X in both cases and in gnomAD.
Coverage filtered variants.

Figure 2.9: Description of coverage generation, annotation and filtration of variants in cases and controls.

2.6.3.3 Annotation and filtering case variants based on alternate allele read depth and frequency

Having removed poorly covered samples and poorly covered individual variants, the next filter applied was to maintain variant quality uniformly for every variant locus and sample. This filter was based on the allelic depths(AD) and the alternate allele read frequency. Specifically, the allelic depth refers to the number of reads supporting each allele at a variant position. The total number of reads is therefore the sum of the allelic depths for every allele at the position. The allelic frequency for every allele is estimated as the allelic depth for the allele/total number of reads at the variant position.

For a heterozygous variant, the allelic read frequency of both the reference and the alternate allele would be expected to be between 0.3-0.6. However, there were some cases where the total number of reads covering a position was greater than the applied threshold (15 reads) but the alternate allele frequency was much lower than 0.3. Such variants are of low quality and would increase the chance of miscalled variant. To account for this, the total number of reads, the allelic depths and the alternate allele read frequencies allele reads were determined at every variant position in the cases. Variants were retained if:

1. The total depth at the locus was at least 15x
2. The allelic depth for the alternate allele was at least 4x.
3. The alternate allele frequency (reads supporting alternate allele / total reads) was at least 0.3.

2.6.3.4 Annotation of cancer gene status

Previously identified familial melanoma genes (including *CDKN2A*, *BAP1* and *POT1*) were also mutated in other cancers, either through somatic or germline mutations. It is therefore expected that any mutation that is involved in the the emergence of melanoma would exist in a gene that is similarly affected in other cancers. A list of genes known to be affected in cancers, the Cancer Gene Census (CGC)[178],was utilized to affix significance to such variants.

The CGC is a regularly updated set of genes with additional data on the types of cancers affected, the type of mutations carried in these cancers. As of the version dated October 22nd 2018, the CGC comprised 719 genes in total. These genes were split into two tiers depending on their impact on cancer development and the evidence available to support the relevance of these genes:

1. Tier 1 - Genes in Tier 1 are considered to be the gold standard for cancer genes and they have compelling, documented evidence to support the relevance of the gene to cancers, including activity that drives cancer and activity that promotes oncogenic transformation. There are 574 genes annotated as Tier 1 genes. Examples include *AMER1*, *ATR*, *BAP1*, *CDKN2A* and *POT1*.

2. Tier 2 - Genes in Tier 2 have limited evidence to strengthen their claim as an important cancer gene but are still considered to play an salient role in cancer development. There are multiple reasons why these genes are not annotated in Tier 1: Lack of sufficient evidence, extremely rare cases, low burden of mutations or genes involved in cancer only through fusion. There are 145 genes annotated to be Tier 2 genes. Examples include *A1CF, CDKN1A*, *FAT3, SKI* and *ZEB1.*

The variants in cases and controls were previously annotated with the affected gene through VEP. These variants were annotated with CGC tier, if present. The tier list of the gene within the CGC was also noted. This was later used to filter the list of genes to focus on cancer genes for a component of the association analysis.

2.6.3.5 Calculation of total number of affected samples in genes

The power afforded by the low number of cases was too small to allow variant by variant association testing. Therefore, variant counts were collapsed for every gene to get gene counts in cases and controls instead. The gnomAD dataset comprised of individual unrelated samples, by contrast the cases consisted of related family members. This meant that the likelihood of two related family members carrying a mutation would be much higher than two unrelated control samples carrying the same mutation. In order to account for this, it was decided that the total number of affected families would be used to count cases, instead of the total number of affected individuals. This was chosen to prevent an overestimation of case counts due to the relatedness of samples. For each variant, every affected family in the cases was counted exactly once. The total number of affected families were estimated for every gene through this process.

For the variants in the gnomAD data set, the provided Genotype Count (GC), defined as the count of individuals for each genotype, was used. The GC was provided for every population subgroup including non-Finnish Europeans (GC NFE) which was used to determine the total number of controls with a mutation. GC NFE counts were summed across all the variants present in a gene to determine the total number of control samples carrying a variant for every gene. This is an approximation which assumes that the same individual does not carry two different variants within the same gene. This is a reasonable assumption since the variants have been filtered for a frequency of 1/1000 which would make it extremely unlikely for the same person to carry two such variants within the same gene.

2.6.4 Statistical testing to determine ranked list of genes

The processed variants from the cases and controls were set up with the following data structure which was used for further analysis:

1. The name of affected gene, its associated HUGO Gene Nomenclature Committee (HGNC) symbol and its corresponding Ensembl stable id.

2. Total number of families with and without a member carrying a filtered variant in the gene (as described in Sections 2.6.2 and 2.6.3).

3. Total number of non-Finnish European samples in gnomAD with and without a filtered variant in the gene (as described in Sections 2.6.2 and 2.6.3).

4. Presence of the gene within the CGC and tier list status, if present.

A 2x2 contingency table was created for every gene using this data. An example of a contingency table is given in Table 2.13.

Contingency table	Cases	Controls
With variants	4	51
Without variants	131	7458

Table 2.13: A 2x2 contingency table as identified for *POT1*.

These tables were used as the input for a Fisher's Exact test. A Fisher's Exact test is a statistical test that determines if there is an association between two categorical variables. The null hypothesis is that the two variables are independent, which in this case would be the variant status (number of people with a variant in the gene vs number of people without a variant in the gene) compared to the disease status (cases vs controls). Deviations from the null hypothesis would indicate that the presence of variants in the gene are associated with the disease status. A p-value is produced as the output of the Fisher's Exact test. As we expect to find an increase in the proportion of members with variants in the cases compared to the controls, a one-sided Fisher's Exact test is more appropriate. Along with the Fisher's Exact test, an odds ratio(OR) is also be estimated. An OR is a measure of quantifying the level of association between two properties, i.e., the ratio of the probability of occurrence of the first property in the presence of the second property compared to the probability of occurrence of the first property in the absence of the second property. Assuming that the disease status and the variant burden are our two properties of interest, the OR could be one of three outcomes:

1. OR<1 : The variant burden in the gene is associated with a lower probability of disease occurrence.

2. OR=1 : The variant burden in the gene is independent of the disease occurrence.

3. OR>1 : The variant burden in the gene is associated with a higher probability of disease occurrence.

For genes with a higher burden of mutations in the cases compared to controls, the OR would be >1 which would associate the presence of variants in the gene with a higher probability of disease occurrence. Thus, a one-sided Fisher's Exact test would show if a variant in a gene was associated with the disease, and the odds ratio would indicate the extent of association. The OR and the p-value are computed for the contingency table for all genes.

Thousands of genes present in alternative scaffolds, pseudogenes and non-coding genes were included in the analysis. These would be very unlikely to play a role in cancer development but would still affect the identification of overburdened genes, particularly when correcting for multiple testing. Two different filters were therefore applied on the list of genes based on protein product : one was restricted to all the genes within the CGC while the other was restricted to all known protein-coding genes as defined on Ensembl. A one-sided Fisher's exact test (coded in RStudio) was then applied on the contingency tables for all of the genes on these lists, thus, producing p-values for every gene. These p-values were corrected for false discovery rate using the Benjamini-Hochberg method. The two lists were then sorted based on the corrected p-value, yielding two ranked lists of genes. These results are discussed in Section 3.3.

2.6.5 Limitations of an association analysis

The use of the gnomAD as a control dataset has helped identify protein-coding genes associated with familial melanoma occurrence. However, this approach has some limitations, which are listed here.

1. Information on individual sample genotypes are not available for gnomAD variants; aggregated variant level information is instead presented across all samples. While this is

sufficient for an association analysis, this comes with the caveat that it is not possible to identify a sample having multiple variants within the same gene. This risk is minimized due to a conservative variant frequency threshold of 1/1000, it is however still present.

2. The total number of cases with a variant in a gene is measured by the total number of families with individuals who have at least one variant in the gene. Families with a higher number of sequenced members have a higher probability of having a mutation. By contrast, the controls consist of unrelated samples. Additionally, a variant that is present in all sequenced members of a pedigree with 11 members would normally be considered to be a much stronger candidate than a variant that is present in a pedigree with 2 members. However, such variants cannot be distinguished in this scenario.

3. Another drawback to determining the number of cases as the number of families carrying a variant in the gene is the loss of information regarding the segregation of variants within the pedigree. For example, in a family with 4 members affected (all of whom have been sequenced), a variant that is present in all four affected members is much more likely to be involved in melanoma development compared to a variant that is only present in only one of the members. An addition of linkage analysis is required to account for this, this is discussed in Section 2.7.

4. While this approach determines an increased burden of mutations in genes, it is restricted to non-synonymous coding region mutations. It does not account for the potential role of non-coding region mutations or structural mutations in the development of familial melanoma. The investigation of non-coding and structural mutations in melanoma development and its resulting outcomes are discussed in Sections 2.11,2.12,3.8 and 3.9.

2.7 Linkage analysis

2.7.1 An introduction to linkage analysis

Association studies, particularly GWAS, are normally aimed at identifying a set of common risk alleles with low impact on the disease. As a result, they do not explain the cause of a disease in a large percentage of cases, particularly for the disorders caused by rare, highly penetrant mutations. This issue can be addressed through linkage analysis, which is primarily used to detect and identify variants with large effect size or impact on the disease. Linkage, in the context of genetics, is defined as the propensity for a group of genetic regions present on the same chromosome to be transmitted together from a parent to an offspring during meiosis. This is represented with a logarithm of odds (LOD) score. A LOD score is a statistical estimate of the likelihood of two regions being inherited through linkage as compared to two regions being inherited through chance; the higher the LOD score, the stronger the linkage. Linkage analysis refers to the set of methods that use linkage to help determine the segments of the chromosome which segregate along with the disease phenotype through affected and unaffected individuals belonging to the same family.

Linkage analysis was originally used in the identification of genomic regions with strong linkage signals. Unlike GWAS, it did not require a comprehensive set of markers and could determine co-segregation of the trait with the marker on a larger scale. Genetic markers for disorders including cystic fibrosis [179] and Huntington's disease [180] were originally determined through linkage analysis. The two most prominent familial melanoma driver genes identified to date, *CDKN2A* and *CDK4* were also initially identified through linkage analysis[181]. Other major cancer genes that were discovered through linkage analysis include *BRCA1* and *BRCA2*[182, 183], which are collectively responsible for 90% of hereditary breast cancers globally[184].

However, linkage analysis did originally have several drawbacks, particularly in the context of identifying a genomic region with strong linkages to a disease where no causal gene could be identified. It could not help in the determination of the exact variant responsible for the presence of linkage signals within a given region. If the study focussed on exonic regions, linkage peaks in regulatory regions present in the non-coding part of the genome could not be identified. Innovations and advances in next-generation sequencing, particularly with cheap sequencing costs of whole genome sequencing have helped resolve these issues and enabled a joint association-linkage approach.

2.7.2 The joint association and linkage approach - pVAAST

Variant Annotation, Analysis and Search Tool (VAAST) is a probabilistic tool developed in 2011 which was aimed at identifying disease causing genes from genome sequences[185]. Originally developed for personal genomes, an updated version of the software called pVAAST (pedigree VAAST) was released in 2014 for the analysis of genetic data from high throughput sequencing of related individuals [186]. pVAAST takes the germline sequences of affected individuals, affected/unaffected relatives of these individuals and unaffected controls as its input. The variants in the cases are compared to variants in the controls using a composite likelihood ratio test (CLRT) [185]. In a CLRT, variants in genes are grouped together along with the information on the frequency of the variants in the cases and controls. A composite likelihood score is then estimated for these variants based on the observed frequencies of the variant in the cases and controls.

Alongside the association, pVAAST also computes linkage scores in the form of LOD scores for pedigrees with more than one affected individual. The estimated LOD scores are unlike typical LOD scores as they compute linkage of entire genes with an associated trait as opposed to individual variants. The LOD scores of each gene are cumulative across all variants present in the given gene across all pedigrees. This helps provide a single linkage peak for each gene across all cases compared to individual linkages for each family for each gene. These LOD scores are combined with the variant frequency information, annotated consequences for each variant using a built-in dataset and the association analysis to provide a prioritized list of genes and variants where the genes are ranked based on their association with the phenotype, shown in Figure 2.10.

In order to run pVAAST on the data, the following input files were required:

i) Variant files in VCF format for cases.

ii) Variant files in VCF format for the controls/background.

iii) A list of genes on which scoring is to be performed in a general feature format (.gff3).

iv) The human reference genome in the Fasta format.

v) The framework file for the different parameters used in running pVAASt in a control format (.ctl).

vi) A pedigree file containing the following information for each affected individual - the family id, the sample id for the individual in each family, the id of father of the individual, the id of the mother of the individual, gender (represented as 1 for male and 2

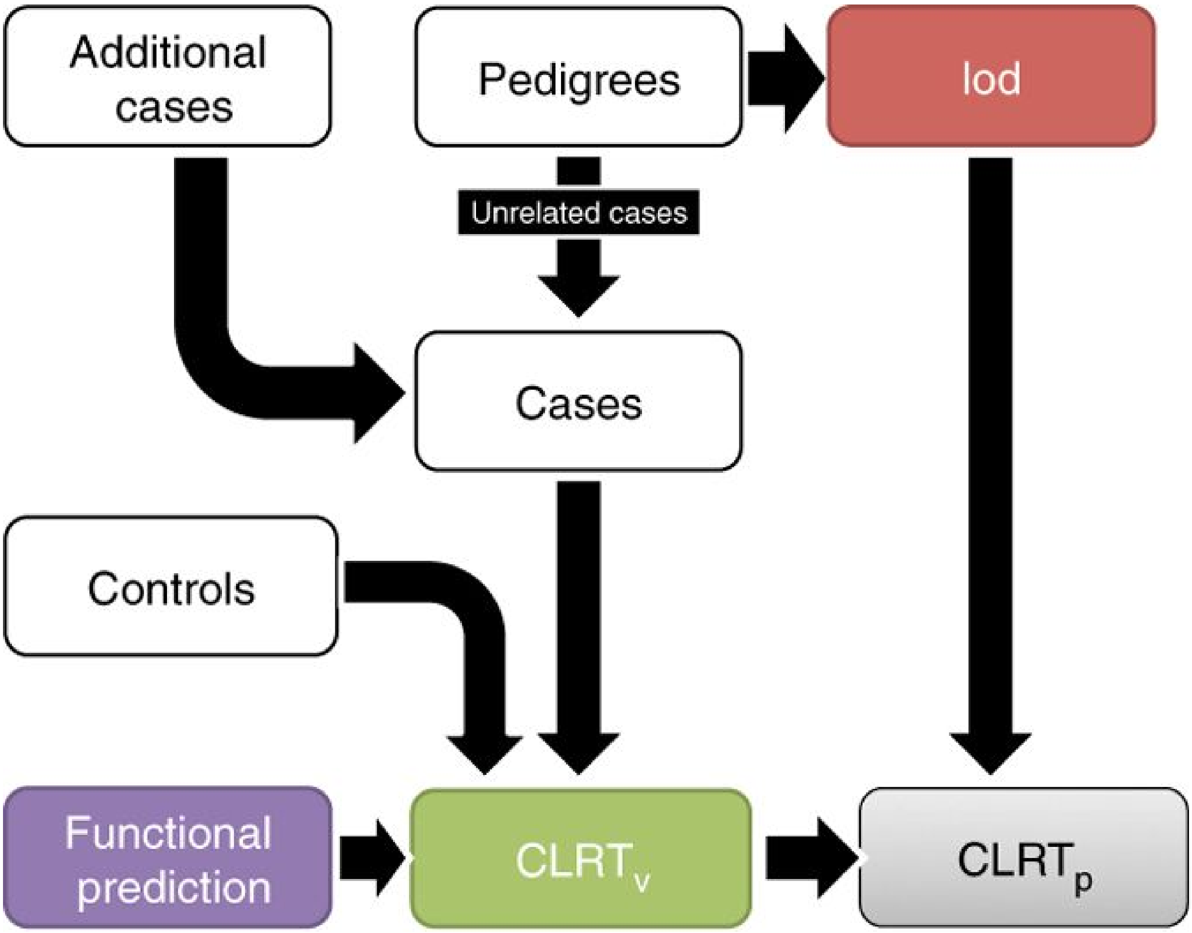


Figure 2.10: A graphical representation of the scoring pVAAST. Reproduced with permission from [186].

for female) and affected status for melanoma (represented as 1 for unaffected and 2 for affected). This file is to be in the pedigree format (.ped).

The following output files are produced by pVAAST during the process of analysing the data:

i) Variant information for each individual in the cases in the Genome Variation Format (.gvf) defined by the Sequence Ontology Group.

ii) Annotated variant files generated for each individual (.vat). This is performed using the inbuilt variant annotator tool and includes information on variant id, the position, the affected gene and the effect of the variant. The format is similar to the .gvf file and the output of this step is usually a .vat.gvf file.

iii) A condenser file containing condensed representation of the variants across (.cdr) with one file for each pedigree, one file for all singleton families combined and one for the background samples.

iv) A Vaast file containing the output of the pVAAST runs including CLRT and LOD scores (.vaast).

In order to maximise computational efficiency, each pVAAST run was performed using variant information from all pedigrees for a single gene. This was repeated for all genes in the cancer gene census list as it was not computationally feasible to run this across the entire genome. This allowed for parallelization of the pVAAST runs as the CLRT scores and the LOD scores for each gene could be directly compared with each other. The generation of the input files, the intermediate files and the output files along with the results are described in Section 2.7.3. The parameters used for each step are given in Supplementary Table 3.

2.7.3 Methods

The filtered list of variants used in the association analysis from Section 2.6.3.3 were used for the joint association and linkage analysis through pVAAST. A bed file containing a list of unique positions from these variants was generated. As pVAAST requires a VCF file as an input; a VCF file with the variants at the filtered positions was generated for the cases. All the files used by pVAAST for annotation and filtration of variants were built on and aligned to GRCh37 reference build. As a result, the VCF file for the variants from the cases was lifted over from GRCh38 to GRCh37 using CrossMap[176]. The new VCF file with the variant positions corresponding to the GRCh37 reference build was sorted and indexed using the

Tabix software. The next step in the procedure involved the generation of .gvf files from the multi-sample VCF file. This was necessary as pVAAST used .gvf files as the primary input for all downstream steps, primarily for variant annotation and condenser file generation. GVF files were generated for every sample in the multi-sample VCF file using the build in vaast converter tool available as part of the pVAAST package. Each GVF file was then sorted in place, meaning that no duplicate files were created in the sorting process.

A PED file for containing the pedigree information for all the families with multiple sequenced members was manually created as a text file. pVAAST has certain inbuilt requirements for the structure of the pedigree : there could be no consanguineous marriages, pedigrees could only have the extended family on one side (either the father or the mother but not both) and only two-generation nuclear families could be analysed in the recessive model. No pedigrees or samples were removed in this process but unaffected extended members of some pedigrees had to be pruned to account for these conditions. Pedigrees with a single sequenced member, or singleton families, were grouped together as unrelated affected individuals as linkage analysis would not have been possible for such individuals. The cases were therefore separated into two groups in this manner.

The sorted GVF files of each individual were then annotated with their impact on the genes that they were present in using another pVAAST program called Variant Annotation Tool. This process required .gvf files, the human reference genome and the list of genes and their positions as the input files. The .gvf files were generated in the previous step of the process while the reference genome file (FASTA format) and the list of genes (.gff3 format) were provided by pVAAST. This produced .vat.gvf files, described in Section 2.7.2, for every affected individual.

In order to run pVAAST across each pedigree, we need to group variants from samples belonging to the same pedigrees together. This is performed through a process called variant selection utilizing the Variant Selection Tool (VST) which is a part of the pVAAST package. It performs set operations on the .vat.gvf files such as intersection, union, complement and difference of variants within a given set of gvfs. The .vat.gvf files for all sequenced members from each pedigree are used as the input for VST. The union of variants across all the samples in the pedigree and the output produced is a condenser file or a .cdr file, described in Section 2.7.2. As the singleton families are considered to be unrelated affected individuals, all singleton families were grouped together and a joint .cdr file was produced for them.

pVAAST provided a set of background genomes as part of its package based on the 1000 genomes dataset comprising of 1,303 sample. This was used as the background population for the first set of pVAAST runs. However, to validate these results, a secondary background dataset with more samples was required. As pVAAST required genotype information for individual samples, data from ExAC or gnomAD could not be used as the background. A set of 4,070 individuals were exome sequenced as part of the INTERVAL study of which the Wellcome Sanger Institute was a collaborative member[187]. These individuals were not enriched for any familial cancers or any other genetic disorders. Due to the presence of high quality sequence data and individual level genotypes, these samples were chosen as the second background dataset. A VCF file with variants from these samples were obtained from our collaborators within the Sanger Institute. These variants were filtered for artefacts and variant frequency similar to the cases using gnomAD. They were then processed similar to the cases to generate a CDR file.

The penultimate step for running pVAAST involved the generation of a parameter file or a .ctl file. This file contains the location of all the .ped files and .cdr files for each pedigree to be considered. The location of the .cdr for the additional cases or the unrelated affected individuals was also provided in this file. The inheritance model for the phenotype of familial melanoma was defined as a dominant model. Additional parameters that were provided within this file include the genotyping error rate, filtering of gene scores based on CLRT and LOD scores, the mode of scoring the gene each gene based on CLRT and LOD scores and the *de novo* mutation rate. This file remained unchanged for every pVAAST run, thus allowing the scores from each run to be compared.

Once the .ctl file had been produced, pVAAST was run across all the pedigrees. The input for each run included the .ctl file for the cases, the background .cdr file for comparing variant frequencies, the region of the genome within which pVAAST calculated the scores and the feature file containing information on all genes present in the genome. Each run of pVAAST resulted in a .vaast output file containing the CLRT and LOD scores for all variants and genes within the specified region. In order to parallelize the process, each pVAAST run was restricted to a single gene. pVAAST was run for all the genes in the Cancer Gene Census as it was not computationally feasible to run it across all the protein-coding genes in the genome. This was done using both the original 1000 genomes project background file and the INTERVAL exomes background file. The results from these runs are given in Section 3.4.

2.7.4 Limitations

A joint association-linkage approach has helped determine novel genes and variants involved in familial melanoma onset which would not have been possible to discern through a straightforward association analysis. However, such an approach still has its limitations:

i) A single run of pVAAST for 135 families in the region of a single gene requires 95-105 hours of computational time to process. As a result, genome wide runs of pVAAST are currently not feasible if there are multiple pedigrees to be analysed.

ii) Extended pedigrees cannot be analysed using this approach as it currently exists. Large pedigrees have to be pruned significantly to be analysed. Additionally, families with multiple affected individuals but low number of sequenced individuals would yield low LOD scores which would impact the scoring of genes.

iii) The annotation provided by the Variant Annotation Tool for the consequences of the variant is not as accurate as the Variant Effect Predictor. This is because VEP is updated more regularly. There is also currently no suitable method for comparing the results of VAT with VEP directly.

iv) Whilst the INTERVAL exomes provide a matching background and were suitable for this approach, a larger dataset with additional samples would provide a much more stringent comparison of variants and provide more accurate results. Additionally, the method has estimated high scores for large genes with multiple functional domains but little to no significance in cancer development such as the *MUC* and *FAT* family of genes. This indicates that the background dataset is not powerful enough to discern and filter out variants in such genes even after stringent quality filters for the variants.

2.8 The search for variants in known driver genes

Previous studies have helped determine several driver genes involved in the development of familial melanoma, as described in Chapter 1. These genes include *BAP1*, *BRCA2*, *CDK4*, *CDKN2A*, *MITF*, *POT1* and *TERT.* The existence of any variants in these genes would explain the presence of familial melanoma in the pedigrees carrying such variants and thereby make it unlikely that these pedigrees also carried other novel, high-penetrant causative variants. Variants in these driver genes were therefore analysed concurrently with the association analysis described in Chapter 3, to determine if they were causative of disease onset in any of the families. Such variants were then annotated with clinical relevance to disease onset, particularly with respect to their connection with hereditary cancer. This information was obtained from ClinVar[188]. The results from this investigation is presented in Section 3.5. Analysis of potential pathogenic variants in all other genes as defined on ClinVar is discussed later in Section 2.10.

2.9 Variants with high segregation within the cases

Variants which are present in all sequenced members of our pedigree are more likely to be causative of the disease for the pedigree. Due to low power, such variants/genes might not be found carrying a higher variant burden in an association study but might still be responsible for the emergence of melanoma within the pedigree. To discern the presence of such variants, variants from Section 2.6.3 were then filtered based on the proportion of samples carrying the variant in the families with the variant. This was represented through a value called the segregation proportion which was defined as: Segregation proportion (SP) = Total number of individuals carrying the variant/Total number of sequenced individuals in pedigrees where the variant is present

A variant is defined as completely segregating within a family if its SP =1, i.e., every sequenced member in the family carries the variant. While complete segregation is ideal for the determination of novel variants, phenocopies are also known to occur in the context of cancers including melanoma [189]. In order to account for phenocopies, variants with high segregation were determined as follows:

i) Variants were removed if there were no affected pedigrees with multiple sequenced members.

ii) Variants were retained if SP ≥ 0.85 , i.e., at least 85% of the samples sequenced in every pedigree containing the variant carried the variant.

iii) Variants were also retained if they were present in a family such that the number of members carrying the variant was at worst one lesser than the total number of sequenced members in the pedigrees. This was to account for cases where the SP would be less than 0.85 but the segregation is still high enough within a single pedigree to warrant further investigation. An example of such a case is the *CDKN2A* missense mutation described in Section 2.8: Three out of four sequenced members carry the mutation; the SP for this variant would therefore be 0.75. However, the variant is still considered interesting as the member without the variant is believed to be a phenocopy.

2.10 Pathogenic variants in ClinVar

A rare-variant association analysis would help in the identification of genes with an increased burden of rare variants in familial melanoma patients. However, there are still a few cases

Clinical significance	Interpretation
Benign/Likely benign	Variants that are not considered to affect disease onset and progression.
Pathogenic/Likely pathogenic	Variants that are considered to affect disease onset and progression.
Uncertain significance	Variants whose impact on the disease are unknown.
Drug response	Variants that disrupt the efficacy of a drug without affecting the disease.
Association	Variants identified in genome-wide association studies.
Risk factor	Variants which contribute to the pathogenicity of a disease without being causative.
Protective	Variants that reduce the pathogenicity of a disease.
Affects	Variants that are not related to disease but are linked to specific disruptive phenotypes.
Conflicting data from Submitters	Variants submitted by a single consortium but with conflicting interpretations of the significance.
Other	Variants that do not fit under any of the above categories such as variants with functional significance but no clinical significance, literature reports with no supporting evidence etc. belong here.

Table 2.14: Classification of clinical significance of variants in ClinVar. The interpretations marked in bold are obtained from the guidelines recommended by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology[190]. Adapted from the online documentation of ClinVar.

where the the development of melanoma within the family is caused due to a single, highly pathogenic variant with a prior role to cancer development. Such mutations might not be identified through a rare variant association analysis and thus, the reason for melanoma development in such pedigrees may be undetected. If there were known phenotypes for the variants observed in our cases, then we may be able to better link the variant with the disease. Data from ClinVar, a repository of variant phenotype relationships[188], was utilized for this purpose. In addition to classifying the clinical significance of variants as being benign or pathogenic, ClinVar also contains meta information on the variant (including information on the protein product and transcript) and the supporting evidence for the classification of the variant. The classification of clinical significance for a variant is shown in Table 2.14 while the different review statuses provided by ClinVar for the variant based on the supporting evidence provided is shown in Table 2.15

Variants from ClinVar were downloaded as a VCF file (version dated September 30th,

Description	Review status
Practice guideline	practice guideline
Reviewed by an expert panel	reviewed by expert panel
Two or more submitters with assertion criteria and evidence provided the same interpretation.	criteria provided, multiple submitters, no conflicts
Multiple submitters provided assertion criteria and evidence (or a public contact) but there are conflicting interpretations. The independent values are enumerated for clinical significance.	criteria provided, conflicting interpretations
One submitter provided an interpretation with assertion criteria and evidence.	criteria provided, single submitter
The allele was not interpreted directly in any submission; it was submitted to ClinVar only as a component of a haplotype or a genotype.	no assertion for the individual variant
The allele was included in a submission with an interpretation but without assertion criteria and evidence.	no assertion criteria provided
The allele was included in a submission that did not provide an interpretation.	no assertion provided

Table 2.15: Review status classification of supporting evidence for variants in ClinVar in descending order of quality. Adapted from the online documentation of ClinVar.

2018). Variants in our cases at locations common with the ClinVar VCF were identified. Sample and family information were annotated to our case variants; the information on the variants from the cases was then merged with the information from ClinVar into a single file, one line per variant. Variants with more than 8 affected families were removed as artefacts, as described in Section 2.6.3. To restrict the analysis to interesting variants, the reference and alternate alleles from ClinVar were compared to the reference and alternate alleles from the cases. Only variants with matching alleles were retained. Finally, the variants were restricted to having one of the following clinical phenotypes from Table 2.14: pathogenic/likely to be pathogenic, risk factor or protective. These were chosen as they were most likely to be disruptive to the protein product and to lead to disease onset. The results from this analysis are discussed in Section 3.7.

2.11 Non-coding variants

2.11.1 Background

2.11.1.1 Introduction

Previous analyses of cancer genomes have been restricted to the coding region due to an interest in the determination of SNPs that disrupt protein function and the restrictive cost of whole genome sequencing[191]. Large scale exome sequencing studies have helped identify several key mechanisms and genes involved in the development of familial melanoma[143][145]. The advent of next-generation sequencing technologies has drastically reduced sequencing costs, from around $14 million in 2006 to $1,500 in 2016, which has enabled cheap sequencing of whole genomes[192]. This has provided new avenues for the investigation of the importance of sequence variation data in disease onset. A facet of analysis that remains unexplored in this context are variations in the non-coding region of the genome. It is estimated that 1.5% of the human genome encodes a gene, which leaves ~98% of the genome as non-coding DNA[193]. While the role of these regions in genetic regulation was unknown for a long time, it is becoming increasingly evident that the non-coding genome encompasses key regulatory elements which play an important role in the transcription and translation of proteins.

Non-coding elements can broadly be classified into cis-regulatory elements and trans-regulatory elements. Cis-regulatory elements (CRE) are generally found in the vicinity of the gene that they modulate and control gene regulation through intramolecular interactions, i.e, the components are active in the same gene[194]. Such elements include promoters, silencers and nearby regulatory elements. Modification of sites in cis-regulatory elements therefore directly impact the activation of a gene. CREs can also be distal, i.e., hundreds of kilobases away from the gene of interest. Enhancers and insulators are examples of distal cis-regulatory elements which activate and repress transcription of the gene respectively[195]. Such elements interact with the promoter/gene in the three-dimensional structure of the genome through chromatin looping regulated by proteins including CTCF and cohesion, as shown in Figure 2.11[196]. Trans-regulatory elements do not directly interact with the gene or the promoter, rely on intermolecular interactions with cis-regulatory elements, and usually encode for transcription factors[194].

2.11.1.2 Transcription factors and sequence logos

Transcription factors (TF) are a family of proteins that bind to the DNA, usually the promoter of a gene, in a sequence specific manner and either activate or repress the transcription of a

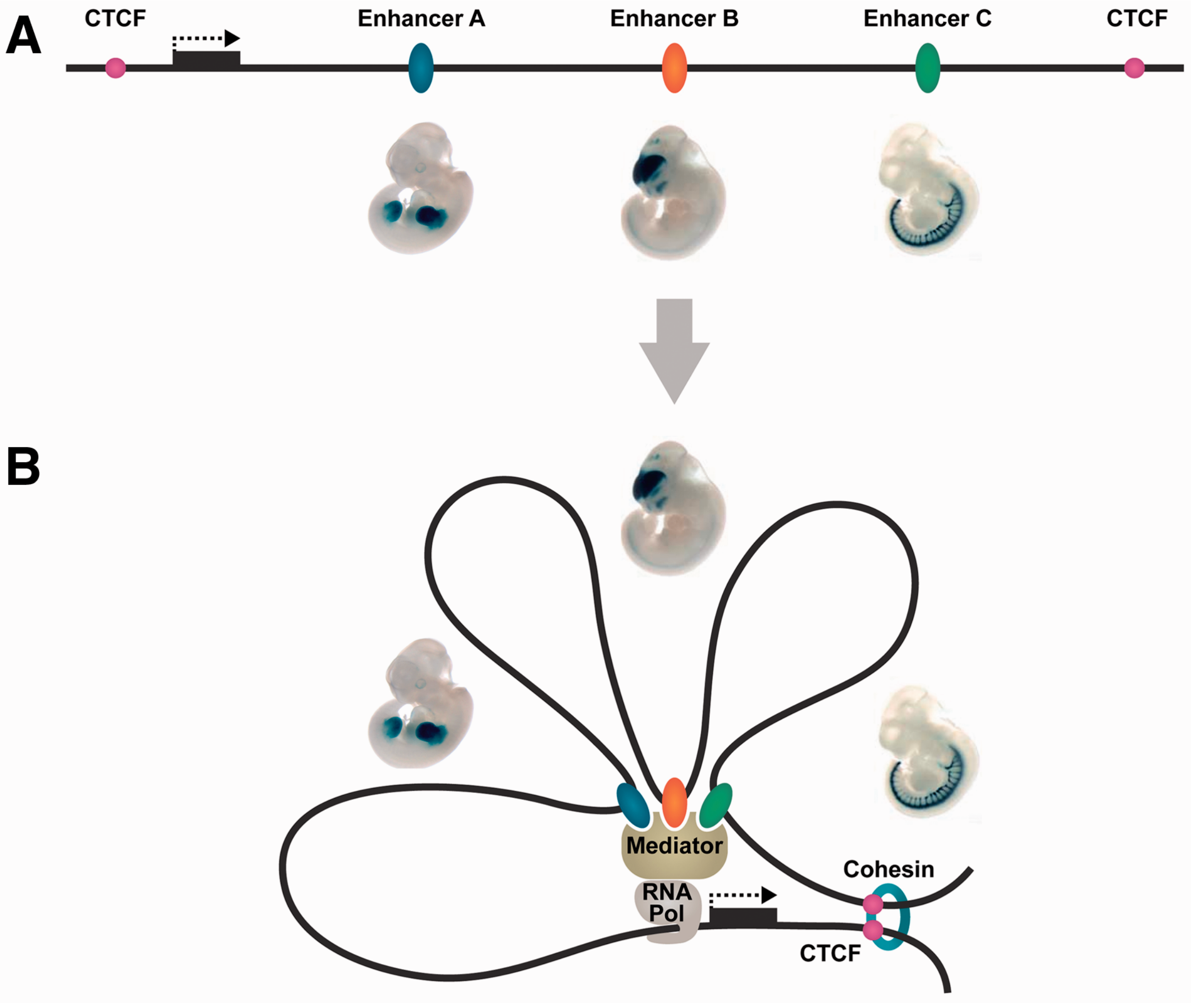


Figure 2.11: Depiction of chromatin looping to show enhancer-promotion interaction moderated through mediators, CTCFs and cohesion. Figure A shows the linear arrangement of proteins and enhancers on the chromatin while Figure B shows the interaction of the different enhancers with the RNA polymerase at the promoter through a mediator protein by chromatin looping. Figure reproduced with permission from reference [196].

gene[197]. The region of the DNA that each transcription factor binds to is defined as a transcription factor binding site (TFBS). A single transcription factor can regulate the transcription of several genes and binds to multiple locations across the genome. Non-coding variants can disrupt normal regulation of transcription by either creating or distorting the interaction between transcription factors and the DNA, usually in the promoter. The conserved sequence that represents the bases across all transcription factor binding sites for a given transcription factor are known as transcription factor binding motifs (TFBM). These motifs are visually represented through sequence logos[198]. An example of a sequence logo is shown in Figure 2.12.

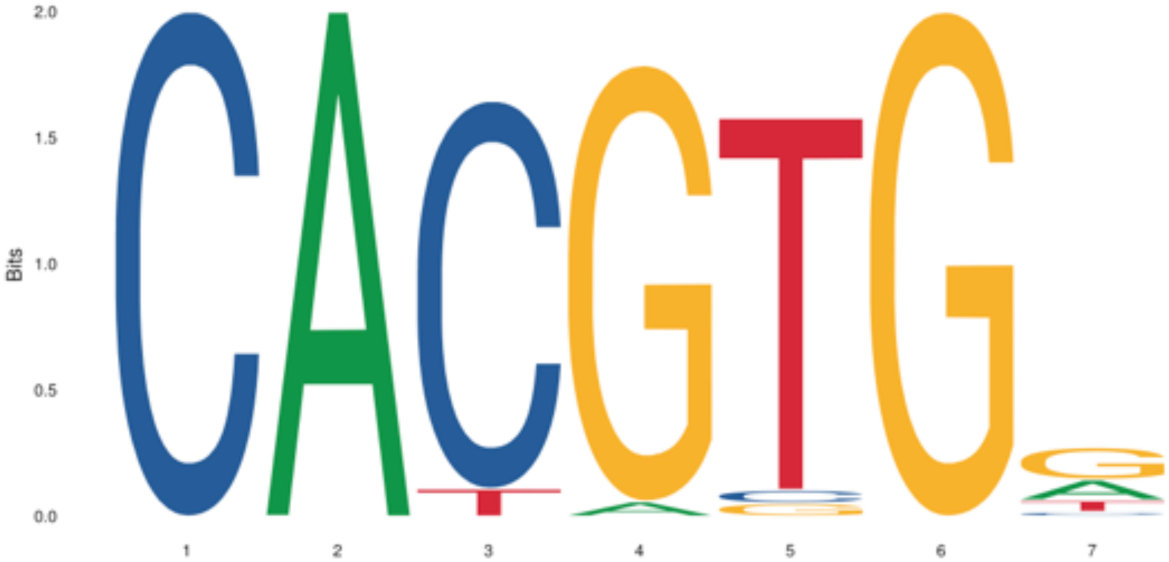


Figure 2.12: An example of a sequence logo for a transcription factor binding motif. Obtained from the JASPAR[199] database.

The x-axis of the sequence logo represents the different bases across the TFBM while the y-axis represents the combined frequency of all nucleotides through bits. A bit measures the total amount of information present at every position of the sequence and is associated with the answer to a binary question[198]. In the example shown in Figure 2.12, the base at the first position of the motif is always a cytosine. This means that two binary questions need to be answered: Is it a purine? If not, is it a cytosine or a thymine? The answers to these questions are represented as bits of information. If there are multiple possible bases at a given position such as position 7 in Figure 2.12, the amount of bits available at the position changes accordingly. The height of every nucleotide at each position constitutes the relative frequency of that nucleotide at that position.

As shown in the figure, not all the positions across a transcription factor binding site are equally conserved, some positions are more conserved than others. If a base is highly conserved across all transcription factor binding sites for a transcription factor, it suggests that the base is essential for the transcription factor to interact with and bind to the DNA. Variants in positions that are highly conserved, such as position 1 in Figure 2.12, would impair the function of a transcription factor more than a variant in a position that is less conserved such as position 7.

2.11.1.3 The role of non-coding variants that modify the function of transcription factors in cancer

Variants in transcription factor binding sites that play a role in carcinogenesis have previously been observed both in the context of familial melanoma and in other cancers. Recurrent variants have been observed in the promoter of *Telomerase Reverse Transcriptase (TERT)* originally in sporadic and familial melanoma[131] and eventually in other cancers[140]. A germline variant observed in familial melanoma was responsible for the creation of a binding motif for the ETS family of transcription factors. This led to the recruitment of TFs including T-cell factors (TCFs) to the promoter region of *TERT* which led to increased expression. The mutation observed in the familial melanoma pedigree was 57 bases upstream of the transcription start site while the three mutations in the sporadic melanoma cases were observed at 124, 138 and 146 base pairs upstream of the transcription start site respectively. A follow-up study determined that such germline TERT promoter mutations were quite uncommon in familial melanoma[132]. Disruption and ablation of transcription factors have also been observed in melanoma. Recurrent promoter mutations in *SDHD* were found to disrupt the TFBS for two ETS transcription factors, GABPA and GABPB1, considered to be key regulators of melanoma driver genes including *TERT*[200].

Murine double minute 2 (MDM2) is a protein encoded by the gene *MDM2*. Promoter variants of *MDM2* result in increased binding affinity to the Sp1 transcription factor which results in the increased expression of *MDM2*[201]. Increased expression of *MDM2* represses the activity of the p53 pathway, accelerating cancer development. Variants in the promoter of *MDM2* have been associated with increased tumour formation in several types of p53-related cancers including Li-Fraumeni syndrome[202] and breast cancer[203].

While the *TERT* promoter mutations are in cis-regulatory elements that are in the vicinity of the gene, variants in distal regulatory elements have also been identified as playing a role in cancer onset. A binding motif for the myeloblastosis family of transcription factors (*MYB*) was created through somatic variants in enhancers upstream of an oncogene called T cell acute lymphocytic leukaemia 1 (*TAL1*). This leads to the overexpression of *TAL1* which results in T-cell acute lymphoblastic leukaemia[204].

2.11.2 Methods

In order to focus on the prospective importance of TFBM disruption in familial melanoma development, the methods and results in the following sections are restricted to the variants within the whole genome sequenced individuals. Only variants that were present within known TFBMs in *Homo sapiens* were utilised for this analysis. For this purpose, the start and end sites of TFBMs along with the JASPAR binding matrices for all known transcription factors in *Homo sapiens* was obtained from the Ensembl Regulation Database v91. Ensembl includes information on transcription factors present in alternative chromosome haplotypes in addition to the normal human chromosomes. To restrict our analysis to relevant variants, only motifs in chromosomes 1-22, X and Y were considered. A bed file comprising the chromosome, the start position of the motif and the end position of the motif was generated from the list of motifs ; the bed file was then sorted on the chromosome and the position. Variants in the whole genome sequenced cases that were located within these TFBM regions were identified. Information on the chromosome, position, reference allele, alternate allele and the consequence of these variants were extracted from the VCF file for the cases and stored independently. After the removal of duplicate variants, this file was sorted based on the chromosome and nucleotide position of the variant. Multi-allelic variants were split into one variant per alternate allele. Information regarding the number of individuals, families and segregation of variant within all sequenced members of the families at each position was annotated to this file. Variants that were present in more than 5 families were removed as they were considered to be sequencing artefacts.

The whole genome non-Finnish European samples from gnomAD were chosen as the control set for this analysis. This comprised summary genotype information for every variant from 7,509 whole genome sequences. Variants from the controls which were present in the binding motifs as determined from Ensembl were identified. Information on the chromosome, position, reference allele, alternate allele and the consequence of these variants were extracted from the control VCF file. This file was sorted based on the chromosome and position of the variants after duplicate variants were removed. Multi-allelic variants were split into one variant per alternate allele. As gnomaD only provided information on summary statistics for samples and not individual genotypes, this resulted in additional duplicate variants as this reported one variant per affected transcript of gene. Every variant at a given position had the same number of affected samples regardless of affected transcript, such duplicate variants were therefore removed. Variants from the cases and controls were then jointly processed.

The mean and median coverage of every position within the cases and controls were determined. Positions with median coverage below 10 reads were removed from both sets of variants. In order to focus on rare variants with a potential impact on the function of TFs, variants in the cases were annotated with the allele frequencies of the variants from gnomAD. If the case variant was present in gnomAD, the frequency of mutation from the gnomAD dataset was annotated to the variant. If the case variant was not present in gnomAD, it was annotated as being absent. Variants with gnomAD frequency > 0.05 were removed from the cases and controls. Variants that were not present in gnomAD were retained in the cases. Several steps that were performed for the association analysis of variants in the coding regions as discussed in Chapter **??**, including the annotation of cancer gene status (Section 2.6.3.4), calculation of total number of affected samples for cases and controls (Section 2.6.3.5) and generation of 2x2 contingency tables based on sample counts for every gene (Section 2.6.4) were replicated for this analysis. P-values were generated for every gene from these tables using the Fisher's exact test, also as described in Section 2.6.4. These values were then corrected for false discovery rate using the Benjamini-Hochberg method. A workflow for these steps is shown in Figure 2.13.

The complete results from this methodology is discussed in Section 3.8.

2.12 Structural variants

2.12.1 Background

2.12.1.1 Introduction

There are two major types of modifications in the human genome that are known to play a role in cancer development. They can classified based on the size of the modification into small variants and large variants.

1. Small variants consist of single base alterations (single nucleotide polymorphisms) and small insertions or deletions of base pairs (indels). Previously, indels were considered to be any variation that were between 1000 bp [205] to 10,000 bp[206] in length but recent studies have identified indels as variants that are less than 50 bp long[207].

2. Large variants comprise of structural variants in the chromosome which change the structure of the affected segment of the genome. They are between fifty to millions of

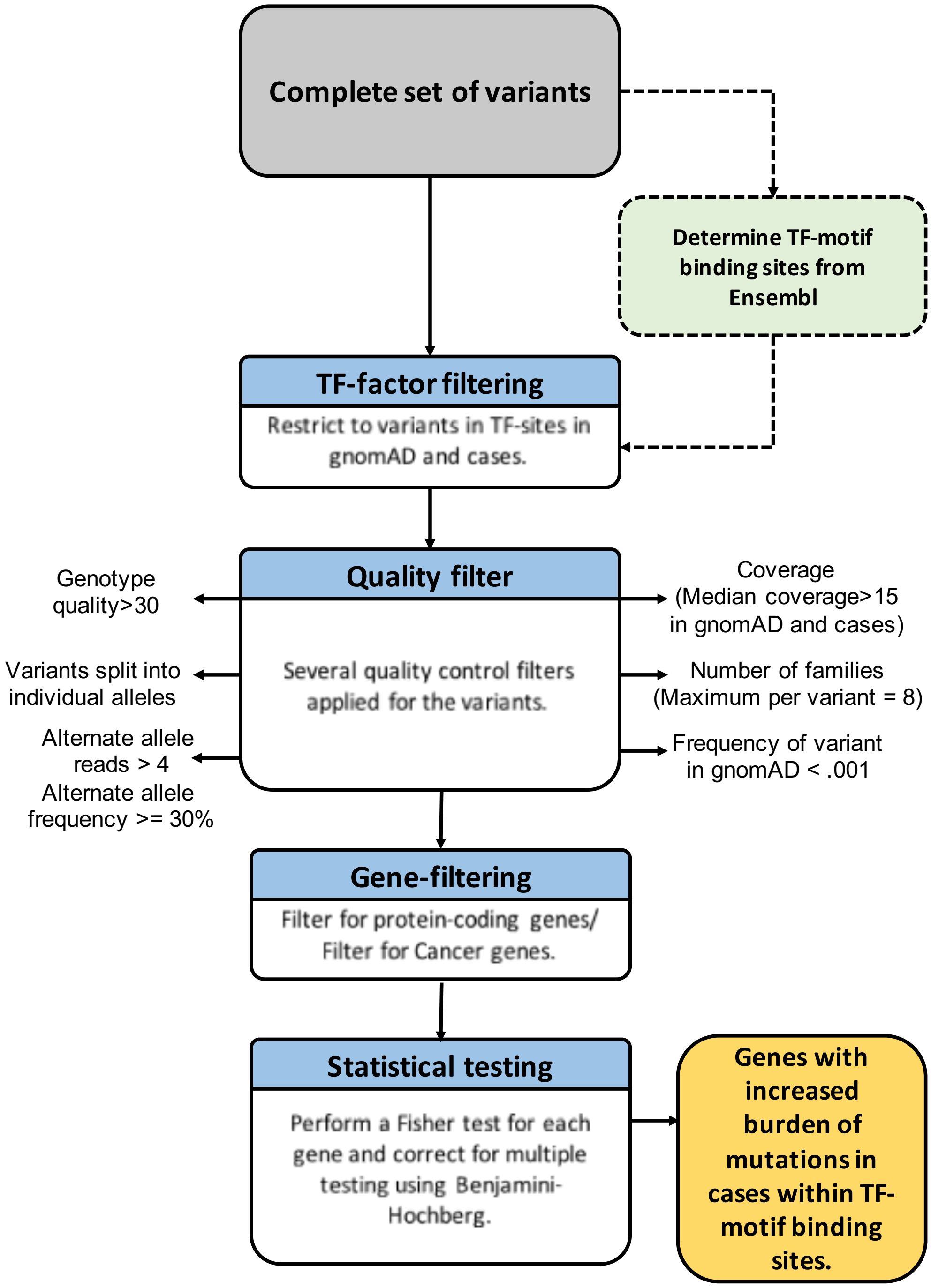

Complete set of variants
Determine TF-motif binding sites from Ensembl
TF-factor filtering
Restrict to variants in TF-sites in gnomAD and cases.
Quality filter
Several quality control filters applied for the variants.
Genotype quality>30
Variants split into individual alleles
Alternate allele reads > 4
Alternate allele frequency >= 30%
Coverage (Median coverage>15 in gnomAD and cases)
Number of families (Maximum per variant = 8)
Frequency of variant in gnomAD < .001
Gene-filtering
Filter for protein-coding genes/ Filter for Cancer genes.
Statistical testing
Perform a Fisher test for each gene and correct for multiple testing using Benjamini-Hochberg.
Genes with increased burden of mutations in cases within TF-motif binding sites.

Figure 2.13: Overview of steps involved in the identification of genes with an increase burden of variants within transcription factor binding motifs in cases through an association analysis.

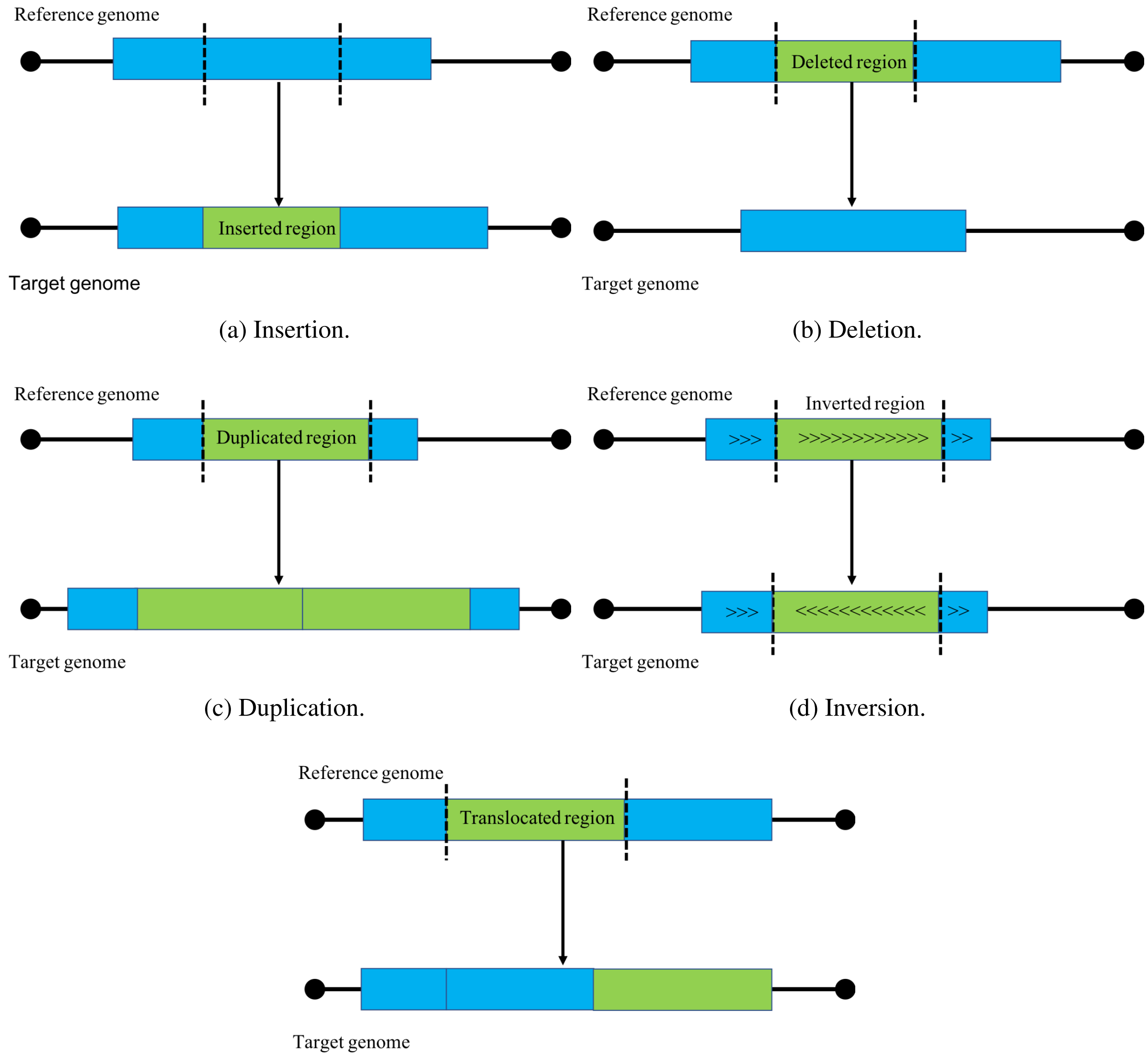
base pairs long. Some structural variant events such as chromothripsis can cluster and disrupt entire chromosomes[208].

There are several types of structural variants that occur within the human genome. The most prevalent types of structural variants are shown in Figure 2.14 and described below:

- Insertion : An insertion is a structural variant caused due to the addition of a segment of DNA between two neighbouring bases in the genome. This is shown in Figure 2.14a.
- Deletion : A deletion is a structural variant caused due to the removal of a segment of DNA between two neighbouring bases in the genome. This is shown in Figure 2.14b.
- Duplication : A duplication is a structural variant where a segment of DNA is replicated and is then inserted alongside the original segment. This is shown in Figure 2.14c.
- Inversion : An inversion is a structural variant that is caused due to the reversal of a segment of the DNA within the genome. This is shown in Figure 2.14d.
- Translocation : A translocation is a structural event where a segment of the DNA is moved to another region of the genome. This is shown in Figure 2.14e. Although Figure 2.14e shows the translocated region to be close to the original position, translocations may occur within or between chromosomes.

2.12.1.2 Structural variants in genetic disorders

Structural variants have been known to play a role in the development of several diseases. The most prominent example of a genetic disorder caused by a structural change is Huntington's disease, encoded by the Huntintgtin gene (*HTT*). A section of the gene comprises trinucleotide repeats of CAG. A normal copy of the gene contains up to 26 copies of the CAG repeats. However, an increase in the number of copies beyond 26 gradually increases the risk and penetrance of the disease, with increased risk of transferring the disease to offsprings as well[209]. A copy of *HTT* with 36-39 CAG repeats have reduced penetrance of disease while copies with 40 or more CAG repeats are considered to be almost completely penetrant[210]. Another notable example of structural disorders is the development of Down syndrome which is a genetic disease caused due to the presence of an additional copy of chromosome 21[211]. Disorders like Down syndrome which are related to extra chromosome copies are defined as trisomy disorders. Other trisomy disorders include Edwards syndrome[212] and Patau syndrome[213].



(a) Insertion.

(b) Deletion.

(c) Duplication.

(d) Inversion.

(e) Translocation.

Figure 2.14: The different types of common structural variants within the human genome. Each sub-figure has the reference genome without the structural variant shown on top. The target genome with the structural variant is shown at the bottom. The black dotted lines on the reference genome indicate the region where breakpoints would be present and should be predicted.

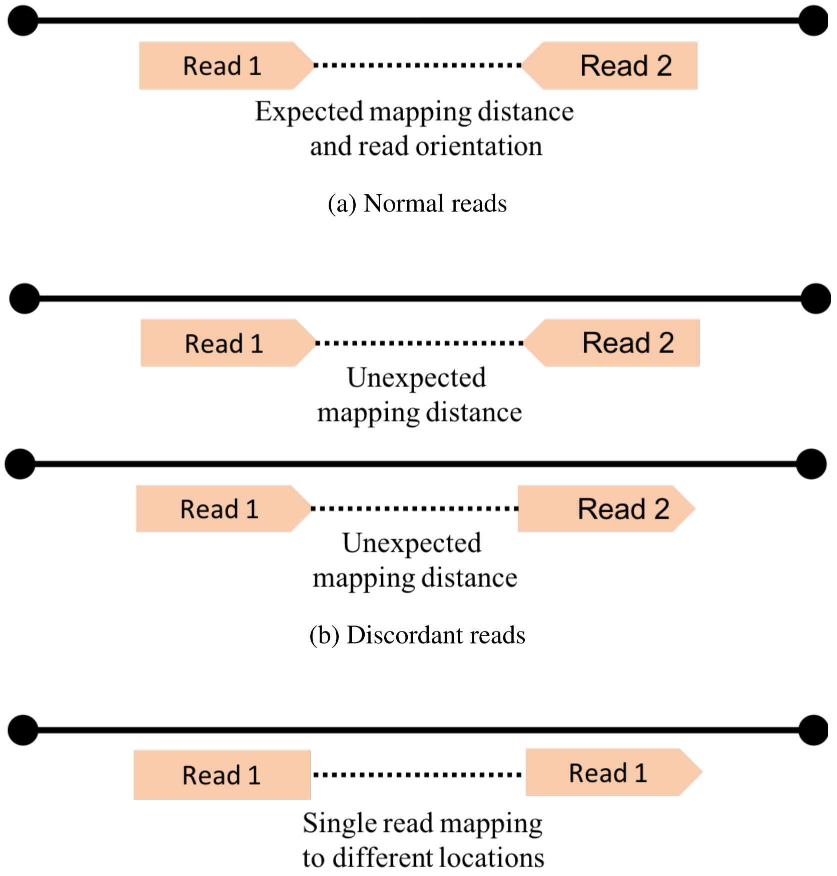
The earliest discovery of structural variants in the context of cancer development was during in the study of cancer cells by Theodor Boveri who associated the growth of cancer cells with observations of segmented chromosomes[214]. Multiple experiments involving fluorescence in situ hybridization (FISH) experiments led to the identification of several gene fusions and amplifications in cancer such as the *BCR-ABL* fusion in chronic myeloid leukemia and *HER2* over-expression in breast carcinomas[215]. The growth and development of microarray technologies furthered the understanding of the role of structural variants in cancer. Comparative genomic hybridization, originally used to identify copy number alterations through FISH, helped in the analysis of amplifications and deletions of genetic regions in solid tumours[216]. SNP genotyping arrays have also been used to determine copy number variations in cancers in several studies such as the Cancer Genome Atlas (https://www.cancer.gov/tcga). In spite of all such improvements, a precise estimation of structural breakpoints was not feasible till the advent of next-generation sequencing technologies. Such technologies also enabled the detection of copy neutral variations such as inversions and translocations.

2.12.1.3 Determination of structural variants in next-generation sequencing data

Structural changes are primarily identified from next-generation sequencing data through errors in the alignment of the target genome reads to the reference genome. Depending upon the type of error, these reads are classified into two types, shown in Figure 2.15:

- Discordant read-pairs. Since the paired-end NGS technique sequences both ends of each DNA fragment with library insert sizes specific to a given library preparation method and size selection procedure, the two paired reads will be generated at an approximately known distance in the sample genome. A signature of a discordant read-pair is formed when the mapping span and/or orientation of the read-pairs crossing the breakpoint are inconsistent with the reference genome. Specifically, both reads of the pair can be mapped to the reference genome, but they may map to different chromosomes or different orientations, or their coordinates may not agree with the insert size.

- Splitting reads. A sequence read that spans a breakpoint in a structural variant is called a splitting read. If both splitting parts of a read can be mapped and its mate is uniquely mapped to the reference genome, the splitting read is further masked as a soft-clipped read by some mapping algorithms such as Burrow-Wheeler Alignment(BWA) tool. Otherwise, it is categorized as an un-mapped read. The splitting reads used by current SV detection tools are all soft-clipped reads, and the term “splitting reads” is generally referred as soft-clipped reads.


Read 1
Read 2
Expected mapping distance and read orientation
Read 1
Read 2
Unexpected mapping distance
Read 1
Read 2
Unexpected mapping distance
Read 1
Read 1
Single read mapping to different locations

(a) Normal reads

(b) Discordant reads

(c) Split reads

Figure 2.15: Different types of read errors used in the identification of structural variants.

The different softwares that are available for the identification of structural variants from next generation sequencing data are distinguished by two factors. The first factor is the type of read error used in the identification of structural variants, i.e., discordant read pairs or split reads. The second factor is the type of sequence that can be analyzed i.e., exome or whole genome sequences. Softwares such as BreakDancer[217], HYDRA[218] and SVDetect[219] use discordant reads for the identification of structural variants while other softwares like CREST[220] use split reads. Recent approaches have also tried to combine the approaches and to use information from both the type of errors in order to determine breakpoints. Examples of such softwares include DELLY[221] and LUMPY[222]. This section of the thesis involves the identification and analysis of structural variants that potentially play a role in cancer development within the sequenced cases.

Parameter	Description
-mw 4	Minimum weight across all samples for a call (number of reads)
-tt 0	Trim threshold
-pe id:sample_id, bam_file:lsample_id.bam	Sample id and its corresponding bam file for a paired-end reads file
histo_file:sample_id.hist	Statistics of insert size across the bam file
mean:461.115724915	Mean insert size
stdev:83.3662594786	Standard deviation of insert size
read_length:151	Read length
min_non_overlap:151	Minimum number of bases that do not overlap, usually equals read size.
min_mapping_threshold:20	Minimum mapping threshold for reads
-sr id:sample_id, bam_file:sample_id_2.bam	Sample id and its corresponding bam file for a split reads file

Table 2.16: Parameters used for the identification of structural variants using LUMPY.

2.12.2 Methods

Cases from the pilot whole genome dataset were considered for the analysis of structural variants. As structural variants consist of large scale changes that affect both the coding and the non-coding region, exomes were considered to be not as informative as whole genome samples. During the time of the analysis, the secondary Leiden whole genome dataset had not yet been sequenced. A similar structural variant analysis approach was eventually performed on these samples by our collaborators in Leiden.

The standard version of Lumpy (v0.2.13)[222] was used to generate structural variants across the 123 whole genome samples from the pilot whole genome dataset. The parameters chosen for this command are given in Table 2.16.

Individual VCF files for each sample were produced as the output from Lumpy, containing information on the structural variants detected in these samples. These VCF files were sorted, zipped and indexed based on the location of the variants. Information regarding the location, length and type of every structural variant were extracted from each sample and saved as individual text files. The number of supporting reads for each variant were also determined and annotated to these files. Variants from each sample were merged into a single joint VCF file. Ensembl contains information on individual haplotypes of chromosomes in addition to entire chromosomes. Variants that were identified as being present in such haplotypes which were not in chromosomes 1-22, X, Y were removed. Generic breakpoints which cannot be classified as inversions, insertions or deletions are marked as "BND" variants by Lumpy. Such variants were also removed as the exact structural variant could not be established. This filtered set of variants was then sorted on the chromosome and the variant position. Some variants started and ended within 250 bp of each other in different samples but largely overlapped with each other. The longest such variants were identified within the overlapping regions and chosen as the representative variant for these regions. The median number of supporting reads were also identified across all samples carrying the variant. A total of 86,697 structural variants were identified in this manner. Variants that were greater than 1 million base pairs were removed as they were considered to be structural aberrations created through sequencing artefacts. This step resulted in the removal of 59,627 variants, resulting in 27,070 variants. The number of samples and families carrying each structural variant were established, with the fraction of samples in each family carrying the variant also being established. Variants that were present in more than five families were removed as they were considered to be artefacts. Only variants that were present in over 75% of the samples in the families with the variant were retained which comprised 773 variants. Depending on the relative location of the structural variant to the closest gene, each variant was annotated with one of six possible orientations, shown in Figure 2.16. They are:

1. Gene contained within the variant (Figure 2.16a).
2. Variant contained within the gene (Figure 2.16b).
3. Variant overlaps with the beginning of the gene (Figure 2.16c)
4. Variant overlaps with the end of the gene (Figure 2.16d)
5. 5' variant (Figure 2.16e)
6. 3' variant (Figure 2.16f)

The strand and orientation of the structural variant and each gene is taken into account for this purpose. In addition to the location of the variant, the distance to the gene was also determined. To focus on structural variants that potentially disrupt cancer pathways, every affected gene was finally annotated with information from the CGC list if they were present in it, including their tier list. Variants that were present downstream of the gene, annotated as being a 3' variant, were removed as the probability of such variants affecting the expression of the gene was lower. This resulted in a final set of 307 variants. Variants in genes associated with melanoma that were identified through this approach are discussed in Section 3.9. The complete list of structural variants is shown in Supplementary Table 10.

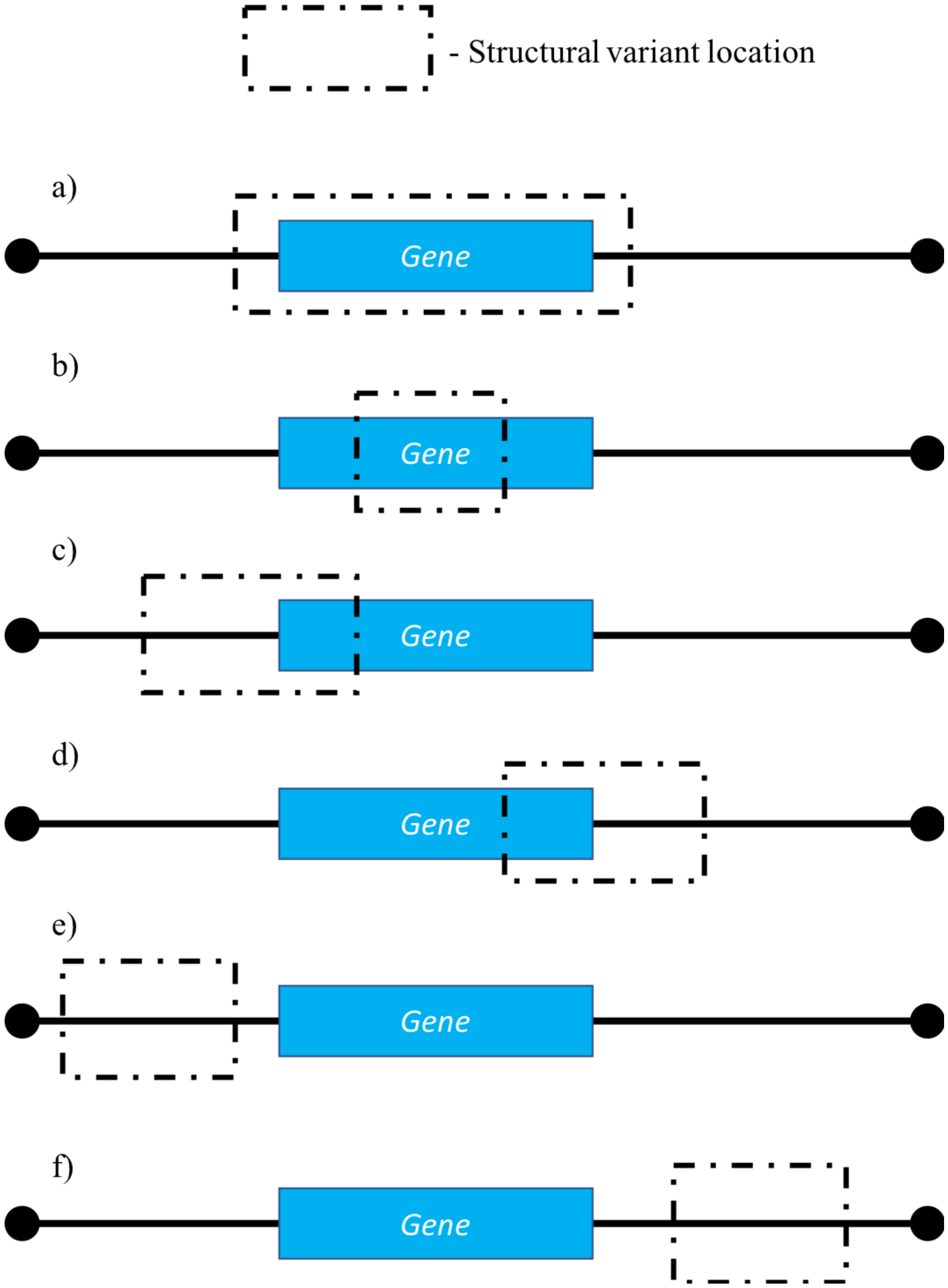


Figure 2.16: The relative locations of structural variant to gene of interest are shown here. a) The structural variant completely encompasses the gene of interest. b) The structural variant is completely contained within the gene of interest. c) The structural variant overlaps with the 5' end of the gene of interest. d) The structural variant overlaps with the 3' end of the gene of interest. e) The structural variant does not overlap with the gene and is present upstream of the gene of interest. f) The structural variant does not overlap with the gene and is present downstream of the gene of interest.

Chapter 3

Results from the analysis of the familial melanoma datasets

3.1 Introduction

This chapter includes the results of all the analysis described in Chapter 2. The first section deals with the description and the implications of the polygenic risk score analysis on the dataset composition. Sections 3.3 and 3.4 describe the novel candidates identified in the association analysis and the joint association-linkage analysis on the coding region variants respectively. The results from the secondary, complementary methods used for discerning novel disruptive variants in addition to the association analyses are discussed in Sections 3.5,3.6 and 3.7. The investigation of non-coding variants and structural variants from the whole genome sequences resulted in novel findings, described in Sections 3.8 and 3.9 respectively. The implications of these results are discussed in Chapter 4.

3.2 Estimation of polygenic risk scores

The distribution of polygenic risk scores from the 123 familial melanoma cases in the pilot study, the sporadic cases and the unaffected controls are shown in Figure 3.1.

From the figure, it can be observed that the risk scores are distributed between -0.6 and 1.2 with the majority lying between -0.5 to 0.6. No discernible differences can be observed between the three groups directly. In addition to the genotype information obtained for the dataset from Leeds, information on the presence of multiple primary melanomas within the samples, early age of onset and the number of members in the family who suffered from

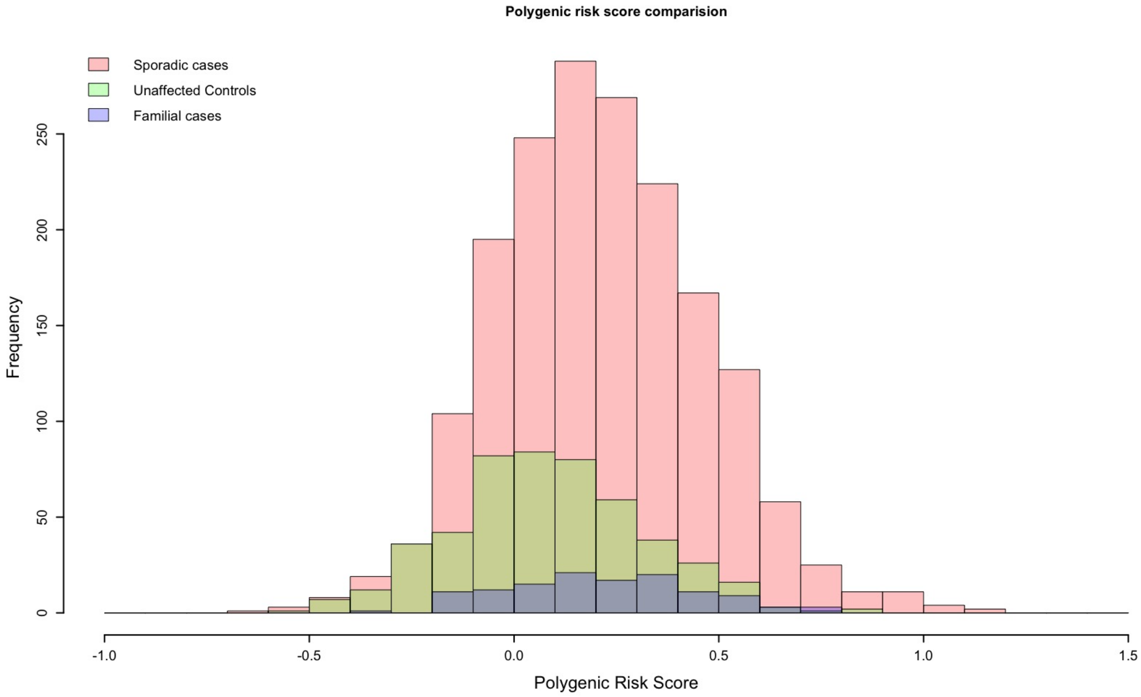


Figure 3.1: Distribution of polygenic risk scores for familial melanoma cases, sporadic melanoma cases and unaffected controls.

melanoma were also provided. The mean and median polygenic risk scores for each of these categories are given in Table 3.1.

The distribution of the risk scores and the mean and median risk scores in each category imply the following:

1. The risk scores are a representation of the burden of risk factors arising from common variants for the development of melanoma within the samples. As the controls do not have melanoma, they would be expected to have lower risk scores as compared to the cases, and this is indeed evident from the histograms, as the distribution of the risk scores for the controls samples lies to the “left” of the distribution of both sporadic and familial cases. This is also borne out by comparing mean and median risk scores between distributions. The controls have a mean risk score of .086 with a median of .082, which is significantly lower than the mean and median risk scores of the Leeds familial cases, the pilot study familial cases and the Leeds sporadic cases. The risk scores of the controls, sporadic values and the pilot study familial values were used as inputs for pairwise t-tests to identify if both the sets in consideration could be obtained from the same distribution. The null hypothesis is that there is no significant difference be-

Group	Mean	Median
Familial melanoma cases from the pilot study	0.2194	0.2203
Sporadic cases	0.2118	0.1993
Leeds familial cases	0.2254	0.1980
Unaffected controls	0.0868	0.0825
Early onset cases (inclusive of sporadic cases and the Leeds familial cases)	0.2249	0.2068
Cases with multiple primary melanomas (inclusive of sporadic cases and the Leeds familial cases)	0.2253	0.2210
Sporadic cases with multiple primary melanomas	0.2212	0.2216
Sporadic early onset cases	0.2125	0.1924
Familial cases from Leeds with multiple primary melanomas	0.2536	0.1277
Familial early onset cases from Leeds	0.3263	0.3579
Early onset sporadic cases with multiple primaries	0.1566	0.1650
Early onset familial cases from Leeds with multiple primaries	0.4039	0.3960

Table 3.1: Mean and median polygenic risk scores for different subgroups of samples.

tween the two sets while the alternate hypothesis is that there is a significant difference. The p-value of the t-test comparing the unaffected controls and the sporadic cases was 2.2*10^{-16}while the p-value comparing the controls and pilot study familial melanoma cases was 3.988*10^{-8}. As the p-values are less than .05, the null hypothesis is rejected, indicating that the controls are significantly different based on their risk scores compared to the sporadic and pilot study familial cases.

2. The distribution of the risk scores of the samples from the pilot study lie within the distribution of sporadic cases from Leeds. Although there is a large difference in the number of sporadic cases as compared to the familial cases, the difference in their mean and median values are not significantly different. The p-value obtained when tested for significant difference between the two groups was 0.7235, indicating that both the categories could be from the same distribution. The risk scores of the pilot study cases are also comparable with the different subcategories of cases from Leeds (inclusive the sporadic and the familial cases) involving early onset and the presence of multiple primary melanomas.

3. Interestingly, the risk scores for early onset familial cases from Leeds, both with and without multiple primary melanomas, are quite high compared to the sporadic risk scores, unaffected control risk scores and the pilot study risk scores. This indicates that these cases contain a high burden of risk factors which predisposes them to the early development of melanoma.

In conclusion, the familial melanoma cases have a higher polygenic risk score on average compared to unaffected controls and a similar risk score to sporadic cases. While this indicates a higher burden of common risk alleles compared to the controls, it does not rule out the presence of a high penetrant allele. None of the familial melanoma cases have an abnormally high risk score to merit their exclusion from the dataset. All samples from the dataset were therefore retained for further analysis.

3.3 The identification of novel variants through association analysis

The complete table of genes and their corresponding p-values both the genes in the Cancer Gene Census and for all protein-coding genes are attached as Supplementary Tables 1 and 2 respectively. The top 10 ranked genes from the Cancer Gene Census is shown in Table 3.2

while the top 10 ranked genes from all protein-coding genes is shown in Table 3.3. Variants from Table 3.2were investigated in detail to determine candidate driver mutations as they had the highest likelihood of playing an important role in familial melanoma development.

Ensembl ID	Gene name	Maximum segregation in families (%)	Fisher's Test p-value	Corrected P-value
ENSG00000145113	*MUC4*	100	1.25E-09	8.50E-07
ENSG00000178104	*PDE4DIP*	100	6.95E-06	0.00196
ENSG00000104517	*UBR5*	100	8.66E-06	0.00196
ENSG00000135333	*EPHA7*	50	4.35E-05	0.00739
ENSG00000163930	*BAP1*	100	7.28E-05	0.00989
ENSG00000087460	*GNAS*	66	0.00018	0.02111
ENSG00000046889	*PREX2*	50	0.00032	0.03090
ENSG00000138448	*ITGAV*	100	0.00036	0.03090
ENSG00000156650	*KAT6B*	54.5	0.00064	0.04645
ENSG00000204713	*TRIM27*	50	0.00071	0.04645

Table 3.2: List of the top 10 genes associated with melanoma within the Cancer Gene Census.

Ensembl ID	Gene name	Maximum segregation in families (%)	Fisher's Test p-value	Corrected P-value
ENSG00000204172	*AGAP10*	75	5.61E-22	1.06E-17
ENSG00000175820	*CCDC168*	75	5.45E-17	5.13E-13
ENSG00000185926	*OR4C46*	75	1.24E-16	7.80E-13
ENSG00000188649	*CC2D2B*	75	1.03E-13	4.83E-10
ENSG00000216937	*CCDC7*	66.6	4.49E-13	1.69E-09
ENSG00000112592	*TBP*	50	1.01E-11	3.15E-08
ENSG00000155495	*MAGEC1*	66.6	4.24E-11	1.14E-07
ENSG00000266714	*MYO15B*	66.6	1.09E-10	2.56E-07
ENSG00000177182	*CLVS1*	100	2.66E-10	5.57E-07
ENSG00000213401	*MAGEA12*	100	9.50E-10	1.74E-06

Table 3.3: List of the top 10 genes associated with melanoma within all protein coding genes.

From Table 3.2, it can be observed that *BAP1,* a gene previously described as a familial melanoma driver gene in Section 1.3.4.5, is ranked fifth on the ordered list of genes. This suggests that the approach in use has the statistical power and capacity to detect other potential familial melanoma driver genes as *BAP1* is a known driver gene. A family from Leiden, the Netherlands carried a *BAP1* variant that completely segregated with this disease. The importance of this variant is further discussed in Section 2.8.

Mucin 4 (MUC4) was the highest ranking gene from the Cancer Gene Census in the association analysis. It a member of the Mucin family, a set of high molecular weight glycoproteins present in the epithelial cells which are responsible for controlling the activity of inflammatory responses. Mucins are classified into two categories: secreted mucins and membrane bound mucins. *MUC4* is a membrane bound mucin responsible for several functions including the activation of an oncoprotein called ERBB2[223]. Two variants segregated completely with the disease in *MUC4 :* A variant at p.R906W (GRCh38 reference build, Chromosome 3, genomic position 195788864, c.2716G>A) in a three member pedigree from Leeds and a p.R468K variant (GRCh38 reference build, Chromosome 3, genomic position 195790177, c.1403C>T) in a three member pedigree from Stockholm. Two other variants also had high but not complete co-occurence of the variant with the disease, with three out of four sequenced members in two pedigrees carrying additional variants. Overexpression of *MUC4* has been observed in multiple cancers including pancreas, gall bladder, ovary, breast and lung carcinomas[224]. It is therefore seen as a potential therapeutic target.

Ubiquitin protein ligase E3 component n-recognin 5 (UBR5) was ranked third on the list of genes. It is an important constituent of the Ubiquitin-Proteasome System (UPS)which is an essential regulator of the DNA damage repair pathway. *UBR5* plays a key role in the development of several forms of cancer, as reviewed in 2015 by Shearer et al*[225].* Two variants segregated with the disease in *UBR5 :* A variant at p.S552I (GRCh38 reference build, Chromosome 8, genomic position 102323440, c.1655C>A) in a two member pedigree from Barcelona and a p.T1721P variant (GRCh38 reference build, Chromosome 8, genomic position 102286414, c.5161T>G) in a two member pedigree from Leeds. However, *UBR5* has not been previously implicated in either sporadic or familial melanoma. The mechanism of activation of melanoma development through *UBR5* disruption is therefore as yet undetermined.

Integrin Subunit Alpha V (*ITGAV*) ranked eighth in the association analysis. *ITGAV* encodes for an integrin membrane protein that regulates angiogenesis and cancer progression. A variant in ITGAV at p.R573Q mutation(GRCh38 reference build, Chromosome 2, genomic position 186656400, c.1718G>A) segregating with the disease was observed in a two member pedigree from Leeds, United Kingdom. Additional variants were observed in *ITGAV*; none

Gene	CLRT_score	P-value	LOD_score
MUC4	656.51	0.000999	5.22
MUC16	221.15	0.000999	1.99
FAM47C	156.23	0.000999	3.542
PDE4DIP	136.42	0.000999	3.67
RNF213	123.8	0.00699	3.22
MLL3	123.68	0.024	0.32
MLL2	117.9	0.0569	2.62
HLA-A_DUP_07	115.23	0.000999	0.49
NOTCH1	106.95	0.000999	0.51
FAT1	106.53	0.138	2.64

Table 3.4: List of top 10 scoring genes from the joint association-linkage analysis using the 1000 genomes dataset as the background dataset.

of which segregated with the disease in any of the pedigrees. Increased expression of *ITGAV* has previously been associated with increased invasion in colorectal cancer tumours[226] and it also facilitates prostate cancer metastasis[227]. Similar to *UBR5*, the role of *ITGAV* in skin disorders, particularly in cutaneous melanoma is currently unknown and warrants further investigation.

Whilst *EPH receptor A7* (*EPHA7*), *GNAS complex locus* (*GNAS*), *phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2* (*PREX2*), *lysine acetyltransferase 6B* (*KAT6B*) and *tripartite motif containing 27* (*TRIM27*) also ranked highly on the association analysis, none of the variants present in these genes segregated with the disease in the affected families. As none of these genes have previously been associated with either cutaneous sporadic melanoma or familial melanoma, the lack of variants segregating with the familial melanoma phenotype suggests that these genes are not relevant in familial melanoma genesis and development.

3.4 The identification of novel variants using a joint association-linkage analysis

The joint association-linkage analysis was performed twice using two different background datasets, one with the 1000 genomes dataset and another with the INTERVAL exomes. Each set focused entirely on the genes in the Cancer Gene Census. The complete results from both sets are attached in Supplementary Tables 4 and 5. The top 10 scoring genes for each of these tables are shown in Tables 3.4 and 3.5 respectively.

Gene	CLRT_score	P-value	LOD_score
MUC4	1203.4	0.000999	15.3
MUC16	373.2	0.000999	1.53
FAM47C	232.85	0.000999	3.64
MLL3	206.77	0.000999	0.32
NOTCH1	157.84	0.000999	0.65
HLA-A_DUP_07	154.91	0.000999	0.49
MLL2	147.9	0.00599	2.62
RNF213	146.99	0.00699	3.22
FAT1	136.62	0.043	2.6
FAT4	127.4	0.005	2.71

Table 3.5: List of top 10 scoring genes from the joint association-linkage analysis using the INTERVAL exomes dataset as the background dataset.

Each individual run for a gene returns two .vaast output files:

i) A simple file that only has the ranked list of genes with the associated p-value, LOD scores and CLRT scores.

ii) A larger file that has variant level information including the samples containing each variant and the filtered variants in each gene that were present in the background.

As each pVAAST run focussed within the region of a single gene, the output file corresponding to each run only has values for the reported gene. The background used was consistent across all runs for each attempt; this allows for the direct comparision of the output values with each other. The CLRT scores corresponding to the association and the LOD scores corresponding to the linkage were obtained for all the genes present within the Cancer Gene Census. The plot of these scores using the 1000 genomes project data as the background dataset is shown in Figure 3.2.

The CLRT scores from the association analysis are plotted on the y-axis. To correct for skewing of results, the logarithm to the base 10 of the LOD scores were estimated. This is plotted on the x-axis. In order to focus on genes with high association and/or linkage, only the genes with $\log_{10}$(LOD) greater than 0 or CLRT scores greater than 50 have been named. *MUC4* and *MUC16* are genes belonging to the Mucin family of proteins which had very high estimated CLRT and LOD scores. This is due to the presence of multiple variants in the cases that were not present in the background and were therefore not filtered out. As the scores of each variant contribute to the overall score of the gene, several low scoring mutations in *MUC4* and *MUC16* cumulatively resulted in the higher ranking of these genes. Mucins are high molecular weight proteins and have been observed to be consistently mutated without playing a role in the development of cancers. To better show the distribution of scores for the

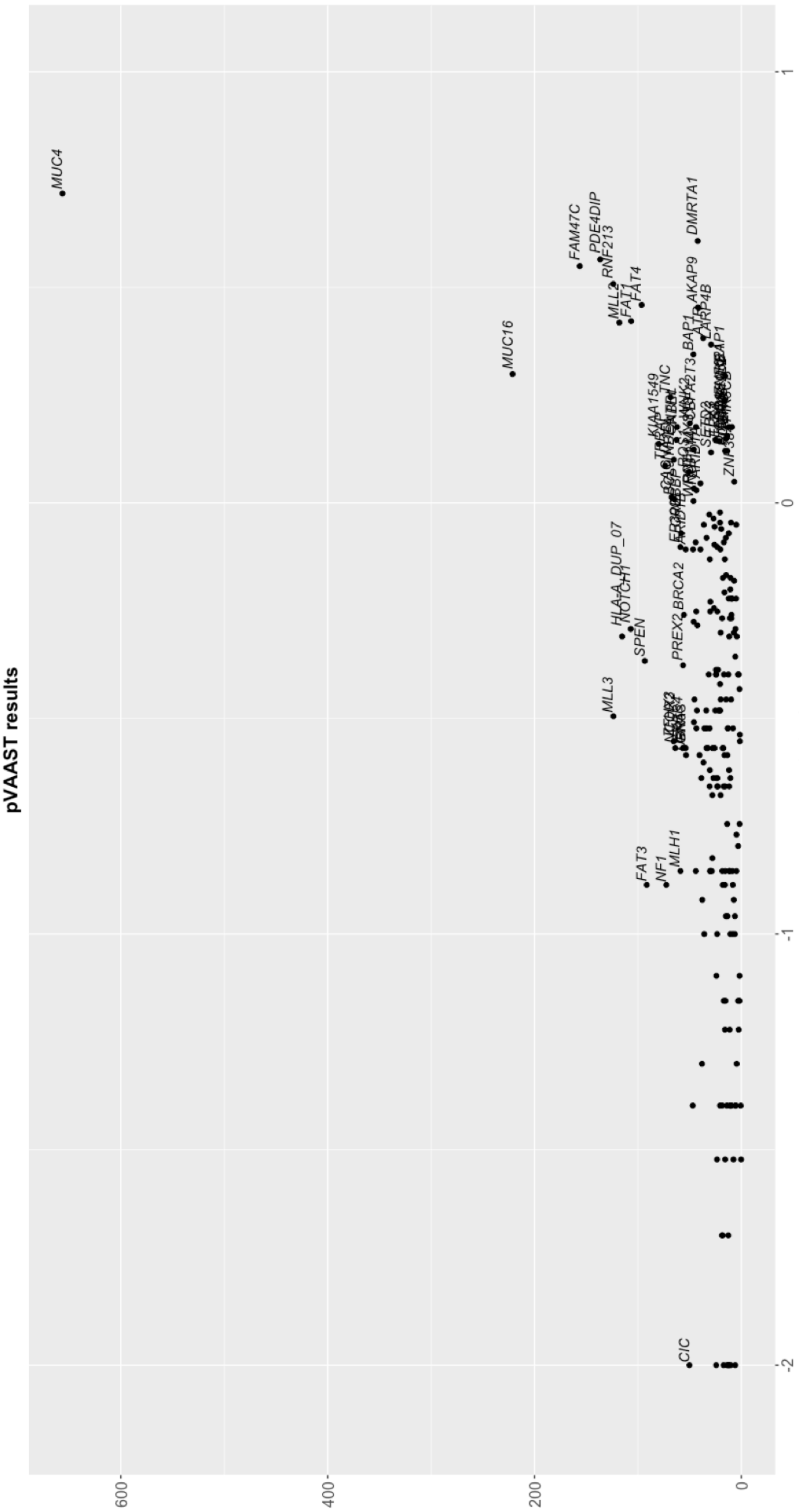


Figure 3.2: Original results from pVAAST for all genes in the Cancer Gene Census. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

other genes, these two genes were removed from the plot and a second plot was generated, shown in Figure 3.3.

In addition to the genes from the Cancer Gene Census, a single additional gene called *Doublesex And Mab-3-Related Transcription Factor A1 (DMRTA1)* was scored using pVAAST. While *DMRTA1* is in itself not a cancer gene, it lies adjacent to *CDKN2A*. A variant segregating in 10 out of 11 sequenced members of a pedigree was observed in *DMRTA1*. This was added to the list of genes as a positive control to ensure that pVAAST estimated LOD scores effectively for all genes as this variant would be expected to have a high LOD score. pVAAST estimated the LOD score *DMRTA1* to be 3.58 which was the highest LOD score estimated for a single variant, indicating that the LOD scores were being determined accurately. The potential implication of this variant in melanoma development is further discussed in Chapters 4 and 5.

Comparing the results from Figure 3.3 to Table 3.2, it can be observed that there are some genes that score well in both lists while others are significantly different. *MUC4* was the highest ranking gene in the association analysis and as discussed above, scores markedly high on pVAAST. *PDE4DIP*, the second highest ranking gene in the association analysis also ranks highly on pVAAST, with a CLRT score of 136.42 and a cumulative LOD score of 3.67 across 4 variants. Other genes that score similarly in both lists include *UBR5* (CLRT=53.4, LOD=0.26), *BAP1* (CLRT=46.34, LOD=2.21) and *GNAS* (CLRT=53.79, LOD=0.79. Interestingly, *UBR5* and GNAS score higher than BAP1 on the association score as they have more rare variants using this background dataset. However, these variants do not segregate through the pedigrees that they are present in unlike the variants in *BAP1* which completely segregate with the disease. As a result, although *BAP1* scores slightly lower on the association score, it would be considered to be more interesting and relevant to disease development compared to *GNAS* and *UBR5*, an important facet which would have been lost in a strict association analysis.

There are also several genes present in this analysis that are not in the top of the association analysis. *FAM47C,* the highest ranking gene that is not a Mucin protein, has a CLRT score of 156.23 and a LOD score of 3.542 obtained from 8 variants. This implies that there are several rare variants in this genes with most of these variants being present in smaller families. Such variants would have low individual LOD scores but cumulatively push the LOD score to greater than 3.5. *FAM47C* ranked 28th on the original association analysis discussed in Section 3.3with an uncorrected p-value of 0.003746353 and a corrected p-value of 0.088908376.

Another gene in a similar scenario is *Ring finger protein 213* (*RNF213*) which is ranked just below *PDE4DIP* with a CLRT score of 123.8 and a combined LOD score of 3.22 across 9

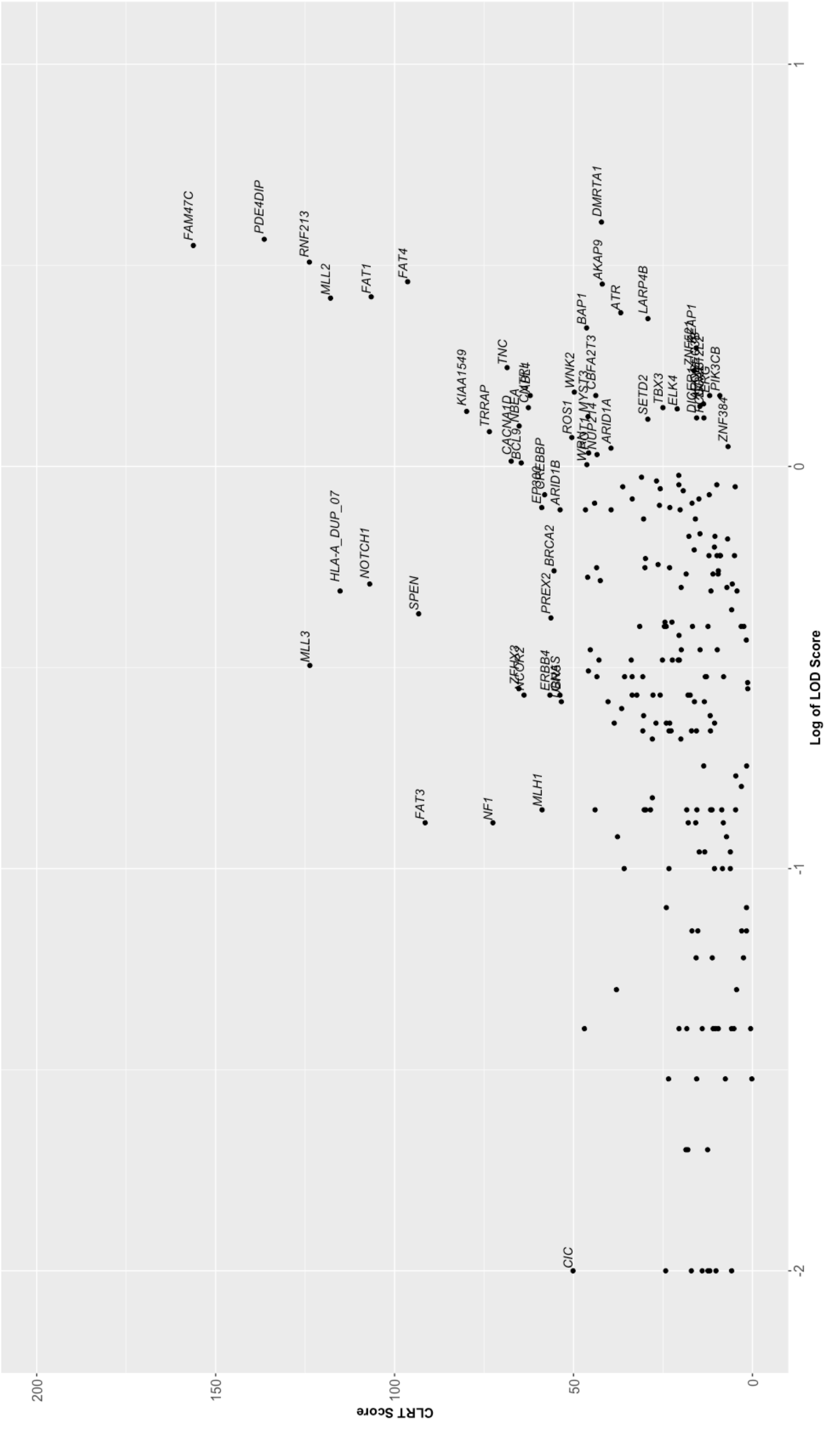


Figure 3.3: Results from pVAAST for all genes in the Cancer Gene Census excluding *MUC4* and *MUC16*. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

variants. The distribution of variants across multiple pedigrees with low number of sequenced members is similar to *FAM47C*. *RNF213*, however, ranks higher than *FAM47C* on the original association analysis discussed in Section 3.3with an uncorrected p-value of 0.001603777 and a corrected p-value of 0.075735829. Thus, while the corrected p-values of these genes might not have been originally significant, the presence of multiple variants segregating with the disease across multiple pedigrees indicate potential roles for *FAM47C* and *RNF213* in familial melanoma development.

An interesting observation from the joint-association linkage results are the presence of multiple members of the Fat atypical cadherin family of proteins. *FAT1* and *FAT4* both score highly on the joint association and linkage analysis, with high CLRT and LOD scores. *FAT3* also scores highly on the CLRT score but not on the LOD score. None of these genes scored highly on the original association analysis with *FAT1*, *FAT3* and *FAT4* ranking 94, 327 and 184 respectively. The presence of such high scores would indicate that these genes and the FAT family of protein would play a vital role in the development of familial melanoma in these pedigrees. However, on closer investigation, it was determined that these genes did not score significantly on the original association analysis as they are several thousand base pairs long and comprise of multiple functional domains. As a result, there were several low frequency variants in gnomAD and the cases leading to a high p-value. The background datasets used for the analysis on pVAAST do not have as many samples as gnomAD, resulting in higher scores for these genes. The functional relevance of the *FAT* family of genes in melanoma development is yet to be determined.

In addition to these high scoring genes, there are several genes with high LOD scores but not high CLRT scores and vice-versa such as *LARP4B* and *CIC* respectively. These genes, seen in Figure3.3, are examples of cases where an association or a linkage analysis on its own might lead to false positives as they might score highly due to the presence of several rare, non-segregating variants or one rare variant completely segregating with the disease but have no part to play in cancer development.

Similar plots and scores were obtained using the INTERVAL exomes as the background dataset. This altered the CLRT scores due to the difference in background variants and occasionally the LOD scores, if the variants were filtered for being in the background. However, the general ranking of genes and their corresponding relative CLRT scores and LOD scores remained the same. The complete scores from this set of pVAAST runs are given in Supplementary Table 5 and the plots from this set of pVAAST run are shown in Supplementary Figures 1 and 2.

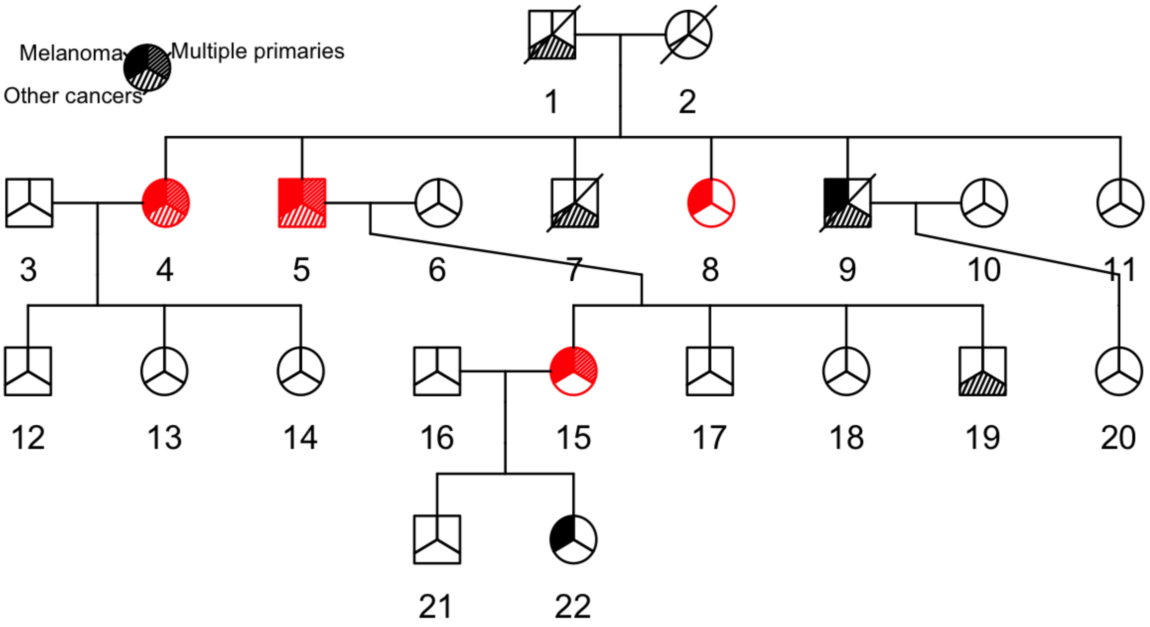


Figure 3.4: Leiden pedigree with the segregating p.Y646Ffs *BAP1* frameshift mutation. The members marked in red were sequenced from the pedigree, all of whom carried the variant.

3.5 The search for variants in known driver genes

1. *BAP1*: A four case pedigree from Leiden, shown in Figure 3.4, carried a frameshift variant encoding a p.Y646Ffs change (GRCh38 reference build, Chromosome 3, genomic position 52402825, c.2408_2409insAA). All four sequenced members carried the variant. Nonsense mutations leading to cancer development have previously been observed at this location: these mutations, however, led to the creation of premature stop-codons as opposed to the frameshift mutations observed in this study[228][229]. This variant was not observed in the gnomAD database.

2. *BRCA2*: A frameshift variant was observed in a single patient from KCL, London encoding a p.Q397Lfs variant (GRCh38 reference build, Chromosome 13, genomic position 32332667, c.1422_1423insTTAG). The frequency of this variant for non-Finnish Europeans in the gnomAD database was 0.000008822; it has not been annotated with phenotypes in ClinVar. A variant that introduced a stop codon resulting in a p.E1415*

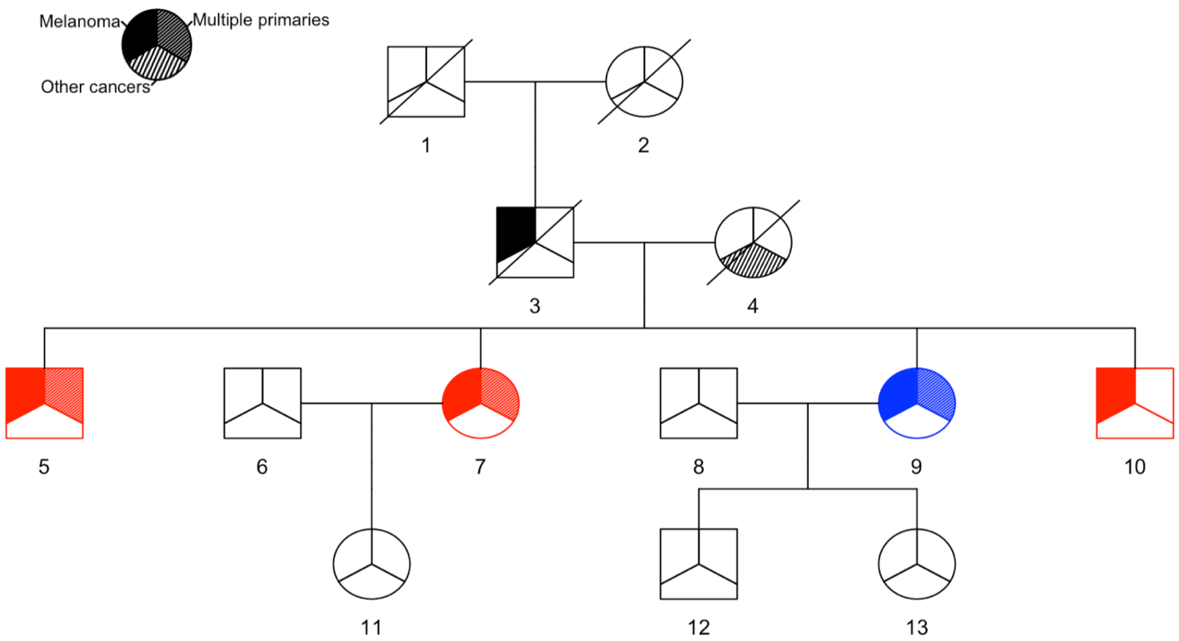


Figure 3.5: Sydney pedigree with the p.I49S missense variant in *CDKN2A* segregating in three out of four sequenced members. Whole genome sequencing for performed for the members shown in red and blue. The members marked in red carry the variant while the member shown in blue did not carry the variant and is predicted to be a phenocopy.

mutation (GRCh38 reference build, Chromosome 13, genomic position 32338598, c.4476G>T) was also observed in a patient from a family in Barcelona. This variant was annotated as being pathogenic on ClinVar for hereditary cancer-predisposing syndrome[230] (dbSNP id rs397507327). This variant was not observed in gnomAD.

3. *CDKN2A*: A p.I49S missense mutation (GRCh38 reference build, Chromosome 9, genomic position 21974682, c.417A>C) was identified in a pedigree from Sydney (shown in Figure 3.5) and was present in three out of four sequenced members. This variant has previously been observed in the context of familial melanoma[231][232] and is predicted to be deleterious based on *CDK4*/*CDK6* binding[233]. The variant is also annotated as being potentially pathogenic for hereditary cutaneous melanoma in ClinVar and was not observed in the gnomAD database.

4. *POT1*: As previously mentioned in Section 2.2.4.1, three novel *POT1* variants predisposing the families to familial melanoma development were previously identified using a dataset comprised of exome sequences of familial melanoma patients from Leeds and Leiden[143]. This dataset was a part of the cases and the samples carrying the *POT1*

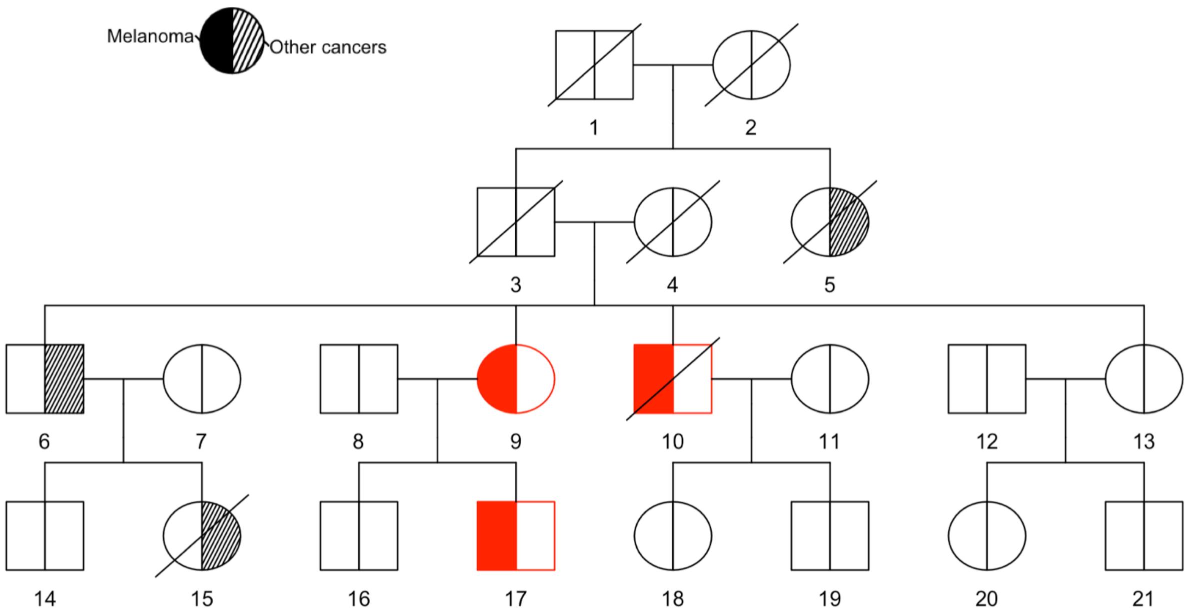


Figure 3.6: Leeds pedigree carrying the p.E312K missense variant in *MITF*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

variants were included in the analysis as positive controls to verify the ability of the applied methods in detecting disruptive variants. Additionally, a novel *POT1* missense variant was observed encoding a p.D185G change (GRCh38 reference build, Chromosome 7, genomic position 124859105, c.1153T>C) in two families: a two-case family and a single case family, both from Leeds. This variant was not observed on ClinVar (dbSNP id rs749741053) and was filtered from the gnomAD dataset.

5. *MITF*: A missense variant in *MITF* responsible for a p.E312K mutation (GRCh38 reference build, Chromosome 3, genomic position 69964940, c.934G>A) was observed to be segregating in a three member pedigree from Leeds, shown in Figure 3.6. This variant was annotated as being a risk factor for cutaneous malignant melanoma and for hereditary cancer predisposing syndrome on ClinVar (dbSNP id rs149617956). The same variant was also present in a one member of a two-member family from Leeds. This variant was also previously observed recurrently in both familial and sporadic melanoma[234] and has an allele frequency of 0.002456 for the non-Finnish European population subgroup in the gnomAD database.

No disruptive mutations were found in *CDK4* and *TERT*.

3.6 Variants with high segregation within the cases

A total of 12,923 variants were obtained after the steps described in Section 3.6. Variants in melanoma driver genes with segregation, as identified in Section 2.8, were observed again in this approach. While variants in *BAP1* and *POT1* were observed, the *MITF* variant was filtered out due to a having a gnomAD allele frequency greater than 10^{-3}. The complete set of variants is attached as Supplementary Table 6. Some of the resulting novel, segregating variants with a link to cancer development are reported here. The variants have been classified into nonsense and missense variants, depending on their effect on the protein product.

3.6.1 Nonsense mutations

1. *Ataxia telangiectasia and Rad3-related protein (ATR)* is a protein kinase that functions as a sensor and transducer for double-strand breaks caused due to to UV radiation[235]. Loss-of-function mutations in *ATR* have previously been shown to play a role in the growth of melanoma tumours[236]. A pedigree from QIMR, Brisbane with 4 sequenced members carried a p.L890*(GRCh38 reference build, Chromosome 3, genomic position 142553363, c.2669A>C) variant that encoded a stop-gain mutation in *ATR*. This variant segregated completely within the pedigree as shown in Figure 3.7. The reported L890* variant was not observed in the gnomAD database or in ClinVar.

2. *tumour protein 53-regulated apoptosis-inducing protein (TP53AIP1)* is a gene that encodes a protein which interacts with *TP53* and plays a role inTP53-mediated apoptosis[237]. Two single-case pedigrees from Leeds and one multi-case pedigree from Brisbane carry a frameshift variant encoding a p.Q22Afs change (GRCh38 reference build, Chromosome 11, genomic position 128937755, c.63_64insG).Three out of four sequenced members in the Brisbane pedigree, shown in Figure 3.8, carry the variant which was previously observed to predispose individuals from two different pedigrees to develop melanoma and is predicted to be an intermediate penetrant risk factor for melanoma onset[238]. This variant was not observed on ClinVar(dbSNP id rs141395772).

3. *Exonuclease 5 (EXO5)* is a single-stranded DNA-specific exonuclease that plays a role in DNA damage repair, with loss-of-function mutations in *EXO5* leading to increased genomic instability[239]. In a study involving the deficiency of DNA damage repair pathways across The Cancer Genome Atlas, *EXO5* was observed to be epigenetically silenced with high frequency, particularly in glioblastoma multiforme and in head and neck squamous cell carcinoma (HNSCC)[240]. A heterozygous p.R344Afs variant

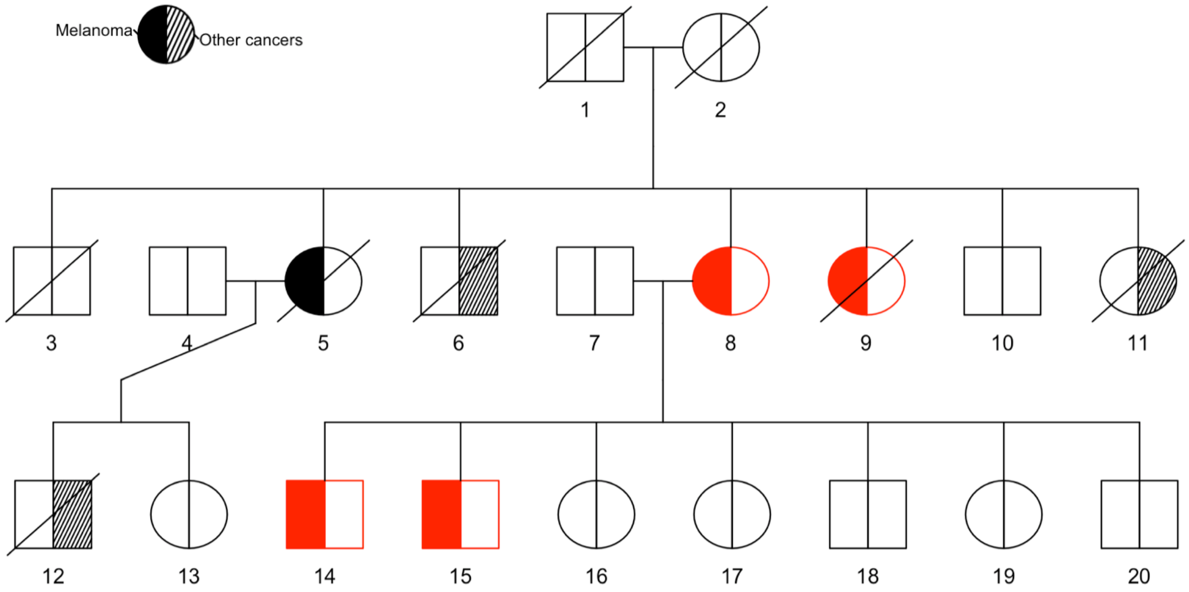


Figure 3.7: Brisbane pedigree carrying the p.L890* stop-gain mutation in *ATR*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

(GRCh38 reference build, Chromosome 1, genomic position 40515573, c.2103_2104insG) was observed in six pedigrees in the cases. Three of these pedigrees were from Leeds with single members sequenced, one was from London, and the final two were pedigrees with multiple sequenced members from Barcelona and Leeds. The Barcelona pedigree had two members sequenced, with one carrying the variant. In contrast, the Leeds pedigree, shown in Figure 3.9, had three members sequenced, all of whom carried the variant. This variant was previously observed in a study of early onset melanoma patients in Poland, where the association of the variant with the increase in melanoma risk was inconclusive[241]. This variant was also identified as a candidate that plays a role in increased susceptibility to testicular cancer[242]. The p.R344Afs was observed in gnomAD with an allele frequency of 0.0190 for non-Finnish Europeans and was not observed in ClinVar(dbSNP id rs150018949).

3.6.2 Missense mutations

1. *DMRTA1:* The variant with the largest number of affected samples in the dataset was a missense variant encoding a p.E383Q (GRCh38 reference build, Chromosome 9, ge-

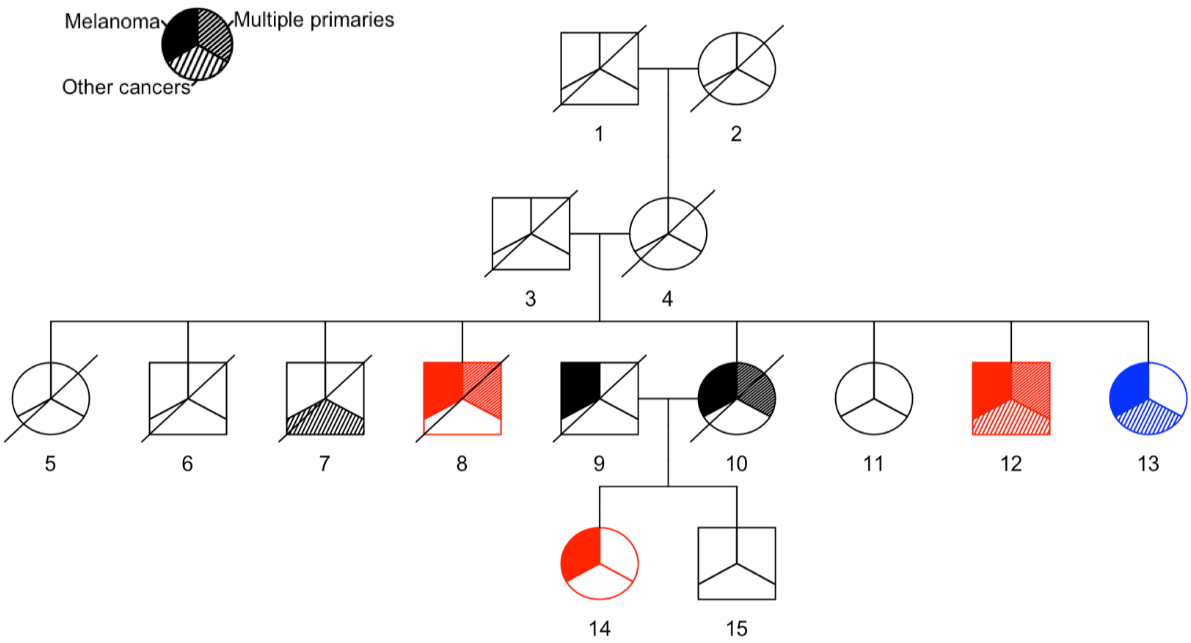


Figure 3.8: Brisbane pedigree with the p.Q22Afs frameshift variant in *TP53AIP1*. Exome sequencing for performed for the members shown in red and blue. The members marked in red carry the variant while the member shown in blue did not carry the variant and is predicted to be a phenocopy.

nomic position 22451543, c.2669G>C) change in *Doublesex- And Mab-3-Related Transcription Factor A1 (DMRTA1)*. Thirteen patients from three different families carried the variant with the following segregation: Two out of three sequenced members in a pedigree from Leiden (exome sequenced), a single sequenced member from a Leeds pedigree (exome sequenced) and ten out of eleven sequenced members in a pedigree from Sydney (whole genome sequenced). Although this variant has been annotated as not being pathogenic on both SIFT and PolyPhen, it lies adjacent to *CDKN2A* in the 9p21.3 chromosomal band. A variant with such a segregation pattern could be indicative of another, more significant variant lying within the same region, potentially be in the non-coding or intergenic region, that regulates the function of *CDKN2A*. A 233,780 base-pair deletion was eventually observed in a eleven member Sydney family 5' of *CDKN2A*. Particularly, this deletion was observed in the same ten members who also carried the *DMRTA1* variant. This is further explored in Section 3.9 in the analysis of structural variants in the whole genome sequences.

2. *AMER1*: A variant encoding a p.D233Y mutation (GRCh38 reference build, Chromosome X, genomic position 64192590, c.970C>A) was detected in a pedigree with a

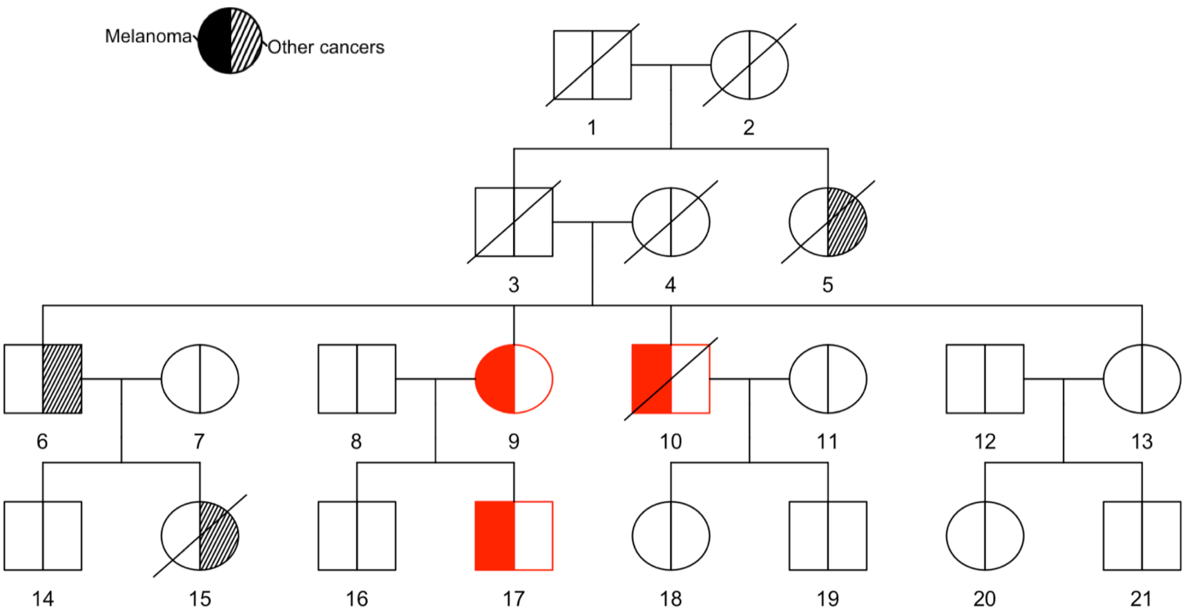


Figure 3.9: Leeds pedigree with the p.R344Afs frameshift variant in *EXO5*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant. This is the same pedigree that carries a moderate risk factor variant in *MITF* as described in 3.5.

single sequenced member from Leeds and a four-case pedigree from Sydney (shown in Figure 3.10) in *APC membrane recruitment protein 1 (AMER1)*. *AMER1*, also known as *WTX*, encodes a tumour suppressor. When mutated, this protein plays a role in the formation of pediatric kidney cancer known as Wilms tumours[243]. Germline variants in *AMER1* are also known to be causative of cranial sclerosis, a developmental disorder[244]. *AMER1* plays a role in the regulation of the WNT signaling pathway[245], a cascade that is involved in the carcinogenesis of several types of cancers including colorectal cancer[246], leukemia[247], melanoma[248] and breast cancer[249]. This variant was observed in the database with an allele frequency of 0.0004596 and was not observed in ClinVar (dbSNP id rs146685042).

3.7 Pathogenic variants in ClinVar

There were a total of 408,919 annotated variants in the ClinVar VCF file. 43,120 of these variants were also found in the cases. 18,205 variants remained after filtering for artefacts. 338 of these variants matched the restricted clinical significances for a wide range of diseases.

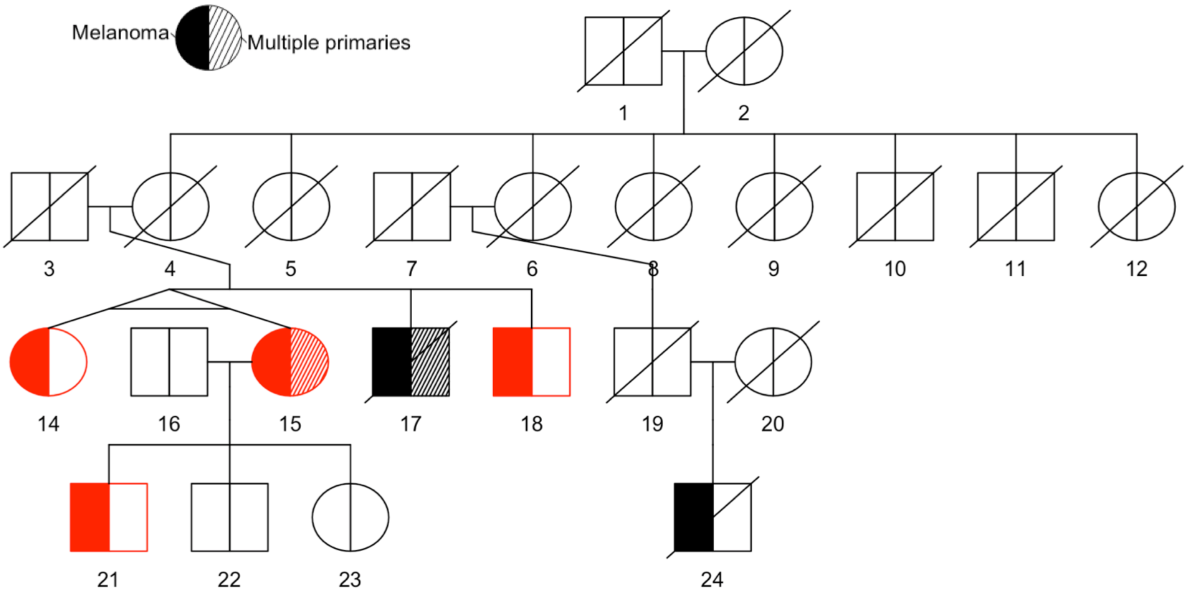


Figure 3.10: Sydney pedigree carrying the p.D233Y missense variant in *AMER1*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

The variants most relevant to cancer and melanoma onset are described below:

1. The two *POT1* missense variants identified by Robles Espinoza et al. and the two disruptive *BRCA2* variants reported in 2.8, marked as pathogenic in ClinVar, were all observed through this approach. The *CDKN2A* p.I49S missense variant was also annotated on ClinVar as being associated with hereditary cutaneous melanoma. However, there were conflicting reports of pathogenicity regarding the clinical significance. As a result, this variant was not observed in our cases as it was filtered out.

2. Several pathogenic variants linked to different types of oculocutaneous albinism were observed. A variant linked to Tyrosinase-negative oculocutaneous albinism encoding a p.Y149C mutation (GRCh38 reference build, Chromosome 11, genomic position 89178399, c.948A>G) affecting the *Tyrosinase (TYR)* gene, was observed in an individual from KCL, London. Three variants were also identified in *oculocutaneous albinism II (OCA2),* a gene previously established as a susceptible locus in a meta-analysis of cutaneous melanoma genome wide association studies (GWAS)[51]. These variants were linked to the development of Tyrosinase-positive oculocutaneous albinism. Previous reports have identified *OCA2* missense mutations in melanoma and indicate a potential role in melanoma predisposition[250]. The location of these variants in the protein is

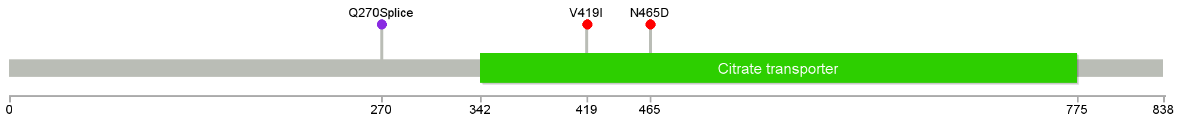


Figure 3.11: Location of pathogenic *OCA2* variants as identified in ClinVar. This plot was generated using Lollipops v1.3.2[251].

shown in Figure 3.11. These variants were as follows:

(a) A homozygous p.N465D missense variant (GRCh38 reference build, Chromosome 15, genomic position 27983383, c.1503T>C) was observed in a two member pedigree from Leiden, shown in Figure 3.12. In addition to the sequenced members, there was another member with melanoma and one more with pancreatic carcinoma. All affected members of the pedigree also had albinism. In total, there were six members in the pedigree who had albinism.

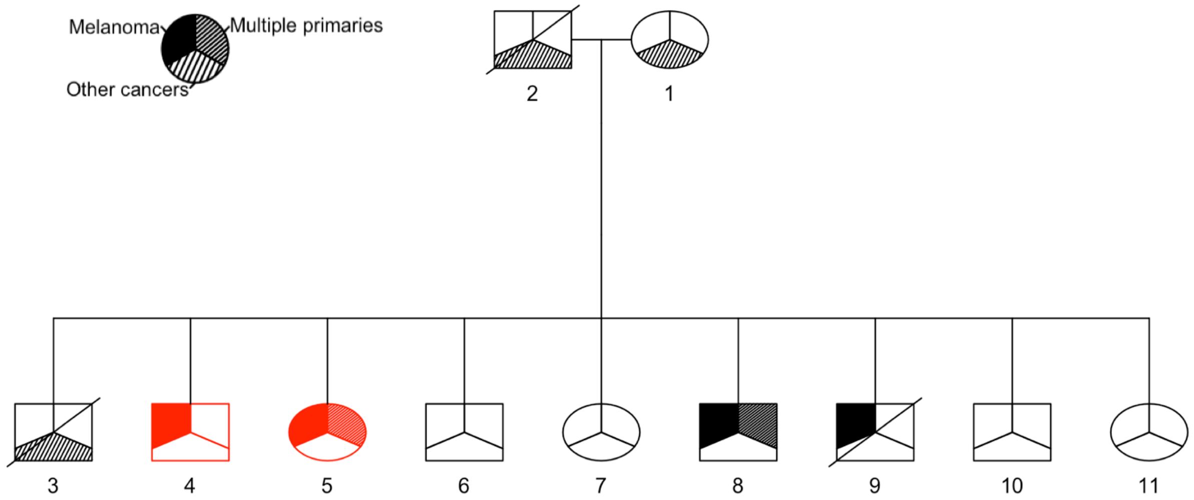


Figure 3.12: Leiden pedigree carrying the p.N465D missense variant in *OCA2*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant. All affected members also had albinism.

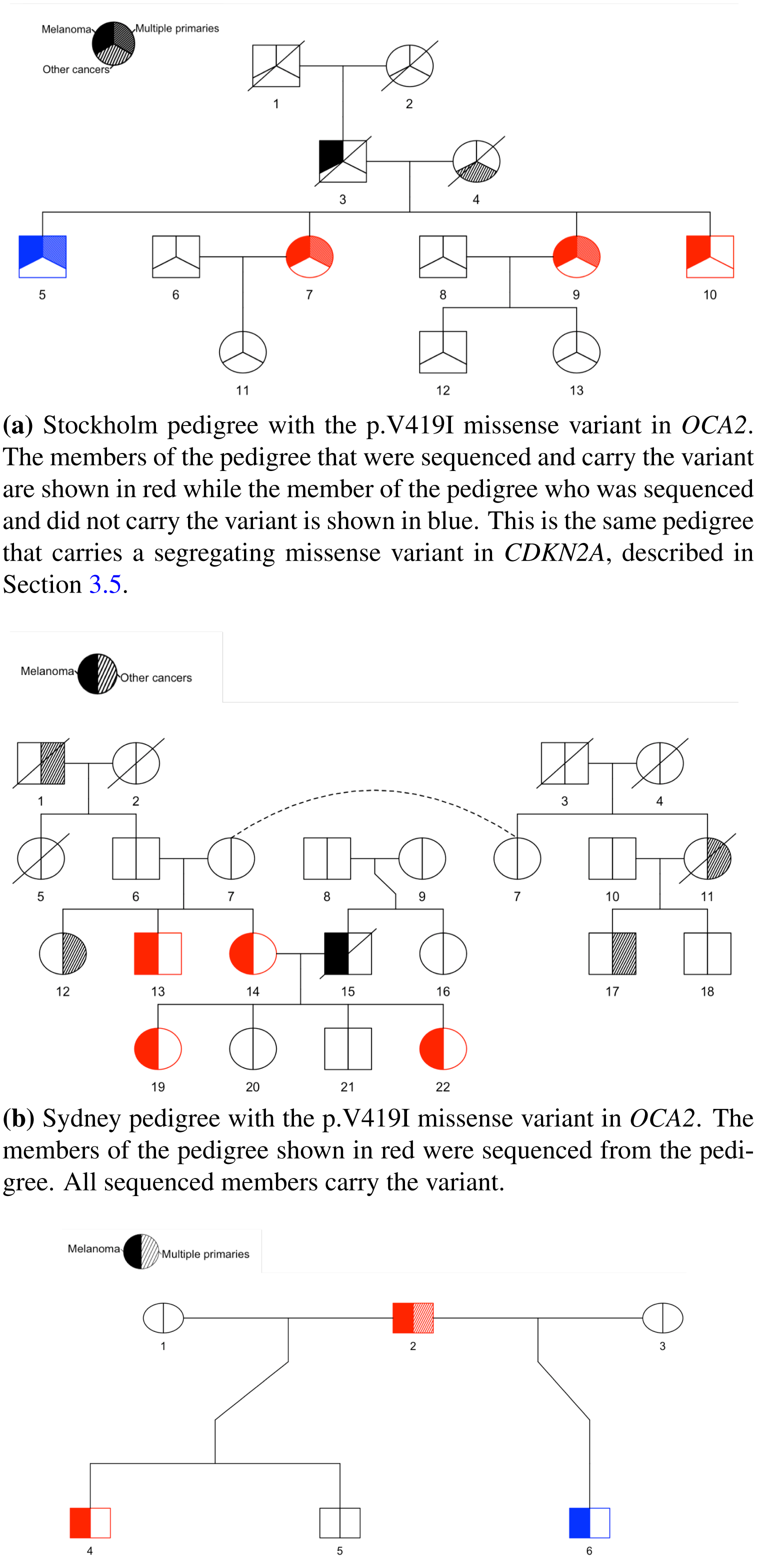
(b) Eleven individuals from five different pedigrees carried a p.V419I missense variant (GRCh38 reference build, Chromosome15, genomic position 27985101, c.1365C>T). Three out of four sequenced members of a Sydney pedigree, four out of four sequenced members of a second Sydney pedigree and two out of three sequenced members of a Stockholm pedigree carried the variant amongst the families with multiple sequenced members. These pedigrees are shown in Figures 3.13a, 3.13b and 3.13c respectively. Two pedigrees from Leeds with single individuals sequenced also carried the variant.

(c) A canonical splice donor variant was identified at the beginning of the intron between exons 7 and 8 (GRCh38 reference build, Chromosome15, genomic position 28018396) in a single member from a Barcelona pedigree where 2 members were sequenced.

3. Multiple variants associated with different types of cancer on ClinVar were identified. Information related to these variants are given in Supplementary Table 7. While these variants are marked as pathogenic and could potentially play a role in the development of cancer in the individuals that they are present in, none of them segregate in a multi-case pedigree, i.e., none of them are present in a pedigree with several sequenced members all of whom carry the variant. This implies that the development of melanoma in the pedigrees carrying these variants cannot be attributed to these variants alone.

3.8 Analysis of non-coding variants affecting transcription factor binding motifs

P-values were estimated and corrected for a total of 537 genes from the Cancer Gene Census. Results for the top 20 genes from the analysis of variants within the transcription factor binding motifs are shown in Table 3.6. The results for the complete set of genes are attached in Supplementary Table 8. Whilst most genes carry multiple variants distributed across several pedigrees, not every sequenced member of each pedigree carry the variant. In order to identify variants with high segregation of variant with the disease, the percentage of segregation of the variant was estimated for every pedigree at every variant. Following this, the pedigree with the maximum percentage of segregation was determined for each gene by comparing the percentages of segregation across all variants in a given gene. These results are also reported in Table 3.6.



(a) Stockholm pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree that were sequenced and carry the variant are shown in red while the member of the pedigree who was sequenced and did not carry the variant is shown in blue. This is the same pedigree that carries a segregating missense variant in *CDKN2A*, described in Section 3.5.

(b) Sydney pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

(c) Stockholm pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree that were sequenced and carry the variant are shown in red while the member of the pedigree who was sequenced and did not carry the variant is shown in blue.

Figure 3.13: Pedigrees with the p.V419I missense variant in *OCA2*.

Ensembl ID	Gene name	Fisher's Test p-value	Corrected P-value	Maximum percentage of segregation in families (%)
ENSG00000184702	*SEPT05*	1.62E-33	8.70E-31	50
ENSG00000164362	*TERT*	2.88E-24	7.73E-22	75
ENSG00000136997	*MYC*	8.50E-21	1.52E-18	67
ENSG00000184640	*SEPT09*	2.58E-19	3.46E-17	83.3
ENSG00000072062	*PRKACA*	1.47E-18	1.58E-16	75
ENSG00000088038	*CNOT3*	1.79E-18	1.60E-16	50
ENSG00000078403	*MLLT10*	3.67E-18	2.81E-16	40
ENSG00000118046	*STK11*	4.57E-17	3.07E-15	83.3
ENSG00000175197	*DDIT3*	1.63E-15	8.78E-14	40
ENSG00000137309	*HMGA1*	1.63E-15	8.78E-14	66
ENSG00000136754	*ABI1*	1.08E-14	4.45E-13	75
ENSG00000141968	*VAV1*	1.08E-14	4.45E-13	100
ENSG00000160957	*RECQL4*	1.08E-14	4.45E-13	66
ENSG00000157933	*SKI*	1.40E-14	5.36E-13	100
ENSG00000071564	*TCF3*	7.52E-14	2.69E-12	40
ENSG00000143970	*ASXL2*	3.00E-13	9.46E-12	75
ENSG00000197122	*SRC*	3.00E-13	9.46E-12	100
ENSG00000157764	*BRAF*	5.07E-13	1.51E-11	60
ENSG00000261652	*C15orf65*	2.46E-12	6.61E-11	33
ENSG00000162367	*TAL1*	2.46E-12	6.61E-11	50

Table 3.6: List of the top twenty genes associated with variants in transcription factor binding motifs within the Cancer Gene Census. The values in the "Maximum percentage of segregation in families" represents the value of the highest percentage of segregation in any pedigree with a variant in the gene where there are at least two sequenced members and two individuals carrying the variant.

Interestingly, previously known driver genes including *TERT* and *MYC* are second and third respectively on the list of genes . *TERT* in particular is prominent due to the history of variants in TFBM within the promoter which lead to the progression of both familial and sporadic melanoma. However, the family with the maximum of percentage of segregation for the variants in these genes are 75% and 67% respectively, indicating that none of the families have a variant that completely segregates with the disease within these genes. Additionally, the variants in *TERT* were investigated to determine if they were present at previously determined promoter variant sites. None of the observed variants were at these positions. Following this, variants with a segregation percentage of 75% or higher were then scrutinized to determine if the TFBMs for any of these genes were disrupted by these variants. The genes with the disrupted motifs and high percentage of segregation of the variant with the disease are described below.

i)Proto-oncogene vav - *VAV1*

Rho GTPases are a set of proteins responsible for several cellular functions including proliferation, adhesion, cellular migration, contraction and secretion[252]. These proteins were originally thought to play a role in the carcinogenesis of several cancer types indirectly through increased expression levels as mutations within them were rare[253]. However, recent studies have identified several direct mutations in RhoGTPases in different cancers[254] such as a recurrent mutation in *RAC1* in melanoma patients[255]. Rho GTPases are mediated and regulated by the Rho Guanine Nucleotide Exchange Factor (Rho GEF) family of proteins. These proteins include Vav1, which is responsible for the activation of Rac1[256]. Expression of wild-type VAV1 has been identified in several types of cancers including neuroblastoma, melanoma, pancreatic, lung and breast cancers[257]. It has also been reported to act as both an oncogene[258] and a tumour suppressor[259] under physiological contexts, similar to p53[260].

An upstream variant in *VAV1* was observed in 4 pedigrees with complete segregation in a Pennsylvania pedigree with 3 sequenced members (GRCh38 reference build, Chromosome 19 genomic position 6772561, G>A). This variant disrupted a binding site for the transcription factor EGR1, indicated in Figure 3.14.

There are 14 positions in the binding motif for EGR1 with varying levels of conservation between them. Positions 6, 8, 9 and 10 are the most conserved with these positions always carrying a guanine nucleotide at these positions. The variant described here changes the guanine at position 9 to an adenine, which is never present at this position. This would ablate the TFBM and disrupt the binding of EGR1 to this location.

ii)The Sloan Kettering Institute protein - *SKI*

The SKI protein, named after the Sloan Kettering institute where it was first described, is a proto-oncogene. *SKI* pathways have previously been reported as being responsible for the activation of β-catenin signalling[261] and for the progression of human malignant melanoma[262]. Following this, the knockdown and deficiency of *SKI* was also reported to reduce the proliferation of human melanoma tumours[263].

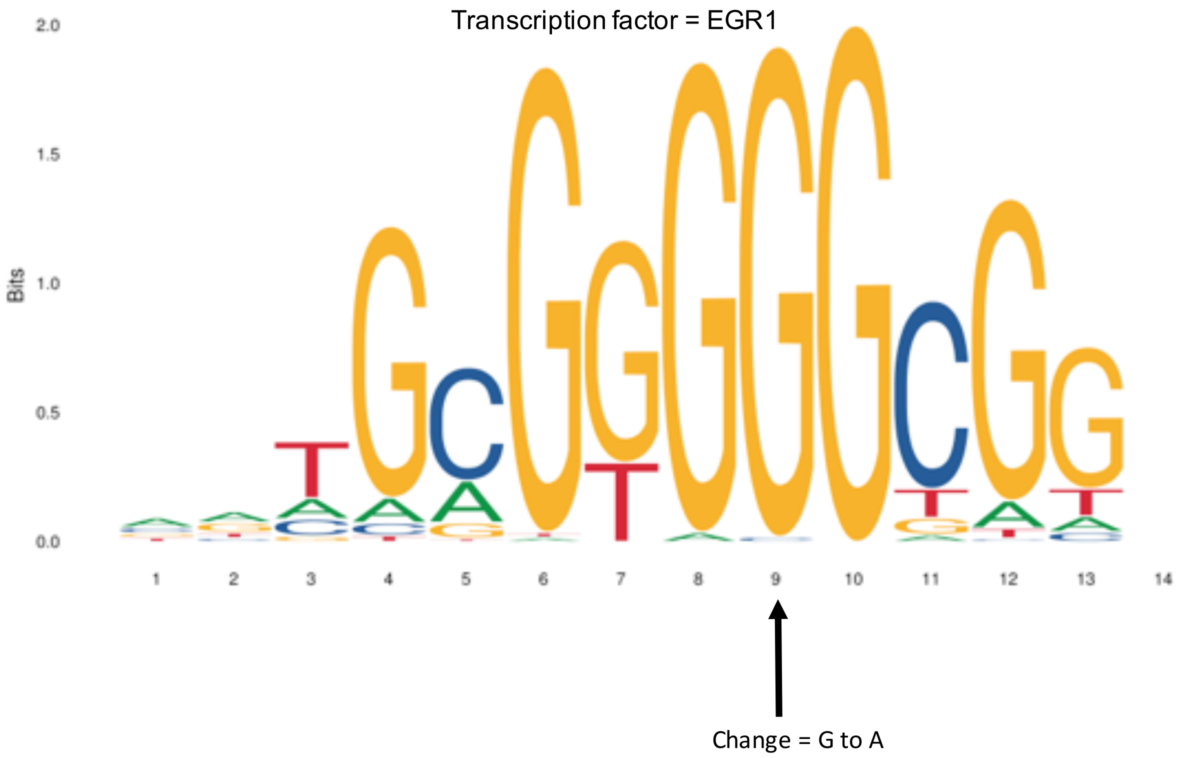


Figure 3.14: Disruption of EGR1 binding motif in *VAV1* as observed in a pedigree from Pennsylvania. Sequence logo obtained from the JASPAR[199] database.

A variant upstream to *SKI* (GRCh38 reference build, Chromosome 1 genomic position 2227654, G>A) was observed in 3 pedigrees. This included complete segregation with the disease in a Sydney pedigree with 4 sequenced members and partial segregation in a Leiden pedigree with 3 out of 6 sequenced members carrying the variant. Figure 3.15 indicates the disruption of the binding motif for GABP by this variant.

The TFBM for GABP consists of 10 bases. Positions 4 to 8 are highly conserved, represented by the base pattern GGAAG. The variant of interest is at position 8, which alters the guanine present at this position to an adenine and ablates the transcription factor binding site. Both GABP ablation and recruitment have been previously been associated with melanoma development. Disruption of GABP binding motif has previously been reported in the context of melanoma in association with recurrent mutations of the subunit D of the succinate dehydrogenase complex *(SDHD)* [200], as discussed in Section 2.11.1.3. Recruitment of GABP has been established in the reactivation of mutant *TERT* with promoter variants, leading to deviant expression of *TERT* in multiple cancer types including melanoma[264].

iii) Proto-oncogene tyrosine-protein kinase - ***SRC***

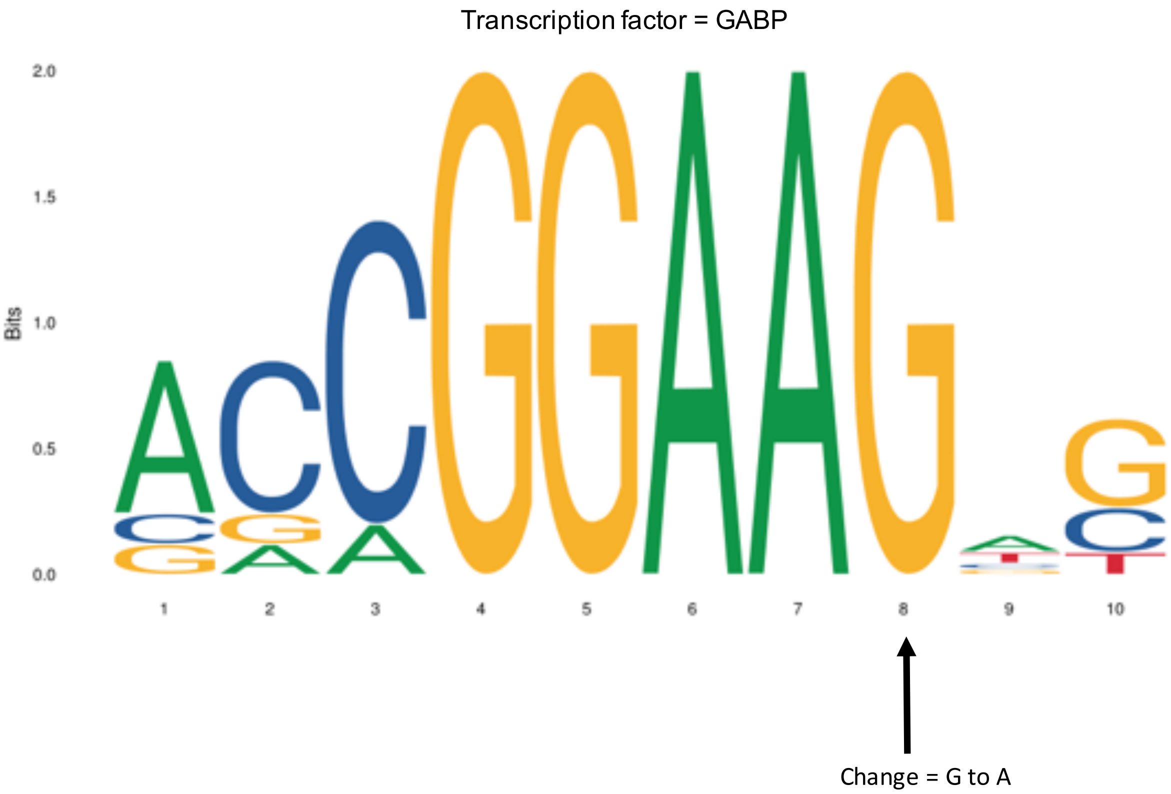


Figure 3.15: Disruption of GABP binding motif in *SKI* as observed in 3 pedigrees. Sequence logo obtained from the JASPAR[199] database.

Proto-oncogene tyrosine-protein kinase is a non-receptor tyrosine kinase protein that in humans is encoded by the *SRC* gene. It is involved in the progression and metastasis of several cancer types including breast, pancreatic, colon and brain cancer[265]. The *SRC* pathway is also active in melanoma[266] and *SRC* inhibitors are seen as potential therapeutic agents in the treatment of melanoma[267] with several studies and clinical trials focussing on *SRC* inhibitors in solid tumours[268]. In particular, a combination of *SRC* and *MEK* inhibition was reported to suppress the growth and invasion of melanoma cells[269].

A pedigree from Leeds with 2 sequenced members carried an intronic variant (GRCh38 reference build, Chromosome 20 genomic position 37396591, G>C) . This variant is shown in Figure 3.16 and affects a binding site for the transcription factor reported to be MSN2 on JASPAR.

The binding motif for MSN2 comprises of 5 positions, with positions 2 to 4 being highly conserved; these positions always have a guanine nucleotide. Adenine is present with the highest frequency in position 1 with guanine and cytosine being present at lower frequencies. The observed mutation changes a guanine at the highly conserved position 2 to a cytosine,

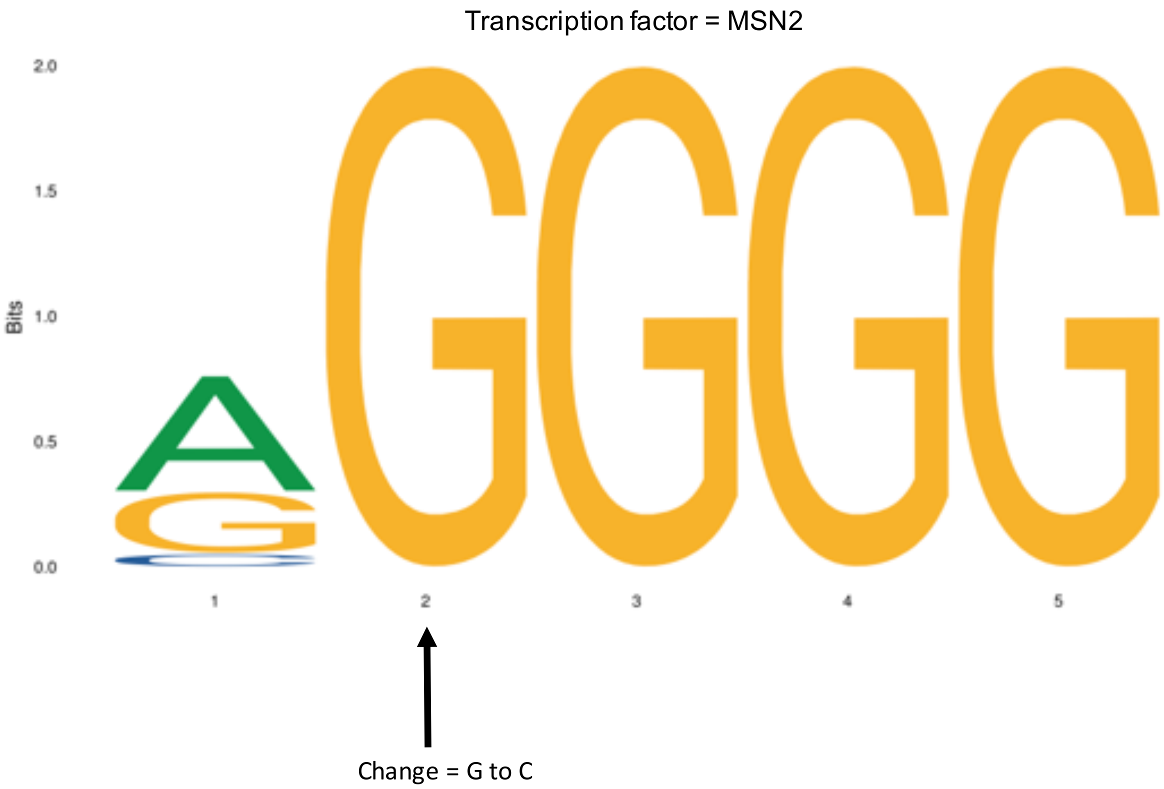


Figure 3.16: Disruption of MSN2 binding motif in *SRC* as observed in a pedigree from Leeds. Sequence logo obtained from the JASPAR[199] database.

potentially disrupting the binding of MSN2. Whilst MSN2 is a transcription factor that plays a role in stress response in yeast, it isn't reported to be active in *homo sapiens*. However, the 5 base binding motif of MSN2 is the same as the central binding motif of EGR1, comprising of several repeating guanine units. This is shown through bases 6 to 10 in Figure 3.14. This mutation in *SRC* would also disrupt and impact this binding motif, indicating that the affected transcription factor could be EGR1 as opposed to MSN2. Additionally, while the binding matrix for this variant refers to MSN2 on JASPAR, Ensembl has annotated this region as having an EGR1 binding motif.

3.9 Structural variant analysis

A 233,780 bp deletion was identified in chromosome 9 (GRCh38 reference build, 22209075-22442855) and is shown in Figure 3.17.

This deletion was detected to be 213,774 bases upstream of *CDKN2A,* a prominent melanoma driver gene (discussed in Section 1.5). The deletion was observed in a large pedigree from

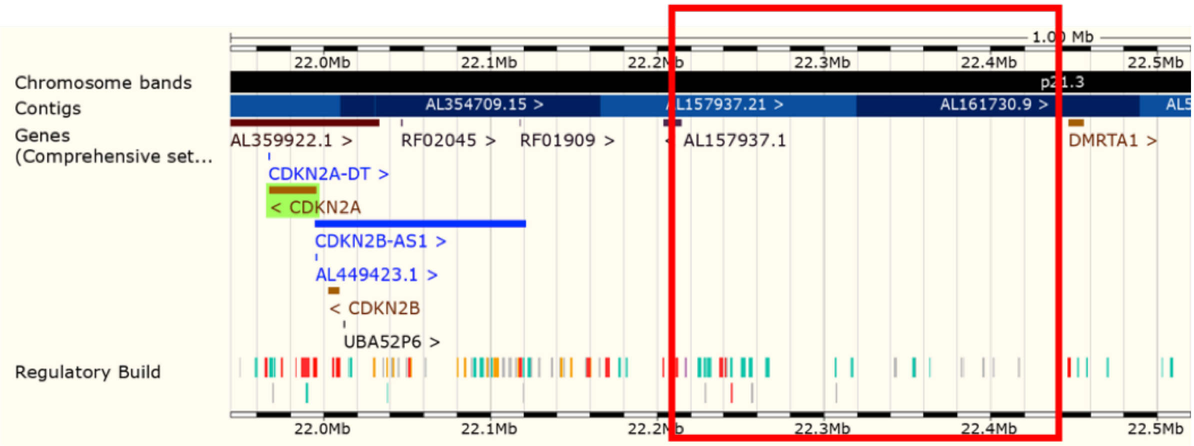


Figure 3.17: A 233,780 bp deletion upstream of *CDKN2A* observed in a Sydney pedigree with 11 sequenced members. The coloured bars indicate the relative locations of genes in chromosome 9 including *CDKN2A*, which is highlighted in green. The red box highlights the deleted region within the chromosome. This lies between *CDKN2A* and *DMRTA1*. Ten out of eleven sequenced members carried the deletion. The remaining sequenced member was confirmed to be a phenocopy by our collaborators in Sydney. Adapted from Ensembl.

Australia, within individuals residing in both Sydney and Brisbane. Twenty individuals with melanoma were identified within the pedigree with four individuals having multiple primary melanomas during their diagnosis. Lung cancer, pancreatic cancer, breast cancer and colon cancer were also observed in individuals from the pedigree. Eleven members with melanoma were chosen to be sequenced within the family, including three out of the four individuals with multiple primary melanomas. Ten out of eleven sequenced members from the pedigree carried the deletion. Other members with melanoma were sequenced by our collaborators in Australia, who confirmed the deletion in these members as well. The high segregation of the disease with the deletion, combined with the knowledge that it potentially disrupts enhancers and transcription factors that regulate expression of *CDKN2A* makes it a highly compelling candidate for a driver variant. Additionally, while roughly 40% of familial melanoma pedigrees carry germline mutations in *CDKN2A,* such variants have only been identified in the context of SNPs affecting the coding region of the gene. The role of structural variants in *CDKN2A* are still unexplored due to a lack of large scale whole genome germline sequencing of familial melanoma patients.

The identification of this deletion led to the investigation of other variants which segregated with the disease in a similar manner within the pedigree. A missense variant was identified in *DMRTA1* which encoded a p.E383Q (GRCh38 reference build, Chromosome 9, genomic position 22451543, c.2669G>C) . This variant was previously discussed in Section 3.6.2 as it was a variant that segregated with the disease in the most number of affected members. The p.E383Q variant was observed in the same ten members who carried the deletion upstream of *CDKN2A*. *DMRTA1* has not been associated with melanoma development previously but is located adjacent to *CDKN2A* on Chromosome 9, as shown in Figure 3.17, bordering the deletion. This indicates that the missense variant is in linkage disequilibrium with the deleted region. In addition to the Sydney pedigree, individuals from two other pedigrees also carried the missense variant. However, these individuals were exome sequenced and not whole genome sequenced. Although the deletion could not be identified within the other members carrying the *DMRTA1* missense variant, the presence of the deletion in the Sydney pedigree implies that the other members could also carry the deletion.

Potentially interesting structural deletions were also observed in *AT-Rich Interaction Domain 1B* (*ARID1B)* and *Cut Like Homeobox 1 (CUX1).* However, mutations in *ARID1B* and *CUX1* have not been previously associated with familial melanoma. Additionally, these variants were present in a large number of individuals within the families that carried the deletions, but they did not completely segregate with the disease. Thus, these deletions may not as impactful or significant as they are for *CDKN2A*. Additionally, the absence of a suitable control dataset for structural variants complicates the process of discovering of novel, rare structural variants. However, the presence of such variants within this dataset warrants further investigation into the importance and role of structural variants in carcinogenesis.

Chapter 4

Discussion

4.1 A summary of the dissertation

Since the discovery of *CDKN2A* as the primary driver gene in familial melanoma, several other driver genes have been established including *BAP1, TERT* and *POT1*. However, the germline mutations responsible for more than half of the individuals affected by familial melanoma globally are still unknown. This dissertation is a description of the work carried out over the duration of my PhD which aimed at identifying novel germline variants that predispose the individuals and the families carrying these variants to develop familial melanoma. A total of 308 familial melanoma patients belonging to 133 families of European descent were selected from 9 different institutions across the world. These individuals were sequenced through a combination of exome and whole genome sequencing. Multiple procedures were implemented for the discovery of the candidate genes. These have been described in this dissertation, with a brief summary of each chapter given below.

To provide context towards the importance of the research question addressed in this dissertation, an understanding of the history and evolution of melanoma research was required. These topics, along with the important mechanisms and genes involved in the development of melanoma, have been described in the background section of the dissertation.

The samples that were selected to be studied as part of this project were split into 4 different datasets: the pilot whole genome dataset, the secondary Leiden whole genome dataset, a primary exome dataset and a secondary exome dataset. The criteria for sample selection and the sequencing methodologies applied were different for each of these datasets. Once the samples had been sequenced, sequence alignment and joint variant calling were performed uniformly across all 4 datasets. A principal component analysis was performed on these samples with data from the 1000 genomes project being used as a control set; this was done to avoid any potential bias in the sequencing due to population stratification[166]. No such bias was found. Additionally, there was an inherent possibility of the presence of melanoma in a few families due to an increased burden of common risk factors. This would indicate that the development of melanoma in such families was not due to the presence of a highly penetrant, low frequency variant but due to an additive burden of multiple low risk alleles. In such a case, these pedigrees would have to be removed from the dataset as they would not contribute to the identification of novel germline variants that predispose to melanoma. Polygenic risk scores were estimated for all samples and compared with a control dataset comprising of sporadic melanoma samples and unaffected individuals. While there was a significant difference between controls and affected individuals, no significant difference was found between the familial cases and the sporadic cases, implying that there was not enough evidence to indicate a higher burden of low risk variants across the cases. As a result, no families were removed from the dataset. Detailed descriptions of sample selection, dataset descriptions, sequencing and variant calling methodologies, population stratification analysis and estimation of polygenic risk scores.

After the completion of the sequencing of the samples and the subsequent variant calling, the next step was to determine rare variants within the dataset. Several quality control filters were applied to remove variants of low quality. In order to effectively define candidate driver genes, a strategy was developed to ascertain genes with an increased burden of mutations. gnomAD were chosen as the control dataset for this purpose. Variants in the cases and the controls were filtered using the same workflow. An association analysis was performed on the variants to recognise genes with an increased burden of mutations in the cases. In addition to *BAP1*, candidate genes including *MUC4, UBR5, ITGAV* and *EPHA7* were discovered. To account for the family structure of the samples in the dataset, a joint association-linkage analysis using pVAAST was also implemented. LOD scores were estimated for families with at least 2 sequenced individuals. These LOD scores were combined with the the CLRT scores from the association analysis to generate CLRTp scores for every gene which was used to rank the genes, resulting in more novel candidate genes.

While the joint association-linkage analysis helped in the identification of novel exonic variants, there were still a few edge cases that this did not account for. Multiple procedures were developed and applied on the exonic variants concurrent to the association-linkage analysis. Variants in known melanoma driver genes were examined to guarantee that the families in the dataset did not carry a nonsense variant in these genes. A mixture of previously known variants and novel variants were discerned in 8 families in genes including *BAP1*, *BRCA2*, *CDKN2A*, *POT1* and *MITF*. In a second method, the proportion of samples in a pedigree carrying a specific variant were estimated for all variants. This was carried out to find variants that segregated with the disease in all sequenced members of a given pedigree and to supplement the linkage analysis. Loss-of-function mutations in *ATR*, *TP53AIP1* and *EXO5* were found in 10 pedigrees using this procedure. Additionally, missense variants in *DMRTA1* and *AMER1* were also found in 3 other pedigrees through the same approach. All of these genes have previously been associated with cancer development and in the case of *TP53AIP1*, *EXO5* and *AMER1*, specifically to melanoma development. The third and final method for the secondary analysis of the exonic region variants focussed on the presence of known pathogenic variants within the cases. ClinVar, a curated database of variants, their estimated pathogenicity, and the associated disorders, was utilised for this purpose[188]. Pathogenic variants in genes associated with oculocutaneous albinism and hereditary cancer syndrome were observed.

A subset of the individuals selected for the project were whole genome sequenced. The availability of variant information across the entire genome allowed for the investigation of both small and large non-coding changes and their potential impact on melanoma oncogenesis, an aspect of familial melanoma research that has been relatively unexplored thus far due to the prohibitive cost of whole genome sequencing. Two complementary workflows were developed for this purpose. These workflows were implemented on the subset of samples that were whole genome sequenced. The first approach focussed on variation in the regions of the genome that contained transcription factor binding motifs. The locations of transcription factor binding motif sites were obtained from Ensembl. Variants within the motifs were filtered in the cases and compared to similar variants in the controls comprising 7509 whole genomes sequenced individuals from the gnomAD dataset. An association analysis, similar to the method utilized for the exonic region variants, was performed to establish genes with an increased burden of transcription factor binding motif variants. *VAV1, SKI* and *SRC* were recognised as potential candidates. The second approach centered on the impact of large scale structural variation on melanoma onset. Insertions, deletions, translocations and duplications were discerned on the 123 whole genome sequenced individuals belonging to the pilot whole genome dataset. An association analysis could not be conducted on these variants due to the lack of a suitable control dataset. Novel structural variants were determined by estimating large overlapping variations that were present in all sequenced members of pedigree. This led to the discovery of a 233,780 base pair deletion upstream of the transcription start site of *CDKN2A*. This deletion was observed in 10/11 members of a pedigree from Sydney. Additional members of the pedigree were sequenced by our collaborators at Sydney who confirmed the presence of the deletion in these members as well. Experiments involving CRISPR induced deletions of the region are currently underway to validate the effect of this deletion on melanoma development.

In summary, a multi-pronged approach was utilized to determine novel germline variants in familial melanoma patients affecting both the coding and the non-coding regions of the genome to identify candidate melanoma driver genes. The merits and demerits of each applied method are also discussed. Contrary to expectations, a single driver gene affecting a large proportion of families, similar to *CDKN2A*, was not identified during this project. However, several candidate genes affecting smaller number of families were discovered across all applied methods.

4.2 Evaluating hypotheses and aims of the project

A list of hypotheses and aims were described in Section 1.7 that were necessary to be achieved in order to fulfill the target goal of the project which was to determine novel variants that predisposed individuals carrying these variants to the development of familial melanoma. This section summarises the work done over the duration of my PhD towards the fulfillment of these aims.

- To obtain the samples of familial melanoma patients from multiple locations/sources and to analyse these samples - through exome or whole genome sequencing.

308 patients diagnosed with familial melanoma from 133 different pedigrees were identified from 9 locations across the world. Samples were collected from these patients and sent by collaborators. These samples were in a mixture of exome and whole genome sequencing - 151 whole genomes and 157 exomes were sequenced.

- To incorporate all the individual datasets sequenced through different methods into a single, consistent dataset.

The 308 patients were sequenced as part of 4 different datasets. These datasets were sequenced at different times using different technologies. The four datasets were then aligned to the same reference genome build and the samples were filtered for a minimum average coverage of 15 across all sequenced positions to create a uniform dataset.

- To perform variant calling uniformly across the dataset and to annotate each mutation with their predicted consequences on protein function.

Variants were called across all samples using GATK haplotype caller. A multisample Variant Calling Format file with all mutations across all samples was generated. Variant Effect Predictor was used to identify the consequences of all variants, both in the coding region and the *non-coding region of the genome. A subset of variants predicted to be deleterious for protein function were identified using these annotations.*

- To perform preliminary analyses on the dataset to eliminate potential pre-existing biases related to an increased burden of common risk factors and population stratification.

Genotype information from the 1000 genomes project was used as a control dataset. A principal component analysis was performed on a filtered set of variants from the cases and controls. The first three principal components were compared and plotted to determine the absence of a bias due to population stratification, which was later confirmed. Additional genotype information for sporadic cases and unaffected controls were obtained for a set of common risk factor variants from collaborators at the University of Leeds. Polygenic risk scores were calculated for all three groups across these common positions. The comparison of polygenic risk scores confirmed the absence of an increased burden of common risk factors within the familial melanoma pedigrees.

- To identify rare, deleterious variants in data from cases and controls by filtering on several criteria.

Variant data from ExAC and gnomAD were chosen, obtained and used as a control dataset. Additional information regarding variant frequency, coverage, alternate allele read depth, alternate allele read frequency and gene status in the Cancer Gene Census were estimated and annotated to the variants from both the cases and the controls. Several quality control filters based on these criteria were applied to discern rare variants.

- To utilise a rare variant association analysis for the identification of genes with a higher mutation burden in cases compared to controls.

Two sets of filtered variants were obtained based on the presence of the reported gene in the Cancer Gene Census or being a protein coding gene. A Fisher's Exact Test was applied on sample counts for every gene in each set to obtain a p-value indicating the increase in mutation burden in the cases compared to the controls. These p-values were corrected for false discovery rate. The corrected p-values were used to obtain a ranked list of genes based on the association of each gene with familial melanoma.

- To design and execute a joint approach combining association analysis and linkage analysis that employs both variant data from the sequencing and the relatedness data from the pedigrees can be utilised in determining novel candidates for familial melanoma development.

Pedigree Variant Annotation, Analysis and Search Tool (pVAAST) was chosen to determine a joint association and linkage score for each gene. The set of genes being analysed was restricted to the Cancer Gene Census to focus on the most probable candidates and for computational feasibility. Suitable background datasets were generated using sequences from the 1000 genomes project and the INTERVAL project. A CLRT score and a LOD score were estimated for each gene in the Cancer Gene Census corresponding to the association and the linkage of the gene with the phenotype of interest, i.e., familial melanoma. CLRT and LOD scores were plotted to identify novel candidates that could not be determined through a pure association analysis.

- To establish methods that can determine variants related to cancer development which cannot be identified through a rare-variant association and linkage analysis.

Three different approaches were identified to determine variants that could potentially be responsible for the onset on melanoma within the pedigrees. The first approach involved the analysis of all non-synonymous variants in known familial melanoma driver genes including BAP1, BRCA2, CDK4, CDKN2A, MITF, POT1 and TERT. The second approach involved the identification of variants with high or complete segregation with disease within the familial melanoma pedigrees and to investigate their role in the development of melanoma. The third approach necessitated the identification of previously known pathogenic variants responsible for several disorders including cancer within the cases. This approach was executed using the ClinVar database as a reference dataset.

- To determine which of these variants have high segregation within our cases and to account for the presence of potential phenocopies within the pedigrees.

A parameter called segregation percentage, representative of the number of sequenced people within a family that carried a particular variant was defined. This parameter was used to find novel variants with high segregation with the disease in pedigrees with several affected members. Different key nonsense and missense variants that were recognized using this approach were reported along with the affected pedigrees.

- To identify variants in known melanoma predisposition genes by annotating their clinical significance using ClinVar and to explore potentially pathogenic variants associated with cancer.

The complete set of variants from ClinVar were obtained and filtered to contain all pathogenic variants. The dataset of variants from the cases were analysed to determine the presence of *such pathogenic variants, including the ones associated with cancer development. Additionally, several variants in key genes associated with the onset of albinism were also discovered and reported.*

- To establish the location of transcription factor binding motifs across the genome.

The location of transcription factor binding motifs in Homo Sapiens were determined and obtained from Ensembl along with their names and corresponding JASPAR motif ids.

- To ascertain rare non-coding variants that lie within transcription factor binding motifs.

The location of transcription factor binding motifs from Ensembl were used to determine the presence of variants within transcription factor binding motifs in the cases.

- To determine a suitable control dataset and to identify genes with increased burden of non-coding variants within transcription factor binding motifs in cases compared to controls and to discern rare variants within the non-coding region of the genome.

Variants from non-Finnish European gnomAD genomes were used as the control dataset for the non-coding region of the genome. A population allele frequency filter of .05 was applied to identify variants that were sufficiently rare within the dataset. Variants were identified within both the cases and controls that were present within transcription factor binding motifs. An association analysis was then performed to determine genes with an increased burden of mutations for variants within the motifs.

- To establish a workflow for the identification of structural variants within the cases.

LUMPY, a software that uses the presence of discordant reads and split reads to recognize breakpoints and report structural changes, was chosen to identify structural changes within the dataset. Variants were then collapsed, filtered and annotated based on several different criteria.

- To identify novel structural variants disrupting known cancer genes.

A large deletion was identified upstream of CDKN2A in a 11 member pedigree with 10 out of 11 sequenced members carrying the deletion. Additional structural variants were observed in ARID1B and CUX1.

4.3 Major findings of the project

Several key results with potential clinical, scientific, therapeutic and technical relevance were determined over the duration of this project. Some of the most relevant results and their prospective importance are discussed in this section.

1. **The relevance of polygenic risk scores in the identification of novel genes in familial studies**: Familial studies in melanoma have so far been restricted to focussing on the identification of key driver genes with high penetrance such as *CDKN2A*, *CDK4* and *POT1*. However, it is increasingly evident that a large percentage of the affected families with an unknown genetic cause are also potentially afflicted due to an increased burden of common low-risk factors. In this study, I used a set of 20 common low risk genetic markers from a previous GWAS study to estimate the polygenic risk scores for a set of whole genome samples. Similar risk scores were also estimated for a set of sporadic and control samples. On comparing the burden of these mutations in the cases, it is clearly apparent that the overall burden of common, low risk markers is significantly higher in both the sporadic and the familial cases compared to unaffected controls. However, there was no difference between the sporadic cases and the familial cases. While it was expected to observe such a burden of mutations in sporadic cases, it was interesting to observe a similar burden of such mutations in familial cases as well. This implies that there is a possibility for other familial melanoma pedigrees to have been affected not due to a mutation in a high penetrance gene like *CDKN2A*, but due to a high polygenic risk score. With continued efforts on large scale GWAS for melanoma such as the study by Landi et al[93], more high-frequency, low-penetrance and low-risk genetic markers can be determined for familial melanoma in the future, which would vastly improve the estimation of polygenic risk scores. Continued efforts on determining the risk scores for familial cases with unknown germline genetic causes could explain the underlying burden that resulted in their predisposition to melanoma.

2. **The identification of *ATM* and *ATR* as potential candidates for melanoma:** *ATM* and *ATR* are highly conserved key regulators of the DNA damage response pathway responsible for maintaining genomic integrity with previously established roles in cancer onset[270]. While *ATR* mutations have been observed previously in melanoma, they have been very rare [236]. The identification of a novel nonsense *ATR* variant that completely segregated with the disease in this study adds credence to the theory of *ATR* being a low frequency, high penetrance gene for melanoma. Additionally, this pedigree had multiple sequenced members who were negative for other known familial melanoma

driver genes, which points at the *ATM-ATR* damage repair pathway being an alternative mechanism for familial melanoma in addition to the well established cyclin dependant kinase pathway and the telomerase maintenance pathway. Following the identification of the *ATR* variant, *ATM* was also investigated for deleterious mutations. While nonsense mutations were not identified, several missense variants that alter the amino acids were determined for ATM, some of which segregated completely with the family. Such strong evidence in the burden of mutations in *ATM* and *ATR* pose an interesting avenue to follow for the identification of other novel pathways for melanoma. The specific role of *ATM* in melanoma onset is currently being investigated jointly with my collaborators in Genoa.

3. **The role of pigmentation genes in melanoma:** Mutations in key pigmentation genes are known to result in an autosomal recessive genetic disorder called oculocutaneous albinism (OCA). Variants in *TYR* cause OCA1 while a milder version of the disorder called OCA2 is caused due to mutations in a gene also called *OCA2[271].* Both *OCA2* and *TYR* have been previously identified as markers in a melanoma GWAS [51]. By using information from Clinvar to specifically focus on disease causing mutations, multiple segregating variants were observed in both *TYR* and *OCA2* that were predicted to be pathogenic on ClinVar. A detailed investigation into one of the affected pedigrees also identified that the individuals affected with melanoma were all affected with albinism and were negative for other driver genes. While most of the mutations we identified were from pedigrees from The Netherlands, similar large scale familial melanoma studies have also identified mutations in pigmentation genes in familial melanoma cases[272]. Knowing the importance of pigmentation on the protection of the skin, this suggests a causal link between the pigmentation pathway that results in oculocutaneous albinism and familial melanoma. Future clinical, therapeutic and diagnostic methods for the treatment of melanoma can potentially be guided by this knowledge as this identifies yet another mechanism of familial melanoma development.

4. **The importance of analysing variants that disrupt transcription factor binding motif:** A reliance on exome sequencing studies over the last decade has resulted in a lack of focus on the relevance of the non-coding region on disease development. The availability of whole genome sequences in this study allowed for the development of a novel approach in studying familial melanoma, which was to focus on transcription factor binding motifs. While studies have previously determined such variants in melanoma such as the identification of *TERT* promoter mutations, such studies have been restricted

to a single pedigree. This project was the first time that a large scale association analysis was performed strictly focussing on variants present in transcription factor binding motifs. Several key oncogenes were identified through this approach including *VAV1, SKI* and *SRC.* As a follow up, the motifs were investigated to see how the variants would affect their structure and in all cases, the conserved positions were disrupted, indicating a potential failure in the binding of transcription factors. This suggests several key approaches for the future of sequencing studies, both for familial melanoma studies and for genetic studies in general. Non-coding variants are still largely under-explored due to the sheer volume of data to be processed and an inherent lack of focus on regions of interest but an approach such as the one utilised in this project makes use of the volume of data while providing clinically meaningful results. Larger sequencing studies and specific research into the non-coding region of the genome will further our understanding of their importance in candidate gene identification for familial melanoma; this study provides a promising start to this idea.

5. **Structural variants and their impact on genetic testing:** This study identified a novel deletion upstream of *CDKN2A* segregating in a large familial melanoma pedigree that also segregated with a missense variant in a gene adjacent to *CDKN2A*. While the impact of this variant is currently being validated experimentally by my collaborators, this discovery already impacts our understanding of what is required for genetic testing. This family was previously tested for genetic variants in *CDKN2A* and tested negatively repeatedly even though they clinically presented phenotypes that represented a disruption of *CDKN2A*. The discovery of this structural variant implies that genetic testing should not just be restricted to SNPs and splice site variations but should increase their scope to include non-coding variants that disrupt enhances, transcription factor binding sites, promoters and larger structural alterations to truly establish the genetic origins of the diseases. This also has serious implications on the therapeutic treatment of patients. Improved sequencing methodologies and other large scale sequencing projects such as this one will refine our understanding of the importance of structural variants in genetic aetiology of cancer and its role in genetic testing in the future.

4.4 Future prospects and conclusion

While the results of the dissertation helped provide further inroads to our understanding of familial melanoma, several questions remain unanswered. The outcomes of the different approaches not only validated the importance of previously discovered driver genes such as *BAP1* but also yielded interesting candidate genes. The biological mechanisms that drive melanoma genesis through these candidate genes and the role of the discovered mutations in this process are however still an enigma. Experimental validation through mouse-models and CRISPR screens are essential to dictate the relevance of these candidate genes. Concurrently, replication of results in other familial studies as shown in the case of the variations in *EXO5* and *TP53AIP1* provide confidence in the relevance of these mutations for melanoma development. Additional sequencing studies are therefore required to determine the general incidence and effect of these variants in familial melanoma pedigrees.

During the initial design of the project, it was intentionally chosen to not sequence normal individuals belonging to the same families as the affected individuals, which is usually the norm in familial studies. The reasoning behind this decision was that it was possible for the unaffected individuals to still carry a causative highly penetrant variant, without developing the disease. The selection of such individuals for the filtering of variants in the dataset could have excluded such driver variants and thus disrupted the detection of novel driver genes. However, over the course of the project, it was apparent that there was no suitable control dataset which enabled large scale association analyses for related individuals. A possible alternative solution for future studies involving familial data would be the creation of a control dataset of related individuals belonging to the same families who are disease-free. Such a set of samples would result in a more accurate and direct comparison of results as opposed to comparison of results with unrelated individuals. This would not only be useful in the context of familial melanoma but for all germline genetic diseases and disorders.

Given a lack of suitable familial control datasets or matched controls within the sequenced pedigrees, gnomAD was chosen as the control dataset. The primary reason for the selection of this dataset was the high number of sequenced samples available. gnomAD v2.0.2, the chosen control dataset, contained variant information for 138,632 individuals which included 15,496 whole genome sequences. This was considerably higher than all other control datasets in contention, such as the 1000 genomes project and the UK10K project. However, gnomAD only provides summary statistics for the genotypes at every variant position, categorized by population. While this still allows for effective comparison of mutation burden with the samples in the cases, the lack of individual-level genotypes means that there is a possibility of a single sample being included multiple times when the number of affected samples is estimated for every gene. This possibility was reduced by filtering variants for rare mutations, thus minimising the likelihood of multiple variants in the same gene in the same sample. The availability of a control dataset with similar number of samples with additional genotype level information for each sample would considerably increase the accuracy of the burden estimation, which would in turn enhance the detection of candidate driver genes.

The dataset generated for the purpose of the dissertation comprised 308 individuals from 133 families and was the largest of its kind over the duration of this project. However, it is increasingly evident that the remaining undiscovered germline driver mutations occur at such low frequencies that additional individuals need to be sequenced to effectively identify them. Other similar studies that were concurrently carried out by the collaborating members of GenoMEL provides credence to this idea. To this end, a new collaborative effort was initiated near the end of my project which aimed at collecting, curating and processing all available sequences (current and future) of familial melanoma individuals and pedigrees. These individuals would be identified and sequenced by the different collaborating members of the GenoMEL consortium, who would then upload the sequences to a universal web portal. Data generated from these individuals would then be processed in a manner similar to the workflow implemented in this dissertation and results would eventually become available to all members of the consortium. This would also help in the harmonisation of different datasets enabling comparison across a much larger number of cases as compared to a single dataset. This collaboration, entitled Bionimbus, has already been established. Work on the creation of the web portal is currently ongoing; the samples used in this dissertation will be part of the first batch of samples uploaded to the portal.

The availability of 151 whole genome sequences enabled the investigation of non-coding region variants and large structural changes as drivers of melanoma. This provided a considerable amount of insight into both, the effects of these variations and the establishment of a workflow for the identification and filtration of such variations. However, we believe that the results of these investigations could be improved with additional whole genome sequences for the cases. Endeavors like the Bionimbus project will be vital for the development and refinement of strategies to further the research of non-coding region variations in melanoma onset.

On a related note, the non-coding variant analyses section of this dissertation focussed on the disruption of transcription factor binding motifs. However, it is already known that the creation of new transcription factor binding motif sites are also responsible for melanoma carcinogenesis, as seen by the effect of germline *TERT* promoter mutations[131]. While a method is in development for the identification of transcription factor binding motifs created by the variants in the samples, this is still in progress and is therefore not included as part of the dissertation. This would be the ideal follow-up for the analysis of non-coding variants and would supplement the results of motif disruption.

The large deletion upstream of *CDKN2A* detected in the structural variant analysis also

various interesting questions in terms of the origins of melanoma and to what is considered as a "genetic variation". Traditional genetic analyses have been restricted to single nucleotide polymorphisms and small indels in the exonic region. This is due to the fact that they would have the most direct impact on the function of the associated gene. With the advent of cheap and improved next-generation sequencing technologies, whole genome sequencing is increasingly feasible for large scale sequencing studies. The availability of additional whole genome sequences should considerably further the investigation of the role of structural variation in cancer progression. Although several novel structural variants were detected in addition to the *CDKN2A* deletion, the lack of a suitable control dataset precluded any further examination as these variants could not be filtered effectively. The availability of such a control dataset containing information on the incidence and frequency of different structural changes across the genome in different populations in normal individuals would facilitate and improve structural variant detection and analysis.

In conclusion, it is possible that all major familial melanoma driver genes affecting a large percentage of familial melanoma pedigrees such as *CDKN2A* have already been discovered. This would suggest that any remaining genes would affect a much smaller proportion of families. For context, there were only 87 *BAP1* mutated probands identified worldwide in 2017 [273] with *POT1* having similarly low rates of incidence in melanoma families[274]. Results obtained from this dissertation seem to replicate such findings of high penetrance variants in limited families for every candidate gene. Hopefully, efforts of pooled data such as Bionimbus can lead to the discovery of more such genes and yield insights not only on the biological roots of familial melanoma but on the central mechanisms of cancer development.

References

[1] Cancer Research UK. Skin cancer statistics. Tech. Rep. (2016).

[2] American Cancer Society. Cancer Facts & Figures 2018. Tech. Rep. (2018).

[3] Cancer Research UK. Melanoma skin cancer survival statistics, accessed on June 21st,2019. (2016).

[4] Urteaga B., O. & Pack, G. T. On the antiquity of melanoma. *Cancer* **19**, 607–610 (1966).

[5] Lane-Brown, M. & Roxanas, M. G. Laennec's melanosis: The first published description of metastatic melanoma. *Australasian Journal of Dermatology* **58**, 234–235 (2017).

[6] Norris, W. Case of Fungoid Disease. *Edinburgh medical and surgical journal* (1820).

[7] Lynch, H. T., Frichot, B. C. & Lynch, J. F. Familial atypical multiple mole-melanoma syndrome. *Journal of Medical Genetics* **15**, 352–356 (1978).

[8] Lancaster, H. O. Some geographical aspects of the mortality from melanoma in Europeans. *Medical Journal of Australia* **1**, 1082–1087 (1956).

[9] Clark, W. H., From, L., Bernardino, E. A. & Mihm, M. C. The histogenesis and biologic behavior of primary human malignant melanomas of the skin. *Cancer research* **29**, 705–27 (1969).

[10] Breslow, A. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Annals of surgery* (1970).

[11] Miller, A. J. & Mihm, M. C. Melanoma. *New England Journal of Medicine* **355**, 51–65 (2006).

[12] Shimizu, K., Goldfarb, M., Perucho, M. & Wigler, M. Isolation and preliminary characterization of the transforming gene of a human neuroblastoma cell line. *Proceedings of the National Academy of Sciences* **80**, 383–387 (1983).

[13] Padua, R. A., Barrass, N. & Currie, G. A. A novel transforming gene in a human malignant melanoma cell line. *Nature* (1984).

[14] Muñoz-Couselo, E., Adelantado, E. Z., Ortiz, C., García, J. S. & Perez-Garcia, J. NRAS-mutant melanoma: Current challenges and future prospect (2017).

[15] Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).

[16] Huebner, K. *et al.* Actively transcribed genes in the raf oncogene group, located on the X chromosome in mouse and human. *Proceedings of the National Academy of Sciences* **83**, 3934–3938 (1986).

[17] Ikawa, S. *et al.* B-raf, a new member of the raf family, is activated by DNA rearrangement. *Molecular and cellular biology* (1988).

[18] Rapp, U. R. *et al.* Structure and biological activity of v-raf, a unique oncogene transduced by a retrovirus. *Proceedings of the National Academy of Sciences of the United States of America* (1983).

[19] Ascierto, P. A. *et al.* The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine* **10**, 85 (2012).

[20] The Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).

[21] Viros, A. *et al.* Improving melanoma classification by integrating genetic and morphologic features. *PLoS Medicine* (2008).

[22] Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature genetics* **11**, 328–30 (1995).

[23] Healy, E. *et al.* Melanocortin-1-receptor gene and sun sensitivity in individuals without red hair. *Lancet (London, England)* **355**, 1072–3 (2000).

[24] Robles-Espinoza, C. D. *et al.* Germline MC1R status influences somatic mutation burden in melanoma. *Nature Communications* (2016).

[25] Kamb, A. *et al.* Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nature genetics* **8**, 23–6 (1994).

[26] Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nature Genetics* **8**, 15–21 (1994).

[27] Goldstein, A. M. *et al.* Features associated with germline CDKN2A mutations: a GenoMEL study of melanoma-prone families from three continents. *Journal of medical genetics* **44**, 99–106 (2007).

[28] Cichorek, M., Wachulska, M., Stasiewicz, A. & Tymińska, A. Skin melanocytes: Biology and development. *Postepy Dermatologii i Alergologii* (2013).

[29] Lin, J. Y. & Fisher, D. E. Melanocyte biology and skin pigmentation (2007).

[30] Bharti, K., Miller, S. S. & Arnheiter, H. The new paradigm: Retinal pigment epithelium cells generated from embryonic or induced pluripotent stem cells (2011).

[31] Plonka, P. M. *et al.* What are melanocytes really doing all day long...? *Experimental dermatology* **18**, 799–819 (2009).

[32] Brenner, M. & Hearing, V. J. The protective role of melanin against UV damage in human skin. *Photochemistry and photobiology* **84**, 539–49 (2008).

[33] Geremia, E. *et al.* Eumelanins as free radicals trap and superoxide dismutase activities in Amphibia. *Comparative Biochemistry and Physiology – Part B: Biochemistry and* (1984).

[34] Nofsinger, J. B., Liu, Y. & Simon, J. D. Aggregation of eumelanin mitigates photogeneration of reactive oxygen species. *Free Radical Biology and Medicine* (2002).

[35] Solano, F. Melanins: Skin Pigments and Much More -Types, Structural Models, Biological Functions, and Formation Routes. *New Journal of Science* (2014).

[36] Haining, R. L. & Achat-Mendes, C. Neuromelanin, one of the most overlooked molecules in modern medicine, is not a spectator. *Neural regeneration research* **12**, 372–375 (2017).

[37] Cui, R. *et al.* Central role of p53 in the suntan response and pathologic hyperpigmentation. *Cell* **128**, 853–64 (2007).

[38] Raposo, G. & Marks, M. S. Melanosomes - Dark organelles enlighten endosomal membrane transport (2007).

[39] Raposo, G. & Marks, M. S. Melanosomes - Dark organelles enlighten endosomal membrane transport (2007).

[40] Delevoye, C. Melanin transfer: The keratinocytes are more than gluttons (2014).

[41] Boyd, K. P., Korf, B. R. & Theos, A. Neurofibromatosis type 1. *Journal of the American Academy of Dermatology* **61**, 1–16 (2009).

[42] Schadendorf, D. *et al.* Melanoma. *The Lancet* **392**, 971–984 (2018).

[43] Bevona, C. *et al.* Cutaneous Melanomas Associated with Nevi (2003).

[44] Duffy, K. & Grossman, D. The dysplastic nevus: From historical perspective to management in the modern era: Part II. Molecular aspects and clinical management (2012).

[45] Bataille, V. *et al.* Risk of cutaneous melanoma in relation to the numbers, types and sites of naevi: A case-control study. *British Journal of Cancer* **73**, 1605–1611 (1996).

[46] Falchi, M. *et al.* Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. *Nature Genetics* **41**, 915–919 (2009).

[47] Shain, A. H. & Bastian, B. C. From melanocytes to melanomas. *Nature Reviews Cancer* **16**, 345–358 (2016).

[48] Grob, J. J. *et al.* Count of benign melanocytic nevi as a major indicator of risk for nonfamilial nodular and superficial spreading melanoma. *Cancer* **66**, 387–95 (1990).

[49] Shitara, D. *et al.* Nevus-Associated Melanomas. *American Journal of Clinical Pathology* **142**, 485–491 (2014).

[50] Patton, E. E. *et al.* BRAF mutations are sufficient to promote nevi formation and cooperate with p53 in the genesis of melanoma. *Current biology : CB* **15**, 249–54 (2005).

[51] Law, M. H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nature Genetics* **47**, 987–995 (2015).

[52] Rivers, J. K., Lewis, A. E. & Tate, B. J. Sunlight: A major factor associated with the development of melanocytic nevi in Australian schoolchildren. *Journal of the American Academy of Dermatology* (1994).

[53] Dodd, A. T. *et al.* Melanocytic nevi and sun exposure in a cohort of Colorado children: Anatomic distribution and site-specific sunburn. *Cancer Epidemiology Biomarkers and Prevention* (2007).

[54] Harrison, S. L., MacLennan, R. & Buettner, P. G. Sun exposure and the incidence of melanocytic nevi in young Australian children. *Cancer Epidemiology Biomarkers and Prevention* (2008).

[55] Goldstein, A. M. & Tucker, M. A. Dysplastic nevi and melanoma. *Cancer Epidemiology Biomarkers and Prevention* (2013).

[56] Tucker, M. A. *et al.* Clinically Recognized Dysplastic Nevi: A Central Risk Factor for Cutaneous Melanoma. *JAMA* **277**, 1439–1444 (1997).

[57] Shain, A. H. *et al.* The Genetic Evolution of Melanoma from Precursor Lesions. *New England Journal of Medicine* **373**, 1926–1936 (2015).

[58] Mocellin, S. & Nitti, D. Cutaneous Melanoma In Situ: Translational Evidence from a Large Population-Based Study. *The Oncologist* (2011).

[59] Ribeiro-Silva, C., Vermeulen, W. & Lans, H. SWI/SNF: Complex complexes in genome stability and cancer (2019).

[60] Reid, A. L. *et al.* Markers of circulating tumour cells in the peripheral blood of patients with melanoma correlate with disease recurrence and progression. *The British journal of dermatology* **168**, 85–92 (2013).

[61] Scott, J. F. & Gerstenblith, M. R. Melanoma of Unknown Primary. In *Noncutaneous Melanoma*, 99–116 (Codon Publications, 2018).

[62] Smith Jr., J. L. & Stehlin Jr., J. S. Spontaneous regression of primary malignant melanomas with regional metastases. *Cancer* (1965).

[63] Dutton-Regester, K. *et al.* Melanomas of unknown primary have a mutation profile consistent with cutaneous sun-exposed melanoma. *Pigment Cell and Melanoma Research* (2013).

[64] Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer (2007).

[65] Wei, Z. & Liu, H. T. MAPK signal pathways in the regulation of cell proliferation in mammalian cells (2002).

[66] Roberts, P. J. & Der, C. J. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer (2007).

[67] Santarpia, L., Lippman, S. M. & El-Naggar, A. K. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy (2012).

[68] McCain, J. The MAPK (ERK) pathway: Investigational combinations for the treatment of BRAF- mutated metastatic melanoma (2013).

[69] Zhang, P. *et al.* Targeting CDK1 and MEK/ERK overcomes apoptotic resistance in BRAF-mutant human colorectal cancer. *Molecular Cancer Research* **16**, 378–389 (2018).

[70] Leung, G. P. *et al.* Hyperactivation of MAPK Signaling Is Deleterious to RAS/RAF-mutant Melanoma. *Molecular Cancer Research* **17**, 199–211 (2019).

[71] Tannapfel, A. *et al.* Mutations of the BRAF gene in cholangiocarcinoma but not in hepatocellular carcinoma. *Gut* (2003).

[72] Cohen, Y. *et al.* BRAF Mutation in Papillary Thyroid Carcinoma. *JNCI Journal of the National Cancer Institute* **95**, 625–627 (2003).

[73] Phipps, A. I. *et al.* BRAF mutation status and survival after colorectal cancer diagnosis according to patient and tumor characteristics. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **21**, 1792–1798 (2012).

[74] Singer, G. *et al.* Mutations in BRAF and KRAS Characterize the Development of Low-Grade Ovarian Serous Carcinoma. *JNCI: Journal of the National Cancer Institute* **95**, 484–486 (2003).

[75] Anglesio, M. S. *et al.* Mutation of ERBB2 Provides a Novel Alternative Mechanism for the Ubiquitous Activation of RAS-MAPK in Ovarian Serous Low Malignant Potential Tumors. *Molecular Cancer Research* (2008).

[76] Pratilas, C. A. *et al.* (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 4519–4524 (2009).

[77] Maurer, G., Tarkowski, B. & Baccarini, M. Raf kinases in cancer-roles and therapeutic opportunities (2011).

[78] Muñoz-Couselo, E., García, J. S., Pérez-García, J. M., Cebrián, V. O. & Castán, J. C. Recent advances in the treatment of melanoma with BRAF and MEK inhibitors. *Annals of translational medicine* **3**, 207 (2015).

[79] Long, G. V. *et al.* Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma. *Journal of Clinical Oncology* (2011).

[80] Karoulia, Z., Gavathiotis, E. & Poulikakos, P. I. New perspectives for targeting RAF kinase in human cancer. *Nature Reviews Cancer* **17**, 676–691 (2017).

[81] Harvey, J. J. An Unidentified Virus which causes the Rapid Production of Tumours in Mice. *Nature* **204**, 1104–1105 (1964).

[82] Kirsten, W. H. & Mayer, L. A. Morphologic responses to a murine erythroblastosis virus. *Journal of the National Cancer Institute* **39**, 311–35 (1967).

[83] Hall, A., Marshall, C. J., Spurr, N. K. & Weiss, R. A. Identification of transforming gene in two human sarcoma cell lines as a new member of the ras gene family located on chromosome 1. *Nature* **303**, 396–400.

[84] Muñoz-Couselo, E., Adelantado, E. Z., Ortiz, C., García, J. S. & Perez-Garcia, J. NRAS-mutant melanoma: Current challenges and future prospect (2017).

[85] Bos, J. L. ras Oncogenes in Human Cancer: A Review. *Cancer Research* **49**, 4682 LP – 4689 (1989).

[86] Fedorenko, I. V., Gibney, G. T. & Smalley, K. S. M. NRAS mutant melanoma: biological behavior and future strategies for therapeutic management. *Oncogene* **32**, 3009–3018 (2013).

[87] Pollock, P. M. *et al.* High frequency of BRAF mutations in nevi. *Nature genetics* **33**, 19–20 (2003).

[88] Bauer, J., Curtin, J. A., Pinkel, D. & Bastian, B. C. Congenital melanocytic nevi frequently harbor NRAS mutations but no BRAF mutations. *The Journal of investigative dermatology* **127**, 179–82 (2007).

[89] Ranzani, M. *et al.* BRAF/NRAS wild-type melanoma, NF1 status and sensitivity to trametinib (2015).

[90] Kiuru, M. & Busam, K. J. The NF1 gene in tumor syndromes and melanoma. *Laboratory Investigation* **97**, 146–157 (2017).

[91] Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications* **9** (2018).

[92] Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nature Genetics* **41**, 920–925 (2009).

[93] Landi, M. T. *et al.* Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nature Genetics* **52**, 494–504 (2020).

[94] Mavrakis, K. J. *et al.* Disordered methionine metabolism in MTAP/CDKN2A-deleted cancers leads to dependence on PRMT5. *Science* **351**, 1208–1213 (2016).

[95] Shen, T. *et al.* Early-onset parkinson's disease caused by PLA2G6 compound heterozygous mutation, a case report and literature review. *Frontiers in Neurology* **10**, 915 (2019).

[96] Bose, A., Petsko, G. A. & Eliezer, D. Parkinson's disease and melanoma: Co-occurrence and mechanisms (2018).

[97] Visconti, A. *et al.* Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. *Nature Communications* **9**, 1–7 (2018).

[98] Hysi, P. G. *et al.* Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nature Genetics* **50**, 652–656 (2018).

[99] Aoude, L. G., Wadt, K. A. W., Pritchard, A. L. & Hayward, N. K. Genetics of familial melanoma: 20 years after CDKN2A. *Pigment Cell & Melanoma Research* **28**, 148–160 (2015).

[100] Serrano, M., Hannon, G. J. & Beach, D. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature* (1993).

[101] Li, J., Poi, M. J. & Tsai, M. D. Regulatory mechanisms of tumor suppressor P16INK4A and their relevance to cancer. *Biochemistry* (2011).

[102] Honda, R. & Yasuda, H. Association of p19(ARF) with Mdm2 inhibits ubiquitin ligase activity of Mdm2 for tumor suppressor p53. *The EMBO journal* **18**, 22–7 (1999).

[103] Chong, S. S. *et al.* Homozygotes for CDKN2 (p16) germline mutation in Dutch familial melanoma kindreds. *Nature Genetics* (1995).

[104] Randerson-Moor, J. A. *et al.* A germline deletion of p14(ARF) but not CDKN2A in a melanoma-neural system tumour syndrome family. *Human molecular genetics* **10**, 55–62 (2001).

[105] Måsbäck, A. *et al.* Clinical and histopathological features of malignant melanoma in germline CDKN2A mutation families. *Melanoma research* **12**, 549–57 (2002).

[106] van der Rhee, J. I. *et al.* Clinical and histologic characteristics of malignant melanoma in families with a germline mutation in CDKN2A. *Journal of the American Academy of Dermatology* **65**, 281–288 (2011).

[107] Zebary, A. *et al.* Somatic BRAF and NRAS mutations in familial melanomas with known germline CDKN2A status: A GenoMEL study (2014).

[108] Staaf, J. *et al.* Primary Melanoma Tumors from CDKN2A Mutation Carriers Do Not Belong to a Distinct Molecular Subclass. *Journal of Investigative Dermatology* (2014).

[109] Ghiorzo, P. *et al.* Characterization of Ligurian melanoma families and risk of occurrence of other neoplasia. *International Journal of Cancer* (1999).

[110] Borg, Å. *et al.* High frequency of multiple melanomas and breast and pancreas carcinomas in CDKN2A mutation-positive melanoma families. *Journal of the National Cancer Institute* (2000).

[111] Lynch, H. T. *et al.* Phenotypic variation in eight extended CDKN2A germline mutation familial atypical multiple mole melanoma-pancreatic carcinoma-prone families: the familial atypical mole melanoma-pancreatic carcinoma syndrome. *Cancer* **94**, 84–96 (2002).

[112] Goldstein, A. M. Prospective risk of cancer in CDKN2A germline mutation carriers. *Journal of Medical Genetics* (2004).

[113] McWilliams, R. R. *et al.* Prevalence of CDKN2A mutations in pancreatic cancer patients: Implications for genetic counseling. *European Journal of Human Genetics* (2011).

[114] Sulong, S. *et al.* A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* (2009).

[115] Carrasco Salas, P. *et al.* The role of CDKN2A/B deletions in pediatric acute lymphoblastic leukemia. *Pediatric hematology and oncology* **33**, 415–422 (2016).

[116] Pacifico, A. & Leone, G. Role of p53 and CDKN2A inactivation in human squamous cell carcinomas (2007).

[117] Saridaki, Z. *et al.* Mutational analysis of CDKN2A genes in patients with squamous cell carcinoma of the skin. *The British journal of dermatology* **148**, 638–48 (2003).

[118] Padhi, S. S. *et al.* Role of CDKN2A/p16 expression in the prognostication of oral squamous cell carcinoma. *Oral Oncology* (2017).

[119] Zhou, C. *et al.* The Association and Clinical Significance of CDKN2A Promoter Methylation in Head and Neck Squamous Cell Carcinoma: A Meta-Analysis. *Cellular Physiology and Biochemistry* (2018).

[120] Burri, N. *et al.* Methylation silencing and mutations of the p14ARF and p16INK4a genes in colon cancer. *Laboratory investigation; a journal of technical methods and pathology* **81**, 217–29 (2001).

[121] Schneider-Stock, R. *et al.* Hereditary p16-Leiden mutation in a patient with multiple head and neck tumors. *American journal of human genetics* **72**, 216–8 (2003).

[122] Smigiel, R. *et al.* Inactivation of the cyclin-dependent kinase inhibitor 2A (CDKN2A) gene in squamous cell carcinoma of the larynx. *Molecular carcinogenesis* **39**, 147–54 (2004).

[123] Helgadottir, H. *et al.* Germline <i>CDKN2A</i> Mutation Status and Survival in Familial Melanoma Cases. *Journal of the National Cancer Institute* **108**, djw135 (2016).

[124] Puntervoll, H. E. *et al.* Melanoma prone families with CDK4 germline mutation: phenotypic profile and associations with MC1R variants. *Journal of medical genetics* **50**, 264–270 (2013).

[125] Lu, W., Zhang, Y., Liu, D., Songyang, Z. & Wan, M. Telomeres-structure, function, and regulation (2013).

[126] O'Sullivan, R. J. & Karlseder, J. Telomeres: Protecting chromosomes against genome instability (2010).

[127] Robles-Espinoza, C. D., Velasco-Herrera, M. d. C., Hayward, N. K. & Adams, D. J. Telomere-regulating genes and the telomere interactome in familial cancers. *Molecular cancer research : MCR* **13**, 211–22 (2015).

[128] Collins, K. & Mitchell, J. R. Telomerase in the human organism. *Oncogene* **21**, 564–579 (2002).

[129] Nandakumar, J. & Cech, T. R. Finding the end: Recruitment of telomerase to telomeres (2013).

[130] Bataille, V. *et al.* Nevus size and number are associated with telomere length and represent potential markers of a decreased senescence in vivo. *Cancer Epidemiology Biomarkers and Prevention* **16**, 1499–1502 (2007).

[131] Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science* **339**, 959–961 (2013).

[132] Harland, M. *et al.* Germline TERT promoter mutations are rare in familial melanoma. *Familial Cancer* **15**, 139–144 (2016).

[133] Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* (2013).

[134] Liu, X. *et al.* Highly prevalent TERT promoter mutations in bladder cancer and glioblastoma. *Cell cycle (Georgetown, Tex.)* **12**, 1637–1638 (2013).

[135] Nonoguchi, N. *et al.* TERT promoter mutations in primary and secondary glioblastomas. *Acta Neuropathologica* (2013).

[136] Liu, X. *et al.* Highly prevalent TERT promoter mutations in aggressive thyroid cancers. *Endocrine-related cancer* **20**, 603–610 (2013).

[137] Tallet, A. *et al.* Overexpression and promoter mutation of the TERT gene in malignant pleural mesothelioma. *Oncogene* (2014).

[138] Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nature communications* **4**, 2218 (2013).

[139] Scott, G. A., Laughlin, T. S. & Rothberg, P. G. Mutations of the TERT promoter are common in basal cell carcinoma and squamous cell carcinoma. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **27**, 516–23 (2014).

[140] Heidenreich, B., Rachakonda, P. S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Current Opinion in Genetics & Development* **24**, 30–37 (2014).

[141] De Lange, T. Shelterin: The protein complex that shapes and safeguards human telomeres (2005).

[142] Li, B., Oestreich, S. & de Lange, T. Identification of Human Rap1: Implications for Telomere Evolution. *Cell* **101**, 471–483 (2000).

[143] Robles-Espinoza, C. D. *et al.* POT1 loss-of-function variants predispose to familial melanoma. *Nature Genetics* **46**, 478–481 (2014).

[144] Shi, J. *et al.* Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nature genetics* **46**, 482–6 (2014).

[145] Aoude, L. G. *et al.* Nonsense Mutations in the Shelterin Complex Genes ACD and TERF2IP in Familial Melanoma. *JNCI: Journal of the National Cancer Institute* **107** (2015).

[146] Carbone, M. *et al.* BAP1 and cancer. *Nature Reviews Cancer* (2013).

[147] Harbour, J. W. *et al.* Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science* (2010).

[148] Testa, J. R. *et al.* Germline BAP1 mutations predispose to malignant mesothelioma. *Nature genetics* **43**, 1022–1025 (2011).

[149] Wiesner, T. *et al.* Germline mutations in BAP1 predispose to melanocytic tumors. *Nature Genetics* (2011).

[150] Popova, T. *et al.* Germline BAP1 mutations predispose to renal cell carcinomas. *American journal of human genetics* **92**, 974–980 (2013).

[151] Walpole, S. *et al.* Comprehensive study of the clinical phenotype of germline BAP1 variant-carrying families worldwide (2018).

[152] O'Shea, S. J. *et al.* A population-based analysis of germline BAP1 mutations in melanoma. *Human molecular genetics* **26**, 717–728 (2017).

[153] Arrangoiz, R. *et al.* Melanoma Review: Epidemiology, Risk Factors, Diagnosis and Staging. *Journal of Cancer Treatment and Research* **4**, 1–15 (2016).

[154] Coleman, W. P. Acral Lentiginous Melanoma. *Archives of Dermatology* **116**, 773 (1980).

[155] Rotte, A. & Bhandaru, M. Melanoma - Diagnosis, Subtypes and AJCC Stages. In *Immunotherapy of Melanoma* (2016).

[156] Allen, A. C. & Spitz, S. Malignant melanoma; a clinicopathological analysis of the criteria for diagnosis and prognosis. *Cancer* **6**, 1–45 (1953).

[157] Petersen, N. C., Bodenham, D. C. & Lloyd, O. C. Malignant melanomas of the skin. A study of the origin, development, aetiology, spread, treatment, and prognosis. *British journal of plastic surgery* **15**, 97–116 (1962).

[158] McNeer, G. & Dasgupta, T. Prognosis in malignant melanoma. *Surgery* **56**, 512–8 (1964).

[159] McGovern, V. J. The classification of melanoma and its relationship with prognosis. *Pathology* **2**, 85–98 (1970).

[160] Tarhini, A., Lo, E. & Minor, D. R. Releasing the brake on the immune system: Ipilimumab in melanoma and other tumors (2010).

[161] Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy (2012).

[162] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). 1303.3997.

[163] Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018).

[164] McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

[165] Ruffier, M. *et al.* Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database : the journal of biological databases and curation* **2017** (2017).

[166] Auton, A. *et al.* A global reference for human genetic variation (2015).

[167] Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics* **9** (2013).

[168] Sacks, D. B. & McDonald, J. M. The Pathogenesis of Type II Diabetes Mellitus: A Polygenic Disease. *American Journal of Clinical Pathology* **105**, 149–156 (1996).

[169] Rao, A. S. & Knowles, J. W. Polygenic risk scores in coronary artery disease. *Current opinion in cardiology* **34**, 435–440 (2019).

[170] Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

[171] Aly, M. *et al.* Polygenic risk score improves prostate cancer risk prediction: Results from the Stockholm-1 cohort study. *European Urology* **60**, 21–28 (2011).

[172] Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience* **18**, 953–955 (2015).

[173] Soura, E., Eliades, P. J., Shannon, K., Stratigos, A. J. & Tsao, H. Hereditary melanoma: Update on syndromes and management: Genetics of familial atypical multiple mole melanoma syndrome. *Journal of the American Academy of Dermatology* **74**, 395–407; quiz 408–10 (2016).

[174] Bishop, D. T. *et al.* Geographical variation in the penetrance of CDKN2A mutations for melanoma. *Journal of the National Cancer Institute* **94**, 894–903 (2002).

[175] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

[176] Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

[177] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–9 (2009).

[178] Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4**, 177–183 (2004).

[179] Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science (New York, N.Y.)* **245**, 1066–73 (1989).

[180] MacDonald, M. E. M. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–8 (1993).

[181] Nancarrow, D. J. *et al.* Confirmation of chromosome 9p linkage in familial melanoma. *American journal of human genetics* **53**, 936–42 (1993).

[182] Easton, D. F., Bishop, D. T., Ford, D. & Crockford, G. P. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **52**, 678–701 (1993).

[183] Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *American journal of human genetics* **62**, 676–89 (1998).

[184] Mehrgou, A. & Akouchekian, M. The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. *Medical journal of the Islamic Republic of Iran* **30**, 369 (2016).

[185] Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Research* **21**, 1529–1542 (2011).

[186] Hu, H. *et al.* A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature biotechnology* **32**, 663–9 (2014).

[187] Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).

[188] Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980–5 (2014).

[189] Helgadottir, H. *et al.* Phenocopies in melanoma-prone families with germ-line CDKN2A mutations. *Genetics in Medicine* **20**, 1087–1090 (2018).

[190] Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**, 405–423 (2015).

[191] Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nature Reviews Genetics* **17**, 93–108 (2016).

[192] Wetterstrand, K. A. The Cost of Sequencing a Human Genome | NHGRI.

[193] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[194] Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**, 59–69 (2012).

[195] Heintzman, N. D. & Ren, B. Finding distal regulatory elements in the human genome. *Current opinion in genetics & development* **19**, 541–9 (2009).

[196] Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics* **25**, R157–R165 (2016).

[197] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10**, 252–263 (2009).

[198] Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. Tech. Rep. 20.

[199] Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46**, D260–D266 (2018).

[200] Zhang, T. *et al.* SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma. *Cancer Research* **77**, 1649–1661 (2017).

[201] Bond, G. L. *et al.* A Single Nucleotide Polymorphism in the MDM2 Promoter Attenuates the p53 Tumor Suppressor Pathway and Accelerates Tumor Formation in Humans. *Cell* **119**, 591–602 (2004).

[202] Ruijs, M. W. G. *et al.* The single-nucleotide polymorphism 309 in the MDM2 gene contributes to the Li-Fraumeni syndrome and related phenotypes. *European Journal of Human Genetics* **15**, 110–114 (2007).

[203] Haupt, S. *et al.* The role of MDM2 and MDM4 in breast cancer development and prevention. *Journal of molecular cell biology* **9**, 53–61 (2017).

[204] Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).

[205] Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7**, 85–97 (2006).

[206] Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research* **16**, 1182–1190 (2006).

[207] Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Experimental & Molecular Medicine* **50**, 98 (2018).

[208] Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

[209] Myers, R. H. Huntington's disease genetics. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics* **1**, 255–62 (2004).

[210] Mccolgan, P. & Tabrizi, S. J. Huntington's disease: a clinical review. *European Journal of Neurology* **25**, 24–34 (2018).

[211] Asim, A., Kumar, A., Muthuswamy, S., Jain, S. & Agarwal, S. Down syndrome: an insight of the disease. *Journal of biomedical science* **22**, 41 (2015).

[212] Edwards, J. H., Harnden, D. G., Cameron, A. H., Crosse, V. M. & Wolff, O. H. A new trisomic syndrome. *The Lancet* **1**, 787–90 (1960).

[213] Patau, K., Smith, D., Therman, E., Inhorn, S. & Wagner, H. Multiple congenital anomaly caused by an extra autosome. *The Lancet* **275**, 790–793 (1960).

[214] Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of cell science* **121 Suppl**, 1–84 (2008).

[215] Ratan, Z. A. *et al.* Application of Fluorescence In Situ Hybridization (FISH) Technique for the Detection of Genetic Aberration in Medical Science. *Cureus* **9**, e1325 (2017).

[216] Kallioniemi, Visakorpi, Karhu, Pinkel & Kallioniemi. Gene Copy Number Analysis by Fluorescence in Situ Hybridization and Comparative Genomic Hybridization. *Methods (San Diego, Calif.)* **9**, 113–21 (1996).

[217] Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2014** (2014).

[218] Lindberg, M. R., Hall, I. M. & Quinlan, A. R. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics* **31**, 1286–1289 (2015).

[219] Zeitouni, B. *et al.* SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)* **26**, 1895–6 (2010).

[220] Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods* **8**, 652–654 (2011).

[221] Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)* **28**, i333–i339 (2012).

[222] Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Tech. Rep. (2014).

[223] Chaturvedi, P. *et al.* MUC4 mucin interacts with and stabilizes the HER2 oncoprotein in human pancreatic cancer cells. *Cancer research* **68**, 2065–2070 (2008).

[224] Kufe, D. W. Mucins in cancer: function, prognosis and therapy. *Nature reviews. Cancer* **9**, 874–885 (2009).

[225] Shearer, R. F., Iconomou, M., Watts, C. K. & Saunders, D. N. Functional roles of the E3 Ubiquitin Ligase UBR5 in cancer (2015).

[226] Waisberg, J. *et al.* Overexpression of the ITGAV gene is associated with progression and spread of colorectal cancer. *Anticancer research* **34**, 5599–607 (2014).

[227] Cooper, C. R., Chay, C. H. & Pienta, K. J. The role of alpha(v)beta(3) in prostate cancer progression. *Neoplasia (New York, N.Y.)* **4**, 191–194 (2002).

[228] Carbone, M. *et al.* Combined Genetic and Genealogic Studies Uncover a Large BAP1 Cancer Syndrome Kindred Tracing Back Nine Generations to a Common Ancestor from the 1700s. *PLOS Genetics* **11**, e1005633 (2015).

[229] Cheung, M. *et al.* Germline BAP1 mutation in a family with high incidence of multiple primary cancers and a potential gene-environment interaction. *Cancer Letters* **369**, 261–265 (2015).

[230] Rahner, N. & Steinke, V. Hereditary cancer syndromes. *Deutsches Arzteblatt international* **105**, 706–14 (2008).

[231] Lal, G. *et al.* Patients with both pancreatic adenocarcinoma and melanoma may harbor germlineCDKN2A mutations. *Genes, Chromosomes and Cancer* **27**, 358–361 (2000).

[232] Goldstein, A. M. *et al.* High-risk Melanoma Susceptibility Genes and Pancreatic Cancer, Neural System Tumors, and Uveal Melanoma across GenoMEL (2006).

[233] McKenzie, H. A. *et al.* Predicting functional significance of cancer-associated p16INK4a mutations in CDKN2A. *Human Mutation* **31**, 692–701 (2010).

[234] Yokoyama, S. *et al.* A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature* **480**, 99–103 (2011).

[235] Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K. & Linn, S. Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints. *Annual Review of Biochemistry* **73**, 39–85 (2004).

[236] Chen, C.-F. *et al.* ATR Mutations Promote the Growth of Melanoma Tumors by Modulating the Immune Microenvironment. *Cell Reports* **18**, 2331–2342 (2017).

[237] Oda, K. *et al.* p53AIP1, a Potential Mediator of p53-Dependent Apoptosis, and Its Regulation by Ser-46-Phosphorylated p53. *Cell* **102**, 849–862 (2000).

[238] Benfodda, M. *et al.* Truncating mutations of <i>TP53AIP1</i> gene predispose to cutaneous melanoma. *Genes, Chromosomes and Cancer* **57**, 294–303 (2018).

[239] Sparks, J. L. *et al.* Human Exonuclease 5 Is a Novel Sliding Exonuclease Required for Genome Stability. *Journal of Biological Chemistry* **287**, 42773–42783 (2012).

[240] Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports* **23**, 239–254.e6 (2018).

[241] Dębniak, T. *et al.* Founder Mutations for Early Onset Melanoma as Revealed by Whole Exome Sequencing Suggests That This is Not Associated with the Increasing Incidence of Melanoma in Poland. *Cancer Research and Treatment* (2018).

[242] Paumard-Hernández, B. *et al.* Whole exome sequencing identifies PLEC, EXON5 and DNAH7 as novel susceptibility genes in testicular cancer. *International Journal of Cancer* **143**, 1954–1962 (2018).

[243] Rivera, M. N. *et al.* An X chromosome gene, WTX, is commonly inactivated in Wilms tumor. *Science (New York, N.Y.)* **315**, 642–5 (2007).

[244] Jenkins, Z. A. *et al.* Germline mutations in WTX cause a sclerosing skeletal dysplasia but do not predispose to tumorigenesis. *Nature Genetics* **41**, 95–100 (2009).

[245] Major, M. B. *et al.* Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. *Science (New York, N.Y.)* **316**, 1043–6 (2007).

[246] Schatoff, E. M., Leach, B. I. & Dow, L. E. Wnt Signaling and Colorectal Cancer. *Current colorectal cancer reports* **13**, 101–110 (2017).

[247] Luis, T. C., Ichii, M., Brugman, M. H., Kincade, P. & Staal, F. J. T. Wnt signaling strength regulates normal hematopoiesis and its deregulation is involved in leukemia development. *Leukemia* **26**, 414–421 (2012).

[248] Pawlikowski, J. S. *et al.* Wnt signaling potentiates nevogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 16009–14 (2013).

[249] Lin, S. Y. *et al.* Beta-catenin, a novel prognostic marker for breast cancer: its roles in cyclin D1 expression and cancer progression. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 4262–6 (2000).

[250] Hawkes, J. E. *et al.* Report of a novel OCA2 gene mutation and an investigation of OCA2 variants on melanoma risk in a familial melanoma pedigree. *Journal of Dermatological Science* **69**, 30–37 (2013).

[251] Jay, J. J. & Brouwer, C. Lollipops in the Clinic: Information Dense Mutation Plots for Precision Medicine. *PLOS ONE* **11**, e0160519 (2016).

[252] Etienne-Manneville, S. & Hall, A. Rho GTPases in cell biology. *Nature* **420**, 629–635 (2002).

[253] Vega, F. M. & Ridley, A. J. Rho GTPases in cancer cell biology. *FEBS Letters* **582**, 2093–2101 (2008).

[254] Porter, A. P., Papaioannou, A. & Malliri, A. Deregulation of Rho GTPases in cancer. *Small GTPases* **7**, 123–38 (2016).

[255] Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nature Genetics* **44**, 1006–1014 (2012).

[256] Bustelo, X. R. Vav family exchange factors: an integrated regulatory and functional view. *Small GTPases* **5**, e973757 (2014).

[257] Katzav, S. Vav1: A Dr. Jekyll and Mr. Hyde protein - good for the hematopoietic system, bad for cancer. *Oncotarget* **6** (2015).

[258] Shalom, B., Farago, M., Pikarsky, E. & Katzav, S. Vav1 mutations identified in human cancers give rise to different oncogenic phenotypes. *Oncogenesis* **7**, 80 (2018).

[259] Robles-Valero, J. *et al.* A Paradoxical Tumor-Suppressor Role for the Rac1 Exchange Factor Vav1 in T Cell Acute Lymphoblastic Leukemia. *Cancer Cell* **32**, 608–623.e9 (2017).

[260] Soussi, T. & Wiman, K. G. TP53: an oncogene in disguise. *Cell death and differentiation* **22**, 1239–49 (2015).

[261] Chen, D. *et al.* SKI Activates Wnt/β-Catenin Signaling in Human Melanoma. *Cancer Res.* **61**, 8074–8078 (2003).

[262] Reed, J. A., Lin, Q., Chen, D., Mian, I. S. & Medrano, E. E. SKI pathways inducing progression of human melanoma. *Cancer and Metastasis Reviews* **24**, 265–272 (2005).

[263] Chen, D. *et al.* SKI knockdown inhibits human melanoma tumor growth in vivo. *Pigment Cell & Melanoma Research* **22**, 761–772 (2009).

[264] Bell, R. J. A. *et al.* Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science (New York, N.Y.)* **348**, 1036–9 (2015).

[265] Irby, R. B. & Yeatman, T. J. Role of Src expression and activation in human cancer. *Oncogene* **19**, 5636–5642 (2000).

[266] Homsi, J., Cubitt, C. & Daud, A. The Src signaling pathway: a potential target in melanoma and other malignancies. *Expert Opinion on Therapeutic Targets* **11**, 91–100 (2007).

[267] Homsi, J. *et al.* Src activation in melanoma and Src inhibitors as therapeutic agents in melanoma. *Melanoma Research* **19**, 167–175 (2009).

[268] Puls, L. N., Eadens, M. & Messersmith, W. Current status of SRC inhibitors in solid tumor malignancies. *The oncologist* **16**, 566–78 (2011).

[269] Ferguson, J., Arozarena, I., Ehrhardt, M. & Wellbrock, C. Combination of MEK and SRC inhibition suppresses melanoma cell growth and invasion. *Oncogene* **32**, 86–96 (2013).

[270] Awasthi, P., Foiani, M. & Kumar, A. ATM and ATR signaling at a glance. *Journal of Cell Science* **128**, 4255–4262 (2015).

[271] Ainger, S. A., Jagirdar, K., Lee, K. J., Soyer, H. P. & Sturm, R. A. Skin Pigmentation Genetics for the Clinic. *Dermatology* **233**, 1–15 (2017).

[272] Nathan, V. *et al.* Germline variants in oculocutaneous albinism genes and predisposition to familial cutaneous melanoma. *Pigment Cell & Melanoma Research* **32**, 854–863 (2019).

[273] Haugh, A. M. *et al.* Genotypic and phenotypic features of BAP1 cancer syndrome: A report of 8 new families and review of cases in the literature. *JAMA Dermatology* **153**, 999–1006 (2017).

[274] Potrony, M. *et al.* POT1 germline mutations but not TERT promoter mutations are implicated in melanoma susceptibility in a large cohort of Spanish melanoma families. *British Journal of Dermatology* **181**, 105–113 (2019).

Appendix A

Supplementary Tables and Figures

Several files that were generated over the course of the dissertation were either too large to be included in print or were descriptions of processes. These files have been included in an accompanying compact disk and are described in this Section.

A.1 Results from association analysis for all genes in the Cancer Gene Census

File name in CD : Supplementary_Table_6_Association_results_for_CGCL_genes.xlsx

This table includes the results from the association analysis for all the genes that are present in the Cancer Gene Census. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene and the false discovery rate corrected p-value of the gene.

A.2 Results from association analysis for all protein coding genes as designated on Ensembl

File name in CD : Supplementary_Table_7_Association_results_for_all_protein_coding_genes.xlsx

This table includes the results from the association analysis for all the genes that are present in all protein coding genes. The complete list of protein coding genes were obtained from Ensembl. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene and the false discovery rate corrected p-value of the gene.

A.3 Parameters used for the different steps involved in the execution of pVAAST for the joint association-linkage analysis

File name in CD : Supplementary_Table_8_parameters_for_running_pVAAST.xlsx

Multiple commands are executed to generate the pVAAST results from the input VCF file containing the jointly called variants across all the samples in the dataset. This file includes all the different steps and softwares that are a part of the pVAAST package which need to be executed for generating the final results. The parameters necessary for running these softwares along with the required input files are also specified in this file.

A.4 Results from pVAAST using the default background file from the 1000 genomes project

File name in CD : Supplementary_Table_9_pVAAST_results_default_1000G_background.xlsx

This table includes the final results from pVAAST using the default background or control file provided by pVAAST which is obtained from the 1000 genomes project. The columns included in the table are the symbol of the gene, the CLRT score generated by the association analysis, the p-value generated during the association analysis and the LOD score generated by the linkage analysis.

A.5 Results from pVAAST using INTERVAL exomes background file

File name in CD : Supplementary_Table_10_pVAAST_results_INTERVAL_exomes_background.xlsx

This table includes the final results from pVAAST using the secondary background file that became available over the duration of the project which was obtained from the INTERVAL exomes project. The columns included in the table are the symbol of the gene, the CLRT score generated by the association analysis, the p-value generated during the association analysis and the LOD score generated by the linkage analysis.

A.6 Complete list of variants with high segregation in cases

File name in CD : Supplementary_Table_11_variants_with_high_segregation_in_cases.xlsx

This table includes the list of variants where the variant almost completely segregates with the diseases in the families each variant is present in.

A.7 Complete list of variants associated with cancer in ClinVAR that are present in the dataset

File name in CD : Supplementary_Table_12_ClinVAR_cancer_variants.xlsx

This table includes all the variants annotated as playing a role in different types of cancer as annotated by ClinVAR which are also present within the cases in the dataset.

A.8 Results from the association analysis for the transcription factor binding motif variants for genes in the Cancer Gene Census

File name in CD : Supplementary_Table_13_Association_results_for_CGCL_genes_TFMOTIF_analysis.xlsx

This table includes the results from the association analysis for all the genes that are present in the Cancer Gene Census where the variants are also present in known transcription factor binding motif locations. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene, the false discovery rate corrected p-value of the gene and a final column showing the number of samples affected in each family to represent the maximum percentage of segregation of the variant with the disease for each gene.

A.9 Parameters used for the generation of structural variants

File name in CD : Supplementary_Table_14_parameters_for_running_LUMPY.xlsx

This file includes the parameters used in the generation of structural variants using Lumpy and their respective descriptions. All required input files are also specified here.

A.10 Complete list of filtered and annotated structural variants

File name in CD : Supplementary_Table_15_Structural_variants_results.xlsx

This table includes the complete list of filtered and annotated structural variants.

A.11 Supplementary Figure 1

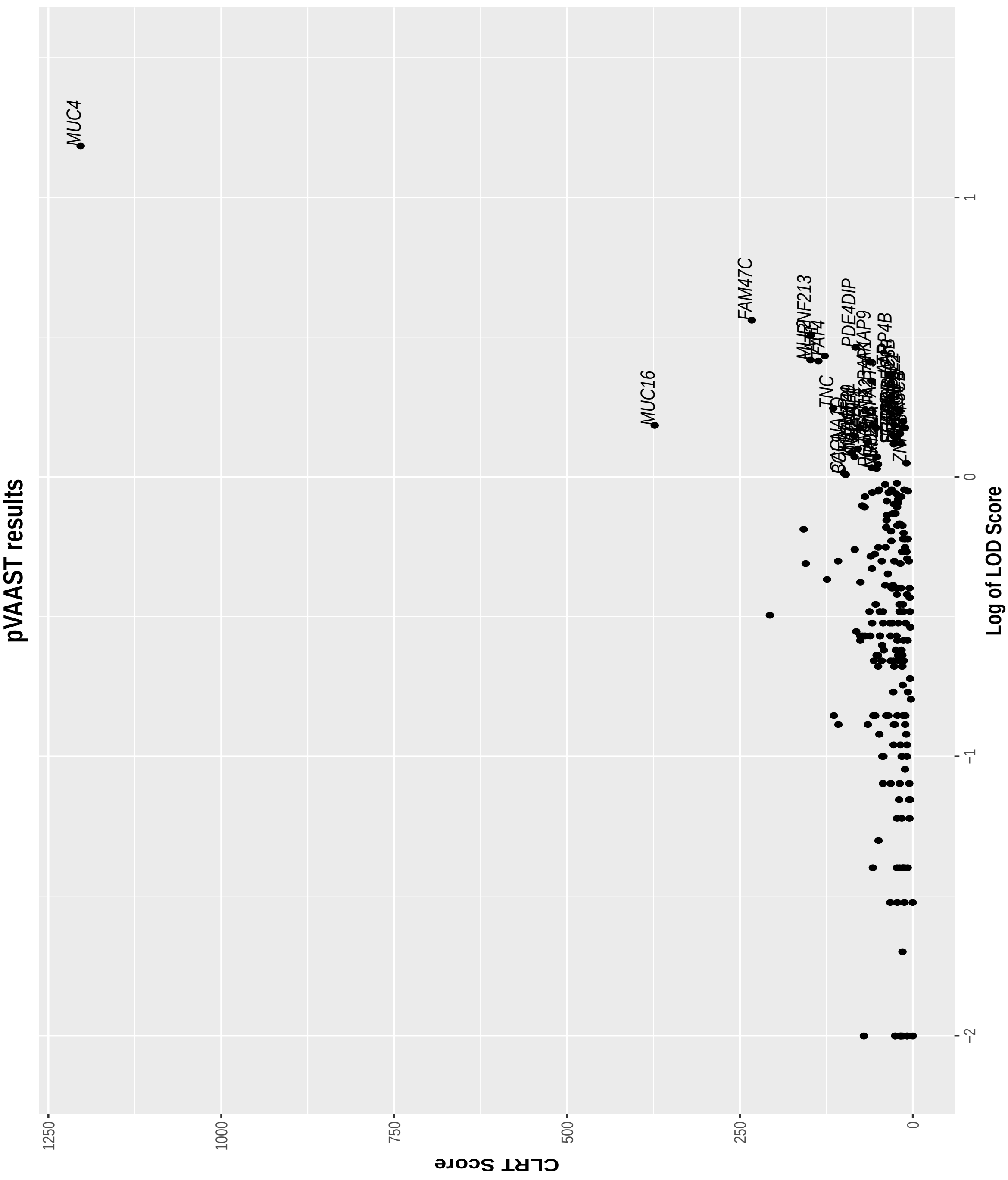


Figure A.1: Original results from pVAAST using the INTERVAL exomes as the background. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

A.12 Supplementary Figure 2

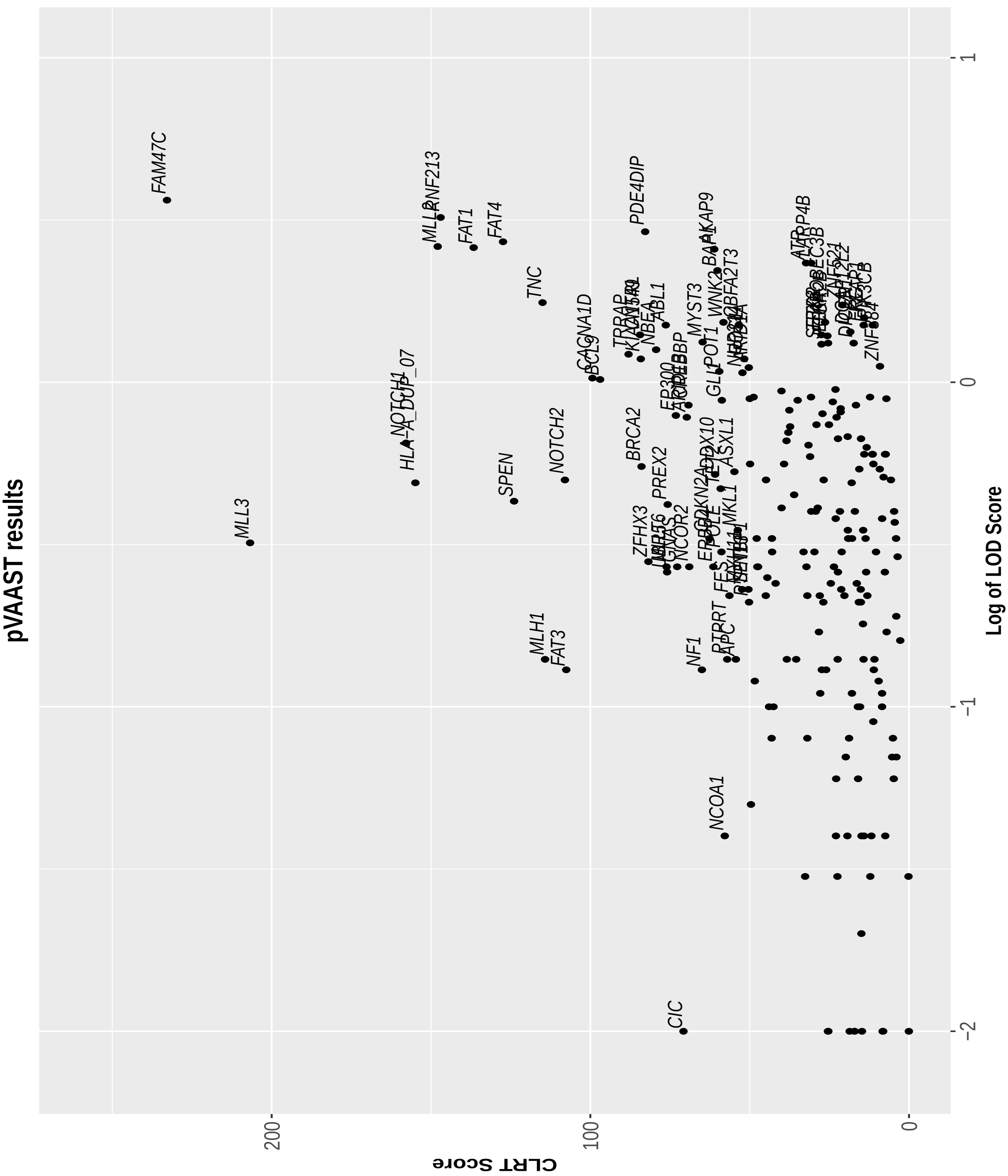


Figure A.2: Results from pVAAST for all genes in the Cancer Gene Census excluding *MUC4* and *MUC16* using the INTERVAL exomes as the background. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.