Drivers of melanoma susceptibility

Carla Daniela Robles Espinoza

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Christ's College

September 2014

Para quienes siempre estuvieron allí:
María Elena, Gabriel, Gabrielito, José y Daniel.

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 as per the requirements of the Degree Committee for the Faculty of Biology.

Carla Daniela Robles Espinoza
September 2014

Acknowledgements

First and foremost, I would like to thank my supervisor Dave Adams for welcoming me into his lab after a rather “interesting” first meeting, and for his kindness and patience over these last few years. I have learnt from him not only things related to my project and science but also important lessons in life. He is a true mentor to me.

I would like to thank all members of the Experimental Cancer Genetics Team at the Sanger Institute for filling every working day with fun and interesting lunch-time discussions, and for their disposition to help whenever I asked: James, Alistair, Mamun, Martín, Clara, Stefan, Marco, Chi, Louise, Rebecca, Marcela, Mia, Gabi, Theo and Richard, and to Ximena who is practically a member of the team as well. I also thank Thomas and Sendu, from the Vertebrate Resequencing team, for their patience and help with analysis pipelines.

Also, none of this work would have been possible without my wonderful collaborators: Julia, Tim and Mark, from the University of Leeds, Nelleke, from the Leiden University Medical Centre, and Nick, Lauren and Antonia, from QIMR Berghofer. I thank them for trusting me with the analysis of their melanoma families, for patiently answering all questions I asked, and for very valuable advice regarding this project. I would also like to thank Andrew, Víctor and Carlos, from the University of Oviedo, for their expertise in protein analyses (and for answering e-mails at strange times during the night!).

I would like to acknowledge Moyra, Yip and Lucy, whom I met the first week I arrived in Cambridge and who have become some of my closest friends. I would like to thank them for keeping me sane, making sure I ate at adequate times and generally taking care of me during these last years. Equally, I owe much of my present happiness to Abi, Rachael and Sarah, whom I met on my first day at Sanger. I thank them for their care, for listening, and for the fun weekend trips we had during these past four years. I would like to thank Carlos, Mariana and Claudia, for their everlasting friendship and support across different time zones.

I would like to acknowledge the Wellcome Trust and the Consejo Nacional de Ciencia y Tecnología of Mexico, my funding bodies, for making sure I was not homeless and hungry during these last years, and for trusting in me. This work would not have been possible without their generous support.

I would like to thank, from the bottom of my heart, those who stayed back home and have not forgotten me: My childhood friends, María and Linda, for awaiting me every time I go back to Mexico, for their life advice, help and understanding. I would like to thank my parents, María Elena and Gabriel, and my little brothers, Gabrielito and José, for their love, their neverending support and enthusiasm.

Finally, with all my love, I would like to thank Daniel. Above all, for understanding my constant travelling and moving to different countries. I will be forever thankful to him for filling every day with laughter, and for his love, support and patience.

Abstract

Cutaneous melanoma is a cancer of melanocytes, the pigment-producing cells in our skin. It is one of the most aggressive human malignancies, constituting only about 2% of all dermatological cancers but being responsible for over 75% of all deaths from skin cancer. It has recently become a major public health problem, as it is now the fifth most common cancer in the United Kingdom after its incidence more than quadrupled in the last three decades. For these reasons, understanding the biological processes that are involved in its development is of great importance for devising novel treatments and for the management of patients in the clinic.

The study of the genetic factors that influence melanoma risk can uncover mechanisms that are relevant in the transition from a benign melanocyte to a malignant melanoma. Approximately 10% of all melanoma cases are familial, and about half of these familial cases can be explained by pathogenetic variants in genes such as cyclin-dependent kinase inhibitor 2A (*CDKN2A*), cyclin-dependent kinase 4 (*CDK4*), breast cancer 2 (*BRCA2*), BRCA1-associated protein-1 (*BAP1*) and in the promoter of the telomerase reverse transcriptase (*TERT*). However, about 50% of all familial melanoma cannot be explained by mutations in known genes. In this dissertation, I detail the methodology I followed in an effort to uncover additional high-penetrance melanoma susceptibility genes.

I analysed exome and genome sequence data from a total of 184 individuals that belong to 105 melanoma-prone families from the United Kingdom, The Netherlands and Australia that did not have any pathogenetic variants in known susceptibility genes. I applied different gene prioritisation strategies and developed novel software tools in order to devise a list of plausible melanoma susceptibility candidate genes; these analyses suggested that genes regulating telomere function could be influencing melanoma risk. After performing functional experimental analyses, our research team was able to determine that carriers of rare variants in the protection of telomeres (*POT1*) gene, a member of the shelterin complex that safeguards telomere integrity, are at high risk for developing melanoma. We successfully described the mechanism by which this

happens, showing that the variants identified either disrupt *POT1* mRNA splicing or abolish the ability of POT1 to bind to telomeres, and lead to increased telomere length in carriers when compared to melanoma cases with wild-type *POT1*.

The main finding of the work described in this dissertation is the identification of telomere dysfunction as an important contributor to the risk of developing melanoma, and possibly other cancers. Our analyses suggest that *POT1* is the second most commonly mutated high-penetrance melanoma susceptibility gene reported thus far, and moreover, that rare variants in this gene constitute the first hereditary mechanism for telomere lengthening in humans.

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

ACD	Adrenocortical dysplasia homolog
BAP1	BRCA1 associated protein-1
BRCA2	Breast cancer 2, early onset
CDK4	Cyclin-dependent kinase 4
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CLL	Chronic lymphocytic leukaemia
GWAS	Genome-wide association studies
IARC	International Agency for Research on Cancer
MC1R	Melanocortin 1 receptor
MITF	Microphthalmia-associated transcription factor
MPM	Multiple primary melanomas
NGS	Next-generation sequencing
OB	Oligonucleotide/oligosaccharide-binding
POT1	Protection of telomeres
QFMP	Queensland Familial Melanoma Project
RB1	Retinoblastoma protein 1
SMG1	SMG1 phosphatidylinositol 3-kinase-related kinase

TERT	Telomerase reverse transcriptase
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Chapter 1

Introduction

1.1 From peas to man: A (very) brief history of genetics

The study of genetics started more than a century ago, when the Austrian monk Gregor Mendel performed his classic experiments with common pea plants and deduced the basic principles of heredity, which he described in a paper written in German in 1866 [1]. Mendel, after careful characterisation of several generations of pea plants, hypothesised that physical characteristics were transmitted from parents to offspring in packets of information he termed 'elements'. The material responsible for carrying these elements was identified about a decade later, albeit unwittingly, by the German biologist Walther Flemming while he was studying cell division [2]. This fibrous material, which he termed 'chromatin', would later be shown to be formed of chromosomes [2], while Mendel's elements would form the basis for the definition of gene [3]. These discoveries, whose importance was unrecognised at the time, would launch a scientific revolution that would completely change the way researchers studied the laws of heredity over the coming century.

The identification of chromosomes as the carriers of heredity in the turn of the nineteenth century by Walter Sutton [4] and Theodor Boveri [5], along with the observation that chromosomes contain nucleic acids by Oskar Hertwig and Albert von Kölliker (reviewed in [6]), sparked huge interest in the scientific community to identify the mechanism by which chromosomes could preserve and transmit information from generation to generation. In 1953, James Watson and Francis Crick proposed that specific nucleotide base pairing in deoxyribonucleic acid (DNA) could provide a mechanism for copying and transmitting information, making it the most likely gene carrier [7]. A few years later, scientists had managed to decipher the genetic code and efforts began to sequence genes and whole genomes.

Between 1977 and 1982, researchers were able to determine the complete genome sequences of the ΦX174 and λ bacteriophages [8, 9], the simian virus SV40 [10] and the human mitochondrion [11]. The success with these relatively small genomes prompted scientists to initiate efforts to sequence the entire human genome, and by 1990 the Human Genome Project had been launched with the participation of numerous research institutions around the globe [12]. This tremendous enterprise yielded the first draft of the human genome (covering about 94% of the total sequence) in 2001 [12], followed by the publication of the gold standard reference covering ~99% of the euchromatic sequence in 2004 [13].

In the few years since the publication of the human genome, sequencing technologies have evolved at an accelerated pace, bringing down the cost and the amount of human resources necessary for generating high-quality genome sequences [14]. As such, the last decade has seen a dramatic increase in the number of organisms that have had their whole genome sequenced, as well as the number of projects attempting to catalogue and understand all existing natural human variation. Examples of these are The 1000 Genomes Project [15], the UK10K Project [16] and the National Heart, Lung and Blood Institute Grand Opportunity Exome Sequencing Project (NHLBI GO ESP) [17]. Other endeavours have focused on charting the different elements encoded by and influencing regulation of the human genome, examples of which are The Encyclopaedia Of DNA Elements (ENCODE) [18], The National Institutes of Health Roadmap Epigenomics Mapping Consortium [19] and BLUEPRINT [20]. All of these interdisciplinary, multinational collaborations have allowed a much deeper understanding of the relationship between the human genome and the regulation of important biological processes, as well as the essential role it plays in health and disease.

1.2 The human genome: Its structure and gene content

In 1956, the long-standing quest to determine the number of chromosomes in human diploid cells was finally settled by two studies authored by the laboratories of Albert Levan and John Hamerton [21, 22]. These researchers were able to improve the karyotyping techniques of the time by adding steps to treat cells with colchicine and an hypotonic solution [23], and finally determined that diploid cells possess 46 chromosomes including one sex-determining pair. Other pre-sequencing era genome measurements included the estimation of the size of the whole human genome, which was approximated to be between 3000 and 3200 megabases (Mb) by physical maps [24], and guesses on the number of genes it contains, which were generally in the range of 50,000-100,000 [25].

The delivery and initial analysis of the human genome helped refine these measurements and allowed an unprecedented description of our genomic landscape. The analysis revealed that the human genome showed marked variation in the distribution of features such as CpG islands and transposable elements, that it had many more segmental duplications than other previously sequenced eukaryotes, and that it seemed to contain only between 30,000 and 40,000 genes, among other remarks [12]. Many of these observations were confirmed with subsequent analyses and studies, whereas some estimates were further refined with the release of the gold standard reference genome and advances in assembly techniques and bioinformatic algorithms. For example, the amount of segmental duplication was estimated at ˜5.3% of the euchromatic sequence as compared to a lower bound of 3.6% in the draft sequence, and the number of protein-coding genes was brought even lower to 20,000-25,000 [13].

The new estimated upper bound on the number of genes left many scientists perplexed. It was comparable to that of the simple nematode *Caenorhabditis elegans* [26], an organism composed of about 1,000 cells, and much lower than those in the single-celled pathogen *Trichomonas vaginalis* [27] and in common crops such as maize [28] and wheat [29]. Where did the perceived complexity in humans come from, if not from our protein-coding genes? Part of the explanation came from experiments that showed that the great majority of human genes are capable of generating more than one protein via alternative splicing [30], which less complex organisms, such as *T. vaginalis*, cannot do as their genes lack introns [27]. Proteins generated by human genes were also estimated to participate in many more interactions with each other than those of organisms of lower perceived complexity, such as *C. elegans* and *Drosophila melanogaster*, with the size of each species' interactome seeming to correlate better with complexity [31]. Other factors that may contribute to explain organism complexity relate to the amount of gene duplication in species such as *T. vaginalis*, maize and wheat [27, 28, 32], the intricate patterns of post-translational protein modification in humans [33], and gene expression regulatory mechanisms [34, 35].

However, protein-coding genes, their regulation, interactions and modifications offer only a small part of the answer. The realisation that the number of genes in an organism and its perceived complexity did not seem to correlate pointed to additional factors, hidden in the non-protein-coding genome, contributing at least as importantly as proteins to biological function. In fact, John Mattick soon noticed that the proportion of protein-coding DNA in an organism decreases as a function of its complexity, with prokaryotes having less than 25% of non-coding sequences and humans having approximately 98.5% [36]. Additionally, it was estimated that the amount of genome that is transcribed into ribonucleic acid (RNA) greatly exceeds the amount that is translated into proteins, significantly adding weight to this hypothesis [37]. Given these observations, efforts began to identify and study the diverse repertoire of biological elements encoded by our genome, beyond protein-coding genes, that participate in structural, regulatory and functional tasks.

Undoubtedly, the ENCODE project has been the largest collaboration, launched in 2004, with the aim of providing a complete catalogue of all the functional components encoded by the human genome [38]. In 2012, the Consortium published their extensive set of analyses in a series of 30 papers published in *Nature*, *Genome Biology* and *Genome Research* (for an overview of these publications, see the Nature ENCODE explorer at `http://www.nature.com/encode/`). By using different approaches such as chromatin immunoprecipitation (ChIP-seq), RNA sequencing (RNA-seq) and mass spectrometry in more than 140 human cell lines, ENCODE was able to assign a biochemical function to the vast majority of the human genome [18]. About 62% of the genome was found to be covered by different RNA-encoding elements, the majority in intronic regions, and included long non-coding (lnc)RNAs and small RNAs, comprising small nuclear (sn)RNAs, small nucleolar (sno)RNAs, micro (mi)RNAs and transfer (t)RNAs [18, 39]. Additionally, about 56% of the genome was found to be enriched for histone modifications, and smaller regions were classified as open chromatin (15%) or transcription-factor-binding sites (8%) [18]. ENCODE also described regions of DNA methylation, long-range physical genomic interactions, and gave an estimate of the amount of human genome under purifying selection (3-8%), among other remarks [18].

Other projects have been initiated with the aim of understanding and cataloguing natural human variation. The 1000 Genomes Project was the first such undertaking, launched in 2008, that sequenced large numbers of individuals in order to discover and haplotype all forms of DNA polymorphisms present in humans [40]. The 1000 Genomes Project Consortium published the results of their pilot phase in 2010, followed two years later by a detailed description of the findings in the genomes of 1,092 individuals from

14 populations [15, 40]. So, although humans are thought to share about 99.9% of their genetic sequence with each other (reviewed in [41]), the Consortium reported around 38 million single nucleotide polymorphisms (SNPs), 1.4 million short insertions or deletions (indels) and more than 14,000 larger structural variants which showed marked differences in allele distribution across populations [15]. In 2012, the NHLBI GO ESP published an analysis of more than 15,000 genes in 6,515 individuals, in which they reported that individuals carry an excess of rare variants, thought to have arisen recently and attributable to explosive population growth [17]. These efforts have provided us with a deeper understanding of the history and migration patterns of human populations, the burden of rare variants typically carried by any single individual, regions of the genome that are essential in determining phenotypic characteristics, and genes that might play an important role in susceptibility to disease.

With other projects currently on-going with the aim of cataloguing all human variation and elucidating the biological function of these elements [16, 19, 20], our understanding of the human genome is likely to increase dramatically in the next years. However, at this moment, the ease of rapidly generating whole-genome or targeted sequencing of hundreds of individuals, coupled with the continued development of bioinformatic tools and the ability to search tens of in-depth catalogues of human variation and diverse encoded DNA elements, has substantially helped the identification of the genomic regions that play an important role in health and disease.

The present dissertation principally deals with the use of the above-mentioned sequencing methodologies, variation catalogues and bioinformatic tools in order to search for melanoma-predisposing genome regions in a cohort of high-density melanoma families from the United Kingdom (UK), The Netherlands and Australia. In the next section, I discuss the basis of cancer aetiology in humans, with special emphasis on the known genetic components of cancer development as it is the main focus of my work. The remaining sections provide a detailed description of melanoma and its risk factors, as it is the phenotype I have studied in the course of the last four years. Finally, I conclude this introduction with a discussion of the unanswered questions in melanoma genetics and the methodology I followed in an effort to answer them.

1.3 The basis of cancer aetiology in humans

Cancer is a disease of the genome [42]. As early as 1890, David von Hansemann examined cancer cells under a microscope and noticed striking aberrations in their chromosomes, such as multipolar mitoses, breakage, asymmetry and altered chromosomal dosage [43, 44]. This observation made him postulate, contrary to beliefs at the time, that cancer was a disease of the internal hereditary material of the cell [44]. Following on these observations, Boveri made yet another landmark contribution to genetics: By studying sea urchin eggs, he hypothesised that tumours might originate as a consequence of cells passing abnormal numbers of chromosomes to their daughters [45, 46]. With their work, von Hansemann and Boveri laid the theory of the genetic origin of cancer.

Other experiments followed to show that cancer was caused by the progressive accumulation of somatic alterations in genomes [47]. By 1930, Katsusaburo Yamagiwa and Koichi Ichikawa had reported that exposing rabbits to coal tar could produce carcinomas with metastatic potential [48], and Ernest Kennaway had identified individual chemical compounds that generated tumours in mice and rabbits [49]. These chemical compounds were later shown to bind covalently to DNA and induce genetic mutations [50–53], thus providing the molecular link between exposure to certain chemicals and cancer development.

This hypothesis could explain well an observation that had been made in the 1770s. The British surgeon Percivall Pott had described that chimney sweeps, who were exposed to coal tar when working, were particularly likely to develop cancer of the scrotum, and ascribed it to be “a disease brought on them by their occupation” [54]. Since then, other industries in which workers face increased cancer risks have been described, such as mining, chemical manufacturing and iron and steel founding [55]. These observations established an important role for environmental risk factors in cancer development.

1.3.1 Environmental risk factors and their role in cancer development

Environmental risk factors, which include all non-genetic aspects such as diet, lifestyle and infectious agents, cause the majority of human cancers [56]. More than 100 agents have been classified as ‘carcinogenic to humans’ by the International Agency for Research on Cancer (IARC), which belongs to the World Health Organisation [57]. These agents, which are known as carcinogens, can have genotoxic or non-genotoxic mechanisms of action, or both. Genotoxic carcinogens can directly cause DNA damage, such as the chemicals in coal tar, whereas non-genotoxic carcinogens can have diverse modes of action and can alter diverse processes such as immune suppression, endocrine modification and inflammatory responses [58]. In this subsection, I review several of the most established or commonly found carcinogens, their modes of action and the malignancies with which they have been causally associated.

1.3.1.1 Asbestos

Asbestos encompasses a series of naturally occurring silicate mineral fibres with desirable industrial properties such as tensile strength and resistance to heat, and that therefore are of commercial interest [59]. Asbestos fibres can be classified in two types: chrysotile, which is the most commonly used type and is flexible and easily breakable, and the amphiboles, which are rigid, sharp and durable [59]. The negative health effects of asbestos exposure have been known for more than a century, when in 1899, Montague Murray diagnosed the first fatal case of asbestosis arising from occupational exposure [60]. However, proper legislation regarding dust control in North American mines was not enacted until 1971 [61]. Because of this, asbestos is still present in buildings, and given the long latency period between initial exposure and disease manifestation (15 to 40 years), it still continues to pose a major health challenge [59]. Distinct modes of action have been described, with some sources noting its genotoxic effects, causing DNA base oxidation, breakage, mutations and deletions and chromosomal aneuploidy (reviewed in [62]) that alter processes such as cell proliferation, cell death and inflammation [63], and other sources describing non-genotoxic effects such as the generation of reactive oxygen and nitrogen species (ROS and RNS, respectively) and also mitogenic and cytotoxic consequences [62, 63]. Given these different effects, which might depend on the studied fibre type, lung clearance, genetics and other characteristics [59], the IARC has included both mechanisms of action in its classification of asbestos as carcinogenic to humans.

Exposure to asbestos can be occupational, as in the case of factory workers, or environmental, in the case of individuals living or working in communities near asbestos mining plants or buildings with a high asbestos content. Workers are mainly at risk of inhaling asbestos fibres when processing materials such as talc or vermiculite that might be contaminated [61], and studies have established that they have a much higher incidence of malignancies such as mesothelioma and lung cancers, among other diseases [61, 64, 65]. The risk for lung cancer can be modified by other factors such as smoking and persistent inflammation (reviewed in [62]). There is also evidence that environmentally-exposed individuals can develop mesothelioma, albeit at much lower frequencies [64, 66, 67].

1.3.1.2 Alcoholic beverages

Alcoholic beverages are pervasive in our society. In surveys conducted in 2012, more than half of people aged 18 or over reported drinking alcohol in the last month in the United States (US) [68], and more than half of adults reported consuming alcohol in the last week in the UK [69]. More than a century ago, French pathologists noticed an increase in the incidence of oesophageal cancer in absinthe drinkers [70], and since then, numerous studies have been conducted that show an association between alcohol consumption and cancers of the oral cavity, pharynx, larynx and oesophagus (reviewed in [71]). Finally, in 1988, the IARC concluded there was sufficient evidence for the carcinogenicity of alcoholic beverages in humans, and that malignant tumours of the oral cavity, larynx, pharynx, oesophagus and liver can be causally associated with their consumption [72].

Ethanol metabolism has been identified as the major mechanism by which alcoholic beverages can cause cancer. Ethanol is metabolised to acetaldehyde by the alcohol dehydrogenase (ADH) enzymes and cytochrome P450, family 2, subfamily E, polypeptide 1 (CYP2E1), and is further converted to acetate by aldehyde dehydrogenases [73]. This in turn causes genotoxic and non-genotoxic effects: Acetaldehyde can bind to DNA and form stable adducts [74], interfering with DNA synthesis and repair, and its production can generate ROS and an activation of other carcinogens present in the environment such as tobacco smoke (reviewed in [73]). Individuals that carry a certain polymorphism in one of the aldehyde dehydrogenase genes, *ALDH2*, have a much less efficient enzyme and thus accumulate higher levels of acetaldehyde, and have therefore been identified as a high-risk group for the development of oesophageal cancer (reviewed in [73]). Other mechanisms by which alcohol may cause cancer relate to an increase in the levels of hormones such as oestrogen, which increase breast cancer risk [75], an increase in the permeability of mouth and throat cells to other carcinogens [76], and a decrease in folate levels [77].

The proportion of all cancer cases and deaths attributable to alcohol intake worldwide has been estimated to be 3.6% and 3.5%, respectively [78]. Therefore, diminishing alcohol consumption has been highlighted as an important and generally underemphasised strategy for cancer prevention when compared to other preventive programs focusing on tobacco usage and genetic screening, among others [79].

1.3.1.3 Coal tar and soot

As discussed previously, coal tar and soot present in chimneys were the first occupational agents to be associated with an increase in cancer risk. In the years since Pott described the link between occupational exposure to coal and scrotal cancer [54], many other reports describing an association between skin cancers and soot were subsequently published (reviewed in [80]). As a result of his observations, Pott recommended that chimney sweeps take a daily bath, and his suggestion was so successful that it managed to greatly reduce the incidence of scrotal cancer in workers that followed his advice compared to those that did not [81, 82]. This evidence, and numerous studies that were conducted subsequently, led the IARC to conclude in 1973 that coal and soot are carcinogenic to humans.

The mechanism of carcinogenicity of coal has been studied extensively. In 1936, Alfred Winterstein reported that he could isolate benzo[*a*]pyrene (BP) from coal tar boiling at high temperature (Winterstein A. Festschrift. Basel: Emil Barell; 1936; quoted by [83]), and subsequent experiments showed that BP applied on mouse skin, even at low doses, was highly carcinogenic and generally produced squamous cell carcinomas [84]. Other researchers were later able to dissect the metabolism of BP: Upon exposure, BP is converted to BP-7,8 epoxide and then further hydrolysed and metabolised, in steps involving the microsomal epoxide hydrolase and cytochrome P450 enzymes, to the carcinogenic diol-epoxide 2 (DE 2) [85]. This species is highly reactive and can bind DNA [85, 86], and it has been shown by several studies that it can cause genetic mutations in both prokaryotes and eukaryotes (reviewed in [87]).

Exposure to coal tar and soot has been associated principally with higher incidences of lung and skin cancer, but limited evidence also exists linking it to bladder cancer [57].

1.3.1.4 Tobacco smoke

As early as 1930, tobacco smoking was proposed to be the underlying cause of the phenomenal rise in lung cancer that was seen after the end of the First World War, a suspicion that was confirmed over the next 20 years as more studies were carried out investigating lung cancer aetiology [88–90]. In numerous experimental systems in which mice, rats, Syrian hamsters, rabbits and dogs were exposed to mixtures of cigarette smoke and air, significant increases in lung tumours as well as emphysema and other cancers were observed (reviewed in [91]). Additionally, cigarette-smoke condensate was found to increase the incidence of tumours when applied to mouse and rabbit skin, as well as acting as a potential co-carcinogen when used in conjunction with other agents (reviewed in [91]). Some of these studies, in conjunction with others, were considered by the IARC when it classified tobacco smoking as a confirmed carcinogen in 1986 [57].

Cigarette smoke contains more than 60 well-known carcinogens, most of which require metabolic activation upon exposure, a step that generally involves the cytochrome P450 enzymes [92]. In addition to BP, which has also been found as a carcinogenic agent in coal tar, other compounds found in tobacco smoke that can form DNA adducts are N-nitrosodimethylamine, N´-nitrosonornicotine and ethylene oxide, among others [92]. Accordingly, in human cell line experiments, hotspots for the formation of BP-DNA adducts have been found in genes important in cancer progression such as Kirsten rat sarcoma (RAS) viral oncogene homolog (*KRAS*) and tumour protein p53 (*TP53*) [93, 94]. In the case of *KRAS*, adducts were found to preferentially affect important codons for kinase activation, and in the case of *TP53*, they matched the mutational signature observed in human lung cancers, implicating the genotoxicity of these carcinogens as a highly likely contributor to their aetiology. Other studies reporting higher levels of adduct formation resulting from exposure to other tobacco carcinogens support the causal relationship between smoking and cancer (reviewed in [92]).

Not only lung cancer has been causally associated with tobacco smoking. According to the IARC, there is sufficient evidence to support a causal role for smoking in the aetiology of many different types of malignancies, including myeloid leukaemia, and colorectal, liver, oesophageal, stomach and bladder cancers, among others. Second-hand smoking has also been linked to lung, laryngeal and pharyngeal cancers (reviewed in [57]).

Tobacco smoking has been identified as the main cause of cancer-related deaths worldwide, being responsible for approximately 16% of all cancers in developed countries (Stewart BW, Kleihues P. World Cancer Report. Lyon, France: IARCPress; 2003; quoted by [95]), and thus has influenced policymaking (for examples, see [96, 97]) and sparked numerous campaigns around the world to reduce its prevalence (for examples, see refs. [98, 99]).

1.3.1.5 Solar radiation

The link between exposure to solar radiation and skin cancer was suspected as far back as 1894, when Paul Unna described what he referred to as ‘seaman’s skin’: chronically sun-exposed skin that presented hyperkeratosis and squamous cell carcinomas (Unna, PG. Die Histopathologie der Hautkrankheiten. Berlin: August Hirschwald; 1894; quoted by [100, 101]). Experiments in diverse model organisms over the next 70 years helped clarify the carcinogenic effects of ultraviolet (UV) light, one of the radiation types that the Sun emits (reviewed in [100]). In 1992, the IARC analysed data from numerous patients and animal studies and concluded that there was sufficient evidence to support a causal role for solar radiation in the development of cutaneous malignant melanoma and non-melanocytic skin cancer in humans [102].

Solar radiation is the combination of UV radiation and visible light that reaches the Earth's surface [103]. Solar UV radiation that reaches Earth is mainly composed of two wavelengths: UVA, which comprises about 95%, and UVB, which contributes the rest [103]. UV radiation seems to have both genotoxic and non-genotoxic effects: it can create cyclobutane pyrimidine dimers (CPDs) and (6,4)-photoproducts (6-4PPs) between adjacent pyrimidine bases in the same DNA strand [104] (Fig. 1.1), and it can also cause other systemic effects such as activation of the cell survival nuclear factor (NF)-$\varkappa$B signalling pathway [105] and generation of ROS [106], which appear to be independent of DNA damage. Although thymine bases in CPDs are not especially mutagenic [107], cytosine bases in these structures are chemically unstable and usually deaminate to generate uracil bases, which may then be incorrectly paired with adenine during replication inducing a C-to-T transition. The same outcome happens if the CPD occurs in a 5-methylcytosine (mC)-containing pyrimidine site, as is frequently the case for solar-induced DNA lesions, because they can rapidly deaminate to form thymine (reviewed in [108]). Therefore, the induction of C-to-T transitions at dipyrimidine sites is the signature consequence resulting from UV radiation.

The incidence rate for malignant melanoma, one of the cancer types for which UV radiation is an important risk factor, has more than quadrupled in the last 30 years in the UK [110]. This dramatic increase has been attributed to lifestyle factors such as the popularity of holidays in lower latitudes and the surge in sunbed usage [111], and has led to increasing numbers of campaigns targeting unsafe exposure to UV radiation (for examples, see [112–114]).

UV radiation, as one of the major causes of malignant melanoma, is revisited in Section 1.6.

1.3.1.6 Other carcinogenic agents

In addition to the carcinogens discussed above, there are about 100 other agents that have been classified in this category by the IARC. These can be chemicals (*e.g.*, aflatoxins or benzene), occupations (*e.g.* aluminium production or painting), metals (*e.g.* arsenic

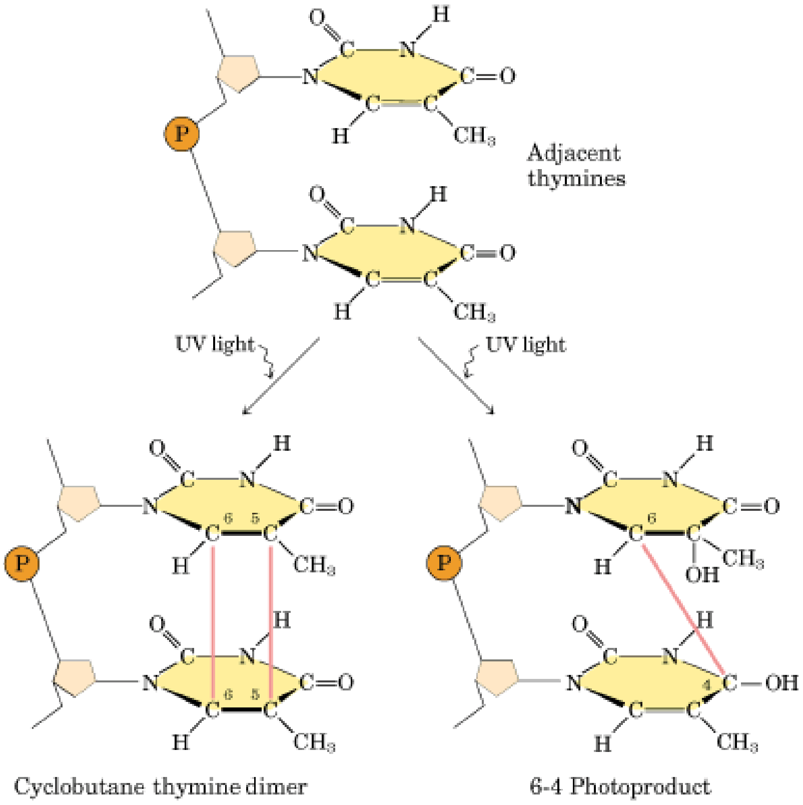

Figure 1.1: **Types of DNA damage caused by UV radiation.** An overview of the structure changes induced by UV radiation in adjacent pyrimidine bases, in this example, thymines are shown. Phosphate and sugar groups forming the DNA backbone are shown in orange and light orange, respectively. In the case of CPDs, the double bonds between carbons 5 and 6 become saturated, forming a four-membered ring. In 6-4PPs, a bond is formed between carbons 6 and 4 of two adjacent pyrimidines. Figure taken from ref. [109]. Copyright 2013 by W.H. Freeman and Company. Used with permission of the publisher.

or nickel compounds), dusts and fibres (*e.g.* leather or wood dust), radiation sources (*e.g.* X or gamma radiation), biological agents (*e.g.* hepatitis B or C or HIV viruses), personal habits (*e.g.* consuming areca nuts or Chinese-style salted fish), or drugs (*e.g.* cyclosporine or tamoxifen) [57]. The value of identifying these carcinogenic agents lies in actionable measures, such as educational campaigns and prevention programmes, aimed at diminishing exposure to them. For example, encouraging results can be seen from efforts of tobacco control campaigns and consequential reductions in lung cancer incidence and mortality in the US [115, 116].

Exposure to carcinogens is an established cause of malignancy, but genetic predisposition also plays an important role in cancer aetiology. In the next subsection, I discuss the genetic factors that have been identified in cancer predisposition, the importance of familial studies and the biological insight we have gained from such analyses.

1.3.2 Genetic risk factors and their influence on cancer development

In 1866, the same year in which Mendel published the basic principles of heredity, French surgeon Paul Broca published perhaps what is the first report on the existence of familial predisposition to malignancy. He constructed the pedigree of his wife's family, who suffered from early-onset breast cancer, and found fifteen cases of breast, liver and uterine cancer spread across four generations of women (Broca P. Traité des tumeurs. Paris: P. Asselin; 1866; pedigree reproduced by [117]) (Fig. 1.2). Then, in 1913, the American pathologist Aldred Warthin published a study of 3,600 cancer cases that had been examined in the University of Michigan in the period from 1895 to 1913. Of all the carcinoma cases where detailed family history was available, he determined that as many as 15% had a familial history of the disease, and thus supported the idea that predisposition to malignancy could be inherited [118]. One of the families in his original study has been followed now for more than a century, and the causes for its cancer-prone phenotype have now been defined molecularly, contributing enormously to our knowledge of the biological processes underlying cancer development [119]. With his work, Warthin provided solid evidence for the importance of genetic factors in cancer aetiology, and thus has been referred to as "the father of cancer genetics" [120].

Studies of other families, in which many different types of cancer seemed to cluster, followed Broca's and Warthin's reports (reviewed in [121]), and epidemiological studies carried out several decades later supported an important role for genetic predisposition to cancer [122]. However, with the exception of these rare malignancies, the origin of common cancers was still considered to have predominantly environmental causes throughout almost all of the twentieth century [123].

Germline transmission of predisposition to certain malignancies, such as the childhood cancer retinoblastoma and syndromes of multiple endocrine neoplasms, had been recognised in the 1980s [124–126]. Several years earlier, scientists had started to regard cancer as a multi-stage process, having a fixed number of rate-limiting steps that should be overcome before malignancy could progress [121]. The American geneticist Alfred Knudson, after meticulously studying 48 cases of retinoblastoma, formulated in 1971 his groundbreaking two-hit hypothesis: he postulated that in the hereditary form of the

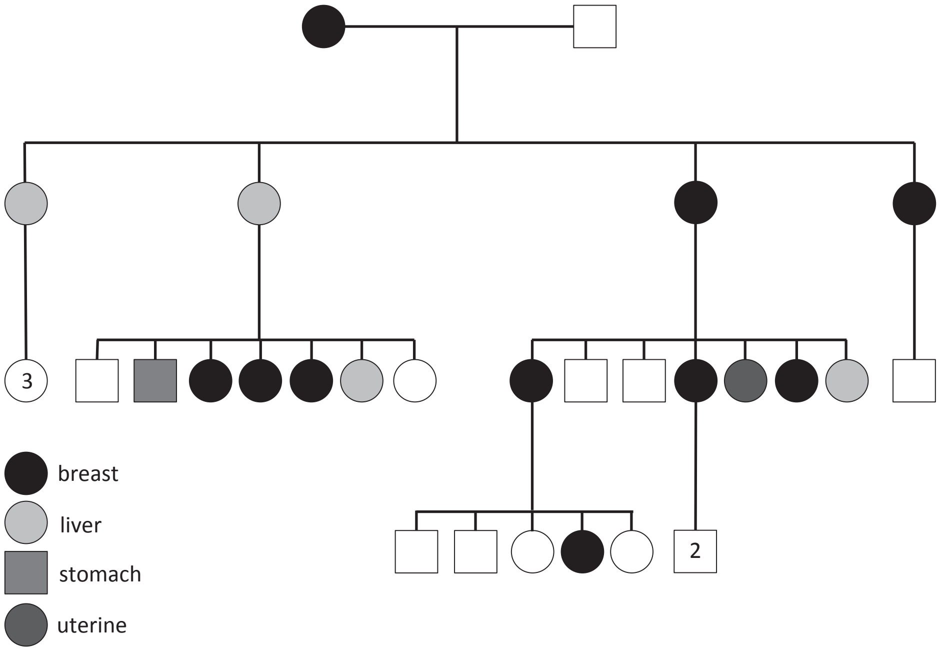

Figure 1.2: **Paul Broca's pedigree evidences familial predisposition to cancer.** This pedigree, possibly the first report on familial cancer, shows fifteen cases of different malignancies spread across four generations of women. Circles represent females, squares represent males. Numbers in symbols refer to numbers of unaffected offspring. Pedigree adapted from ref. [117].

disease, one mutation was inherited while the other happened somatically, whereas in the sporadic form both happened somatically [127]. Knudson's hypothesis was confirmed in 1986 with the successful isolation of the gene responsible for retinoblastoma [128], now known as *RB1*, and the characterisation of its aberrations in retinoblastoma DNA samples.

Since these landmark discoveries, many other genes have been discovered that, when mutated, predispose their carrier to the development of malignancies. Almost all of these have been categorised as tumour suppressor genes, and echo Knudson's two-hit hypothesis in the mechanism by which they contribute to cancer development [129]. Tumour-suppressor genes normally have functions inhibiting cellular growth or genomic instability, and thus inactivation of both copies of the gene in a cell leads to the loss of these protective functions. Genes that contribute positively to cell proliferation can also be affected by germline mutations, although this happens much less frequently. These genes are referred to as proto-oncogenes, and inherited mutations render the encoded proteins constitutively active [129], so a single copy of the gene harbouring the mutation is sufficient to promote tumourigenesis. Because of their modes of action, tumour suppressor and proto-oncogenes show distinct patterns of germline and somatic alterations, with tumour suppressors displaying truncating or disruptive mutations distributed throughout the encoded protein and proto-oncogenes displaying clusters of mutations in key residues that activate them.

The discoveries mentioned above, along with other genetic models in epidemiological studies of cancer clustering, supported the hypothesis that a substantial proportion of cancer incidence might be accounted for by genetic effects [123], and thus obligated researchers to revisit the theory for the origins of common cancer. The importance of genetic effects in the aetiology of common cancers was further supported, in 1990, by the discovery of a genomic region linked to early-onset familial breast cancer [130]. Four years later, breast cancer 1 (*BRCA1*), the gene responsible for the disease, was identified by positional cloning methods [131]. It is now recognised that between 5 and 10% of all cancers are inherited, and that the interaction between genetic factors, or gene-environment interactions, might play an important role in a further 15% [132]. These numbers give us an idea of the importance of genetic make-up in cancer aetiology.

However, identifying cancer susceptibility genes is not an easy task. Although a strong genetic component can be suspected in families where there are multiple affected individuals across generations or individuals that present with multiple primary cancers or with an early age of onset, various confounding factors can greatly complicate its isolation. For example, the allele might show incomplete penetrance, in which case some carriers will not be affected by the disease. It can also be the case that sporadic forms of the disease, termed phenocopies, arise in non-carrier family members. In these situations, a genotype comparison between affected and unaffected family members will not be informative [133]. Despite these and other difficulties, such as small family size or uncertain family history when assessing affected pedigrees [133], hundreds of genomic regions have been found to influence cancer susceptibility over the last 30 years.

The advent of next-generation sequencing (NGS) technologies and other analysis methodologies has facilitated the discovery of cancer-predisposing genes and their relative contributions to disease development. Over 100 high-penetrance genes have been identified to date [129], and hundreds of other genomic regions have been associated with cancer risk [134], deepening our understanding of the biological processes that underlie normal cell proliferation and maintenance. In this subsection, I review the different models of cancer risk inheritance that have been proposed, the most established tumour suppressor genes and proto-oncogenes and the technologies that have been developed in their identification.

1.3.2.1 Mendelian inheritance of elevated cancer risk

Germline variants with strong cancer-predisposing effects can be inherited in a simple Mendelian fashion, and generally underlie an early-onset of cancer and the appearance of multiple primaries [135]. High-penetrance mutations in cancer-predisposing genes can show different modes of transmission, with the majority displaying autosomal dominant inheritance but some others requiring germinal bi-allelic inactivation to promote cancer progression [136]. Syndromes that show autosomal dominant inheritance in families are generally recessive at the molecular level, with one inactivated allele being transmitted germinally from generation to generation and the other one being inactivated by subsequent somatic mutations [135].

In general, rare germline variants with large effects (for example, underlying lethal childhood cancers) tend to be purged rapidly by natural selection, because the carriers generally do not reproduce [135]. In contrast, variants that predispose to later-onset cancer, such as those in *BRCA1*, are removed more slowly from the population's gene pool. Therefore, rare alleles with large effects will tend to be younger than common alleles with smaller effects, and will generally not be associated with any polymorphisms in the genomic vicinity [135]. This distinction is important because it impacts the methodologies used for gene identification, with whole-genome or exome sequencing (WGS and WES, respectively) being better suited for the identification of rare alleles with large effects and genome-wide association studies (GWAS) being more useful in the identification of common alleles with weak effects. The genes described in this Subsection belong to the first category, and have been identified by linkage and candidate gene studies.

1.3.2.1.1 *RB1* and retinoblastoma: The prototypical example of an inherited cancer-predisposition syndrome The gene underlying almost all cases of childhood retinoblastoma, *RB1,* was identified in 1986 by Stephen Friend and colleagues [128] following Knudson's hypothesis [127]. Accordingly, the majority of inherited cases present with bilateral retinoblastoma, whereas this proportion is much lower among sporadic cases [137], evidencing the different numbers of rate-limiting mutational steps required for cancer development in mutation carriers versus non-carriers. *RB1* is the prototypical tumour suppressor gene, displaying about 90% penetrance and showing autosomal dominant inheritance in families [137]. It has since been found mutated in other cancers, such as oesteosarcoma and small-cell lung carcinoma [138].

The biology of *RB1* has been studied extensively in the three decades since its discovery. Shortly after its isolation, two research teams found that its gene product was the target of two virus-encoded oncogenes, adenovirus E1A and SV40 large T antigen [139, 140], implicating that tumour-promoting and suppressor genes could act in the same biological pathway. Further studies established that the RB1 protein acted as an inhibitor of cell proliferation given that its over-expression caused cell-cycle arrest at the G1 phase, and, conversely, its deficiency caused fast G1 progression (reviewed in [138]). In the following years, researchers reported that RB1 was able to exert its molecular functions by inhibiting members of the E2F family of transcription factors and cyclin-dependent kinases (CDKs), which promote cell cycle entry (reviewed in [141]).

Retinoblastoma is a rare cancer, with an incidence of only about 12 cases per million young children in the US, similar to several European countries [142]. Although more than 500 distinct mutations in the *RB1* gene have been described [143], novel mutations are constantly being reported as most of the germline mutations underlying hereditary retinoblastoma arise *de novo* [144, 145].

1.3.2.1.2 *BRCA1* and *BRCA2*: High risk for breast and ovarian cancer

The first report of breast cancer might be a 1500 BC Egyptian papyrus which described tumours of the breast and their palliative treatment, although this is disputed [146, 147]. In any case, it is clear that breast cancer has been recognised for centuries. It is by far the most common cancer in women, and indeed, the most prevalent cancer in the majority of countries worldwide [148]. Broca's pedigree (Fig. 1.2) showed breast cancer segregating as an autosomal dominant trait, and accordingly, modern analyses have estimated a 96% chance that Mrs. Broca harboured a *BRCA1* or *BRCA2* mutation, and that the other cancers seen in the family represent metastatic disease [117].

Familial studies in the mid-1990s led to the isolation of *BRCA1* and *BRCA2* [131, 149], the two genes underlying the great majority of multi-case breast cancer families [150]. A meta-analysis considering several estimates of the penetrance of mutations in these genes reported that the breast and ovarian cancer risk was around 60% and 40% for *BRCA1*, and 50% and 18% for *BRCA2* mutation carriers, respectively [151]. However, although these genes confer high cancer risks and might explain a large proportion of multiple-case breast cancer families, they only account for about 2-6% of all breast cancer cases [152]. This is the reason why other genes and susceptibility regions have been intensively searched for using other methods, such as GWAS and WES.

The BRCA proteins interact with several regulatory proteins, and have an important role in the transcriptional regulation of the DNA damage response and thus in its associated cell cycle checkpoints. Upon DNA damage, BRCA1 is phosphorylated by several kinases, including ataxia telangiectasia-mutated (ATM) and ATM-related (ATR), and participates in double-strand (ds) DNA break (DSB) repair (reviewed in [153]). BRCA1 has also been shown to transcriptionally co-activate NF-$\varkappa$B-regulated genes, which might contribute to maintain genomic stability [154]. It is thought that *BRCA1*-haploinsufficient cells need to accumulate further genomic damage before inactivating the wild-type allele, because although tumours display BRCA1 loss-of-function, its bi-allelic inactivation in normal cells results in cellular lethality [155]. The mechanism by which *BRCA1* haploinsufficiency predisposes to breast carcinogenesis might therefore represent a modification of the two-hit hypothesis, with an additional third hit (genomic instability) required before the inactivation of the wild-type allele [155].

BRCA2 may also play an important role in DSB repair, as shown by the chromosomal breakage and abnormal mitotic exchanges observed in *BRCA2*-deficient cells (reviewed in [153]). Through its binding to the RAD51 recombinase, its synergistic interactions with P53 and its association with CDKs, BRCA2 promotes genomic stability and the response to radiation-induced DNA damage [156, 157]. Loss of the *BRCA2* wild-type allele is commonly seen in tumours from germline mutation carriers [158, 159].

Other methodologies have been used to find important genes in breast cancer susceptibility. For example, WES has identified germline mutations in the DNA repair genes Fanconi anaemia complementation group C (*FANCC*) and Bloom syndrome (*BLM*) [160], and GWAS has found a few other low-penetrance loci influencing breast cancer risk (reviewed in [161]).

1.3.2.1.3 DNA mismatch repair genes, *APC* and familial colon cancer syndromes One of the families in Warthin's original report, Family G, had several members that presented with colon, stomach and abdominal cancers across two generations [118]. This family has been followed regularly, initially by Warthin and his colleagues and subsequently by other researchers, for more than a century now (reviewed in [162]). Their cancer-prone phenotype is now known as Lynch syndrome (in honour of one of the physicians that has been studying it for several decades) or hereditary nonpolyposis colon cancer [132]. The molecular basis for their phenotype remained elusive throughout the twentieth century, but in 2000, Hai Yan and colleagues tested one member of this family and identified a T to G transversion at a splice acceptor site in the mutS homolog 2

(*MSH2*) gene [163], encoding a protein important for DNA mismatch repair [164]. Since then, mutations in any of the five DNA mismatch repair genes (*MSH2, MSH6*, mutL homolog 1 [*MLH1*], PMS1 postmeiotic segregation increased 1 [*PMS1*] and *PMS2*) have been found to underlie this syndrome [132]. Because of these defects, a type of DNA replication error called microsatellite instability is the mutational hallmark of colorectal tumours in Lynch syndrome [132]. It presents as an autosomal dominant disease with early onset, and has a typically high penetrance (~85%) [132, 164]. The development of cancer in these individuals follows the two-hit hypothesis, with the wild-type allele often found inactivated by deletion, somatic mutations or epigenetic alterations [165].

The other major subtype of hereditary colorectal cancer is known as familial adenomatous polyposis (FAP). Individuals with this disease, in contrast to Lynch syndrome, present at a young age with hundreds to thousands of adenomatous polyps in the colon and rectum, and have a greatly increased risk of developing colorectal cancer [132]. The gene underlying this syndrome was discovered by linkage analysis in 1987 [166], and is now known as adenomatous polyposis coli (*APC*). The APC protein behaves as a tumour suppressor, having a myriad of important cellular functions in cell cycle control, migration, differentiation and apoptosis that it exerts through its modulation of the integration/wingless (Wnt) signalling pathway [132]. Abrogation of APC function is not only important in tumours from individuals with hereditary disease, as mutations in *APC* are the earliest genetic alteration found in sporadic cases as well [167, 168]. Malignancy then ensues with the acquisition of additional somatic mutations in genes such as *KRAS* and *TP53* [168].

The susceptibility to colon cancer has also been linked to other genes such as mutY homolog (*MUTYH*), axin 2 (*AXIN2*), mothers against decapentaplegic homolog 4 (*SMAD4*), bone morphogenetic protein receptor, type IA (*BMPR1A*) and phosphatase and tensin homolog (*PTEN*), among others (reviewed in [169]). GWAS have also uncovered several different loci associated with the risk of colorectal cancer and adenomas (reviewed in [161]).

1.3.2.1.4 *TP53* and the multi-cancer Li-Fraumeni syndrome In 1969, Frederick Li and Joseph Fraumeni reported four families in which children and young adults presented with a high frequency of different types of malignancies, including soft-tissue sarcomas and breast cancer, and often with multiple primaries [170]. Some years later, two groups identified germline mutations in the tumour suppressor *TP53* in six families presenting with this syndrome [171, 172]. Since then, many other mutations in *TP53*, both in coding and non-coding regions, have been described. These mutations are sometimes associated with inhibition of growth arrest, apoptosis, or transcriptional activation, and their activity can be influenced by genetic background, accumulation of further genetic alterations or other epigenetic or environmental factors (reviewed in [173]). Although the germline alterations result in *TP53* loss of function, experiments made *in vitro* suggest that missense mutations harbour oncogenic potential (reviewed in [173]). This observation could help explain why only a fraction of tumours present loss of the wild type allele (following Knudson's hypothesis) [174].

TP53 encodes an important tumour suppressor gene that integrates signals from multiple pathways and controls cell cycle progression and fate (reviewed in [175]). Mutations in this gene have been found in a fraction of virtually every sporadically-occurring malignancy [173], and given its pivotal role in maintaining genomic stability, has been referred to as "the guardian of the genome" [176].

Because not all families with Li-Fraumeni syndrome have segregating mutations in *TP53*, other genes have been searched in connection with the phenotype. Genes in the P53 pathway, such as checkpoint kinase 1 (*CHEK1*), *CHEK2,* cyclin-dependent kinase inhibitor 2A (*CDKN2A*) and *PTEN* have all been considered as candidates, however, their involvement in the aetiology of the disease remains unconvincing (reviewed in [176]).

1.3.2.1.5 *RET*: A proto-oncogene activated by germline mutations Multiple endocrine neoplasia (MEN) syndromes are a group of diseases in which endocrine glands present with two or more tumours, and in which ectopic hormone production is common (reviewed in [124]). MEN syndromes are generally divided into three main types: MEN1, MEN2 and MEN4, which are characterised by different combinations of endocrine tumours [177]. MEN2 is further divided in MEN2A, MEN2B and familial medullary thyroid carcinoma (FMTC), that have medullary thyroid carcinoma as their principal clinical characteristic [178]. In 1993, two groups studying MEN2A and FMTC families discovered germline mutations in a proto-oncogene called rearranged during transfection (*RET*), and most of these mutations affected the same cysteine residue in the protein [179, 180]. A year later, MEN2B was also attributed to germline mutations in the same gene [181]. Since then, numerous studies have shown that >95% of MEN2A patients carry mutations affecting any one of 6 different cysteine residues that result in constitutive RET signalling, and that the MEN2B phenotype is almost exclusively associated with two different methionine and alanine substitutions that result in protein conformational changes (reviewed in [182]). In contrast to tumour suppressors, which normally require inactivation of both copies of the gene before tumourigenesis can progress, proto-oncogenes require mutation of only one copy, as it is enough to render them constitutively active. Since the discovery that certain cancer-prone families carry activating mutations in *RET*, only a handful of other genes have been found that predispose to cancer through gain-of-function mutations [129].

RET encodes a single-pass transmembrane receptor tyrosine kinase (RTK) that is highly evolutionarily conserved and participates in spermatogonial stem cell maintenance, kidney induction, neural crest cell migration among other biological processes [182]. Upon co-receptor- and ligand-binding, RET autophosphorylates and activates the mitogen-activated protein kinase (MAPK)-RAS-rapidly accelerated fibrosarcoma (RAF) and the phosphatidylinositol-3 kinase (PI3K) pathways [183], providing a biological explanation for the observed cancer-prone phenotype upon its constitutive activation. *RET* is notorious not only for being the first proto-oncogene discovered to be implicated in a cancer predisposition syndrome, but also because variable penetrance has been observed depending on the amino acid affected in germline mutation carriers (reviewed in [132]). Interestingly, *RET* germline loss-of-function mutations have also been observed in patients with Hirschsprung's disease, which is a common congenital malformation characterised by functional intestinal blockage [132]. A small number of patients have been observed with both Hirschsprung's and MEN2, as they seem to carry mutations that lead both to a decrease in *RET* expression and constitutive kinase activation (reviewed in [184]).

RET has also been implicated more broadly in diverse sporadic tumour types, such as pancreatic ductal carcinoma, invasive breast cancer and acute myeloid leukaemia (AML), among others (reviewed in [182]). Because of its mode of action and its possible implication in several cancer types, RET constitutes an important therapeutic target. Several small-molecule RTK inhibitors that target RET, such as vandetanib and cabozantinib, have been tested on patients with thyroid cancers, with varying results depending on *RET* germline status and the identity of the mutations (reviewed in [182]).

1.3.2.1.6 *CDKN2A* and *CDK4*: Melanoma risk increased by mutations in both a tumour suppressor and a proto-oncogene By 1994, the tumour suppressor CDK4 inhibitor (INK4A, also called p16, one of the proteins encoded by the *CDKN2A* locus) had emerged as an important player in human cancer development. INK4A is an inhibitor of CDK4 and CDK6, and thus prevents G1/S transition by inhibiting RB1 hyperphosphorylation (reviewed in [185]) (Fig. 1.3a,b). Deletions in this gene had been originally reported at high frequencies in a number of cell lines, including those derived from melanoma, lung, bladder, leukaemia and brain cancers, among other malignancies [186, 187]. Only two years before, a potential melanoma predisposition locus, named MLM, had been mapped via linkage studies in melanoma-prone kindreds to the same chromosomal location, 9p21 [188]. Several studies then ensued that demonstrated that MLM and INK4A were the same gene and delineated some of the predisposing mutations, which appeared to be missense in its majority [189–192]. These mutations were later shown to impair the ability of INK4A to inhibit CDK4 and CDK6, thus providing a rationale for the cancer-predisposing syndrome observed in carriers [193] (Fig. 1.3c).

However, not all the families that showed disease linkage to chromosome 9p21 had mutations in INK4A, and furthermore, not all melanoma-prone families showed linkage to 9p21. *CDKN2A* not only encodes INK4A, but also the structurally unrelated alternate reading frame (ARF) (Fig. 1.3a). ARF stabilises P53 signalling and thus suppresses growth by blocking mouse double minute 2 homolog (MDM2), an important negative regulator of P53 (reviewed in [185]) (Fig. 1.3b). Although almost half of the *CDKN2A* germline alterations in melanoma-prone kindreds affect the shared exon 2 and thus disrupt both gene products (Fig. 1.3c), mutations affecting only ARF have been described [196, 197], and a role for this tumour suppressor in melanoma susceptibility is further supported by studies in mice (reviewed in [185]). Both INK4A and ARF behave like classic tumour suppressors, with loss of the wild type allele frequently seen in tumour cell lines and, some times, in uncultured tumours (reviewed in [198]). These observations evidence the importance of both products of the *CDKN2A* locus in melanoma susceptibility.

In some of the families that did not show linkage to the *CDKN2A* locus, germline mutations were found in the gene encoding one of the INK4A-binding partners, *CDK4*, two years later [199]. Mutations were found to cluster in the arginine at position 24, displaying the properties of a proto-oncogene. Arginine 24 was shown to be important for INK4A binding, but not for its kinase activity [199], and thus this mutation results in constitutional CDK4 activation by preventing INK4A-mediated inhibition (Fig 1.3d). These observations established that the functions of both a tumour suppressor and a proto-oncogene, that cooperate in the same pathway, can be altered by germline mutations that result in the same phenotype.

Genetic susceptibility to familial melanoma, as the subject of this dissertation, is

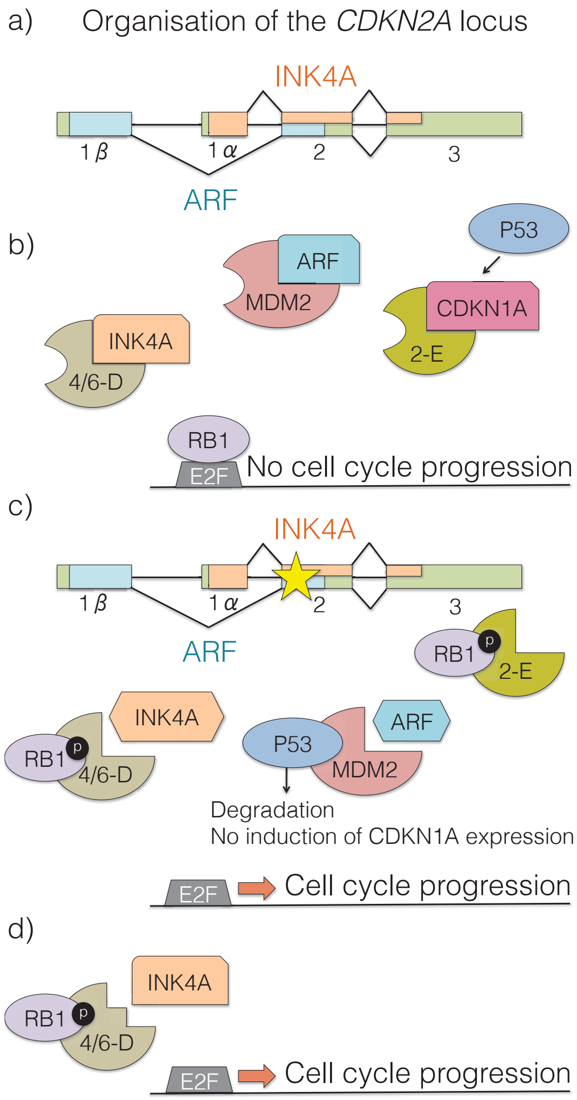

Figure 1.3: **Known melanoma susceptibility genes and their functions.** a) Organisation of the *CDKN2A* locus and its alternatively spliced products INK4A and ARF. Both proteins are structurally unrelated, but they share exon 2. The green rectangles depict untranslated regions (UTRs). b) Normal physiological function of the INK4A and ARF tumour suppressors. INK4A prevents the Cyclin D-CDK4/6 (depicted as '4/6-D') complex from phosphorylating RB1, and thus E2F-mediated cell cycle progression is inhibited. CDKN1A (also known as p21), whose expression is induced by P53, also inhibits RB1 phosphorylation by binding the Cyclin E-CDK2 complex (shown as '2-E'). ARF prevents MDM2-mediated ubiquitylation and degradation of P53, and thus contributes to its stability. c) Effect on the cell cycle of a disruptive mutation affecting INK4A and ARF. Mutations that affect exon 2 (depicted by a star) alter the function of both INK4A and ARF. If these proteins are incapable of binding their targets, P53 degradation ensues and hyperphosphorylation of RB1 (symbolised by 'p') occurs via direct interaction with the cyclin complexes, allowing progression of the cell cycle. d) CDK4 is a proto-oncogene that has been found mutated in a subset of melanoma-prone families. Mutations affect the INK4A binding site, but not its kinase activity, so CDK4 constitutively phosphorylates RB1. Adapted from refs. [194, 195].

further discussed in Section 1.6.

1.3.2.1.7 Other high-penetrance cancer-predisposing mutations and autosomal recessive inheritance of cancer predisposition Many other cancer-predisposing genes have been discovered by diverse methodologies such as linkage studies, candidate gene analyses and genome-wide mutation analyses (reviewed in [129]). These genes have different inheritance patterns, penetrances, mechanisms of disease (requiring biallelic inactivation or showing haploinsufficiency or dominant-negative effects, or constitutive activation in the case of proto-oncogenes) and participate in a diverse array of biological processes, such as cell cycle regulation and DNA repair (reviewed in [129]). Other examples of autosomal-dominant inheritance of cancer predisposition are Cowden syndrome, caused by germline mutations in *PTEN*, von Hippel-Lindau disease, caused by alterations in the eponymous *VHL*, and juvenile polyposis, caused by mutations in the *SMAD4* and *BMPR1A* genes (reviewed in [132]).

There are other cancer predisposition syndromes that are inherited in an autosomal-recessive manner. Examples of these are ataxia telangiectasia (A-T), caused by germline mutations in *ATM*, Nijmegen breakage syndrome (NBS), caused by mutations in nibrin (*NBN*), and the Bloom (BS) and Werner (WS) syndromes, caused by mutations in the DNA helicase genes *BLM* and *WRN*, respectively. All of these syndromes are characterised by telomere abnormalities and genomic instability (reviewed in [200]).

1.3.2.2 Polygenic model of cancer risk inheritance

Studies of cancer-prone families carrying deleterious alleles of some of the genes mentioned in Section 1.3.2.1 suggested that additional genetic components might be involved in the aetiology of their cancers, albeit with lower or modifier effects. For example, although specific mutations in *RET* can distinguish MEN2A and FMTC from MEN2B, they cannot distinguish between MEN2A and FMTC, suggesting that other genes might play a role in predisposition to these syndromes [133, 182]. Additionally, only about a fifth of the total clustering of familial breast cancer is observed in families carrying a predisposing *BRCA1* or *BRCA2* mutation, but diverse twin and familial studies suggest that genetic components are significant in the aetiology of the remaining families [123]. In fact, a polygenic risk model was found to best describe the distribution of breast cancer not due to *BRCA1* or *BRCA2* mutations in a population-based series, with the quintile most at risk showing 40-fold higher risk than the quintile least at risk [201]. Similar conclusions have been reached in other studies (reviewed in [123]) (Fig. 1.4).

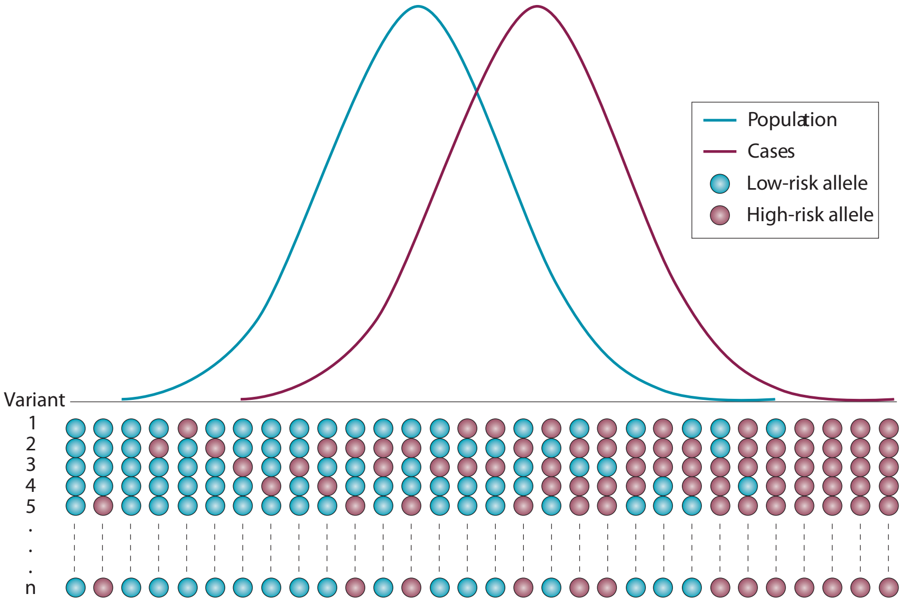

Figure 1.4: **Polygenic model of the distribution of cancer risk in the population and in individual cases.** Models of cancer risk that take into account several predisposition loci with small effects have been found to fit observed population-based data. Cases show an enrichment of high-risk alleles when compared to the rest of the population. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer ([122]), copyright (2010).

Two models for polygenic predisposition to cancer have been formulated. The first of these, comprising common low-penetrance alleles predisposing to common cancers, is practically impossible to explore using linkage or candidate gene approaches, and thus GWAS have been carried out to compare allele frequencies in cases and controls. These common alleles are thought to have arisen once during evolution, and thus can be marked and genotyped in large numbers of cases and controls with high-density SNP arrays [123]. GWAS have been successful at pinpointing more than 200 low-penetrance genomic regions that influence cancer risk (reviewed in [134]), but identifying the causal alterations within these regions remains a daunting task. The second model involves rare alleles with moderate effects that are thought to have arisen recently or multiple times during evolution, and thus cannot be marked by SNPs in the vicinity. Several examples of these alleles have been found principally in breast cancer, such as variants in *CHEK2*, *BRCA1*-interacting protein C-terminal helicase 1 (*BRIP1*) and partner and localiser of BRCA2 (*PALB2*) (reviewed in [122]).

In conclusion, human cancers originate from complex interactions between environmental and genetic components. Environmental factors can be behavioural in nature, such as alcohol drinking and smoking, occupational, such as exposure to asbestos or coal tar, or naturally prevalent, such as solar radiation and outdoor air pollution. Although the majority of cancers arise from environmental effects, it is estimated that up to a quarter of all human cancers have a major genetic component. Genetic predisposition to cancer can display Mendelian inheritance, when single genes account for the majority of the inherited cancer risk. Autosomal dominant conditions require only one mutated allele to confer an elevated cancer risk, and these can be caused by mutations in tumour suppressor genes (*e.g.*, *TP53* or *CDKN2A*) or in proto-oncogenes (*e.g.*, *RET* and *CDK4*). Autosomal recessive cancer-predisposing conditions also exist, such as A-T, NBS, BS and WS. However, the majority of the genetic risk for cancer is thought to arise from the additive or multiplicative effects of many genes with small effects, and in support of this hypothesis, polygenic models of cancer predisposition have been found to explain observations made in population-based series.

In the remaining sections of this introduction, I focus on the particular cancer I studied during my PhD, melanoma. I cover its biology, classification and presentation, its risk factors, the melanoma genome, and unanswered questions in melanoma genetics.

1.4 Melanoma: Facts, origin and biology

Melanoma is a malignancy arising from melanocytes, the pigment-producing cells in our skin. Even though it accounts for less than 2% of all dermatological cancer cases, it causes more than 75% of skin cancer-related deaths [202]. Over the last thirty years, its incidence has increased more rapidly than that of any other common cancer in the UK (Fig. 1.5) [110]. Some of this effect is perhaps due to better surveillance and improved detection methods and diagnosis criteria, but the majority is thought to be real and linked to changes in sun-related behaviour, and has been reflected, for example, in the increase in popularity of holiday packages in lower latitudes and sunbed usage [110]. This increase means that malignant melanoma is now the fifth most commonly diagnosed cancer in the UK, and is projected to become as common in the US by 2030 [203, 204].

The great majority of melanomas are diagnosed at an early stage, in which case the 5-year survival rates are high (~95% for young adults and ~90% for older people) [206, 207]. However, melanoma diagnosed at a metastatic stage is highly aggressive and resistant to chemotherapeutic treatment [208, 209], and the 5-year survival rates decline

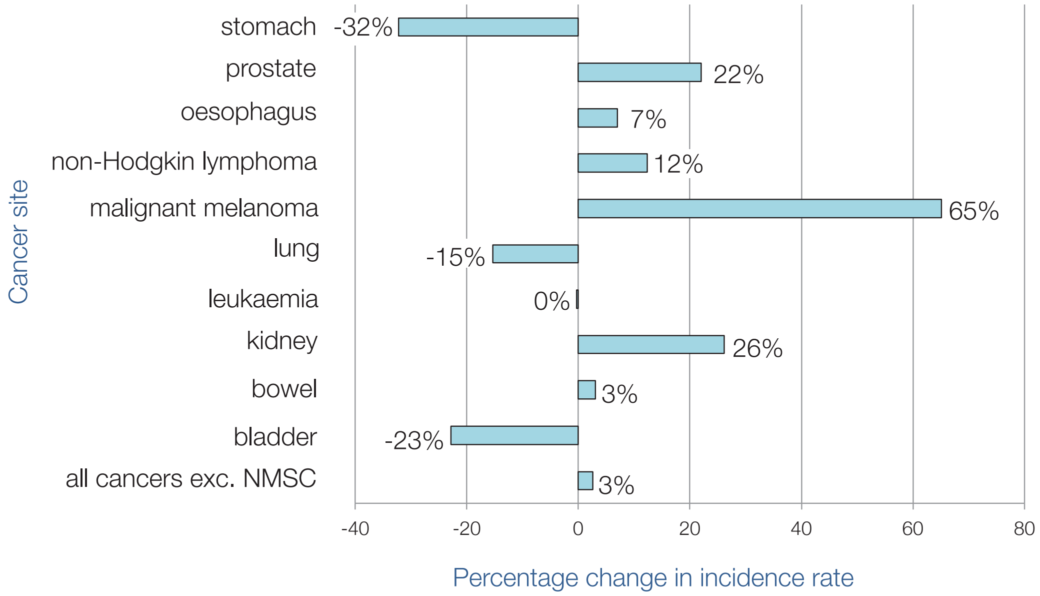

Figure 1.5: **Percentage change in European age-standardised three-year average incidence rates, in the UK, for the ten most common cancers.** European age-standardised three-year average incidence rates, UK, between 1999-2001 and 2008-2010. The information shown displays only incidence in males, but is representative of both sexes (malignant melanoma is, in both sexes, the fastest increasing common cancer in incidence). Information from ref. [205].

to ~60% and ~15% for regional and distant metastatic disease, respectively [202]. These effects are thought to arise from particular melanocyte biology properties, which I review in the next subsection.

1.4.1 Melanocytes: How they originate, where they reside, what they do

Melanocytes are specialised cells that synthesise melanin, a biopolymer that has various roles in cellular protection against damaging stimuli. Almost all melanocytes are derived from the neural crest, which is a transient embryonic cell population that can migrate extensively and give rise to numerous cell types, including much of the peripheral nervous system (reviewed in [210]). The exception are those melanocytes that constitute the retinal pigment epithelium (RPE), which are derived from the neuroepithelium [211]. The neural crest is formed during the embryonic process of neurulation, when the neuroectoderm (at the outermost of the three germ cell layers) is transformed into the neural tube (Fig. 1.6a). Cells at the border between the neural and non-neural ectoderm dissociate and induce the epithelial-mesenchymal transition (EMT), allowing them to migrate out of the neuroepithelium [210]. These migrating cells are initially multipotent but they become gradually lineage-restricted depending on anatomical location (reviewed in [212]). The moment at which neural crest cells become committed to the melanocytic lineage (after which point they are called melanoblasts) is open for discussion [213], but they do so before reaching their final destination in the developing embryo [214]. Melanoblasts migrate mainly dorsolaterally, and settle principally in the epidermis, dermis, hair follicles and the inner ear cochlea [215].

During their migration process, melanoblasts rely on the Notch and Wnt signaling pathways for lineage commitment [212], and express several genes essential for their survival, such as paired box 3 (*PAX3*), sex determining region Y-box 10 (*SOX10*), microphthalmia-associated transcription factor (*MITF*) and v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (*KIT*) [217, 218] (Fig. 1.6b). In line with their functions in the developing melanocyte, defects in these genes underlie piebaldism and the Waardenburg and Tietz syndromes, which have in common pigmentation defects (reviewed in [219]).

Having reached their target site, melanoblasts differentiate into melanocytes by up-regulating genes important for the production of melanin, such as tyrosinase (*TYR*), tyrosinase-related protein 1 (*TYRP1*) and dopachrome tautomerase (*DCT*), mainly driven by *MITF* (reviewed in [214]) (Fig. 1.6b). Although the production of melanin (known as melanogenesis) is widely considered to be the most important function of melanocytes, it is by no means the only one, as these cells also have important sensory and immunological roles [220, 221]. They also participate in eye organogenesis and vision, hearing and possibly cardiac functions, depending on their anatomical location [211]. As such, melanocytes can be described as “classical”, which are those found in the skin and contribute to its pigmentation, and “non-classical”, which are those found in all other parts of the body [211].

1.4.1.1 Classical melanocytes

The skin is divided into three layers: The hypodermis, which consists of fatty tissue and connects the skin with underlying tissues, the dermis, which consists of connective tissue and fibroblasts and houses the lymphatic, neural, vascular and secretory systems in the skin, and the epidermis, the outermost layer [222]. The epidermis can be further

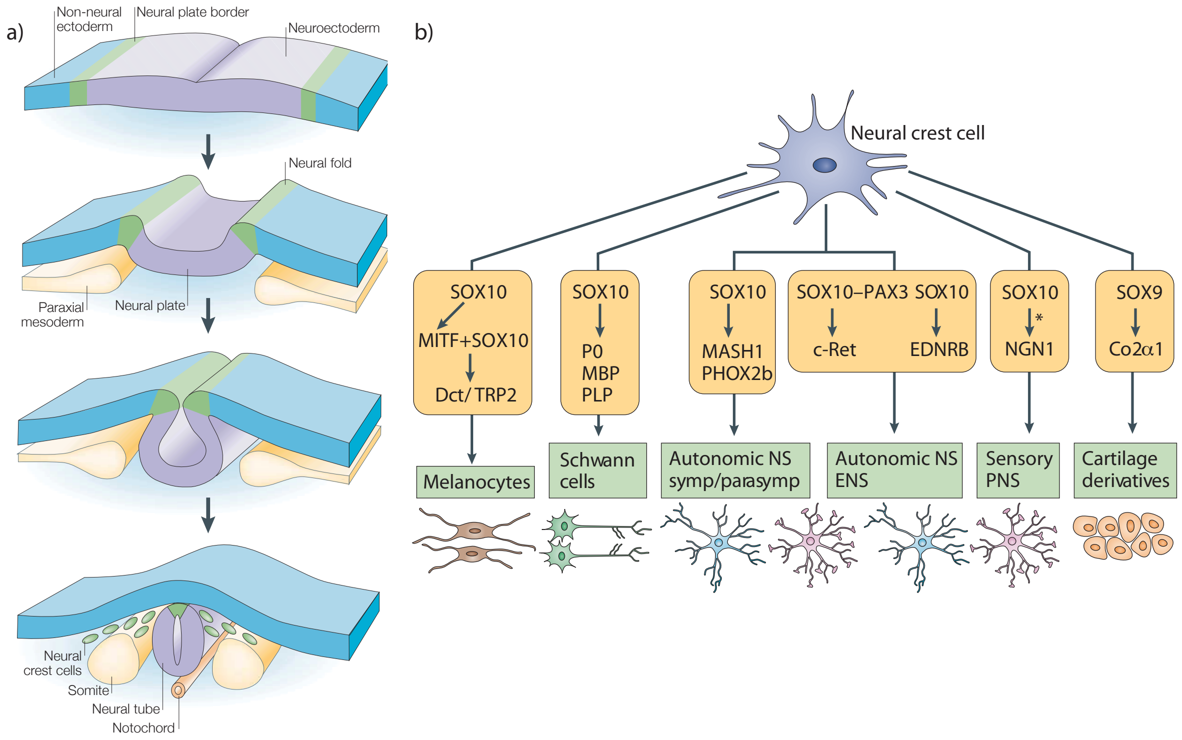

Figure 1.6: **Origin of the melanocytic lineage.** a) The neurulation process. In this embryonic stage, the neural plate rolls into the neural tube, and neural crest cells, localised at the border between the neural and non-neural ectoderm (green), delaminate from the neural folds and induce EMT, at which point they are able to migrate out of the neuroepithelium. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Neuroscience [216], copyright (2003). b) Neural crest cells can differentiate into numerous cell types, including much of the peripheral nervous system. Coordinated expression of *SOX10* and *MITF*, in conjunction with genes important for melanin synthesis such as *DCT*, determine the melanocytic fate. Important genes for other neural crest lineages are indicated. NS: autonomic nervous system; ENS: enteric nervous system; PNS: peripheral nervous system. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [210], copyright (2008).

divided into four layers, which, in order from innermost to outermost, are known as the basal, spinous, granular and cornified layers [222]. Classical melanocytes can be found in the dermis or in the basal layer of the epidermis (Fig. 1.7).

Epidermal melanocytes are dynamic cells, with dendritic projections that they use to communicate with keratinocytes and Langerhans cells, thus mediating their pigmentation and immune functions [211, 224]. The interactions between epidermal melanocytes and keratinocytes constitute the basis of the epidermal melanin unit, a complex that coordinates the production and distribution of melanin and is capable of responding

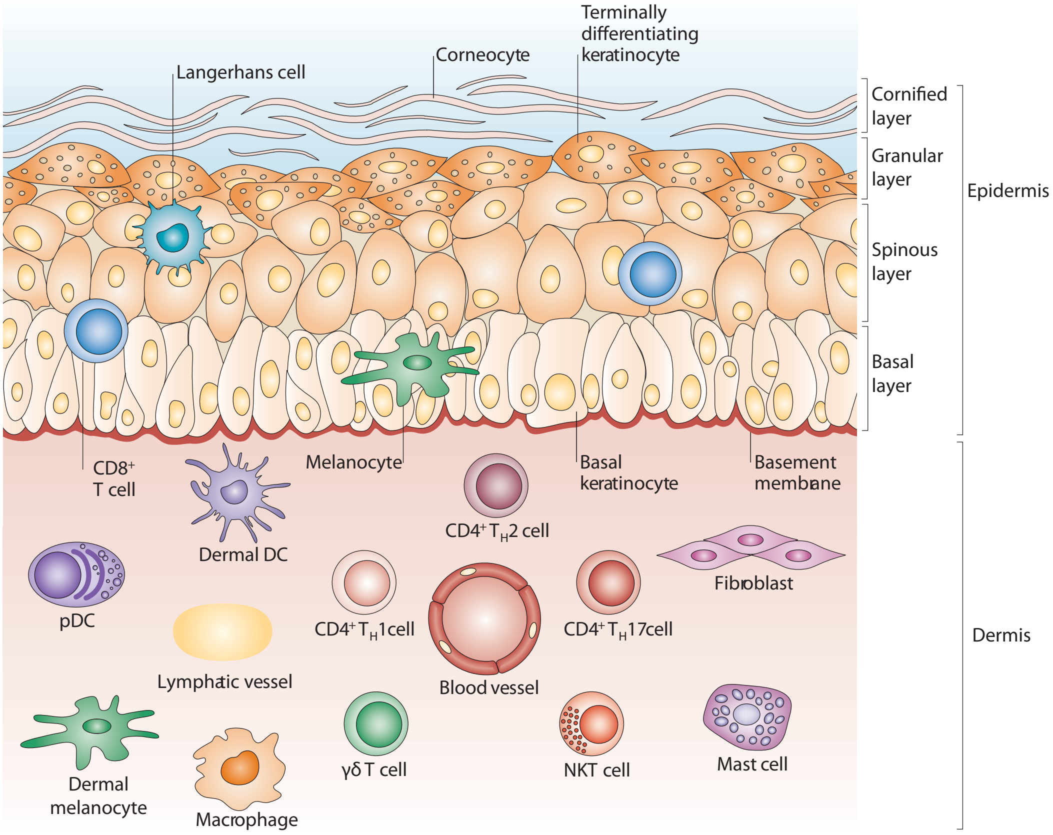

Figure 1.7: **Skin structure and melanocyte location.** The two outermost layers of the skin, dermis and epidermis, are shown. The four layers of the epidermis are indicated. Melanocytes (in green) are located in the dermis and the basal layer of the epidermis. Other types of cells they cohabitate and interact with, such as keratinocytes, Langerhans cells and fibroblasts, are depicted. The dendritic nature of melanocytes is also depicted. DC: dendritic cell, pDC: plasmacytoid DC, NKT: natural killer T, T_H: T helper. Modified and reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Immunology [223], copyright (2009).

rapidly to a large array of environmental stimuli [222]. It has been calculated that each epidermal melanocyte is in contact with about 30 keratinocytes [211], thus being able to distribute melanin upon a large area when required.

Within the epidermis, melanocytes can be located in the papillary bulbs of hair follicles or can be distributed in inter-follicular space. Inter-follicular melanocytes rarely undergo cell division [225], but those in hair follicles are able to proliferate and are involved in hair pigmentation and in the purging of toxic byproducts of melanin production [211].

Dermal melanocytes do not interact with keratinocytes, as they are surrounded by fibroblasts. They are present in very small numbers, and although they are capable of producing melanin, little is known about their function [211].

1.4.1.2 Non-classical melanocytes

Non-classical melanocytes are present in many different locations, such as the eye, the inner ear, the heart, the brain and in adipose tissues, and do not generally contain large amounts of melanin, although some exceptions exist [211]. Melanocytes in the eye can be part of the RPE, in which case they play an essential role in retinal function and visual acuity, or be located in the uvea, where they contribute to its development and are responsible for iris coloration [211]. Melanin produced by the RPE is crucial for neural retina development, and can protect it from ROS, thus preventing age-related macular degeneration, whereas that produced by uveal melanocytes contributes to protection from oxidative damage related to their location in a densely vascularised section of the eye [211]. Otic melanocytes can be found in the cochlea or the vestibular organ, in which they are necessary for normal hearing functions and might also participate in balance perception [211, 226]. Cardiac melanocytes have only recently been described [227]; they do not appear to transfer melanin to surrounding cells [228] and are not essential for normal cardiac function, so their contribution in the heart remains unclear [211]. Brain melanocytes are mainly located in the leptomeninges, which are the innermost tissue layers that cover the brain [229]. It has been suggested that these melanocytes may aid in sequestering toxic compounds from the circulation and participate in neuroendocrine functions, although their physiological role has not been clarified [211, 226]. Melanocytes that reside in adipose tissue have been found to synthesise melanin at much higher levels in obese people than in controls, and it is thought that they might help to neutralise ROS and cellular fat deposition present in these patients [211].

In summary, the majority of melanocytes originate from the neural crest, a multipotent embryonic cell population that also gives rise to much of the peripheral nervous system; however, melanocytes in the RPE differ in that they originate from the neuroepithelium. Melanocytes can be classified according to their anatomical location, with "classical" melanocytes being those that populate the skin (and thus have been extensively studied) and "nonclassical" melanocytes being those residing in all other parts of the body. Melanocytes contribute importantly to organ development and function, but melanin biosynthesis is widely considered to be their most important task. In the next paragraphs, I review the melanin biosynthetic pathway, as well as the specialised organelles in which it takes place, the melanosomes.

1.4.1.3 Melanosomes and melanin biosynthesis

Melanins are important biopolymers that determine the most obvious phenotypic characteristic in human and other vertebrates, skin colour, and have an essential role in protecting the body against harmful UV radiation and other environmental challenges [222]. Melanins can be divided in two main groups: eumelanins, which are dark brown or black, and phaeomelanins, which are lighter or yellowish. Eumelanins are insoluble and participate in various protective functions, being able to oxidise and reduce other molecules, bind diverse metal ions, and absorb the most hazardous components of solar radiation (reviewed in [230]). Phaeomelanins have poor protective properties as they are photolabile at physiological conditions, and thus might even increase phototoxicity [231].

The pathway for melanin synthesis is the same for all melanocytes, although some types of melanocytes display particularities. For example, although melanin synthesis generally occurs in specialised organelles called melanosomes, it happens in the cytosol of melanocytes that reside in adipose tissue [211]. Additionally, dopaminergic neurons are able to synthesise a pigment related to melanin, termed neuromelanin, that appears to have similar properties but may arise from a different non-enzymatic synthesis pathway [232]. Also, the timing of melanosome generation and melanin biosynthesis can differ significantly, for example, skin melanocytes produce melanosomes throughout life but the RPE synthesises melanin only during embryonic and early post-natal life [211]. Nonetheless, the majority of melanocytes carry out the synthesis of melanin in melanosomes and express a common set of catalysing enzymes.

Melanosomes are membrane-bound, lysosome-related organelles that possess the conditions and proteins required for melanin synthesis, and that prevent toxic byproducts associated with its production from harming other cellular components [233]. Melanosome development is generally divided into four stages, termed I-IV (Fig. 1.8). Stage I melanosomes probably originate from the endoplasmic reticulum, and possess an amorphous matrix and internal vesicles. Premelanosome protein (encoded by *PMEL*) is then transferred to stage I melanosomes and aids in transforming them into fibrillar, elongated organelles with TYR activity, now termed stage II melanosomes. Pigment synthesis starts in stage III melanosomes, which accumulate melanin in their fibrillar matrix. Finally, when they become fully pigmented they are termed stage IV melanosomes, at which point they are ready for transfer to neighbouring cells (reviewed in [226, 233, 234]).

Proteins necessary for melanosome structure and function are delivered to these

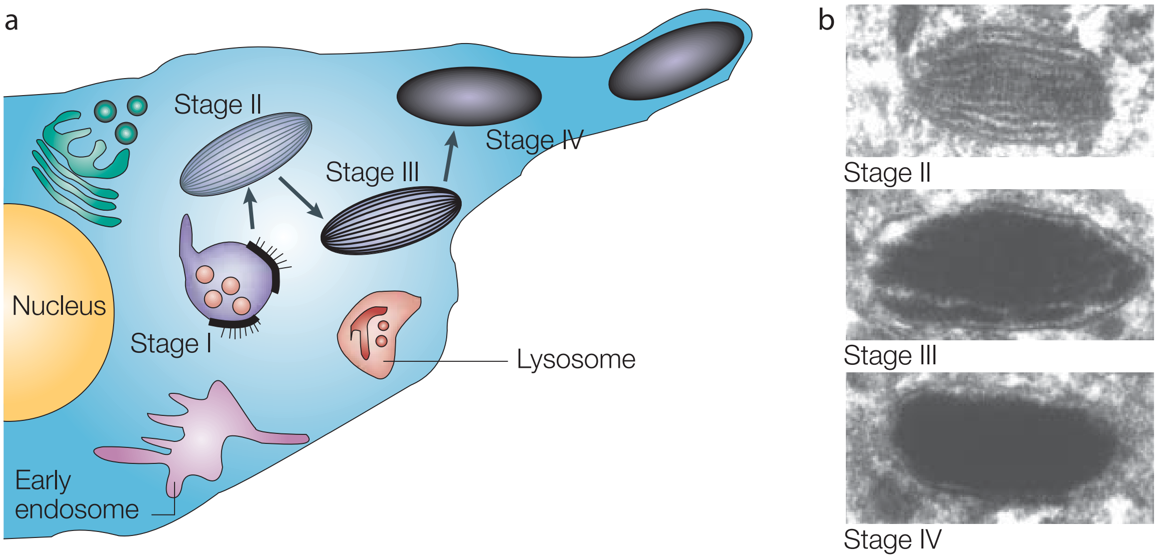

Figure 1.8: **Stages of melanosome maturation.** a) Diagram of a portion of a melanocyte body and dendrite. All stages of melanosome maturation are shown, alongside other relevant organelles. The degree of melanisation is indicated by black. b) Electron micrograph of melanosomal stages II, III and IV are shown. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [235], copyright (2001).

organelles throughout their maturation process, PMEL early and TYR, TYRP1 and DCT during the late stages [233]. The last three enzymes are required for melanin production, and therefore their malfunction or absence underlies distinct oculocutaneous albinism syndromes [222].

Biosynthesis of both eumelanins and phaeomelanins start with the hydroxylation of L-tyrosine to form L-3,4-dihydroxyphenylalanine (L-dopa), a reaction catalysed by TYR, and which is the rate-limiting step in melanin synthesis. The same enzyme then is able to oxidise L-dopa to L-dopaquinone, the last common step in the synthesis of dark and light pigments (see proximal phase, Fig. 1.9).

In eumelanin synthesis, L-dopaquinone yields L-dopachrome in a series of spontaneous chemical reactions, which then can form different precursors of dark pigments, such as 5,6-dihydroxyindole (DHI) and DHI-2-carboxylic acid (DHICA). These compounds then undergo spontaneous polymerisation reactions to form black and insoluble eumelanin (DHI-melanin) and brown, somewhat soluble eumelanin (DHICA-melanin). DCT and TYRP1 participate in the processing of L-dopachrome to these eumelanin precursors (reviewed in [236]) (see distal phase, Fig. 1.9).

Phaeomelanin synthesis is preferred if compounds such as L-cysteine or reduced glutathione are present. In this case, L-dopaquinone reacts with the thiol group in these compounds to form a complex mixture of intermediates, of which 5-cysteinyldopa is thought to be the most abundant. This compound then is hypothesised to undergo dehydration and several structural rearrangements to yield an alanyl-hydroxy-benzothiazine monomer, which is thought to be the monomeric subunit of phaeomelanin (Fig. 1.9). However, the phaeomelanin synthesis pathway is poorly understood, and thus could proceed via different and more complex intermediaries [236].

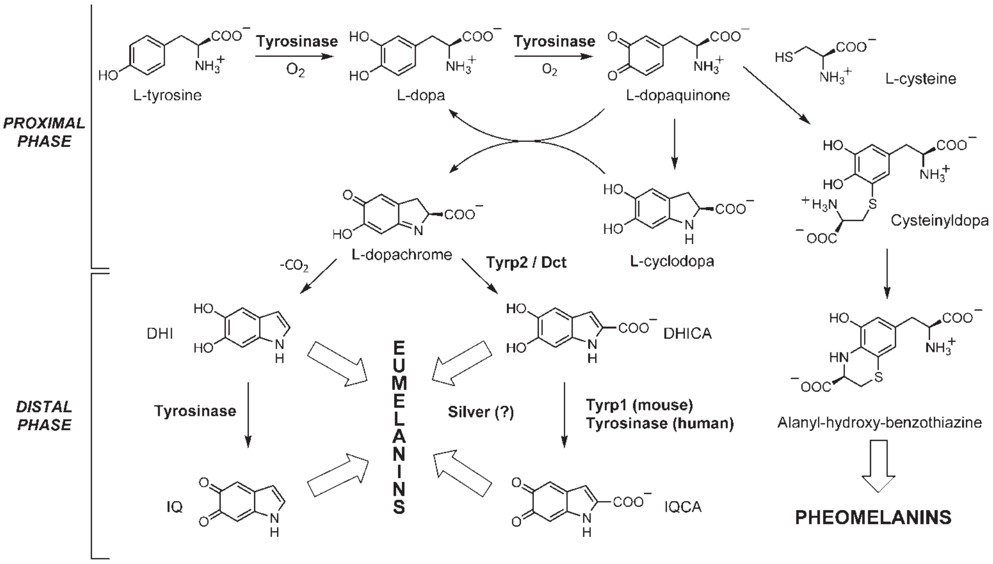

Figure 1.9: **Melanin biosynthesis pathway.** For simplicity, the pathway can be divided into the proximal phase, that contains the tyrosinase (TYR)-catalysed reactions, and the distal phase, that contains the subsequent reactions. In eumelanin synthesis, L-dopaquinone may form L-cyclodopa before this compound is spontaneously oxidised to L-dopachrome. IQ: 5,6-indolequinone and IQCA: indole-2-carboxylic acid-5,6-quinone are oxidation products from DHI and DHICA, respectively, and can be precursors for eumelanins. The silver locus, named after the mouse phenotype, encodes in humans the PMEL protein. Reproduced by permission from John Wiley & Sons, ref. [236], copyright (2011).

Melanocytes respond to different environmental challenges by producing melanin. Upon UV radiation, both melanocytes and keratinocytes increase their expression of the proopiomelanocortin (POMC) protein, which is then processed to different peptides, including α-melanocyte stimulating hormone (α-MSH) and adrenocorticotropin (ACTH). These peptides can bind the melanocortin 1 receptor (MC1R) on the surface of melanocytes and increase MITF expression through a pathway involving cyclic adenosine monophosphate (cAMP), protein kinase A (PKA) and the cAMP response element binding protein (CREB) [222, 237]. The epidermal melanin unit can also respond to other environmental stimuli such as physical pressure, growth factors or cytokines [222, 238]. Keratinocytes also produce and secrete other factors that stimulate melanocyte dendricity and melanin synthesis, such as endothelin 1 (EDN1), colony-stimulating factor 2 (CSF2), prostaglandins E2 and F2α, leukaemia inhibitory factor (LIF), KIT ligand (KITLG), fibroblast growth factor 2 (FGF2) and the hepatocyte and nerve growth factors (HGF and NGF respectively) (reviewed in [222, 225]). These factors bind to different receptors in the surface of melanocytes and signal down different but cross-talking biological pathways, stimulating melanogenesis, dendrite formation, proliferation and differentiation [222] (Fig. 1.10).

1.4.1.4 Implications of the melanocyte lineage in melanoma treatment

The embryonic origin of melanocytes and their function offer some clues as to the aggressiveness and resistance of melanoma. For example, it has been shown that human primary melanocytes become metastatic when transfected with a specific set of genes, but not other cell types, indicating that at least a part of the characteristic aggressiveness of melanoma can be attributed to melanocyte lineage-specific factors [240]. In human benign naevi, the transcription factor snail family zinc finger 2 (SNAI2), that strongly contributes to EMT during neural crest migration [241], has been shown to be expressed along with other neural crest cell migration-associated genes [240]. Additionally, other studies have shown that melanoma cells exploit lineage-specific RTKs or transcription factors to promote their plasticity and metastatic potential [242, 243]. These observations might indicate that melanocytes are predisposed to acquiring invasive properties, thus necessitating fewer alterations to metastasise than other tissue types.

Additionally, melanocyte function may help explain why melanoma is highly resistant to chemotherapeutic treatment. Many chemotherapeutic drugs approved or being tested for treating melanoma, such as dacarbazine [244], temozolomide [245] and cisplatin [246], exert at least part of their function by inducing DNA damage and thus triggering cell cycle arrest or cell death [247, 248]. Given the crucial role that melanocytes have in protecting our skin against damage caused by UV radiation, they have developed powerful anti-apoptotic mechanisms that contribute to their intrinsic resistance to DNA damage and cell death [215]. Lineage-specific factors, such as MITF and KIT, are thought to play a role in activating anti-apoptotic genes such as B-cell chronic lymphocytic leukaemia (CLL)/lymphoma 2 (*BCL2*), although the mechanisms of melanoma chemoresistance

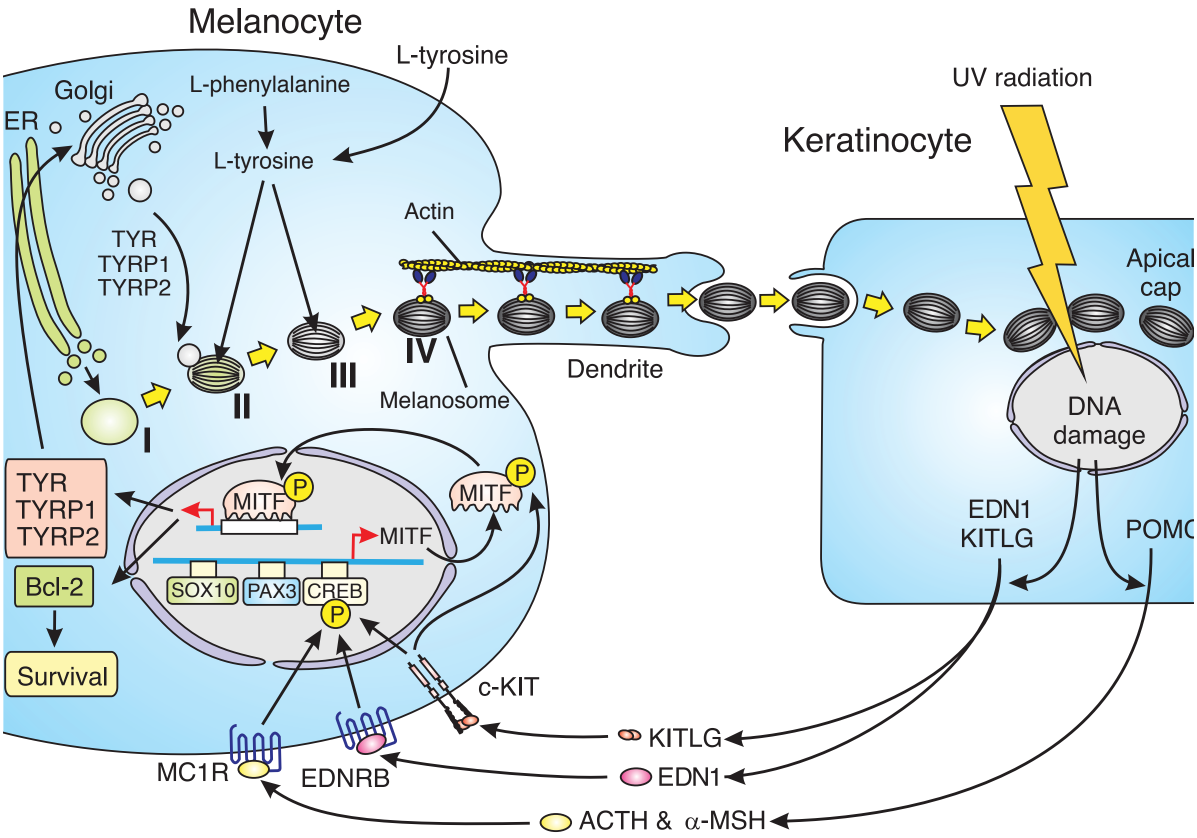

Figure 1.10: **Interactions between keratinocytes and melanocytes.** Schematic diagram of a melanocyte (left) and a keratinocyte. Upon UV radiation, keratinocytes produce factors that stimulate melanin production in melanocytes. These factors can, via signalling pathways, stimulate the trascription of genes important for survival and melanin synthesis. These gene products are then shipped to melanosomes, which upon maturation are transferred to keratinocytes where they contribute to DNA protection. Figure reproduced, with some modifications, from ref. [239].

remain controversial [215, 249]. Some authors have described melanocytes as cells that are "born to survive" [249, 250], and indeed, this seems to be the case given the low success rate of treatments for metastatic melanoma, making it one of the most challenging cancers to treat [251].

But how can a benign melanocyte give rise to one of the most aggressive and resistant human cancers? In the following paragraphs, I address the different models and theories that attempt to explain the transition from a benign melanocyte to malignant melanoma.

1.4.2 From melanocytes to melanoma: Different models of disease progression

It has been estimated that about three quarters of all melanomas arise from *de novo* melanocyte transformation, and the rest from pre-existing naevi [215, 252]. In both of these distinct paths to malignancy, changes in skin morphology are associated with a series of underlying molecular events.

1.4.2.1 The Clark model: Naevi become malignant

In 1984, Wallace Clark Jr. and colleagues described a step-wise model for the development of malignant melanoma from common naevi [253]. Naevi in the skin normally have an activating mutation in either v-raf murine sarcoma viral oncogene homolog B (*BRAF*) or neuroblastoma RAS viral (*NRAS*), which promotes growth by hyper-activation of the MAPK signalling pathway [254, 255] (Fig. 1.11). However, naevi rarely progress towards cancer, as oncogene-induced senescence is triggered and tumour suppressors such as INK4A (Fig. 1.3) and PTEN are up-regulated [256, 257]. However, this oncogene-induced senescence can be overridden by additional mutational events, sometimes in INK4A or PTEN themselves, and thus benign naevi can progress towards a dysplastic state (Fig. 1.11). As discussed in Paragraph 1.3.2.1.6, families with inactivating mutations in INK4A are predisposed to familial melanoma, perhaps because naevi progress more easily to the dysplastic state, whereas in sporadic cases the loss of PTEN is a more common alteration [258].

Additional mutational events might then affect MITF, although its role in melanoma progression is varied. In 10-20% of samples, MITF is found amplified, an event that is associated with worse prognosis. Although MITF is associated with melanocyte differentiation, it is thought that it confers a growth advantage to cells by cooperating with *BRAF* activation to transform them, thus being able to function as an oncogene [259]. Other samples have been found where MITF and its targets are down-regulated, suggesting that there are different subsets of melanoma with different MITF activities (reviewed in [260]). Amplification of cyclin D1 (encoded by *CCND1*) is also a marker of some acral melanomas, in which its inhibition has been shown to cause apoptosis [258] (Fig. 1.11). This phase, referred to as the radial growth phase, is characterised by an ability of melanocytes, now immortalised, to proliferate intradermally. The transition to the vertical growth phase, in which melanocytes gain the ability to invade the dermis and form tumours, is characterised by a loss of cell adhesion markers and the expression

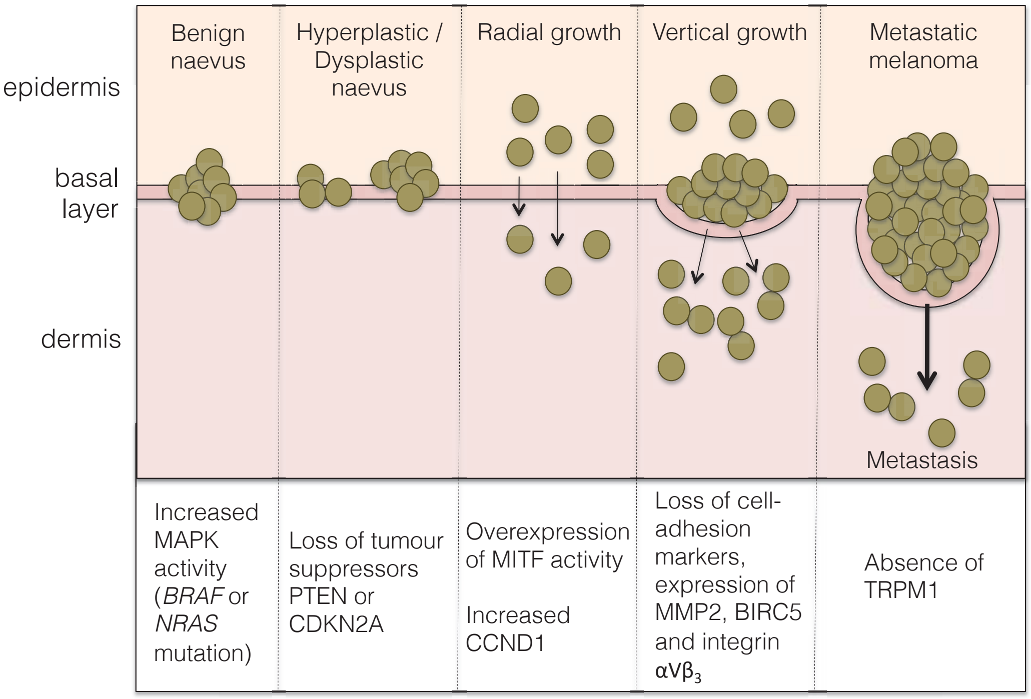
Benign naevus
Hyperplastic / Dysplastic naevus
Radial growth
Vertical growth
Metastatic melanoma
epidermis
basal layer
dermis
Metastasis
Increased MAPK activity (*BRAF* or *NRAS* mutation)
Loss of tumour suppressors PTEN or CDKN2A
Overexpression of MITF activity
Increased CCND1
Loss of cell-adhesion markers, expression of MMP2, BIRC5 and integrin $\alpha V\beta_3$
Absence of TRPM1

Figure 1.11: **The Clark model for the progression from naevi to malignant melanoma.** The hyperplastic and dysplastic phases (originally described as separate stages) are drawn together in this figure. Molecular events associated with each stage are in the bottom panel. Figure adapted from ref. [258].

of integrin αVβ3, baculoviral inhibitor of apoptosis domain repeat containing 5 (BIRC5) and matrix metallopeptidase 2 (MMP2), which contribute to survival and degradation of the collagen in the basal layer of the epidermis (reviewed in [258]). It has been also observed that the level of telomerase activity correlates well with tumour clinical stage, with higher levels of tumour cell penetration having higher telomerase activity [261]. Finally, melanocytes that have gained the ability to colonise other tissues, thus giving rise to malignant melanoma, commonly have a reduction or absence of expression of transient receptor potential cation channel, subfamily M, member 1 (TRPM1), a target of MITF. Although its function is not completely clear, this gene has been hypothesised to function as a tumour suppressor by comparison to other members of its family [262] (Fig. 1.11).

Many models of melanoma progression have been based on Clark's initial description of the melanoma stages, and thus is widespread in melanoma research community [263].

However, this view has been challenged by the observation that melanocytes within a naevus are not monoclonal as the Clark model would suggest, and the realisation that most melanomas arise in normal skin, not in association with an existing naevus [263]. Thus, an alternative model for the origin of malignant melanoma, aiming to explain how the majority of these cancers arise, has recently been put forward.

1.4.2.2 *De novo* progression from melanocytes to melanoma

Recently, Minoru Takata and colleagues proposed an alternative model for melanoma progression that does not pass through the dysplastic naevus stage and thus is different from Clark's [263]. They proposed that a yet unidentified hit might cooperate with the inactivation of INK4A or over-expression of cyclin D1 or CDK4, therefore allowing the cell to bypass oncogene-induced senescence. The acquisition of an activating mutation in *BRAF* would come after this first hit, resulting in clonal proliferation and thus yielding the *BRAF* mutation heterogeneity observed in naevi. Activation of telomerase and additional hits in late stages might be similar to those in cancers arising from benign naevi. Therefore, the order of mutational events might determine whether a melanoma arises from a pre-existing naevus or *de novo* [263]. This hypothesis is supported by observations made in acral and mucosal melanoma, in which cyclin D1 amplification arises frequently as a founder event, followed by oncogene activation, normally *KIT* [263].

At least one mouse model has been developed in an effort to recapitulate *de novo* disease progression. Mayuko Kumasaka and colleagues achieved *de novo* malignant melanoma formation in endothelin receptor B (Ednrb)-heterozygous mice constitutively expressing the RET oncoprotein [264]. This model could reproduce some characteristics of the human disease, such as the late age of onset, poor prognosis and high percentage of metastases, and, coupled with the observation that melanoma risk is increased in patients with *EDNRB* loss-of-function mutations [265], could provide a molecular rationale for the development of *de novo* malignant melanoma.

In summary, there seem to exist different paths to malignancy, with the majority of melanomas arising *de novo* but some arising from pre-existing naevi. This distinction might arise from the order in which important players for melanoma progression are somatically altered, but the lesions involved in both cases (*e.g.* senescence bypass, oncogene and telomerase activation and down-regulation of adhesion markers) are thought to be similar.

1.4.3 Clinical classification and staging systems

The models described above attempt to describe how melanocytes progress from a benign to a malignant state. Researchers have also classified melanoma in different types depending on the microscopic growth patterns of these lesions and have created diverse staging systems to aid treatment choice when diagnosing patients. In this Subsection, I review the recognised types of melanoma and the diverse systems used in diagnostic settings.

1.4.3.1 Types of melanoma

Clark not only defined the stages that a benign naevus had to progress through to become a malignant tumour, but also classified melanoma according to the microscopic growth patterns of these lesions [266]. He and his team divided it into four broad main types: superficial spreading melanoma (SSM), lentigo maligna melanoma (LMM), nodular melanoma (NM) and acral lentiginous melanoma (ALM) [263, 267] (Fig. 1.12).

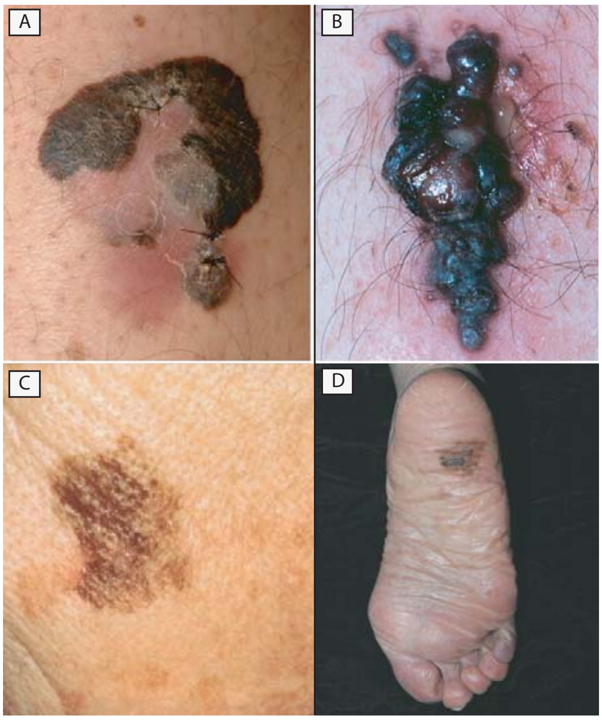

Figure 1.12: **Types of melanoma.** a) Superficial spreading melanoma, b) Nodular melanoma, c) Lentigo maligna melanoma and d) Acral lentiginous melanoma. Panels a, b and d reproduced from ref. [209] under a Creative Commons CC BY-NC 4.0 license. Panel c taken from the Skin Cancer Foundation, and reproduced by [268].

SSMs constitute about 70% of all invasive melanomas, and are characterised by melanocytes expanding upward within the epidermis and radially in its basal layer. They display variation in pigmentation and the majority occur *de novo* [267, 269]. LMMs evolve from lentigo maligna, a lesion that occurs on sun-exposed skin of older patients, although the progression is rare and slow. They are considered to be at the earliest stage of melanoma (see Paragraph 1.4.3.2.3), and are characterised by a proliferation of abnormal melanocytes along the basal layer of the epidermis [267, 270]. NMs are the second most common subtype of melanoma, comprising 10-15% of cutaneous melanomas, and are more aggressive than SSMs. They occur more commonly in sun-exposed areas of the skin, and present as expanding, darkly pigmented nodular lesions [271, 272]. ALMs present in palms and soles, and are the most common manifestation of melanoma in people of Black ancestry. They are highly aggressive, and present as darkly pigmented patches with varying degrees of pigmentation that expand rapidly [267, 273, 274].

Other types of melanoma have been identified that do not conform to the typical classification system. The IARC recognises, apart from the types mentioned above, desmoplastic melanoma (DM), melanoma arising from blue naevi (BN), melanoma arising from giant congenital naevi (GCN), childhood melanoma, naevoid melanoma and persistent melanoma [275]. DMs are superficial melanocytic lesions in which malignant cells are separated by fibrous stroma or collagen fibres, and can present cellular abnormalities. Melanomas arising from BN, which owe their colour to their deep position within the epidermis, are exceedingly rare and seem to be highly metastatic. GCN are large lesions that commonly cover more than 2% of the body surface, and are direct precursors to melanoma. Indeed, about 6% of GCN progress to malignancy, and the tumours are commonly sharply demarcated and characterised by asymmetry. Childhood melanoma comprises melanomas that develop in individuals before they reach puberty. Their incidence is quite low, and their clinical features are similar to melanomas arising in adults. Naevoid melanomas are also rare, and are distinctive because they resemble common intradermal naevi, but with the potential to metastasise. Persistent melanomas are those that grow out of primary lesions that have been excised, and normally have the same epidemiological and etiological characteristics as the primary tumour they are derived from [275].

Because of the heterogeneity in melanoma lesions and the existence of tumours that do not conform to the above classification system, John Curtin and colleagues proposed in 2005 an alternative classification in which melanoma types could potentially be distinguished not by histological characteristics but by molecular signatures [276]. As such, characteristics such as the number of DNA copies, gains in the *CCND1* locus or mutations in *BRAF* or *NRAS* in a sample were able to accurately distinguish among acral melanomas, mucosal melanomas or melanomas on skin with or without chronic sun-induced damage [276]. Melanomas with and without chronic sun-induced damage roughly correspond to LMM and SSM, respectively, and acral melanoma to ALM in Clark's classification [263].

These distinct types of melanoma and their associated molecular events reveal a glimpse of the complexity of melanoma aetiology, indicating that diverse biological pathways can be altered in different ways in melanoma development, depending on characteristics such as anatomical site, amount of sun exposure, age and ancestry.

1.4.3.2 Melanoma staging systems

Different staging systems have been developed to describe tumour depth and the amount of spreading to other parts of the body. These are routinely utilised by medical doctors and researchers when diagnosing a patient and recommending treatment procedures.

1.4.3.2.1 Clark levels Tumour depth had been known to inversely correlate with prognosis, which inspired Clark to propose, in 1969, a 5-level classification system to score tumour depth [277] (Table 1.1).

Table 1.1: **Clark levels used to score tumour depth.**

Clark level	Description
Level I	All tumour cells are above the basal layer of the epidermis (melanoma *in situ*)
Level II	Tumour cells have broken into the papillary dermis
Level III	Tumour cells have reached the junction between the papillary and reticular dermis
Level IV	Tumour cells have invaded the reticular dermis
Level V	Tumour cells have spread into the subcutaneous tissue

Clark levels were used as the main tumour staging determinant in the American Joint Committee on Cancer (AJCC) staging system of 1997, but following studies that demonstrated that they only held independent predictive value for thin melanoma lesions (T1 stage, see Table 1.2), the AJCC revised their classification system in 2001 to restrict their assessment to this type. This recommendation has stayed the same for the last release of their classification system, in 2009 [278, 279].

1.4.3.2.2 Breslow thickness Around the same time that Clark published his classification system, the pathologist Alexander Breslow also published a paper arguing that melanoma lesions had to be scored on tumour thickness as this factor, along with Clark's level, seemed to correlate well with prognosis [280]. Tumour thickness is measured by determining the depth of invasion from the skin surface to the farthest malignant cell, usually by a pathologist examining a lesion under the microscope. Breslow thickness is still used as the main determinant of tumour staging, according to the 2009 AJCC melanoma staging system [279] (Table 1.2).

Table 1.2: **Breslow thickness measurements used by the 2009 AJCC melanoma tumour staging system.** Tis stands for tumour *in situ*, where cells are confined to the top layer of the epidermis. Modified from ref. [279].

Tumour stage	Breslow thickness (millimetres)
Tis	Not applicable
T1	$\leq$ 1.00
T2	1.01-2.00
T3	2.01-4.00
T4	$>$ 4.00

1.4.3.2.3 TNM scale The TNM classification is the main cancer staging system in use, and the most recent guidelines for melanoma were published by the AJCC in 2009 [279]. T stands for tumour size, N for the number of metastatic nodes, and M for the number of distant metastases. Each of these can be further subdivided to indicate characteristics such as whether ulceration is present, the amount of nodal metastatic burden and the site where metastases are present (Table 1.3). As such, this system utilises the Clark levels and Breslow thickness to aid in defining tumour stage and prognosis. The TNM stages include the most important parameters that retain individual prognostic value, and thus are the most important indicator of probability of survival [279] (Fig. 1.13).

1.4.3.2.4 Melanoma diagnosis: The ABCDE acronym In 1985, the mnemonic "ABCD" was created and published by three medical doctors to aid both healthcare professionals and the public in the early diagnosis of malignant melanoma [281]. Its main idea is that naevi that present the ABCD criteria should be examined closely; A stands for asymmetry, B for border irregularity, C for colour variegation and D for diameter larger than 6 millimetres (mm). More recently, a fifth criterium, E (evolving),

Table 1.3: **Most recent pathologic stage categories for cutaneous melanoma as determined by the AJCC.** Tumour thickness (T) stages are as described in Table 1.2, with the letters a and b indicating the absence and presence of ulceration, respectively. The number of metastatic nodes (N) are scored as follows: N0, not applicable, N1, 1 metastatic node, N2, 2 or 3 metastatic nodes, and N3, 4 or more. The letters a, b and c indicate micrometastases, macrometastases and in transit metastases, respectively. M0 and M1 refer to the absence and presence of distant metastases, respectively. Reprinted with permission from ref. [279]. © (2009) American Society of Clinical Oncology. All rights reserved.

Stage	T	N	M
0	Tis	N0	M0
IA	T1a	N0	M0
IB	T1b	N0	M0
	T2a	N0	M0
IIA	T2b	N0	M0
	T3a	N0	M0
IIB	T3b	N0	M0
	T4a	N0	M0
IIC	T4b	N0	M0
IIIA	T1-4a	N1a	M0
	T1-4a	N2a	M0
IIIB	T1-4b	N1a	M0
	T1-4b	N2a	M0
	T1-4a	N1b	M0
	T1-4a	N2b	M0
	T1-4a	N2c	M0
IIIC	T1-4b	N1b	M0
	T1-4b	N2b	M0
	T1-4b	N2c	M0
IV	Any T	Any N	M1

was added to recognise the dynamic nature of melanoma and to aid in identifying lesions that might not conform to the ABCD rule, such as NMs [282]. The ABCD rule has been found useful in recognising early lesions [283, 284], and due to its simplicity, continues to be widely used and promoted.

As described in this Section, the different types of melanoma, their progression through different disease stages and their resistance mechanisms are associated with genomic lesions such as activation of *KIT*, *NRAS* or *BRAF*, or inactivation of *CDKN2A* or *PTEN*, among others. In the next section, I review the current knowledge regarding the human melanoma genome, including known driver mutations (*i.e.*, those that confer

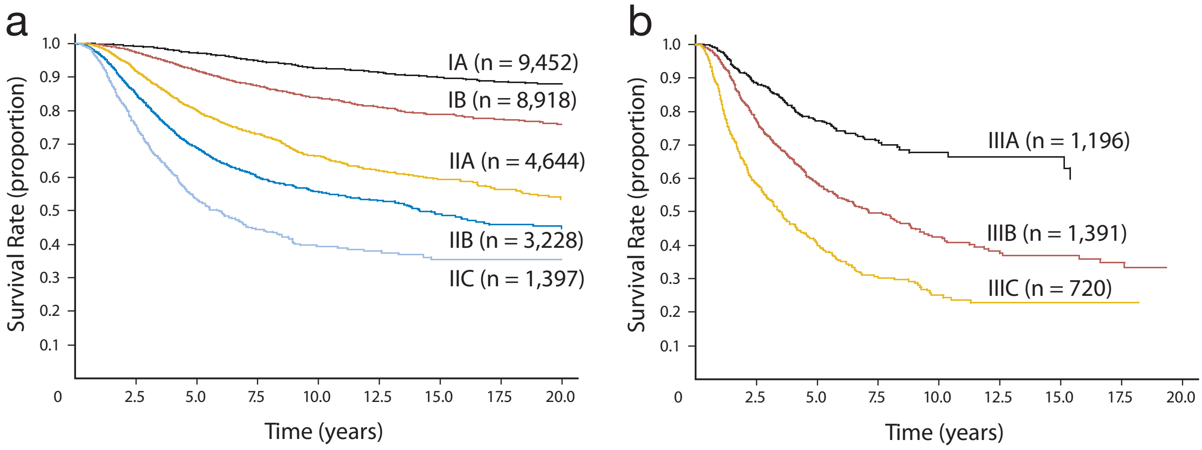

Figure 1.13: **Survival curves from the AJCC Melanoma Staging Database grouped by TNM stage.** The AJCC analysed data from 30,946 patients with Stage I-III melanoma, survival curves grouped by TNM stage over a 20-year period are shown. a) Stages I and II. b) Stage III. Reprinted with permission from ref. [279]. © (2009) American Society of Clinical Oncology. All rights reserved.

a fitness advantage to the carrier cell) and mutational signatures.

1.5 The human melanoma genome

Since the discovery of *BRAF*-activating mutations in almost 70% of human melanomas [285], somatically-acquired aberrations in the MAPK signalling pathway have been intensely studied. Overall, the great majority of melanomas have MAPK pathway dysregulation, with 15% of melanomas acquiring activating *RAS* mutations in a *BRAF* mutually-exclusive manner [258, 286]. Curtin and colleagues found that activating mutations of the MAPK pathway seemed to be less common in melanomas with chronic sun damage and non-cutaneous melanomas, and in general, that distinct subsets of genetic alterations were present depending on the type of melanoma [276]. Additionally, traditional *BRAF* or *NRAS* activating alterations are not C-to-T transitions, which are typically induced by UV radiation (see Subsection 1.3.1.5) [287]. These observations indicate that UV exposure significantly affects melanoma progression, possibly through the mutation of other oncogenic pathways.

In the last five years, the human melanoma genome has been intensely investigated using WGS, WES and exon capture methodologies. Two large studies analysing the somatic mutation frequency in more than 25 cancer types agreed that melanoma has the highest mutational burden across all of them, principally due to the ubiquitous presence of C-to-T transitions [288, 289] (Fig. 1.14). This elevated mutational rate therefore can act as a confounding factor when attempting to uncover melanoma driver mutations.

Despite these difficulties, several studies have described many genomic events contributing to melanomagenesis or tumour progression. The majority of these mutations, that have been classified as drivers according to different criteria (*e.g.*, mutation frequency across samples, reduction of tumour fitness upon gene silencing or tumour progression upon ectopic gene expression), affect signalling pathways important for cell proliferation and migration. As such, somatically-aquired mutations resulting in MAPK pathway activation other than *BRAF* or *RAS* have been identified, such as gain-of-function events in glutamate receptor metabotropic 3 (*GRM3*) [290], MAPK kinase 1 (*MAP2K1*), *MAP2K2* [291], and v-erb-b2 avian erythroblastic leukaemia viral oncogene homolog 4 (*ERBB4*) [292]. Additionally, loss-of-function events in MAPK kinase kinase 9 (*MAP3K9*) and *MAP3K5* have been found to decrease MAPK signalling but increase drug resistance [293].

The PI3K pathway also plays a key role in melanoma progression. Activating mutations in *NRAS* and loss of function on *PTEN*, which result in activation of this signalling pathway, have been found in 15% and 20-30% of melanomas, respectively, in a mutually-exclusive manner [294]. Other significantly mutated loci might also activate the PI3K pathway, such as phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2 (*PREX2*) [295], phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*) [287], v-akt murine thymoma viral oncogene homolog 3 (*AKT*3) [294], and *KIT* (which also signals down the MAPK pathway) [287, 296].

Expectedly, other common somatic alterations involve pathways important for cell cycle control and genome maintenance. Copy-number gains or somatic activation of the telomerase reverse transcriptase (*TERT*) gene are a frequent event in melanoma, as point mutations in its core promoter leading to increased telomerase expression were found in about 70% of examined tumours [297, 298]. These mutations were all found to be C-to-T transitions, providing a mechanism for UV radiation-induced tumourigenicity. Additionally, copy-number gains in *CCND1* are common in the acral subtype of cutaneous melanoma [299, 300], where its encoded product cyclin D1 contributes to tumour development by binding the CDK4/6 complex and by phosphorylating RB1 (see Paragraph 1.3.2.1.6). Recently, two independent studies identified protein phosphatase 6, catalytic subunit (*PPP6C*) as frequently mutated exclusively in melanoma samples with *BRAF* or *RAS* mutations [287, 301]. *PPP6C* encodes a protein able to block entry

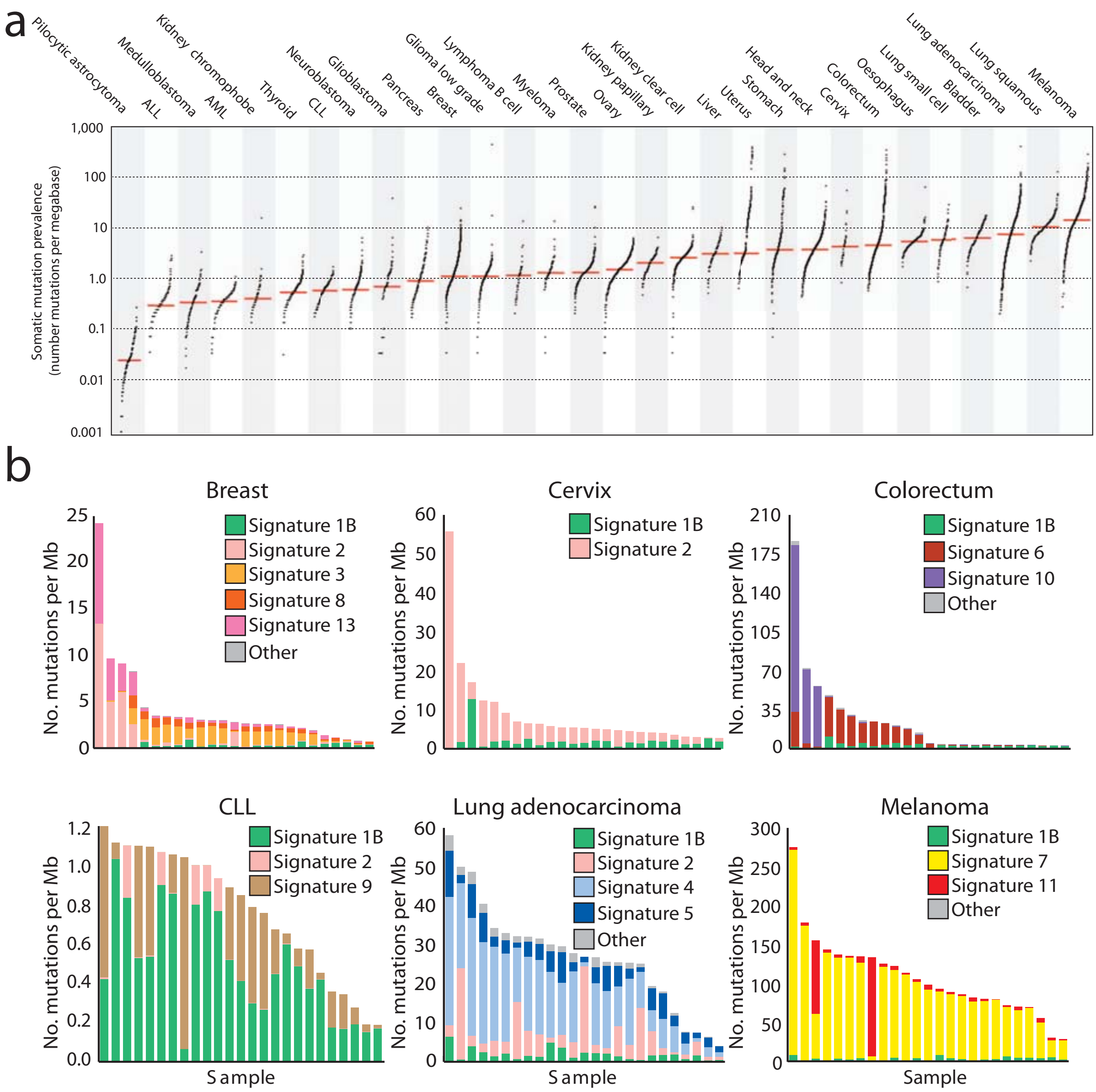

Figure 1.14: **Frequency and type of somatically-acquired mutations in melanoma genomes.** a) Prevalence of somatic mutations in human cancers. Each dot represents one sample, and the red line represents the median number of mutations per cancer type. The vertical axis is log-scaled and shows the number of somatic mutations per megabase (Mb), whereas the horizontal axis has the cancer types ordered by the number of mutational events. Melanoma is the type of cancer with the highest mutational burden overall. b) The contributions of somatic signatures to individual cancer types. Bars represent selected samples from each cancer type whereas the vertical axis denotes number of mutations per Mb. Whereas other cancer types have distinct contributions from different mutational processes, melanoma tumours are dominated by C-to-T transitions (Signature 7) and temozolomide-induced mutations in treated cases (Signature 11). Signature 1B has been linked to ageing, and is present across different cancer types. Reprinted by permission from Macmillan Publishers Ltd: Nature [288], copyright (2013).

into S phase, suppress the levels of cyclin D1 and reduce RB1 phosphorylation [302]. However, the two studies did not agree on the nature of the mutations: One identified probable loss-of-function events [301], whereas the other one identified clusters of variants in highly conserved residues that might represent oncogenic mutations [287].

Although the majority of cancers inactivate the P53 pathway at the level of *TP53* itself, mutations in *TP53* are not too frequent in melanoma, generally found in less than 20% of samples [209, 287, 301]. However, loss of this pathway's functions seems to be achieved, in a mutually-exclusive manner, by mutating *RB1* or the INK4A-coding region of *CDKN2A*, or amplifying *CDK4* or *MDM2* [209, 287, 303] (see Paragraph 1.3.2.1.6). Interestingly, more than half of the mutations in *TP53* in one large systematic review could be attributed to UV stress, implicating this locus in UV-induced disease progression [304]. Another study identified a synonymous C-to-T somatic mutation in the BCL2-like 12 (*BCL2L12*) gene, which caused accumulation of its messenger (m)RNA and protein product due to differential miRNA targeting [305]. This accumulation was found to cause increased binding to P53, decreased induction of UV-induced apoptosis, and a reduction in the transcription of endogenous P53 target genes.

Other pathways that might be commonly targeted for somatic alterations are those important in melanocyte development. The MITF pathway is altered in about 20% of metastatic tumours, with *MITF* amplification being a frequent event [259, 306]. Copy-number gains in *MITF*, in conjunction with oncogenic *BRAF*, were shown to be able to transform melanocytes [259], and the *MITF* regulator *SOX10* was found to be mutated in about 10% of primary melanomas and to be inactivated in some metastatic tumours [306].

Recently, the glutamate receptor pathway has also been found to be somatically deregulated in melanoma cells. Mutations in *GRM3* not only activate the MAPK signalling pathway, but also alter glutamate receptor function [290]. Additionally, mutations in the glutamate receptor, ionotropic, N-methyl D-aspartate 2A (*GRIN2A*) have been found in about a quarter of melanomas, and phospholipase C, beta 4 (*PLCB4*), a downstream effector from GRIN2A, has also been found to be recurrently mutated [307]. Glutamate receptor dysfunction has been noted in neuronal tumours, with gliomas that have high glutamate release being more aggressive [308]. Given the neural crest origin of melanocytes, it might indicate that shared molecular mechanisms between neurons and melanocytes allow melanoma cells to take advantage of neurophysiological receptors [209]. Although it is unclear from the pattern of mutations if GRIN2A acts as a tumour suppressor or an oncogene, mutant *GRM3* was found to harbour a mutation hotspot and to confer increased anchorage-independent growth and migration, possibly indicating that it acts as an oncogene [290, 307].

Many other significantly mutated loci have been described. Recently, ras-related C3 botulinum toxin substrate 1 (*RAC1*) was described as a new oncogene in two independent studies, as both found UV-induced recurrent mutations affecting the same codon [287, 301]. RAC1 is an important player in cell proliferation and cell migration, and when mutated, it could provide a migratory advantage to melanocytes through ERK activation [301]. Other candidate genes include deleted in colorectal carcinoma (*DCC*), dynein cytoplasmic 1 intermediate chain 1 (*DYNCI1I*), RPTOR independent companion of MTOR, complex 2 (*RICTOR*), AT rich interactive domain 2 (*ARID2*), serine/threonine kinase 19 (*STK19*), sorting nexin 31 (*SNX31*) and transforming acidic coiled-coil containing protein 1 (*TACC1*) [287, 301].

Although somatic mutations are required to drive melanomagenesis, the acquisition of mutations is facilitated by certain factors, such as induction of C-to-T transitions upon UV radiation exposure, or having a preponderance of phaeomelanin over eumelanin, which can generate oxidation-induced DNA damage. In addition, germline mutations in some of the genes mentioned above have been found to predispose to melanoma with high penetrance. In the next section, I discuss these and other risk factors for melanoma development.

1.6 Risk factors for melanoma development

Several risk factors for the development of melanoma have been identified. These can be environmental, such as exposure to UV radiation, physical characteristics, such as having fair skin, eyes or hair, or genetic, such as carrying mutations in genes that control melanogenesis or more general cell integrity processes. In this Section, I review established risk factors for melanoma and their modes of action.

1.6.1 Environmental factors

1.6.1.1 UV radiation

It has been estimated that about 86% of all melanomas can be attributed to exposure to UV radiation, whether from the sun or from UV-emitting devices such as sunbeds [110], and as such, it is considered the main risk factor for developing melanoma (for mutagenic mode of action, see Subsection 1.3.1.5). However, different patterns of sun exposure carry different risks. For example, a meta-analysis of 57 studies and a total of 38,671 patients identified that intermittent and intense sun exposure significantly increased melanoma risk when compared to chronic sun exposure [309]. Although a history of sunburn might appear to increase melanoma risk, this effect may arise because people are more likely to remember it, but a causal relationship cannot be excluded [309].

Additionally, the relationship between the use of sunbeds and melanoma has been analysed in several studies, and at least two meta-analyses concluded that melanoma risk was increased in individuals that had ever used sunbeds, with risk increasing if the exposure was at a young age [310, 311]. As such, the IARC decided to raise the carcinogenicity level of sunbed usage to the top category in 2009 [312]. In 2011, legislation came into force in England and Wales that made it illegal for people under 18 years of age to use sunbeds [313], and other countries, such as France, Brazil, Austria and Germany, have also taken steps to limit the exposure of younger individuals [314].

1.6.1.2 Other environmental factors

Diverse studies have identified other factors that might play a role, albeit probably limited, in melanoma susceptibility. Melanomagenesis has been linked to arsenic exposure [315], regular swimming (probably water chlorination) [316], exposure to residential magnetic fields [317], and industries such as electronics, metal and transport and communications [318]. However, these studies analysed only a limited number of people and reported small odds ratios (ORs) with wide confidence intervals, so further research is necessary to establish whether a causal relationship exists between these environmental factors and melanoma formation.

1.6.2 Physical characteristics and medical history

1.6.2.1 Skin, hair and eye colour

Skin and hair colour are primarily determined by the amount and types of melanins produced, although other pigments such as haemoglobin and dietary carotenoids also contribute to them [319]. The absence of both eumelanin and phaeomelanin results in white hair (albinism), higher eumelanin production results in dark hair and skin, and higher phaeomelanin production results in blond or red hair and fair skin [319]. Therefore, given the crucial role of eumelanin in skin protection against UV radiation, it is expected that individuals with darker skin and hair would be better protected than blond or red-haired individuals when exposed to UV radiation [226] (see Subsection 1.4.1.3).

Although the molecular details underlying hair and skin colour are still incompletely understood, some important players have been elucidated. The switch between eumelanin and phaeomelanin production within a melanocyte is controlled by *MC1R*, which is highly polymorphic in the human population [319]. Variants that impair MC1R binding to α-MSH and/or subsequent cAMP production, and therefore affect eumelanin production, are over-represented in redheads and fair-skinned people [320]. Furthermore, *MC1R* mutation carriers are not only more susceptible to UV radiation-induced damage, but may also have increased oxidation-induced DNA damage arising from phaeomelanin synthesis, contributing to melanomagenesis [321].

MC1R variants play a role not only in red-haired individuals, but also account for inability to tan in individuals without red hair [322], and increase the risk of developing freckles independently of skin and hair colour [323]. As such, an inability to tan and the presence of freckles represent independent melanoma risk factors [324, 325]. More recently, a mutation altering *KITLG* expression levels was found to underlie classic blond hair colour in Europeans [326], and given that this pathway seems to play an important role in melanocyte number, size and dendricity [327], it will be interesting to see whether this variant contributes to melanoma susceptibility in this population, and if so, whether the mechanism involves the KIT pathway.

Eye colour is a polygenic trait, but two main contributing loci have been identified: oculocutaneous albinism II (*OCA2*) and homologous to the E6-AP carboxyl terminus (HECT) and regulator of chromosome condensation 1-like domain (RLD) containing E3 ubiquitin protein ligase 2 (*HERC2*) [328]. OCA2 is involved in melanosome transport and maturation, and thus variations in this gene affect the type and quantity of melanin produced [328]. Polymorphisms within *HERC2* probably act in the same pathway, as they have been found to affect *OCA2* expression levels [329]. Therefore, eye colour reflects genetic variants that affect melanin distribution, and therefore has been identified as an independent melanoma risk factor [324, 330].

1.6.2.2 Presence of multiple naevi

The presence of an elevated number of naevi is considered one of the most important risk factors for melanoma development. Naevi are benign melanocytic tumours, and although the majority are stable and will not progress to melanoma, some will become malignant (see Subsection 1.4.2.1). A meta-analysis carried out in 2005 identified a 7-

fold higher melanoma risk in people with more than 100 common naevi compared with those that have fewer than 15, and individuals with five dysplastic naevi presented a 6-fold higher risk compared with individuals with no dysplastic naevi [331].

Although the majority of melanomas arise *de novo* (see Subsection 1.4.2.2), it has been calculated that around 20% of all melanomas originate from a dysplastic naevus (reviewed in [332]), and, at least within *CDKN2A* mutation-carrier families, individuals with dysplastic naevi are more likely to carry the mutant gene compared with family members without dysplastic naevi [333]. Therefore, the presence of multiple common or dysplastic naevi might indicate the presence of genetic or environmental factors that favour melanoma progression, such as carrying a mutation that accelerates melanocyte proliferation, or having a history of sun exposure.

1.6.2.3 Presence of a previous melanoma

Individuals with a previous diagnosis of melanoma have a 8-12-fold increased risk of developing a subsequent melanoma if they do not have a family history of melanoma, and 300-500-fold increased risk if they do and have dysplastic naevi when compared to population incidence rates [334, 335]. As is the case with dysplastic naevi, the development of a single melanoma might be indicative of the presence of underlying risk factors, but it can also happen that cells from a primary tumour that failed to be surgically excised proliferate and originate a secondary tumour. This event is referred to as local recurrence or persistent melanoma (discussed in Subsection 1.4.3.1).

1.6.2.4 Other medical conditions

Immunocompromised patients, such as organ transplant recipients, are at an increased risk of developing many types of cancer, with skin cancer including melanoma being one of the most common [336, 337]. Additionally, Parkinson's disease patients also seem to have a significant increase in melanoma incidence, although the reason remains unknown [338]. Perhaps not surprisingly, individuals with other types of cancer, such as non-Hodgkin's lymphoma, leukaemia and renal and gastrointestinal tumours also have increased melanoma risk [339–342]. Additionally, several studies have reported that retinoblastoma patients that survive their disease are also at high risk of developing melanoma [209, 343, 344].

1.6.3 Genetic risk factors

Another important melanoma risk factor is having a family history of the disease, which could be indicative of underlying genetic alterations that predispose to its formation. Meta-analyses have shown a ~2-fold increase in risk for individuals with a family history compared to those without [345, 346]. Several loci have been discovered, through methodologies such as linkage studies, WGS, WES or GWAS, to influence cutaneous melanoma risk (Fig. 1.15). These genes participate in cell cycle control or genome stability, and have varying degrees of penetrance.

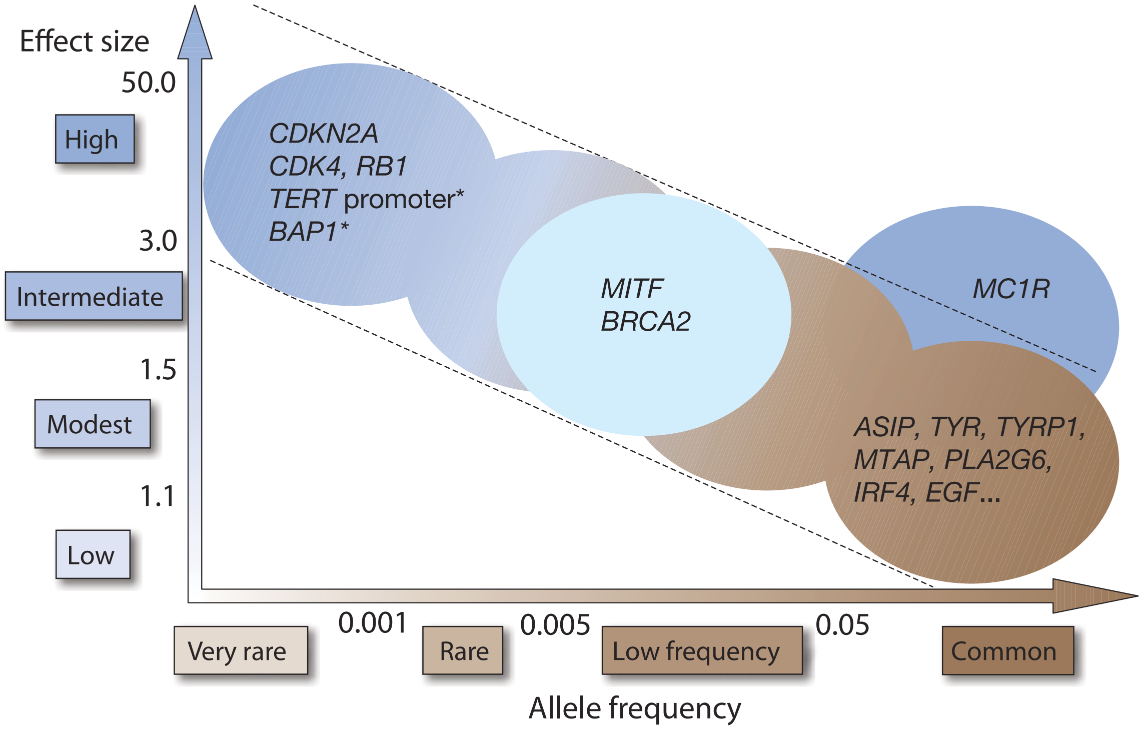

Figure 1.15: **Effect sizes and allele frequencies of loci influencing melanoma risk.** The dotted line indicates where the highest emphasis is when searching for predisposition loci. Loci with asterisks have been estimated, as very few families have been studied to ascertain their true effect size and allele frequency. The position of *BRCA2* has been estimated from refs. [347, 348]. Remaining loci positions are from ref. [215]. Modified and reprinted by permission from Macmillan Publishers Ltd: Nature [349], copyright (2009).

1.6.3.1 High-penetrance loci

The biological function of the two classical high-penetrance melanoma susceptibility loci, *CDKN2A* and *CDK4*, is discussed in Paragraph 1.3.2.1.6. About 40% of melanoma-prone families carry mutations in *CDKN2A*, whereas very few kindreds with pathogenetic *CDK4* mutations have been described worldwide [350, 351]. For familial cases, the overall penetrance of *CDKN2A* mutations has been estimated at 30% by 50 years of age and 67% by 80 years of age, although it can be modified by residence location [352]. For sporadic cases, it has been calculated at 14% by 50 years of age and 28% by 80 years of age [353], indicating that other shared environmental or genetic risk factors might influence familial melanoma risk.

RB1 mutation carriers that survive retinoblastoma (see Paragraph 1.3.2.1.1) are at a 4-80-fold risk of developing melanoma when compared to population incidence levels [209]. Other genes have been reported recently that might be highly penetrant, although estimating their penetrance has been hampered by the small numbers of carrier individuals identified. In 2011, Thomas Wiesner and his team identified deleterious mutations in BRCA1-associated protein 1 (*BAP1*) co-segregating with melanocytic tumours in two families, with tumours showing loss of heterozygosity of the wild-type allele [354]. Two years later, Susanne Horn and colleagues identified a mutation in the promoter of *TERT* co-segregating with the disease in a large melanoma-prone German family [297]. This mutation was found to increase *TERT* expression levels, and since then, similar somatically-acquired mutations have been found in a variety of other cancers [355].

1.6.3.2 Medium-penetrance loci

As discussed in Subsection 1.6.2.1, individuals that carry variants in *MC1R* have a higher risk of developing melanoma. ORs for different variants within the gene range from 1.42 to 2.45 [356]. More recently, a single nucleotide variant (SNV) affecting codon 318 in MITF was found to disrupt a conserved small ubiquitin-like modifier (SUMO)-ylation site, thus altering the spectrum of MITF target genes and leading to melanomagenesis. The OR for this allele was calculated to be between 2.7 and 4.78, consistent with an intermediate penetrance effect, but its allele frequency is much rarer than that of *MC1R* variants [357, 358]. *BRCA2* mutation carriers have more than 2.5-fold risk of developing the disease when compared to population incidence levels [347].

1.6.3.3 Low-penetrance loci

Many low penetrance loci, with high allele frequencies, have been identified through GWAS. Many of these genes affect pigmentation functions or skin sensitivity to UV-induced DNA damage [215]. For example, variants in agouti signalling protein (*ASIP*), an antagonist of MC1R, *TYR* and *TYRP1* have been associated with cutaneous melanoma with ORs ranging from 1.15 to 1.45 [359], and similar ORs were obtained for variants in the naevus-associated genes methylthioadenosine phosphorylase (*MTAP*), phospholipase A2, group VI (*PLA2G6*) and interferon regulatory factor 4 (*IRF4*) [360]. Variants in epidermal growth factor (*EGF*) and its receptor (*EGFR*) have also been associated with melanoma and its progression [361, 362]. Other loci potentially implicated are caspase 8 (*CASP8*), *CCND1*, solute carrier family 45, member 2 (*SLC45A2*), *ATM*, *OCA2*, myxovirus resistance 2 (*MX2*), among others (reviewed in [194, 209, 215]).

As we can see from the above paragraphs, melanoma is a complex disease arising from the interaction of diverse environmental factors, such as UV radiation and potentially other carcinogens, with the underlying genetic make-up of an individual. It is possible that dozens or maybe hundreds of genes, contributing to pigmentation, skin sensitivity, cell cycle control and genome stability, influence the genetic risk of this disease. In the next section, I discuss melanoma clustering in families, as the subject of study of this dissertation.

1.7 Familial melanoma

The English general practitioner William Norris first reported a case of familial melanoma in 1820 [363]. He speculated, given that his patient's tumour had originated from a naevus, that his children and brothers had many naevi, and that his father had also died of the disease, that melanoma was hereditary. This observation is notable because it was made nearly half a century before Mendel published his treatise on genetics [1].

Familial melanoma can be characterised by multiple melanoma cases across several generations on one side of the family, or by the presence of multiple primary melanomas in a single individual, or by an early age of onset [209]. In 1978, Lynch coined the term "familial atypical multiple mole-melanoma syndrome" (FAMMM) for families with a clustering of multiple large, dysplastic naevi of variable colour with pigmentary leakage, one of the main risks for the development of melanoma [364] (see Subsection 1.6.2.2). However, tumours from individuals with FAMMM do not show any histopathologic differences when compared to sporadic cases, and they might not show all typical FAMMM characteristics, so their analysis is not useful when diagnosing the familial condition [365]. Individuals with FAMMM caused by mutations in *CDKN2A* also have been found to have a higher incidence of other malignancies, especially pancreatic cancer [365, 366].

Since Norris's report, it has become clear that although the majority of melanomas can be attributed to UV radiation exposure [110], about 10% of all melanoma cases have a family history of the disease [194]. However, heritability of melanoma is high, as it has been estimated that about 55% of variation in liability to melanoma is due to genetic effects [367]. This discrepancy between high heritability and low familial melanoma rate might indicate that a large proportion of the risk arises from common, low-penetrance variation [368], although some rare high-penetrance loci might still remain undiscovered. This could be because individual variants in a locus might have a very low allele frequencies and/or might not be obviously disruptive, which precludes systematic identification of potential candidates.

The definition of familial melanoma varies depending on geographical location, as some regions, such as Australia, have a higher prevalence of melanoma and thus a higher probability of seeing clusters of melanoma within a family caused by non-genetic reasons [352, 369, 370]. In the UK and Europe, familial melanoma is generally defined as a cluster of two or more first- or second-degree relatives with melanoma [370, 371], whereas in Australia it constitutes a diagnosis of one invasive melanoma and two or more other cancer diagnoses among first- or second- degree relatives [370]. Additionally, patients might be referred for genetic testing if they live in a region with low melanoma incidence and present with at least two primary tumours, or if they present with three or more primary tumours and they live in a high melanoma incidence region [370].

1.8 Unanswered questions in melanoma genetics

Since the discovery of *CDKN2A* and *CDK4* about 20 years ago [192, 199], only a handful of other high-penetrance genes in very rare families have been described, namely *BAP1*, *RB1* and *TERT* (see Subsection 1.6.3.1). Collectively, these genes explain only about 50% of all familial melanoma cases. This fact begs the question: Are there any more high-penetrance genes remaining to be discovered, or can all unexplained familial risk be attributed to many lower-penetrance alleles?

Additionally, these known high-penetrance genes encode proteins that participate in cell cycle control and genome maintenance pathways, with *CDKN2A*, *CDK4*, *RB1* and *BAP1* participating in G1/S cell cycle progression ([372], see Paragraph 1.3.2.1.6), and *TERT* having a paramount role in genome stability after each cell division [373]. If other high-penetrance loci remain to be discovered, do these participate in the same biological pathways, or are there any other processes important in melanoma predisposition?

During my PhD, I endeavoured to help answer these questions. The dramatic drop in the costs in sequencing costs over the past decade [14] has allowed the use of this technology to explore large numbers of affected individuals in the search for common affected loci or biological pathways. When planning this project, we, as a team, decided to use WES in multiple members of several melanoma-prone families for four main reasons:

- Unbiased projects (*i.e.*, assessing the whole genome or exome rather than sets of candidate genes) to date studying genetic susceptibility to melanoma are very few, and have been carried out in three or fewer individuals [354, 358]. This means that the possibility of rare, high-penetrance alleles in coding regions has not been properly addressed.

- The cost of sequencing an exome is ~20× less than that of sequencing a whole genome, which would allow us to study multiple individuals from a greater number of affected families. This is helpful when attempting to obtain statistical support for the co-segregation of candidate mutations.

- Variation in protein-coding regions can be directly interpretable, *e.g.* causing an amino acid change or introducing a premature stop codon in a protein, which then can be assessed by functional, structural or conservation scores.

- The possibility that deep catalogues of human genetic variation would be developed and made publicly available during the course of this project, which would allow a better estimate of the allele frequency of candidate variants in human populations.

1.8.1 Overview of whole-exome sequencing as a tool for providing answers

In order to analyse the protein-coding regions of the genome, the WES method involves, firstly, the enrichment of the DNA sample for exonic regions (or any other custom regions, in the case of candidate gene sequencing). Generally, this is achieved by solid- or liquid-phase hybridisation. In the first case, probes complimentary to the target regions are fixed in a solid surface, such as a microarray, which can then be washed and captured regions eluted; in the second case they can be hybridised to biotinylated DNA or RNA probes and captured with streptavidin beads [374, 375] (Fig. 1.16). Captured DNA regions can then be massively sequenced, mapped onto the human genome reference assembly, and potential variants can be called and filtered in the search for candidate melanoma predisposition genes.

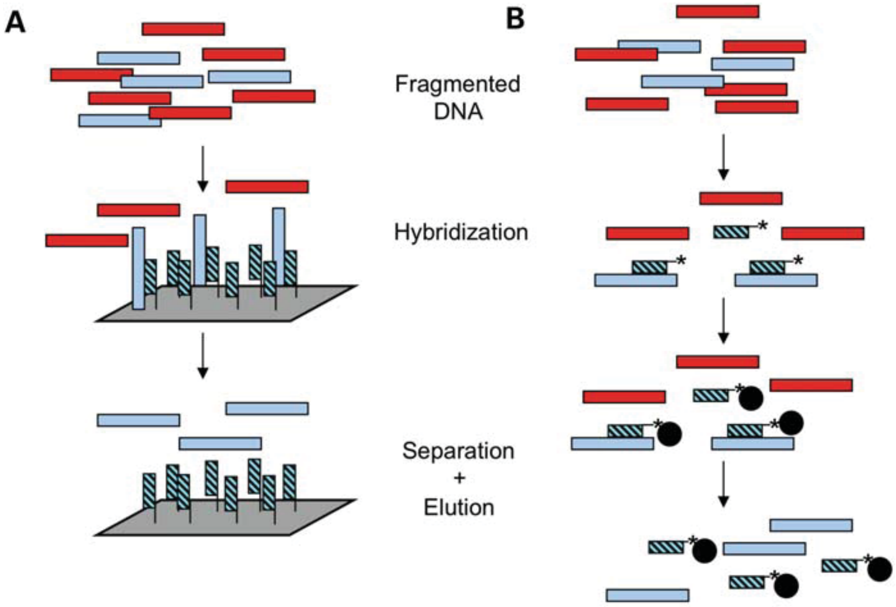

Figure 1.16: **Hybridisation methods for target DNA capture.** A) Solid-phase. Baits complementary to the targeted sequence (blue with black stripes) are synthesised on a microarray. Upon fragmentation, target DNA (blue) hybridises with the microarray baits, the array is washed and the captured DNA is eluted. B) Liquid-phase. Baits complementary to the targeted sequence (blue with black stripes) are synthesised and biotinylated, indicated by an asterisk. Fragmented DNA hybridises with the baits, and these are captured with streptavidin beads (black). The bead-bait complexes are washed and the target DNA eluted. Target DNA is then sequenced. Reprinted from ref. [374] by permission of Oxford University Press.

1.8.2 Overview of the methodology followed to study melanoma predisposition genes

I have organised the chapters in this dissertation roughly according to the sequential order in which I carried out the different steps in the study (Fig. 1.17). In order to search for high-penetrance melanoma susceptibility genes, I used three main familial melanoma collections, originating from the UK, The Netherlands, and Australia. These datasets were not all available at the start of the project and thus the initial analyses were only performed with the UK and Dutch datasets (discovery and replication phases, covered in Chapter 2). Then, with the availability of the Australian cohort, I could perform an integrative analysis and identify candidate susceptibility genes (covered in Chapter 3). Finally, I cover the mechanistic investigation of the mode of action of these candidate genes in Chapter 4. Finally, in Chapter 5, I discuss the relevance of the results presented here, as well as the future directions of this project.

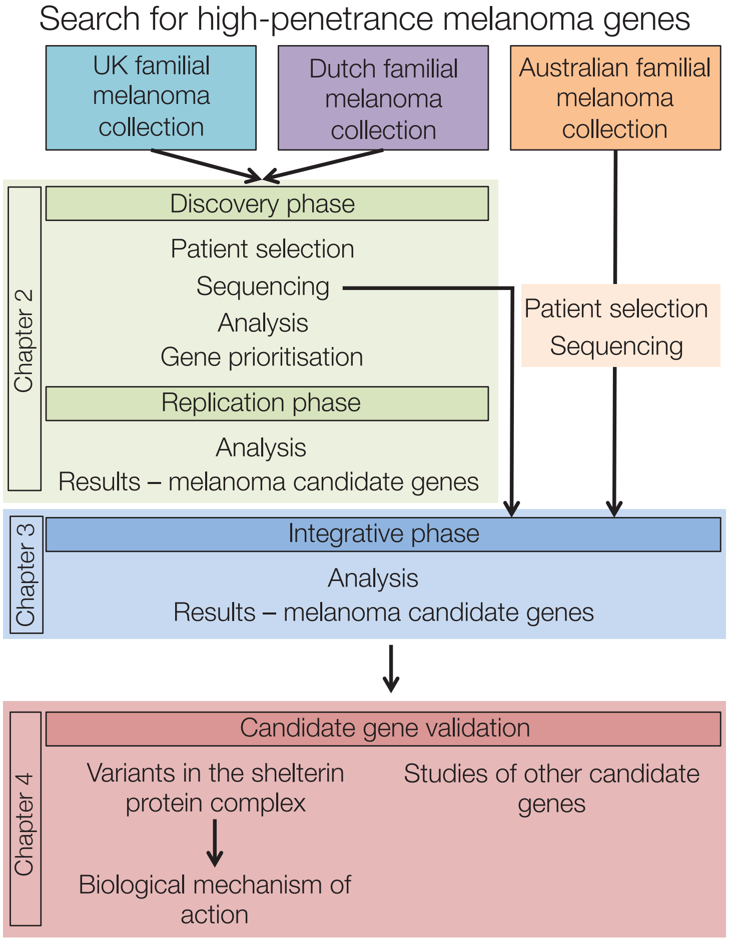
Search for high-penetrance melanoma genes
UK familial melanoma collection
Dutch familial melanoma collection
Australian familial melanoma collection
Chapter 2
Discovery phase
Patient selection
Sequencing
Analysis
Gene prioritisation
Replication phase
Analysis
Results – melanoma candidate genes
Patient selection
Sequencing
Chapter 3
Integrative phase
Analysis
Results – melanoma candidate genes
Chapter 4
Candidate gene validation
Variants in the shelterin protein complex
Biological mechanism of action
Studies of other candidate genes

Figure 1.17: **Methodology followed during my PhD to study melanoma susceptibility genes.** The chapter in which each step is covered is indicated at the left. An arrow indicates a contribution of that dataset to the next analysis.

Chapter 2

Familial melanoma sequencing: European phase

Some methods in this chapter have been published (ref. [376]). Some parts of the text have been reproduced from this reference; I confirm I have ownership of copyright for reproduction in this work.

In an attempt to identify high-penetrance germline variants that contribute to melanoma development, we exome-sequenced and analysed affected members from predisposed families. For this phase, I had access to an extensive set of samples collected by clinicians and scientists over past decades from families from the UK and The Netherlands. This chapter explains the rationale for patient selection, the sequencing methodology, data processing and the gene prioritisation analyses performed.

This phase is divided into two stages, discovery and replication. Briefly, we sequenced whole exomes from high-priority families (*i.e.*, those with a higher number of cases, early age of onset and/or MPM) and compiled a list of candidate genes. For the replication phase, we targeted these genes for sequencing in additional families to search for supporting evidence of the involvement of the identified genes in melanoma predisposition. Finally, we developed novel gene prioritisation strategies combining evidence from both the discovery and replication phases in order to proceed to biological validation. An overview of all the steps explained in this Chapter is depicted in Fig 2.1.

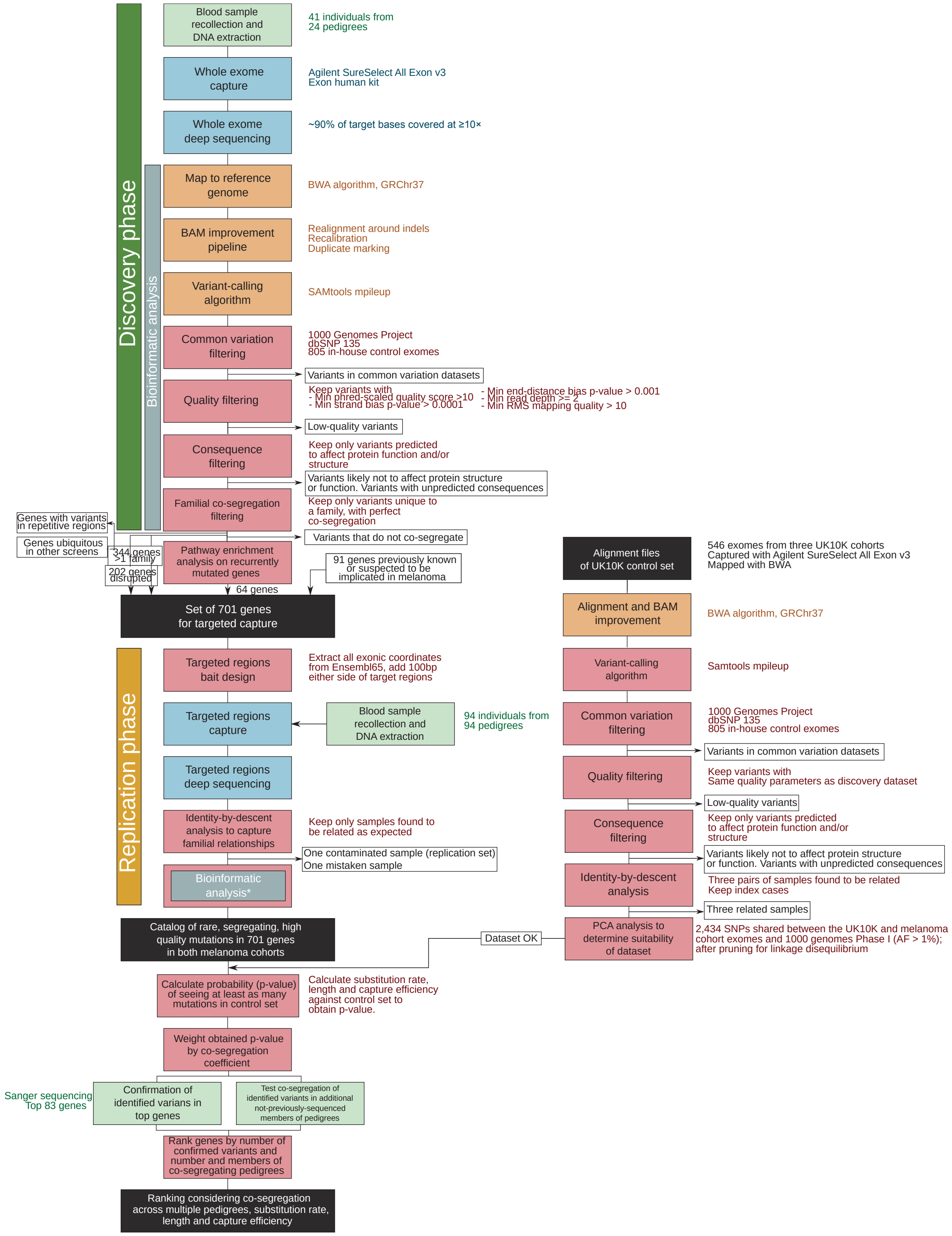
Discovery phase
Bioinformatic analysis
Blood sample recollection and DNA extraction
41 individuals from 24 pedigrees
Whole exome capture
Agilent SureSelect All Exon v3 Exon human kit
Whole exome deep sequencing
~90% of target bases covered at ≥10×
Map to reference genome
BWA algorithm, GRChr37
BAM improvement pipeline
Realignment around indels
Recalibration
Duplicate marking
Variant-calling algorithm
SAMtools mpileup
Common variation filtering
1000 Genomes Project
dbSNP 135
805 in-house control exomes
Variants in common variation datasets
Quality filtering
Keep variants with
- Min phred-scaled quality score >10
- Min strand bias p-value > 0.0001
- Min end-distance bias p-value > 0.001
- Min read depth >= 2
- Min RMS mapping quality > 10
Low-quality variants
Consequence filtering
Keep only variants predicted to affect protein function and/or structure
Variants likely not to affect protein structure or function. Variants with unpredicted consequences
Familial co-segregation filtering
Keep only variants unique to a family, with perfect co-segregation
Genes with variants in repetitive regions
Genes ubiquitous in other screens
Variants that do not co-segregate
344 genes >1 family
202 genes disrupted
Pathway enrichment analysis on recurrently mutated genes
91 genes previously known or suspected to be implicated in melanoma
64 genes
Set of 701 genes for targeted capture
Replication phase
Targeted regions bait design
Extract all exonic coordinates from Ensembl65, add 100bp either side of target regions
Targeted regions capture
Blood sample recollection and DNA extraction
94 individuals from 94 pedigrees
Targeted regions deep sequencing
Identity-by-descent analysis to capture familial relationships
Keep only samples found to be related as expected
One contaminated sample (replication set)
One mistaken sample
Bioinformatic analysis*
Catalog of rare, segregating, high quality mutations in 701 genes in both melanoma cohorts
Alignment files of UK10K control set
546 exomes from three UK10K cohorts
Captured with Agilent SureSelect All Exon v3
Mapped with BWA
Alignment and BAM improvement
BWA algorithm, GRChr37
Variant-calling algorithm
Samtools mpileup
Common variation filtering
1000 Genomes Project
dbSNP 135
805 in-house control exomes
Variants in common variation datasets
Quality filtering
Keep variants with
Same quality parameters as discovery dataset
Low-quality variants
Consequence filtering
Keep only variants predicted to affect protein function and/or structure
Variants likely not to affect protein structure or function. Variants with unpredicted consequences
Identity-by-descent analysis
Three pairs of samples found to be related
Keep index cases
Three related samples
PCA analysis to determine suitability of dataset
2,434 SNPs shared between the UK10K and melanoma cohort exomes and 1000 genomes Phase I (AF > 1%); after pruning for linkage disequilibrium
Dataset OK
Calculate probability (p-value) of seeing at least as many mutations in control set
Calculate substitution rate, length and capture efficiency against control set to obtain p-value.
Weight obtained p-value by co-segregation coefficient
Sanger sequencing
Top 83 genes
Confirmation of identified varians in top genes
Test co-segregation of identified variants in additional not-previously-sequenced members of pedigrees
Rank genes by number of confirmed variants and number and members of co-segregating pedigrees
Ranking considering co-segregation across multiple pedigrees, substitution rate, length and capture efficiency

Figure 2.1: **Flowchart of analysis steps followed in the search for melanoma susceptibility genes, European phase.** Steps are colour-coded depending on the place where these were done, green: Leeds or Leiden, blue: Sanger Sequencing Facility, orange: Sanger Vertebrate Resequencing team, red: Sanger Experimental Cancer Genetics team. Black indicates a ready dataset. Arrows indicate datasets entering or exiting the pipeline. Details of each step are annotated at their right.

2.1 Discovery phase

Initially, we decided to sequence the whole exomes of 41 patients from 24 melanoma-prone families that did not harbour pathogenetic variants in previously known genes. We then prioritised and captured resulting candidate genes from this phase in an extended set of 94 patients.

2.1.1 Patient selection

The families selected for sequencing all had three or more cases of melanoma. Additionally, families were preferentially sequenced if DNA was available from multiple members, if they had members that presented with multiple primary melanoma (MPM) or if melanoma presented at an early age (before the fourth decade of life) (Table 2.1 and Figure A.1.1). The families sequenced were recruited to a UK Familial Melanoma Study directed by the Section of Epidemiology and Biostatistics, University of Leeds (Leeds, UK), and the Leiden University Medical Centre (LUMC, Leiden, The Netherlands). All cases were found to be negative for pathogenetic variants in *CDKN2A* and *CDK4* at the institution of origin. 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 peripheral blood using standard methods. This work was carried out by Prof. Julia A. Newton-Bishop, Prof. D. Timothy Bishop and Dr. Mark Harland at the University of Leeds for Leeds cases, and by Assoc. Prof. Nelleke A. Gruis at LUMC for Leiden cases.

2.1.2 Exome sequencing

DNA was supplied by the institutions of origin to the Sequencing Facility at the Wellcome Trust Sanger Institute (hereinafter referred to as “Sanger”). DNA libraries were prepared from 5 µg of genomic DNA, and exonic regions were captured with the Agilent SureSelect Target Enrichment System, 50 Mb Human All Exon kit, which is a liquid-phase hybridisation method. Paired-end reads of 75 base pairs (bp) were generated on the HiSeq 2000 platform and mapped to the reference GRCh37/hg19 human genome assembly using the Burrows-Wheeler Aligner (BWA) [377] (for software versions and parameters see Table A.1). Reads were duplicate-marked using Picard [378] and were recalibrated and realigned around indels using the Genome Analysis Toolkit (GATK) package [379] (Table A.1). Exome capture and sequencing resulted in an average of 90.8% of target

Table 2.1: **Pedigrees sequenced as part of the discovery phase.** NL: The Netherlands.

Pedigree ID	Origin	Num. melanoma cases in pedigree	Num. sampled cases	Presence of MPM in pedigree	Age of diagnosis of first melanoma
UF1	Leeds, UK	4	2	Yes	28
UF2	Leeds, UK	5	1	Yes	57
UF3	Leeds, UK	4	1	Yes	37
UF4	Leeds, UK	5	1	Yes	27
UF5	Leeds, UK	4	1	Yes	25
UF6	Leeds, UK	4	1	Yes	42
UF7	Leeds, UK	4	2	Yes	36
UF8	Leeds, UK	4	1	Yes	34
UF9	Leeds, UK	5	1	Yes	25
UF10	Leeds, UK	3	3	Yes	35
UF11	Leeds, UK	4	1	Yes	18
UF12	Leeds, UK	4	1	Yes	44
UF13	Leeds, UK	4	1	Yes	42
UF14	Leeds, UK	4	2	Yes	35
UF15	Leeds, UK	8	2	No	22
UF16	Leeds, UK	4	2	Yes	50
UF17	Leeds, UK	6	1	Yes	16
UF18	Leeds, UK	5	1	Yes	25
UF19	Leeds, UK	6	2	Yes	27
UF20	Leeds, UK	5	3	Yes	21
UF21	Leeds, UK	3	2	Yes	41
NF1	Leiden, NL	4	3	Yes	25
NF2	Leiden, NL	4	2	No	42
NF3	Leiden, NL	5	4	Yes	27

bases being covered $\geq$10× across the autosomes and sex chromosomes. Genomic variants were then called using SAMtools mpileup [380] (Table A.1). This work was done by the Sequencing Facility and by pipelines written by the Vertebrate Resequencing Team at Sanger.

2.1.3 Data processing

Under the rationale that potential disease-causing germline mutations are not commonly found in human populations, I removed known variants in common variation datasets from further analyses. For this step, I used The 1000 Genomes Project, October 2011 release database [15], and The Single Nucleotide Polymorphism database (dbSNP), release 135 [381]. Additionally, I also removed all variants found in 805 in-house control exomes that belonged to either the 500 Exome Project, developed by the Metabolic Disease Group, or a set of control exomes part of the Cancer Genome Project at Sanger.

One of the best-known problems of NGS and variant-calling algorithms is the high false-positive rate of resulting variant calls [382, 383]. For this reason, it is important to remove variants for which confidence is low, a concept that is captured by the base and mapping quality scores assigned by the Illumina platform and the BWA algorithm, respectively [377, 382]. These quantities are represented in the variant's quality score, calculated by SAMtools mpileup [384], and which is given by $-10log_{10}P(\text{call is erroneous})$ [385]. In an attempt to remove false positives but at the same time keep any potentially disease-causing mutation that could be affected by low local coverage or alignment errors, I decided to remove all variants with a quality below 10. This filter ensures that we keep only those variants whose probability of being wrong is less than 1 in 10, but at the same time, this low quality cut-off warrants subsequent confirmation by Sanger sequencing before proceeding to biological validation.

Additionally, I applied other standard variant quality filters to control for known causes of false positives, such as removing variants observed in one strand more than expected by chance (P-value$\leq$0.0001), variants called predominantly close to the end of reads, where quality is known to drop (P-value$\leq$0.0001), variants supported by two or less reads, and variants with a root mean square mapping quality lower than 10. This left a total of 316,097 mutations across all samples for further analyses (Fig. 2.2).

As we have sequenced only exonic regions, I decided to keep only variants resulting in protein-altering changes. In order to predict the consequences of each variant, I used the Ensembl project's Variant Effect Predictor (VEP) tool, version 2.1 (Ensembl release 63) [386]. The types of consequences kept for further analyses are shown in Table 2.2. The code I wrote to perform these analyses, with some modifications, has been published [387].

Finally, when we sequenced more than one member of a pedigree, I retained only variants co-segregating with melanoma that were unique to that pedigree, whereas I considered all variants unique to an individual from pedigrees in which only one affected family member was sequenced. We decided to keep only variants unique to a pedigree in an effort to reduce any systematic biases arising from the sequencing, processing and variant calling methodologies [388], but have also done a separate analysis examining

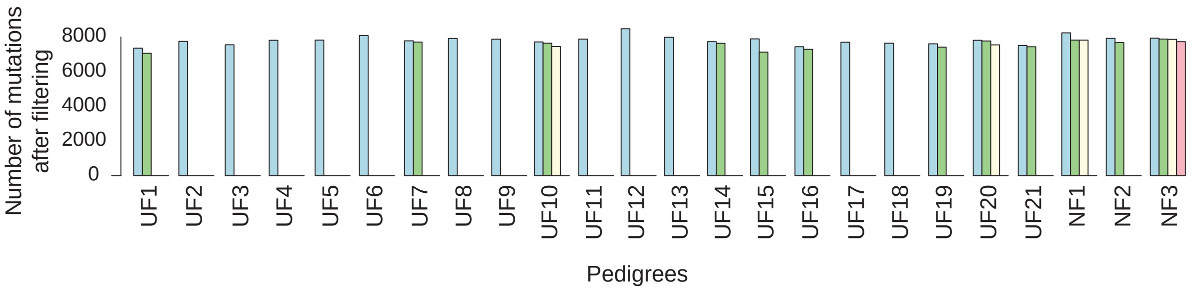

Figure 2.2: **Number of variants per exome remaining after filtering for common variation and quality, discovery phase.** Pedigree IDs are indicated below each set of bars, each bar represents one exome. Colours are used only to distinguish between different members within the same family. Similar numbers of variants were called across all samples.

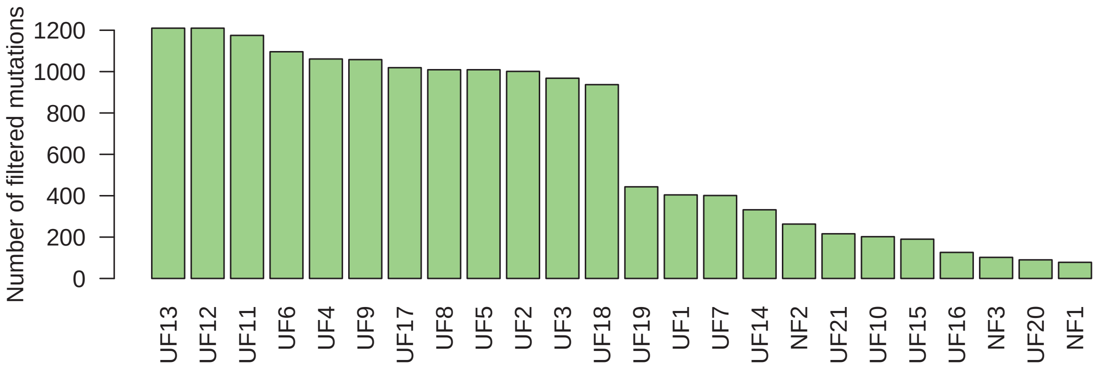

Figure 2.3: **Number of variants per pedigree remaining after filtering for co-segregation and protein-altering changes, discovery phase.** Pedigrees are ordered from the highest to the lowest number of variants passing all filters. In general, pedigrees with more members have less variants passing the filtering criteria due to the co-segregation requirement.

variants that did not pass this filter (see Subsection 2.3.4.1). This left a total of 15,600 mutations for further analyses (Fig. 2.3).

2.1.4 Gene prioritisation for replication phase

In order to define a list of candidate genes for sequencing in additional melanoma families, I decided to retain only those genes that were mutated in two or more pedigrees, which revealed 344 recurrently mutated genes after manual removal of genes likely to be false positives (see Subsection A.2.1). Genes mutated in 3 or more families for which co-

Table 2.2: **Consequences of variants kept for further analyses, discovery phase.** Table reproduced from refs. [389, 390]

Ensembl 63 term	Sequence Ontology term	Sequence Ontology description
Essential splice site	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
Stop gained	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
Frameshift coding	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
Stop lost	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
Non synonymous coding	initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript
	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence
	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved
Splice site	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron

segregation information exists are shown in Table 2.3, and the full list can be seen in Table A.1.2.

Table 2.3: **Top recurrently mutated genes, discovery phase.**

Gene name	Num. families (num. members per family)	Coding length in kilobases (kb)	Number of mutations per kb
RNF213	5 (3, 2, 2, 2, 1)	24.435	0.204624514
C10orf93	5 (2, 2, 1, 1, 1)	10.638	0.47001316
SMG1	4 (3, 2, 2, 1)	16.112	0.248262165
ADAMTS7	4 (3, 2, 1, 1)	5.497	0.727669638
ARID1A	4 (2, 1, 1, 1)	9.225	0.433604336
PRUNE2	4 (2, 1, 1, 1)	14.81	0.270087779
NPHP4	4 (2, 1, 1, 1)	13.136	0.304506699
EP400	4 (1, 1, 1, 1)	13.501	0.29627435
PLEC	4 (1, 1, 1, 1)	16.941	0.236113571
ANKK1	3 (4, 1, 1)	2.589	1.158748552
KIF26B	3 (3, 2, 1)	8.091	0.370782351
FAT2	3 (3, 1, 1)	14.514	0.206696982
ZFC3H1	3 (3, 1, 1)	9.636	0.311332503
NEBL	3 (2, 2, 1)	13.683	0.219250164
AGBL1	3 (2, 2, 1)	3.294	0.910746812
SH3TC2	3 (2, 2, 1)	15.027	0.199640647
MPHOSPH9	3 (2, 2, 1)	8.663	0.346300358
XDH	3 (2, 1, 1)	5.999	0.500083347
MYO5C	3 (2, 1, 1)	10.002	0.299940012
FN1	3 (2, 1, 1)	17.274	0.173671414
SYTL5	3 (2, 1, 1)	4.876	0.615258409
ANKRD17	3 (2, 1, 1)	12.001	0.249979168
DGKQ	3 (2, 1, 1)	4.945	0.606673407
SCN7A	3 (2, 1, 1)	8.164	0.367466928
SLC26A8	3 (2, 1, 1)	6.017	0.498587336
KIF26A	3 (2, 1, 1)	6.741	0.445037828
NCKAP5	3 (2, 1, 1)	9.681	0.309885342
PRKG1	3 (2, 1, 1)	7.637	0.392824407
RP1L1	3 (2, 1, 1)	8.875	0.338028169

To recapitulate all filtering steps so far, these 344 genes present with at least two different variants in at least two different pedigrees. Each of these variants is shared by all sequenced members of the pedigree and is absent from all other pedigrees, and is also likely to affect protein function based on its predicted consequences. We reasoned that these criteria ensure that the variant segregates with the disease while discarding any systematic mapping errors or SNPs not present in the common variation filtering sets.

Next, we reasoned that we should investigate whether there were any overrepresented biological pathways in these 344 candidate genes, and if there were, include all pathway gene members in the subsequent screening in additional families as they would represent plausible candidates. In order to do this, I ran hypergeometric tests on the list of candidates against all biological pathways in the curated gene sets in the Molecular Signatures Database (MSigDB), version 3.0, from the Gene Set Enrichment Analysis (GSEA) [391]. The hypergeometric test is a statistical tool that is able to calculate the probability of drawing k successes, out of n total draws without replacement, from a specified population of size N containing exactly K successes. Therefore, it is able to assign a P-value to the event of observing k genes belonging to a pathway of size K in $n = 344$ draws (our list of recurrently mutated genes) from $N =$ the total number of human genes. The 833 biological pathways taken as reference were annotated either by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [392] (186 pathways), Reactome [393] (430 pathways) or BioCarta [394] (217 pathways) databases, and the reference gene universe were all annotated genes in the Ensembl database, release 65 [395] ($N = 19,975$ genes).

I performed the hypergeometric tests with a custom R script, facilitated by Dr. Alistair G. Rust, from the Experimental Cancer Genetics team at Sanger, and which uses functions from the package HTSanalyzeR, version 2.8.0 [396]. After correcting for multiple tests (using Benjamini and Hochberg's method [397]), three pathways remained with an adjusted P-value≤ 0.05: the ABC transporters and the pantothenate and coenzyme A biosynthesis pathways, annotated by KEGG, and the linkage to MAPK signalling for integrins, annotated by Reactome (Table 2.4, refer to Table A.1.3 for the full list of pathways). The number of GSEA-annotated genes belonging to these pathways is 64 in total, which were added to the set of genes to capture in the replication set.

In addition to the 344 recurrently mutated genes (Table A.1.4, "*direct evidence*") and the 64 genes belonging to overrepresented pathways in these (Table A.1.4, "*ABC transporter*", "*Pantothenate and CoA biosynthesis pathway*", "*Linkage to MAPK signalling for integrins*"), we also decided to include all genes that presented with obviously disruptive consequences (essential splice site, stop gained and frameshift-coding, Table 2.2), regardless of the number of pedigrees in which these were found (202 genes) (Table A.1.4, "*disruptive consequence*"). Additionally, 91 genes that had been found previously involved in melanoma or cancer development were also included for further screening

Table 2.4: **Significantly overrepresented biological pathways in set of recurrently mutated genes, discovery set.** For all these comparisons, $n = 344$ and $N = 19,975$.

Biological pathway name	Pathway size (K)	Expected hits	Observed hits (k)	P-value	Adjusted P-value	Human gene names
ABC transporters (KEGG)	44	0.7577	6	9.37E-06	0.0061	*ABCB6*, *ABCA12*, *ABCA6*, *ABCA1*, *ABCC3*, *ABCA7*
Pantothenate and CoA biosynthesis (KEGG)	16	0.2755	3	0.0001	0.0219	*ENPP3*, *PANK4*, *BCAT1*
GRB2 and *SOS* provides linkage to MAPK signalling for integrins (Reactome)	15	0.2583	3	0.0001	0.0219	*FN1*, *FGG*, *FGB*
P130CAS linkage to MAPK signalling for integrins (Reactome)	15	0.2583	3	0.0001	0.0219	*FN1*, *FGG*, *FGB*

(Table A.1.4, "*previous evidence for involvement in melanoma/cancer*"). Therefore, the gene set to be captured in the replication cohort was composed of a total of 701 genes.

2.1.5 Custom probe design

Having the list of genes to be evaluated in additional samples, I then made a design to capture these regions by liquid hybridisation (Figure 1.16). For this step, I extracted the genomic coordinates of all exons per gene as annotated in Ensembl release 65, adding 100 bp on each side to capture potential variants in splice regions, and sent this design to Dr. Bram Herman, from Agilent Technologies, who ran scripts to generate efficient complementary probe sequences. The finished design targeted ~5.6Mb of the human genome and covered ~99.23% of the target exons. After manufacturing and sending back to Sanger, these probes were ready to be used in an extended set of melanoma patients for assessment.

2.2 Replication phase

DNA from an additional 94 cases, each from a different melanoma pedigree without *CDKN2A* or *CDK4* variants, was sent to the Sanger for targeted exon sequencing. These families constituted the remaining pedigrees with two or more melanoma cases in the Leeds and Leiden collections.

2.2.1 Targeted sequencing and data processing

Samples from the replication phase were exon-captured using the Agilent SureSelect XT custom kit, and sequenced on the Illumina HiSeq 2000 platform generating 75bp paired-end reads. Alignment, duplicate marking, recalibration, indel realignment and variant calling were done by the Sanger Sequencing Facility and Vertebrate Resequencing Team as described above (see Subsection 2.1.3; for specific parameters see Table A.1).

In order to capture the familial relationships in all samples to corroborate that sample exchanges had not occurred, a pairwise identity-by-descent (IBD) analysis was performed on the 135 melanoma samples (41 from the discovery phase and 94 from the replication) based on the polymorphic sites on these 701 genes, as annotated in dbSNP 135 [381] (14,820 positions). The rationale behind this analysis is that pairs of first-degree relatives, such as parents and offspring or siblings, are expected to share about 50% of their genome (meaning a pairwise IBD score of 0.5, approximated by the amount of shared SNPs); second-degree relatives are expected to have a pairwise IBD score of 0.25, and so on, and thus this score is able to discriminate between related and unrelated samples. So, after keeping only positions with allelic frequency > 0.05 and $r^2 < 0.05$ (an estimator of pairwise linkage disequilibrium [LD]), only 722 SNPs remained for further analyses. These filters are necessary to satisfy assumptions of the distribution of allele frequencies in the population, and thus to calculate accurate IBD results. All expected familial relationships, which were all anticipated to have a pairwise IBD score of 0.25 or higher, were captured with a cut-off value of 0.14. This low value might be due to the small number of SNPs used for the analysis, as this fact is expected to increase background noise (Fig. 2.4). With this analysis, we were able to detect contamination in one sample. This sample and another one that was found during the course of this analysis to be an unaffected sibling from a melanoma patient were excluded, leaving 92 samples in the replication phase. The IBD analysis was performed by Jimmy Z. Liu at the Sanger, using the genome-wide complex trait analysis (GCTA) tool [398].

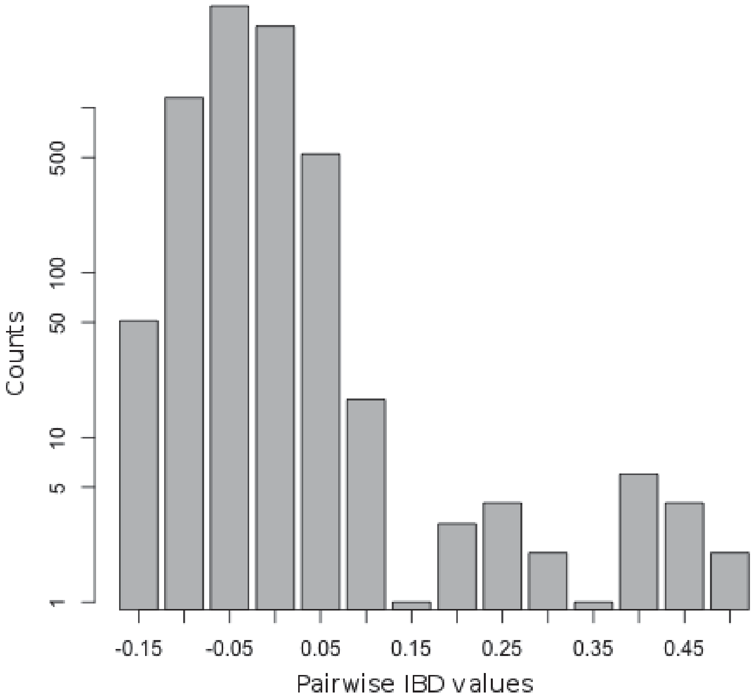

Figure 2.4: **Distribution of pairwise IBD values for samples in the discovery and replication sets.** The x-axis shows pairwise IBD values, whereas the y-axis shows counts of pairwise comparisons (out of $\begin{pmatrix} 135 \\ 2 \end{pmatrix}$). Values are centered around 0.45 for first-degree relatives, around 0.25 for second-degree relatives, and around 0 for unrelated individuals. The wide distributions observed are due to the low number of SNPs used for the analysis.

I then applied the same variant filtering steps as in the discovery set analysis explained above (common variation, quality and variant consequence filters).

Table 2.5: **Control cohorts used in the melanoma gene prioritisation stage, European phase.** Explanations are taken from the UK10K website (`www.uk10k.org`).

Cohort name	Description	Number of samples
UK10K Neuro Muir	Sample consists of subjects with schizophrenia, autism, or other psychoses all with mental retardation (learning disability)	167
UK10K Neuro IOP Collier	Sample consists of samples from subjects with schizophrenia, psychotic symptoms, or bipolar disorder. Set is of UK origin.	112
UK10K Neuro Aberdeen	Sample set comprises cases of schizophrenia with additional cognitive measurements, collected in Aberdeen, Scotland.	267

2.3 Gene prioritisation strategy

2.3.1 Gene ranking methodology

We then decided to devise a method to prioritise genes based on the likelihood of observing the number of mutations found in melanoma patients when compared to a set of controls. The resulting strategy takes into account the number of non-synonymous variants detected, the coding length of the gene and the exonic capture efficiencies in both the discovery and replication phases and the controls.

2.3.1.1 Choice of control exomes

As controls, we decided to use all samples from three neurodevelopmental cohorts from the UK10K Sequencing Project [399], release 14/03/2012, consisting of a total of 546 exomes (Table 2.5). These samples were chosen because, in addition to presenting with a phenotype unrelated to cancer, they were captured with the same Agilent SureSelect exome probes as those used for the melanoma cases described above, and were also sequenced on the Illumina HiSeq 2000 platform.

Control exomes were aligned, filtered for duplicates, recalibrated and realigned around indels as described above. I then called variants and filtered for common variation, quality and consequences with the same tools and parameters as the melanoma cohort. As there were three pairs of siblings across the three cohorts, I decided to keep one index case from each of these, thus ensuring the samples were not related. This left 543 exomes for further analyses.

2.3.1.2 Principal component analysis to ensure that cases and controls are matched by ancestry

Before being able to compare allele frequencies in cases and controls, it is necessary to ensure that these two groups are matched by ancestry. Not accounting for population structure is a frequent source of false-positive results and reduced power in genetic studies [400]. In order to ensure this, I decided to perform a principal component analysis (PCA) on the melanoma cases and the UK10K controls to group individuals on the basis of their genomic variation. For this, I obtained a set of SNPs, with an allelic frequency higher than 1% in The 1000 Genomes Project dataset, that were shared between the sequenced melanoma cases (discovery and replication sets) and the UK10K controls. This allele frequency filter is necessary, as otherwise populations can be grouped on particular chromosomal segments instead of on genome-wide population structure [400]. This filter left 2,434 bi-allelic positions spread across the 701 genes that were captured.

Then, with a custom R script supplied by Mamunur Rashid at the Experimental Cancer Genetics team at Sanger, I was able to plot the cases and controls according to their genetic variation. Briefly, the program converts genotypes to 0 if the individual is homozygous for the reference allele, 1 if heterozygous, and 2 if homozygous for the variant allele, and then normalises the resulting matrix. Then, it obtains the pairwise covariance matrix and estimates its eigenvalues and eigenvectors. It obtains the loadings for each SNP based on The 1000 Genomes dataset, and then applies them to the melanoma cases and controls. The resulting plot shows that, to the best resolution we could obtain with the small number of shared SNPs, cases and controls are ancestry-matched as they are all European (Fig. 2.5).

2.3.1.3 Gene prioritisation

Having shown that cases and controls are matched by ancestry, I was able then to compare the number and types of germline variants detected in both datasets. I thank Drs. Jeroen de Ridder from the Delft University of Technology and Kees Albers, from the Sanger, for very useful discussions in devising the following prioritisation strategy.

For each gene g we calculate $\bar{\mu}_g^{nuc}$, which is the average per-nucleotide variant rate for gene g in the control set. $\bar{\mu}_g^{nuc}$ is calculated by counting the number of non-synonymous variants detected in an individual, and then dividing it by the number of exonic bases

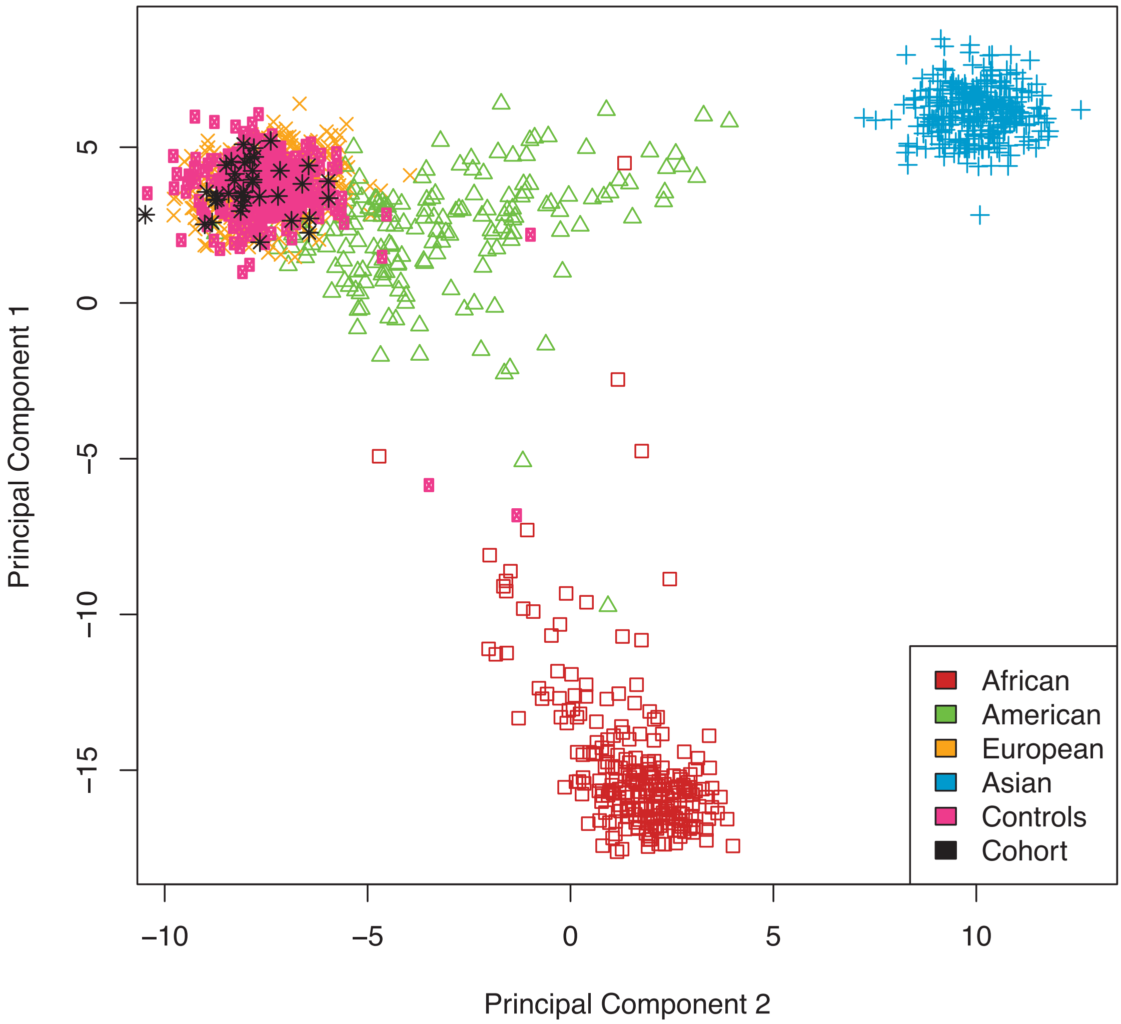

Figure 2.5: **Principal component analysis plot showing that cases and controls are matched by ancestry, European phase.** Plot showing the first and second principal components. Ancestry was estimated using the 1000 Genomes Project individuals and then projected onto the melanoma (black) and UK10K control (pink) cohorts.

in that individual that were captured with a coverage of at least 2. Therefore,

$$\bar{\mu}_g^{nuc} = \frac{1}{n} \sum_{k=1}^{k=n} \frac{m_{g,k}}{b_{g,k}},$$

where n is the number of control individuals ($n = 543$ in this case), $m_{g,k}$ is the number of non-synonymous mutations detected in gene g in individual k, and $b_{g,k}$ is the number of bases with a coverage of at least 2 in gene g in individual k.

We can then use $\bar{\mu}_g^{nuc}$ to calculate $\mu_{g,s}$, which may be interpreted to be the rate at which we find at least one nucleotide variant in gene g in study phase s, if the mutation events are independent:

$$\mu_{g,s} = 1 - (1 - \bar{\mu}_g^{nuc})^{L_{g,s}},$$

where $L_{g,s}$ is the average captured coding length of gene g in nucleotides in study phase s, and $s \in S$ where $S = \{\text{discovery, replication}\}$. The length of gene g is not taken as a constant, as it can change across different study phases because of variations in target capture efficiency. Then, the probability that at least $X_{g,s}$ out of Y_s individuals have at least one variant in gene g is:

if $X_{g,s} \neq 0$,

$$P\text{-value}_{g,s} = 1 - \sum_{j=0}^{j=X_{g,s}-1} \begin{pmatrix} Y_s \\ j \end{pmatrix} \mu_{g,s}^j (1 - \mu_{g,s})^{Y_s - j},$$

else,

$$P\text{-value}_{g,s} = 1.$$

Finally, we obtain a P-value for gene g as

$$P\text{-value}_g = \prod_{s \in S} P\text{-value}_{g,s}.$$

The P-value calculated from the steps above attempts to capture the likelihood of observing as many variants as we detect in the melanoma exomes, or more, when compared to a control dataset, and utilise an index case from each pedigree in the calculations as we are assuming that mutation events are independent (which means that $Y_{\text{discovery}} = 24$ and $Y_{\text{replication}} = 92$). However, it does not capture any co-segregation information. In order to take into account this information, we then decided to correct this P-value for the likelihood that multiple members in a pedigree share each variant in gene g:

$$\text{score}_g = P\text{-value}_g \times \prod_{p \in P} C_p,$$

where C_p is the co-segregation coefficient for pedigree p, and P is the set of all pedigrees where gene g was found with at least one mutation in the discovery phase. C_p captures the probability that, given a pedigree structure, a given pair of relatives share a variant. Therefore, $C_p = 0.5$ for a pair of first-degree relatives, $C_p = 0.25$ for a pair of second-degree relatives, and so on. In general, $C_p = \frac{1}{2^m}$, where m is the highest number of meioses separating any two members of pedigree p.

The above methodology will assign $\text{score}_g = 0$ to genes that only have one detected mutation in either the replication or the discovery phase if no mutations are detected in the control exomes. So while they might be interesting, we regard genes with only one detected variant as uninformative. The top 30 genes resulting from this ranking after removing uninformative genes are shown in Table 2.6, and the full list can be consulted in Table A.1.5.

Before proceeding to biological validation, it is of paramount importance that the variants detected in these genes are confirmed as real. This is because we decided to keep variants that have a probability of being wrong of up to 1 in 10 given sequencing data (discussed in Subsection 2.1.3).

2.3.2 Validation of next-generation sequencing-detected variants

In order to validate detected variants in candidate melanoma susceptibility genes, polymerase chain reaction (PCR) primers were designed against all variants in the top 83 genes according to the ranking in Table A.1.5 (the top 30 genes are shown in Table 2.6). Genomic DNA from carriers was amplified and capillary-sequenced in order to confirm the presence or absence of NGS-detected germline variants. An overall confirmation rate of almost 86% was achieved, as 522 out of 608 putative tested variants were detected in the original sample (Fig. 2.6). The PCR validation work was done by Dr. Mark Harland at the University of Leeds.

It would seem that a lower-than-expected confirmation rate was achieved, as only about ~18% and 45% of the variants with quality scores between 10 and 20 and those between 20 and 30 were confirmed, respectively. However, it is important to take into account that the quality score gives the likelihood of observing the reported genotype

Table 2.6: **Top 30 genes after prioritisation, European phase.** P-values were not corrected for multiple tests as we did not use them to assess significance of mutational events but for producing a ranked list of candidates.

g	P-value$_g$	score$_g$
MAGEB1	0	0
SPINK2	0	0
LCN1	0	0
BEST1	1.32E-06	6.62E-07
NFE2L3	3.98E-06	1.99E-06
NEBL	4.03E-05	5.04E-06
CACNA1E	1.25E-05	6.27E-06
WBP11	6.44E-05	8.05E-06
C1orf93	1.08E-05	1.08E-05
PASK	3.03E-05	3.03E-05
ZNF160	0.000290599	3.63E-05
CTSA	0.000154824	3.87E-05
SOX17	0.000316849	7.92E-05
MPHOSPH9	0.001369208	8.56E-05
SYTL5	0.000205992	0.000102996
KLHDC8A	0.001758527	0.000109908
C6orf25	0.000529058	0.000132264
NECAB3	0.000154336	0.000154336
BIN1	0.000164977	0.000164977
NBR2	0.000178824	0.000178824
RNF213	0.026272146	0.000205251
TMC2	0.000875866	0.000218967
CRIP2	0.000228709	0.000228709
NAT10	0.000275362	0.000275362
PCDH15	0.000700415	0.000350207
SMG1	0.023773426	0.00037146
ITIH5	0.000390115	0.000390115
PDZD7	0.000858102	0.000429051
AGAP3	0.00044007	0.000440073
SCLT1	0.00046525	0.000465246

given the sequencing data and their quality, and that no systematic biases are considered in its calculation. Therefore, it is unable to predict the experimental validation rate.

Based on the experimental validation rate, a modified prioritisation list was compiled from Table 2.6. This re-ranked list takes as basis the methodology performed above, as only variants in previously highly-ranked genes were tested. Additionally, we were

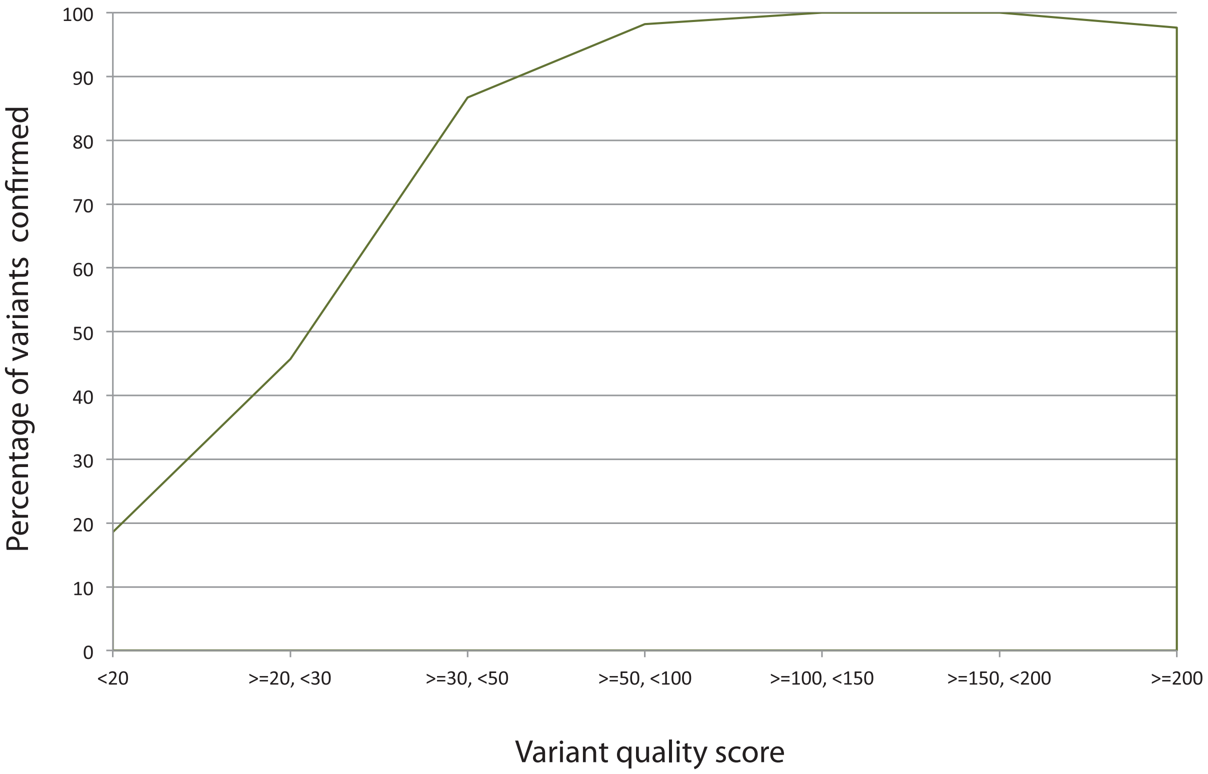

Figure 2.6: **Percentage of variants confirmed by PCR to be real grouped by quality score.** Overall, 608 variants were tested, and 522 were confirmed to be real.

able to test for co-segregation in additional members of several pedigrees, that were not sequenced in the discovery or replication phases. The final candidate list from this prioritisation strategy, ranked by the number of variants co-segregating in pedigrees, is shown in Table 2.7, along with the Gene Ontology terms from each gene.

2.3.3 Development of a visualisation tool for germline variants

The gene ranking methodology described above takes into account the number and types of mutations detected in a gene, the likelihood of seeing as many as those mutations in a control dataset, and the probability that those variants co-segregate in a given pedigree. However, it does not take into account the positions within a protein where these mutations lie, or if they are found in other common variation datasets that we did not use in the filtering steps. For example, a gene with three different variants disrupting one functional domain might be more biologically relevant than one with five mutations scattered throughout the protein.

In order to take these concerns into account, I wrote a programme capable of taking a list of genomic variants (specifying only chromosome, variant position, base change and strand) and outputting a schematic diagram showing where these mutations lie in protein context. The programme plots all translatable transcripts per gene along with their functional domains, and shows variants alongside a colour code indicating whether they are found in any user-specified variation datasets (specified in variant call format [VCF] when running the program). Optionally, the user can specify distinct variation files for distinct study phases, which the programme draws in different colours.

I wrote this piece of software in the Perl programming language, using the Graphic Design (GD) module for plotting [401], the Ensembl VEP for variant prediction [386] and the Ensembl Perl Application Programme Interface (API) [395] for obtaining information about transcripts and protein structure. An illustrative example can be seen in Fig. 2.7.

With this tool, I could easily and quickly generate plots for manual inspection, for all 701 genes. Examination of PCR-validated variants in the top 25 genes, however, revealed no evident mutational patterns.

2.3.4 Other analyses

2.3.4.1 Founder mutation analysis

In the discovery phase and subsequent gene prioritisation, we only considered genes that had two or more variants in different pedigrees. Additionally, we required these variants to be unique to a pedigree. We reasoned that this strategy would highlight genes with multiple, rare and potentially causal variants in the melanoma cohort while discarding any systematic biases arising from the sequencing, processing and variant calling methodologies [388].

However, it can also be the case that a causal variant that would normally be rare has become more common with small population isolation and interbreeding, as has been shown for variants in *BRCA1* and *BRCA2* [403]. Such variants would have been missed from the above analysis. In order to address this issue, we performed the same filtering steps described above but we lifted the requirement for the variant to be unique, as well as the filter requiring it to co-segregate with melanoma in more than one pedigree.

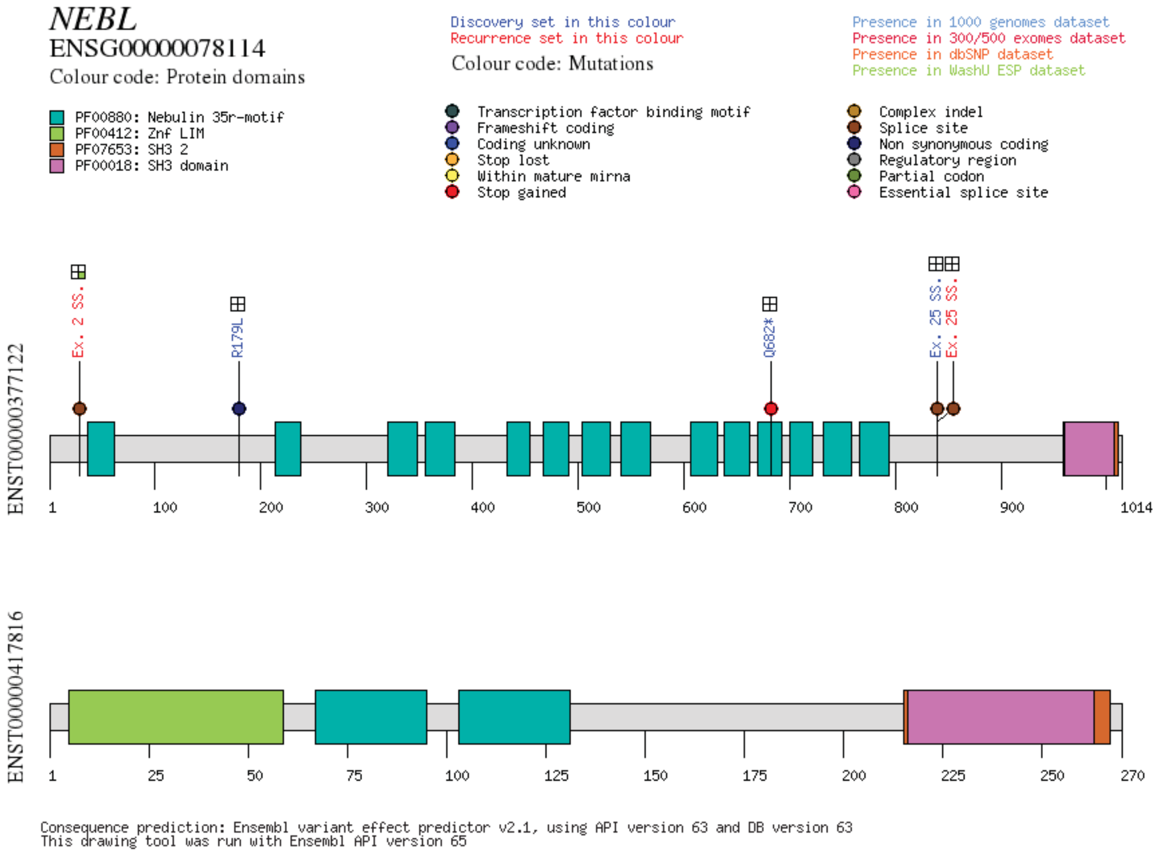
NEBL
ENSG00000078114
Colour code: Protein domains
PF00880: Nebulin 35r-motif
PF00412: Znf LIM
PF07653: SH3 2
PF00018: SH3 domain
Discovery set in this colour
Recurrence set in this colour
Colour code: Mutations
Transcription factor binding motif
Frameshift coding
Coding unknown
Stop lost
Within mature mirna
Stop gained
Presence in 1000 genomes dataset
Presence in 300/500 exomes dataset
Presence in dbSNP dataset
Presence in WashU ESP dataset
Complex indel
Splice site
Non synonymous coding
Regulatory region
Partial codon
Essential splice site
Ex. 2 SS.
R179L
Q682*
Ex. 25 SS.
Ex. 25 SS.
ENST00000377122
1
100
200
300
400
500
600
700
800
900
1014
ENST00000417816
1
25
50
75
100
125
150
175
200
225
250
270
Consequence prediction: Ensembl variant effect predictor v2.1, using API version 63 and DB version 63
This drawing tool was run with Ensembl API version 65

Figure 2.7: **Example of a protein plot showing NGS-detected variants.** The *NEBL* gene was chosen for illustrative purposes. For brevity, only two transcripts are shown, although the original plot depicts five. The gene name and associated Ensembl ID are shown at the left, followed by a list of domains (in this case, from the Pfam database [402]). Types of consequences are shown at the right, along with the colour code for different phases in the study. Four common variation datasets were fed into the program (shown at the top right corner in different colours). The four squares next to a variant indicate whether it is found in any of the variation datasets: If it is, the square is filled with the colour corresponding to that dataset, otherwise it is blank. One variant (Ex. 2 S.S.) is found in the NHLBI GO ESP dataset [17] (represented by a green square). The three variants detected in the discovery phase (Table 2.3) are depicted in blue, whereas two variants detected in the replication phase are shown in red. Ensembl transcript IDs are indicated at the left of each plot. Versions of software used to run the programme are indicated at the bottom. Ex: Exon, SS: Splice site.

Table 2.7: **Gene prioritisation after validation of NGS-detected variants.** The numbers of variants that were found to co-segregate and not co-segregate in melanoma-prone pedigrees are indicated, as well as the number of variants successfully tested. These variants include those detected in the replication phase. Note that some pedigrees might have additional members that were not sequenced but were PCR-tested. Some variants failed at the confirmation stage and could not be assessed. Only genes from The Gene Ontology (GO) terms in this table are representative from the full list, and were extracted from Ensembl release 65.

Gene	Number of pedigrees with confirmed co-segregating variants (number of members in each pedigree)	Confirmed variants not co-segregating	Variants successfully tested	GO terms
SMG1	4 (4, 3, 3, 2)	0	6	DNA repair, response to stress, nucleotide binding, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, protein serine/threonine kinase activity, protein binding, ATP binding
RNF213	4 (3, 3, 2, 2)	5	9	Nucleotide binding, protein binding, ATP binding, zinc ion binding, nucleoside-triphosphatase activity
PDZD7	3 (3, 2, 2)	2	6	Protein binding, nucleus, cilium
NFE2L3	3 (2, 2, 2)	0	3	Sequence-specific DNA binding transcription factor activity, transcription from RNA polymerase II promoter
CTSA	3 (2, 2, 2)	0	3	Serine-type carboxypeptidase activity
KLHDC8A	2 (3, 3)	0	2	Protein binding
C6orf25	2 (3, 3)	0	2	Receptor activity, endoplasmic reticulum, heparin binding
ZNF160	2 (3, 2)	0	2	DNA binding, regulation of transcription, zinc ion binding, hemopoiesis
MPHOSPH9	2 (3, 2)	0	2	M phase of mitotic cell cycle, Golgi membrane

CACNA1E	2 (2, 2)	0	3	Behavioural fear response, regulation of heart rate, voltage-gated calcium channel activity, visual learning, sensory perception of pain, sperm motility
WBP11	2 (2, 2)	0	2	Single-stranded DNA binding, RNA processing, protein phosphatase type 1 regulator activity
PCDH15	2 (2, 2)	0	2	Photoreceptor outer segment, startle response, morphogenesis of an epithelium, calcium ion binding, cell adhesion, visual perception, locomotory behavior
NECAB3	1 (4)	0	1	Calcium ion binding, Golgi cis cisterna, oxidoreductase activity, antibiotic biosynthetic process, regulation of amyloid precursor protein biosynthetic process
BEST1	1 (2)	0	3	Chloride channel activity, visual perception
NEBL	1 (2)	1	2	Structural constituent of muscle, regulation of actin filament length
PASK	1 (2)	0	3	Nucleotide binding, protein kinase activity, signal transducer activity, ATP binding

SOX17	1 (2)	0	3	Negative regulation of transcription from RNA polymerase II promoter, angiogenesis, vasculogenesis, cardiogenic plate morphogenesis, negative regulation of Wnt receptor signalling pathway involved in heart development, sequence-specific DNA binding transcription factor activity
SYTL5	1 (2)	0	1	Intracellular protein transport, Rab GTPase binding, metal ion binding

We found 60 variants predicted to affect protein sequence that were present in more than one pedigree for which co-segregation information was available (Table A.1.6). Unsurprisingly, many of these seem to be systematic errors after removal of common SNPs in common variation databases, especially those found in multiple pedigrees (see Subsection 2.1.3) (Fig. 2.8). Consistent with these types of error, almost all of these variants are insertions or deletions, with only two missense variants not involving these types of mutational events (Table A.1.6). These two variants are located in WAS protein family homolog 6 pseudogene (*WASH6P*) and Golgin A6 family-like 10 (*GOLGA6L10*). *WASH6P* was found mutated in pedigrees UF15, UF16 and UF19 (Fig. A.1.1), and is located in chromosome X. So far, although somatic inactivation of X-linked tumour suppressors has been demonstrated (*e.g.*, for forkhead box P3 [*FOXP3*] and APC membrane recruitment protein 1 [*AMER1*]), no examples of germline inactivation of X-linked tumour suppressors have been found. This might be because these genes seem to be homozygous lethal (reviewed in [404]). That it acts as a proto-oncogene cannot be discarded; however, this is unlikely given that *WASH6P* is classified as a pseudogene [405]. There are no reports for the function of *GOLGA6L10*, although its annotation has disappeared from the most current Ensembl database release (76), indicating that the gene model has changed dramatically from the Ensembl version used to call variants (65).

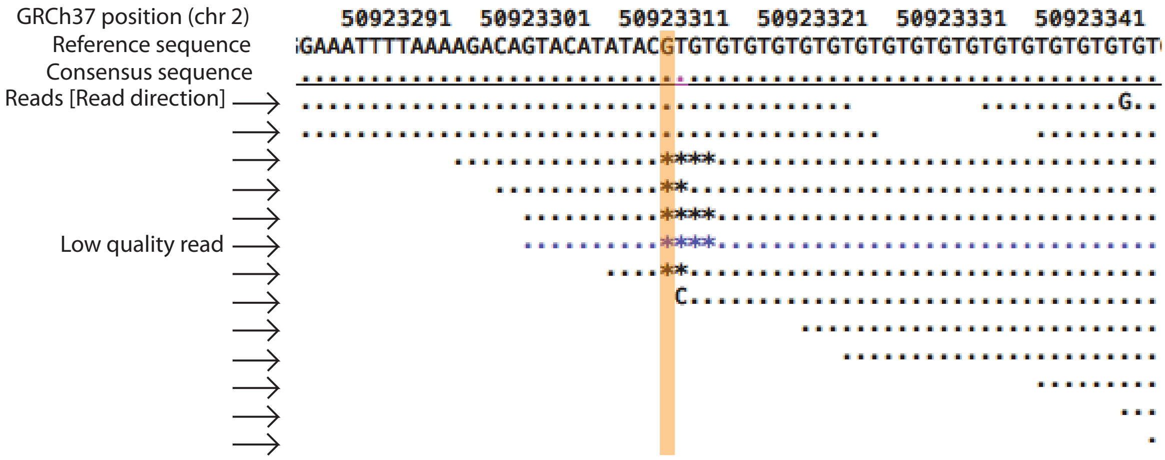

Figure 2.8: **Example of a putative founder variant found in melanoma pedigrees.** This variant is found co-segregating in 10 different melanoma pedigrees, the same or more than any other variant detected in this analysis. The position of the variant is indicated in orange. Dots indicate that the base matches the reference, asterisks indicate deletions. Note that the variant is found in a low-complexity region (a long stretch of GT repeats). These variants are probably mapping errors, as three different genotypes are observed (four or two bases deleted, and no deletion). Chromosome position, genome assembly, consensus sequence as calculated by SAMtools (before variant calling), and read orientation are indicated. Note there are only forward reads in this region; a low-quality read is indicated. Sequence data is displayed with SAMtools' tview tool [380].

There were 441 variants found to co-segregate in only one pedigree, corresponding to 318 genes that were absent from the prioritisation stage (Table A.1.7). Some of these might represent interesting candidates, but additional information is needed to assess their significance (such as biological function). Some of these are covered in more detail in the Discussion.

2.3.4.2 Mutations in known loci

During the course of these analyses, we were also interested to see whether these families had any novel variants in other known melanoma susceptibility genes that were not assessed at the institution of origin. As explained in Subsection 2.1.1, patients recruited to this study were previously found to be negative for any pathogenetic variants in *CDKN2A* and *CDK4*, and we were able to corroborate this when examining their NGS data. However, other high- and medium-penetrance susceptibility genes have been described, such as *RB1*, *BAP1*, the promoter of *TERT*, *BRCA2*, *MITF* and *MC1R* (discussed in Subsection 1.6.3).

BAP1 was part of the genes captured in the replication set, however, no missense, essential splice site, frameshifts or gains of stops were detected in any family in the whole cohort (representing a total of 116 pedigrees). One variant in *BRCA2*, encoding a proline to leucine change at position 2107 (P2107L), that passed our filters, was detected in one individual part of the replication dataset. This variant seems to be novel, as it is not found in the Universal Mutation Database for *BRCA2* [406] or the Leiden Open Variation Database (LOVD), from the IARC [407]. *RB1* was not captured in the replication dataset, and no germline variants passing our filters were found in any of the pedigrees in the discovery dataset.

To analyse variants in the *TERT* promoter that were not captured in the exome dataset, Mia Petljak from the Sanger and Dr. Mark Harland from the University of Leeds PCR-amplified and capillary-sequenced the genomic region described in the original study in all families, using the same primers [297]. We were able to identify one family, UF19, carrying the same germline variant in both members for whom DNA was available (Fig. A.1.1). An analysis of SNPs within the haplotype comprising this region in both the German family described originally and UF19 suggested that this variant arose independently, as no common variants were found (data not shown).

As the *MITF* medium-penetrance variant E318K was found not to co-segregate perfectly with melanoma [358], we decided to lift the co-segregation requirement when searching for this variant in the melanoma pedigrees. All three members of family UF10 and one member of family UF16 were found to be carriers (Fig. A.1.1), as well as three individuals from the replication dataset.

Three different variants in *MC1R* were found in individuals from the discovery set: The individual part of UF4 is heterozygous for R213W, a polymorphism that has been observed in previous melanoma studies, although no concrete association exists between this variant and melanoma risk [408, 409]. Other variants we identified in the discovery phase were R109W, in one member of the four-member pedigree NF3, and the frameshift variant F179ins c.537_538insC in one member of UF1, that has been observed previously but for which no risk association has been found [410]. The variants found in samples from the replication set are shown in Table 2.8. These have been found previously to be risk factors for melanoma development [411, 412].

Table 2.8: **List of *MC1R* variants found in samples from the replication dataset.**

Genomic variant	Protein change	Number of samples
16:89985750, C/CA	c.85_86insA	4
16:89985844, G/T	V60L	22
16:89985918, C/A	D84E	4
16:89985940, G/A	V92M	14
16:89986091, G/A	R142H	5
16:89986117, C/T	R151C	30
16:89986130, T/C	I155T	2
16:89986135, A/C	T157P	2
16:89986144, C/T	R160W	18
16:89986154, G/A	R163Q	7
16:89986252, T/C	F196L	1
16:89986546, G/C	D294H	11

2.4 Summary and conclusion

During this phase of the study, we tested different methodologies in order to reduce vast amounts of genomic data into a set of plausible melanoma susceptibility candidate genes for biological testing. We considered diverse criteria, such as the number and types of mutations found in a gene, the allelic frequency of these, the likelihood of finding those variants in a matched control population, the probability that members within an affected pedigree share the variant, the occurrence of different mutations within functionally-relevant portions of a protein, and the biological function of the gene (as given by GO terms). We reasoned that a high score in these attributes might be predictive of the involvement of a gene in familial susceptibility to melanoma. We also developed novel gene prioritisation strategies, including a software tool to graphically assess the impact and novelty of variants detected by NGS on protein structure.

The gene at the top of our list, SMG1 phosphatidylinositol 3-kinase-related kinase (*SMG1*), was found to have rare variants co-segregating in four different pedigrees, one with four tested members. This is the main reason why this gene is the first candidate in our ranking, as not all variants detected in genes scoring higher (Table 2.6) were validated by PCR and capillary sequencing (Table 2.7). Additionally, it is a biologically plausible candidate given that it participates in DNA repair and response to UV-induced DNA damage [413]. Follow-up experiments on SMG1 are described in Chapter 4. Various variants in known melanoma loci were also detected, especially a novel variant in *BRCA2*, a recently reported causal variant in the promoter of *TERT*, and several variants in the medium-penetrance loci *MITF* and *MC1R*.

During the course of this phase of the study, an extensive set of samples from Australian pedigrees became available for analysis. The number of individuals in this dataset more than doubles the number of UK and Dutch samples used in this phase, and therefore, we performed a different set of analyses to study them. This new, integrative phase, is described in the next chapter.

Chapter 3

Familial melanoma sequencing: Integrative phase

Methods and results of this chapter have been published or are accepted for publication (refs. [376] and [414]). Some parts of the text have been reproduced from these references; I confirm I have ownership of copyright for reproduction in this work.

While I was studying the European samples discussed in Chapter 2, a large dataset of multi-case Australian samples became available for analysis, which more than doubled the initial dataset. This Chapter describes the clinical characteristics of the pedigrees and samples in this new dataset and explores the integrative analysis rationale we followed, as well as the potential melanoma susceptibility candidates that were uncovered. An overview of the steps followed in this phase is depicted in Fig. 3.1.

3.1 Patient selection

Australia is the country with the highest melanoma incidence in the world [369]. For this reason, criteria for the collection of melanoma families in the UK and Australia vary, as melanoma risk factors such as the presence of atypical naevi are much more common in the Australian population [415]. This suggests that sunlight-induced phenocopies occurring in melanoma-predisposed families might confound the search for high-penetrance susceptibility genes.

To account for this factor, the great majority of pedigrees sequenced for this phase had five or more melanoma cases, which reduces the probability that the aggregation of cases observed within families was due to clustering of sporadic cases. The additional

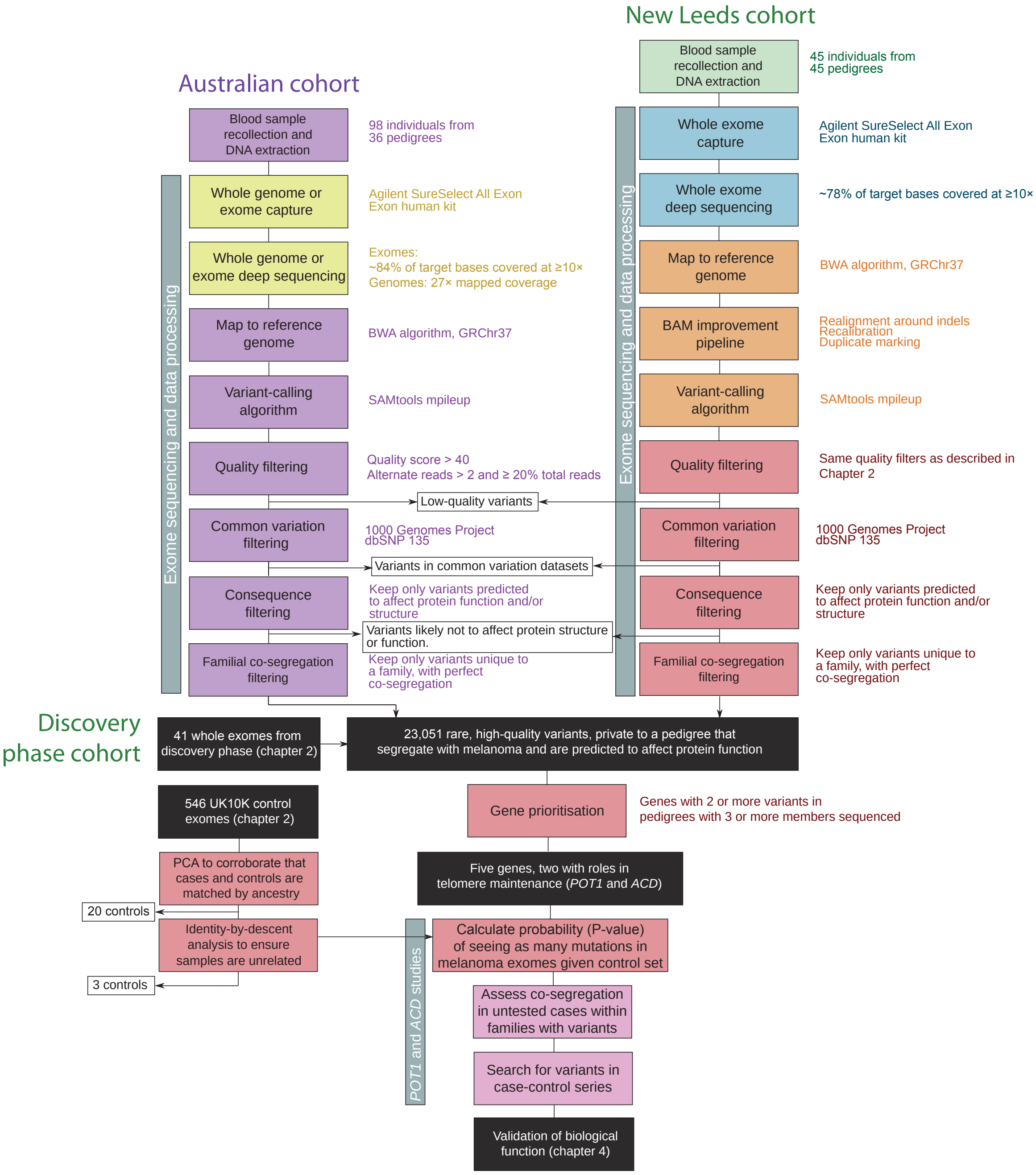
New Leeds cohort
Blood sample recollection and DNA extraction
45 individuals from 45 pedigrees
Whole exome capture
Agilent SureSelect All Exon Exon human kit
Whole exome deep sequencing
~78% of target bases covered at ≥10×
Map to reference genome
BWA algorithm, GRChr37
BAM improvement pipeline
Realignment around indels
Recalibration
Duplicate marking
Variant-calling algorithm
SAMtools mpileup
Quality filtering
Same quality filters as described in Chapter 2
Common variation filtering
1000 Genomes Project
dbSNP 135
Consequence filtering
Keep only variants predicted to affect protein function and/or structure
Familial co-segregation filtering
Keep only variants unique to a family, with perfect co-segregation
Exome sequencing and data processing
Australian cohort
Blood sample recollection and DNA extraction
98 individuals from 36 pedigrees
Whole genome or exome capture
Agilent SureSelect All Exon Exon human kit
Whole genome or exome deep sequencing
Exomes:
~84% of target bases covered at ≥10×
Genomes: 27× mapped coverage
Map to reference genome
BWA algorithm, GRChr37
Variant-calling algorithm
SAMtools mpileup
Quality filtering
Quality score > 40
Alternate reads > 2 and ≥ 20% total reads
Low-quality variants
Common variation filtering
1000 Genomes Project
dbSNP 135
Variants in common variation datasets
Consequence filtering
Keep only variants predicted to affect protein function and/or structure
Variants likely not to affect protein structure or function.
Familial co-segregation filtering
Keep only variants unique to a family, with perfect co-segregation
Exome sequencing and data processing
Discovery phase cohort
41 whole exomes from discovery phase (chapter 2)
23,051 rare, high-quality variants, private to a pedigree that segregate with melanoma and are predicted to affect protein function
546 UK10K control exomes (chapter 2)
Gene prioritisation
Genes with 2 or more variants in pedigrees with 3 or more members sequenced
PCA to corroborate that cases and controls are matched by ancestry
Five genes, two with roles in telomere maintenance (*POT1* and *ACD*)
20 controls
Identity-by-descent analysis to ensure samples are unrelated
Calculate probability (P-value) of seeing as many mutations in melanoma exomes given control set
3 controls
Assess co-segregation in untested cases within families with variants
Search for variants in case-control series
Validation of biological function (chapter 4)
POT1 and *ACD* studies

Figure 3.1: **Flowchart of analysis steps followed in the search for melanoma susceptibility genes, integrative phase.** Steps are colour-coded depending on the place where these were done, purple: QIMR Berghofer, yellow: Macrogen, Inc., green: Leeds or Leiden, blue: Beijing Genomics Institute, orange: Sanger Vertebrate Resequencing team, red: Sanger Experimental Cancer Genetics team, pink: Leeds and QIMR Berghofer. Black indicates a ready dataset. Arrows indicate datasets entering or exiting the pipeline. Details of each step are annotated at their right.

samples included in this study were recruited to the Queensland Familial Melanoma Project (QFMP) [416], and informed consent was obtained from the Human Research Ethics Committee of the QIMR Berghofer Medical Research Institute. Genomic DNA was extracted from peripheral blood using standard methods. This work was done by Lauren G. Aoude, Antonia L. Pritchard, Jane M. Palmer, Judith Symmons and Prof. Nicholas K. Hayward at QIMR Berghofer, Queensland, Australia.

In addition to this dataset, we decided to sequence the whole exomes of 45 additional samples from Leeds: 31 that had a family history of the disease and 14 single cases. Twenty-one of the samples with a family history had been already included in the replication set, but we only had sequence for 701 genes in these samples. Single cases were selected for sequencing if they presented with either MPM or an early age of onset ($\leq$40 years of age), which strongly suggests a genetic component to their disease.

Therefore, our new, integral dataset comprises 184 samples in which the whole protein-coding genome has been sequenced: 41 from the discovery set, 98 Australian samples, and 45 additional Leeds samples, belonging to a total of 105 pedigrees. These samples were processed in different institutes and sequencing centres (Tables 3.1 and A.1.8).

3.2 Exome sequencing and data processing

Sixteen samples from the Australian cohort underwent whole genome sequencing, and the rest were whole exome-sequenced (Table A.1.8). In both cases, DNA libraries were prepared from 5µg of genomic DNA. For whole exomes, exonic regions were captured with the Agilent SureSelect Target Enrichment System, 50 Mb Human All Exon kit, and for whole genomes, libraries were prepared using the standard Illumina library preparation protocol. Then, 100bp paired-end reads were generated on the Illumina HiSeq2000 platform. These samples were processed at Macrogen, Inc. The additional samples from the Leeds cohort were sequenced at the Beijing Genomics Institute (BGI), and were also captured with the Agilent SureSelect Target Enrichment System, 50 Mb Human All Exon kit and sequenced in the Illumina HiSeq2000 platform, generating 90bp paired-end reads. Finally, both sample sets were mapped to the reference GRCh37/hg19 human genome assembly using the Burrows-Wheeler Aligner (BWA) [377], and Leeds samples were further recalibrated and realigned around indels using the GATK package [379].

For Leeds samples, exome capture and sequencing resulted in an average of almost

Table 3.1: **Summary of pedigrees sequenced by Institute and sequencing centre, integrative phase.** Cases marked with an asterisk were selected for an absence of phenotypic risk markers (self-reported sun sensitivity and/or low mole count). The case marked with two asterisks was selected because they presented with three different primary cancers, one of which was melanoma. BGI: Beijing Genomics Institute.

Institutes	Leeds (UK)		Leiden (NL)	QFMP (Australia)	Total
Sequencing centres	Sanger	BGI	Sanger	Macrogen	
Familial pedigrees					
5+ cases	8	0	1	25	34
4 cases	11	1	2	5	19
3 cases	2	12	0	3	17
2 cases	0	18	0	2	20
Total	21	31	3	35	90
Single cases					
MPM*	0	4	0	0	4
Early age of onset* (≤40 years of age)	0	10	0	0	10
Different primaries	0	0	0	1**	1
Total	0	14	0	1	15
Total by Institute	66		3	36	105

78% of target bases being covered by ≥10× across the autosomes and sex chromosomes, rising to 82% when the whole set of exomes, including the Australian samples, were considered. Whole genomes were sequenced to at least 27× mapped coverage. Variants were then called using SAMtools mpileup [380] and filtered for quality. Then, we removed common variants found in the 1000 Genomes Project, October 2011 release [15] and the dbSNP 135 release [381] (as described in Subsection 2.1.3).

We also decided, as before, to take forward only protein-changing consequences. However, as we used an updated version of the Ensembl database (70) and the VEP (2.8), we also updated the consequences kept for further analyses (Table 3.2). We also kept only variants co-segregating in all affected members of a pedigree, while considering all variants called in pedigrees for which a single individual was sequenced. The processing of the Australian samples was done by Peter Johansson and Mitchell S. Stark, based at the Oncogenomic Laboratory in QIMR Berghofer. Leeds samples were aligned, improved and variant-called by pipelines written by the Vertebrate Resequencing group at Sanger, while I performed all the variant filtering steps as described in Subsection 2.1.3. After these filtering steps, 23,051 non-polymorphic variants private to a single pedigree remained for downstream analysis, which were filtered for quality, co-segregation in all affected members, and were predicted to affect protein structure or function.

3.3 Gene prioritisation strategy

With the addition of the Australian families, we now had access to a much more extensive set of pedigrees in which multiple members have been sequenced (Table A.1.8). Therefore, we decided to search for genes that had variants, passing the above filtering criteria, co-segregating with melanoma in pedigrees for which we had sequence information for 3 or more members. We found 320 such genes (Table A.1.9); however, we only found 5 genes that had variants co-segregating with melanoma in more than one of these pedigrees (Table 3.3).

It would seem that there is a discrepancy between this list and the one presented in Table 2.7, as genes *RNF213*, *KLHDC8A* and *C6orf25* were found to harbour co-segregating variants in two or more pedigrees for which information was available for 3 or more members in the European phase of this study. However, the list compiled in Table 2.7 included capillary sequencing of additional members that were not exome-sequenced. In particular, *RNF213* had co-segregating variants in UF10, UF14, UF21 and NF2, of which only UF10 had three members sequenced by NGS (Fig. A.1.1). Similarly, *C6orf25* had co-segregating variants in UF10 and one individual from the replication cohort, for which two additional family members were available for testing. As such, these genes are included in Table A.1.9. *KLHDC8A*, however, was found with co-segregating variants in two pedigrees that had only two members sequenced by NGS, with an additional member per pedigree available for PCR testing (UF15 and UF16) (Fig. A.1.1). A novel variant in *SMG1* was found to co-segregate with melanoma in an Australian pedigree for which 3 members were sequenced (discussed in Chapter 4).

This prioritisation strategy, which adds additional weight to variants co-segregating in multiple pedigrees for which we have more information, identified two genes with similar biological roles (as indicated by GO terms) in telomere maintenance, protection of telomeres (*POT1*) and adrenocortical dysplasia homolog (*ACD*) (also known as *TPP1*, *TINT1*, *PTOP* and *PIP1*) (Figs. 3.2a,b and 3.3, and Table 3.3). We then extended our

Table 3.2: **Consequences of variants kept for further analyses, integrative phase.** Table reproduced from refs. [389, 390]

Sequence Ontology term	Sequence Ontology description
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature
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
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
initiator_ codon_variant	A codon variant that changes at least one base of the first codon of a transcript
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved
transcript_amplification	A feature amplification of a region containing a transcript
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
incomplete_terminal_ codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site
TF_binding_site_variant	A sequence variant located within a transcription factor binding site
feature_elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence
feature_truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence

Table 3.3: **Genes with co-segregating variants in more than one pedigree with three members sequenced.** The number of co-segregating pedigrees is shown alongside their pedigree IDs and number of sequenced members. GO terms are extracted from Ensembl release 70.

Gene name	Pedigrees with co-segregating variants	GO terms
NEK10	2 (NF1: 3, AF11: 4)	Protein serine/threonine kinase activity, ATP binding, protein phosphorylation, transferring phosphorus-containing groups, metal ion binding
POT1	2 (UF20: 3, AF1: 3)	Telomere maintenance via telomerase, nuclear telomere cap complex, single-stranded DNA binding, positive regulation of DNA strand elongation, protein binding
ACD	2 (AF19: 3, AF10: 3)	Telomere maintenance, nuclear telomere cap complex, negative regulation of telomere maintenance via telomerase, positive regulation of single-stranded telomeric DNA binding
MPDZ	2 (AF19: 3, AF4: 3)	Tight junction, cell adhesion, protein C-terminus binding, postsynaptic density, apical plasma membrane, virus-host interaction, myelination
SMG1	2 (UF20: 3, AF3: 3)	DNA repair, response to stress, nucleotide binding, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, protein serine/threonine kinase activity, protein binding, ATP binding

search for other variants falling within these genes in the remaining pedigrees that were not considered in the prioritisation strategy. We identified an additional two pedigrees in which we had sequenced only one individual, UF23 and UF31, also harbouring missense variants in *POT1* (Fig. 3.2c,d), and a single case that presented with MPM carrying a predicted splice site variant (UN4) (Table A.1.8). We could not identify other co-segregating variants in *ACD* (Table 3.4). All of these variants were confirmed by capillary sequencing by Dr. Mark Harland, for Leeds samples, and Lauren G. Aoude, for Australian samples.

In order to obtain a bioinformatic estimation for the pathogenicity of the variants

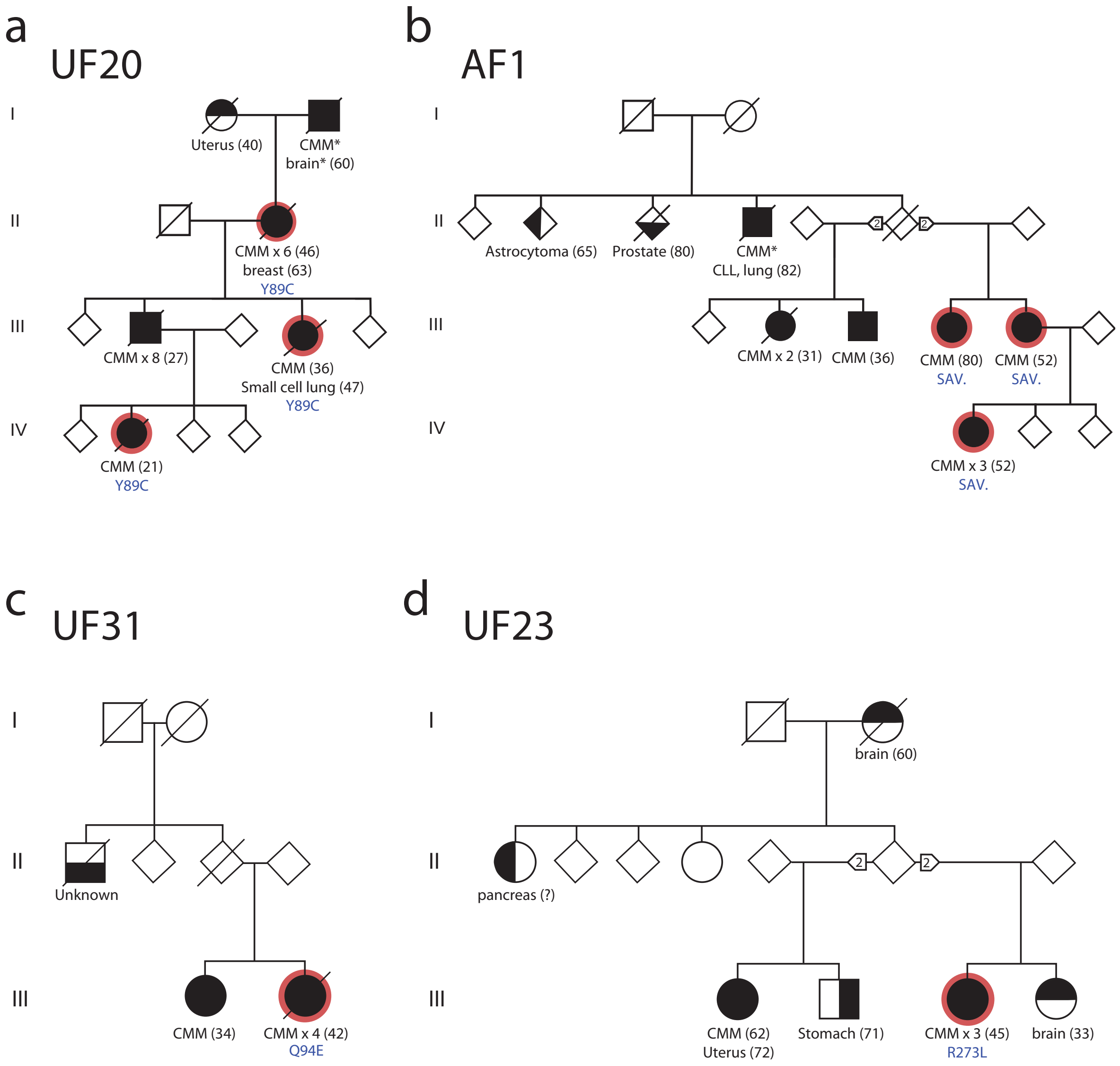

Figure 3.2: **Pedigree structure for families with *POT1* variants.** These pedigrees were sequenced as part of the European phase of this study (UF20), or were additional Australian (AF1) or Leeds (UF23 and UF31) pedigrees included in the integrative phase of the study. Individuals that had their whole exome sequenced are shown with a red outline. Types of cancers are indicated under each symbol, along with the number of primaries and the age of onset (if known) in parenthesis. Circles represent females, squares represent males, diamonds represent individuals of undisclosed sex. A line through the symbol means the individual is deceased. Filled symbols represent individuals with cutaneous malignant melanoma (CMM), and other cancers are indicated by half-filled symbols. All melanomas were confirmed by histological analysis, with the exception of two cases (marked by asterisks). Note that pedigrees have been adjusted to protect the identity of the families without loss of scientific integrity. a) Pedigree UF20, carrying the Y89C variant. b) Pedigree AF1, that was found to carry the splice acceptor variant. c) Pedigree UF31, that was found to be a carrier of the Q94E variant. d) Pedigree UF23, which carries the R273L variant.

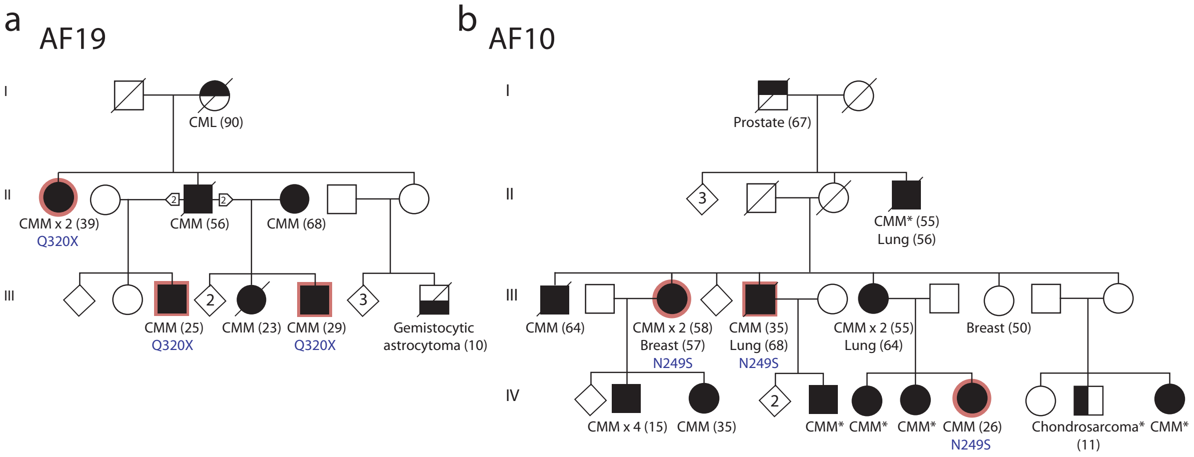

Figure 3.3: **Pedigree structure for families with *ACD* variants.** These pedigrees were additional Australian pedigrees included in the integrative phase of the study. Individuals that had their whole exome sequenced are shown with a red outline. Types of cancers are indicated under each symbol, along with the number of primaries and the age of onset (if known) in parenthesis. Circles represent females, squares represent males, diamonds represent individuals of undisclosed sex. A line through the symbol means the individual is deceased. Filled symbols represent individuals with cutaneous malignant melanoma (CMM), and other cancers are indicated by half-filled symbols. All melanomas were confirmed by histological analysis, with the exception of the cases marked by asterisks. Unaffected siblings are indicated by a diamond with the number of siblings shown in the centre of the symbol. Note that pedigrees have been adjusted to protect the identity of the families without loss of scientific integrity. CML: Chronic myeloid leukaemia. a) Pedigree AF19, that was found to be a carrier of the Q320X variant. b) Pedigree AF10, which carries the N249S variant.

Table 3.4: **Variants in *POT1* and *ACD* identified in melanoma pedigrees.** Pedigrees are shown with the number of individuals that were sequenced. The reference transcripts are ENST00000357628 for POT1 and ENST00000393919 for ACD, which are the only ones per gene for which both the Ensembl automated annotation and the manual Havana team annotation agree.

Gene	Pedigree	Genomic position	Consequence prediction	SIFT [417]	PolyPhen-2 [418]
POT1	UF20 (3)	7:124503684, T/C	Y89C	Deleterious	Probably damaging
POT1	AF1 (3)	7:124465412, C/T	Splice acceptor variant	-	-
POT1	UF23 (1)	7:124493077, C/A	R273L	Deleterious	Probably damaging
POT1	UF31 (1)	7:124503670, G/C	Q94E	Tolerated	Probably damaging
POT1	UN4 (1)	7:124467262, A/C	Splice variant	-	-
ACD	AF19 (3)	16:67692665, G/A	Q320X	-	-
ACD	AF10 (3)	16:67693137, T/C	N249S	Tolerated	Benign

identified in this study, I used the Sorting-Intolerant-From-Tolerant (SIFT) [417] and PolyPhen-2 [418] algorithms. These tools utilise sequence conservation in protein families and multiple-sequence alignments to determine the functional impact that a given substitution is likely to have. Almost all the variants identified in these genes are either obviously disruptive or are predicted to have a high functional impact (Table 3.4).

Using the visualisation tool described in Subsection 2.3.3, I could analyse the positions where these variants lie in the protein structures (Fig. 3.4). None of the variants detected in *POT1* are found in the common variation datasets that I used in this study, and are also not found in the set of ~6,500 exomes released by NHLBI GO ESP [17]. Of the two variants predicted to affect splicing, one falls in a splice acceptor site and the other one affects a base in close proximity to a splice donor site, both in the intron between exons 17 and 18, between amino acids 562 and 563 (Fig. 3.4). Interestingly, all missense variants in *POT1* lie within two repeats of a functional domain annotated in the SUPERFAMILY database [419] as a nucleic-acid binding, oligonucleotide/oligosaccharide-binding (OB) fold (Fig. 3.4).

Of the variants identified in *ACD*, one is obviously disruptive, introducing a premature stop codon, and the other one affects an amino acid in a functional domain. This second variant, N249S, was found at an allelic frequency of 0.0012 in the 500 Exome Project from the Metabolic Disease Group at Sanger, and at an overall frequency of

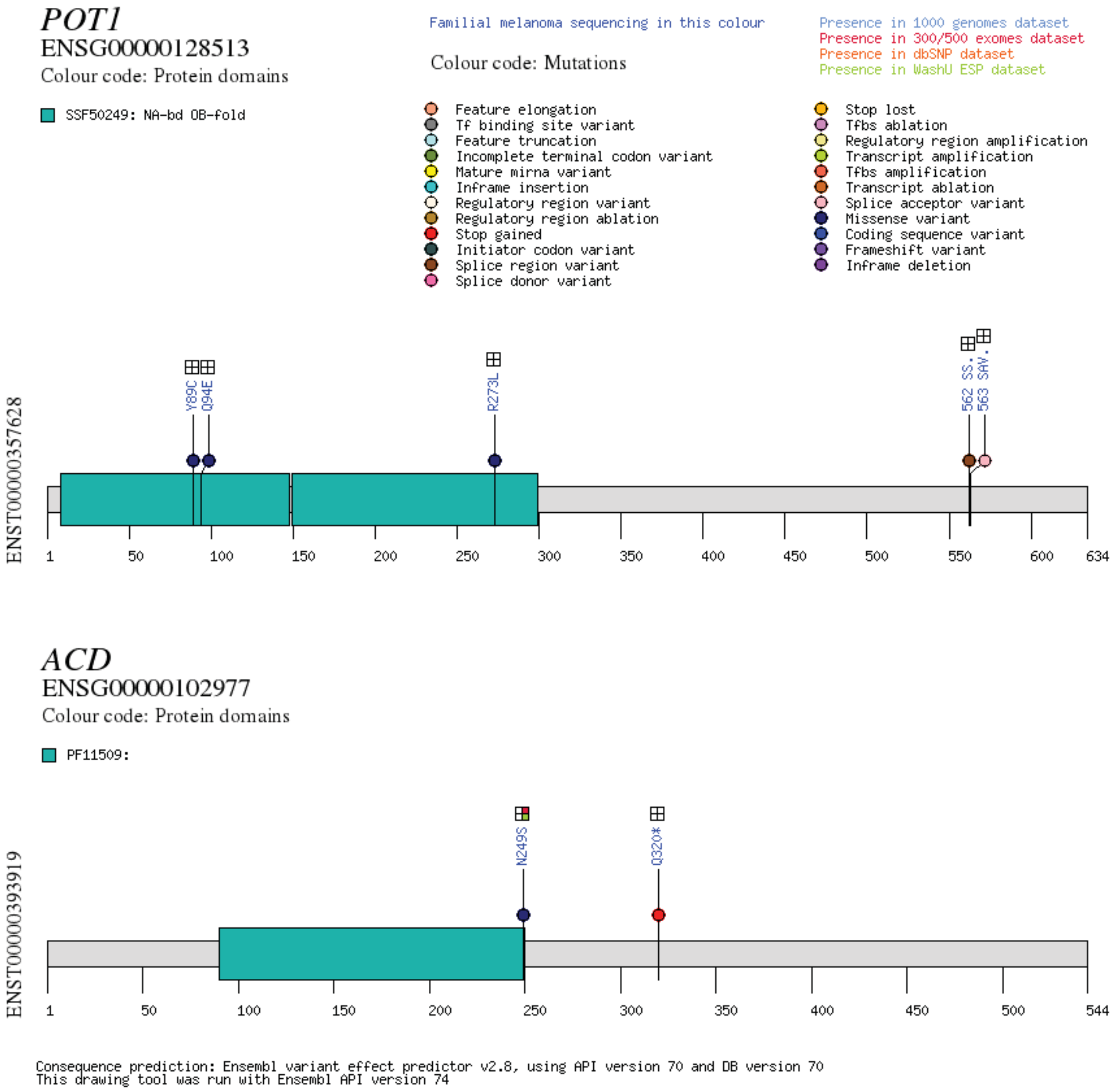

Figure 3.4: **Plot of the *POT1* and *ACD* variants found in familial melanoma pedigrees.** This image shows the variants detected in melanoma pedigrees in protein context. The programme was amended to display Sequence Ontology terms, as per the new terminology of Ensembl release 70. The "300/500 exome dataset" label refers to the 805 in-house exomes used in filtering steps during the European phase of the study. "WashU ESP dataset" refers to 6,500 exomes released by the NHLBI GO ESP [17]. Domains in POT1 were retrieved from the SUPERFAMILY database [419] and those in ACD from Pfam [402]. SS: Splice site, SAV: Splice acceptor variant.

0.0002 in the set of exomes released by NHLBI GO ESP [17]. The functional domain, annotated in the Pfam database [402], is described as having POT1-binding function [420, Note: The Pfam entry for PF11509, which is the domain in the figure, was recently merged with PF10341.].

3.4 Frequency of *POT1* and *ACD* variants in a control dataset

In order to gather further evidence for an association between variants in *POT1* and *ACD* and familial melanoma, I compared the representation between variants in our familial melanoma cases with variants in controls. The controls chosen for this comparison were the same three neurodevelopmental cohorts from the UK10K Project [399] described in Subsection 2.3.1.1 (546 exomes) (Table 2.5). Previous to making this comparison, however, it was necessary to ensure that cases and controls are matched by ancestry, as was performed in Subsection 2.3.1.2. It is necessary to perform this step again, because we have now both an extended set of cases and increased resolution to determine ancestry (with whole genomes instead of a small targeted capture).

To ensure that the controls were matched by ancestry to the melanoma cohort, I performed a PCA using 1,092 individuals across 14 populations from the 1000 Genomes Project, October 2011 release dataset [15]. I took forward for analysis a subset of high-quality variant positions (quality score >10, minimum mapping quality >10, strand bias *P*-value >0.0001, end distance bias *P*-value >0.0001) that were common to the melanoma cohort and the UK10K controls, as well as the 1000 Genomes Project data set. I also excluded SNPs with a minor allele frequency <0.05 or that were in linkage disequilibrium with another SNP (pairwise $r^2 > 0.1$) in the 1000 Genomes Project data set, or that had a Hardy-Weinberg *P*-value $<1 \times 10^{-5}$ in the UK10K controls. After filtering, 7,196 SNPs remained that were spread across all autosomes. Then, using the R package SNPRelate [421], I estimated the first ten principal components using the 1000 Genomes Project individuals and then projected them onto the melanoma cohort samples and UK10K controls. I then removed controls lying greater than 2 standard deviations (s.d.) from the mean scores for principal component 1 or 2, calculated using only European individuals in the 1000 Genomes Project data set ($n = 20$). This left a total of 523 individuals in the control set, and the result of this analysis is shown in Fig. 3.5. I thank Jimmy Z. Liu, from the Sanger, for his time and valuable advice on performing this analysis.

Finally, to ensure that individuals within the UK10K cohort were not related, an IBD analysis was performed using the PLINK toolset [422] and the same set of variants that were used for the PCA. For each pair of individuals with an estimated IBD >0.2, one individual was removed at random ($n = 3$). This filtering left 520 exomes for comparison against the melanoma cohort. The IBD analysis was done by Jimmy Z. Liu, from the Sanger.

I then filtered variants in this collection of 520 UK10K control exomes as described above (keeping positions within exonic regions and removing all variants in Phase 1 of the 1000 Genomes Project, October 2011 release [15] and the dbSNP 138 release [381]). I then predicted and filtered consequences as described above.

3.4.1 Variants in *POT1*

Because I used an updated version of dbSNP (138) for the UK10K filtering step, I assessed whether the *POT1* variants found in this phase of the study passed this filter. After ensuring that they did, I counted variants that passed our filters in the UK10K cohort. I found three variants in *POT1* in different individuals of the UK10K cohort: two predicted to affect splice regions falling in introns and one rare, missense variant outside the OB domains (Q539H).

To estimate the probability of observing as many variants as I found in the melanoma cohort, I performed a two-tailed Fisher's exact test comparing the 4 out of 105 families with melanoma, excluding a discovery pedigree, to three individuals out of 520 controls carrying rare variants in *POT1*, yielding a *P*-value of 0.01703. The detected variants in *POT1* are also not found in the 805 in-house control exomes (Fig. 3.4).

The low *P*-value obtained by this comparison suggests that we find rare *POT1* variants in our melanoma cohort at a higher-than-expected frequency when compared to a control population matched by ancestry. Additionally, this *P*-value is likely to be an underestimate, as it does not consider the fact that some of these variants are shared across multiple individuals within a pedigree.

We also decided to search for these variants in 2,402 population-matched controls belonging to the Leeds Melanoma Case-Control study. This control set includes 499 population-matched control DNA samples, 370 family controls (family members of melanoma cases without a diagnosis of melanoma) and 1,533 DNA samples from the Wellcome Trust Case Control Consortium. All 2,402 samples were wild-type for the *POT1* variants. Moreover, when we genotyped 1,739 population-based melanoma cases that were

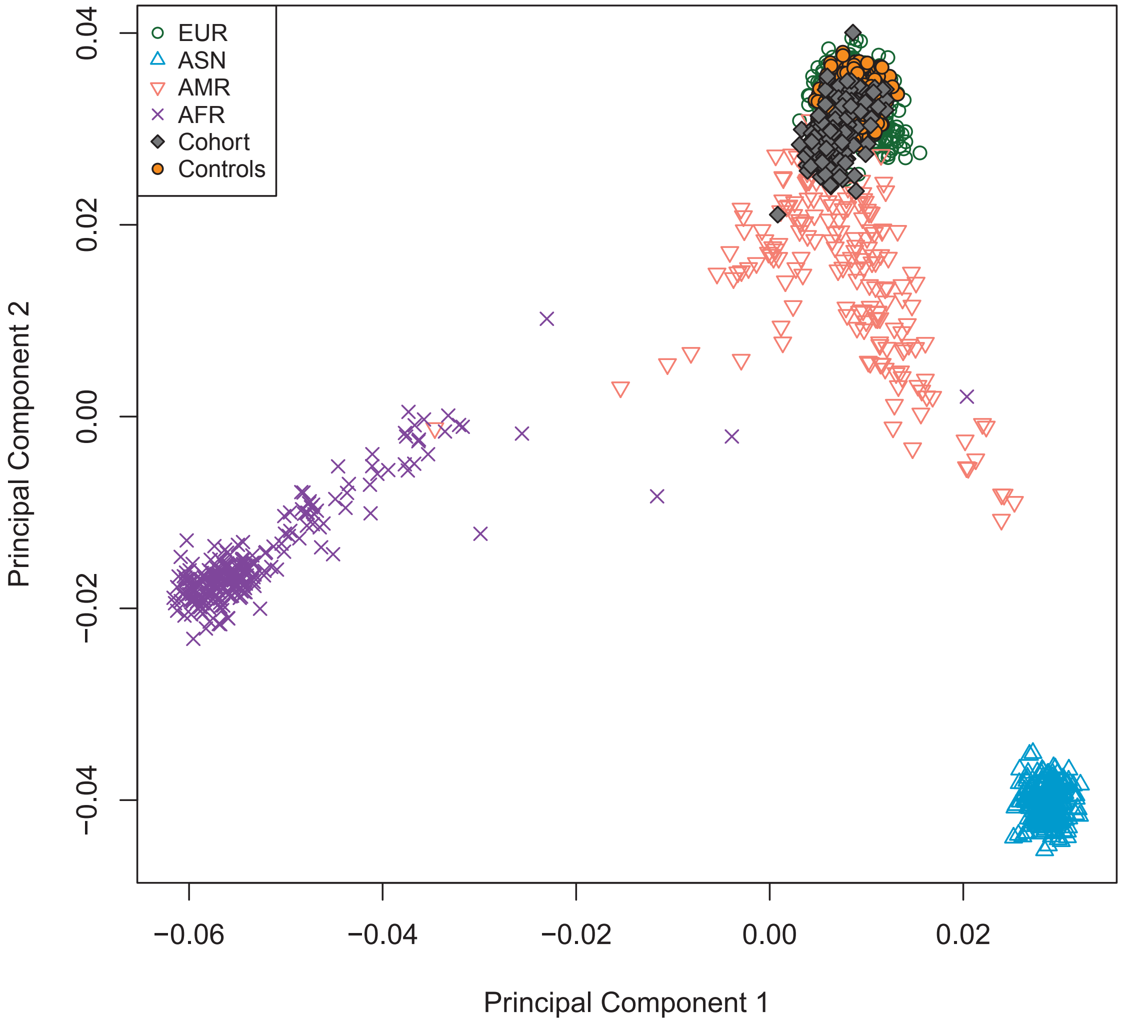

Figure 3.5: **Principal component analysis plot showing that cases and controls are matched by ancestry, integrative phase.** Plot showing the first and second principal components. Ancestry was estimated using the 1000 Genomes Project individuals and then projected onto the melanoma (gray) and UK10K control (orange) cohorts. Note that controls lying greater than 2 s.d. from the mean principal component 1 or 2 scores, calculated using only European individuals in the 1000 Genomes Project data set, are not shown in this plot and were not considered in subsequent analyses. I did not depict three individuals from the QFMP cohort for whom the zygosity of the called genotypes was not available.

recruited from across Yorkshire, UK, as part of the same study, we found one case who carried the R273L variant. This individual presented with MPM with early onset (at 48 years of age), similar to the phenotype presented by the familial cases. This variant was confirmed by PCR sequencing. The genotyping work I describe here was performed by Dr. Mark Harland at the University of Leeds.

DNA was available from two individuals that were not whole-exome sequenced, one from pedigree UF20 (predicted to be an obligate carrier, individual III-2, Fig. 3.2a) and another one from pedigree UF23 (individual III-1, Fig. 3.2d). Unsurprisingly, the obligate carrier tested positive for the presence of the variant, whereas the additional member from pedigree UF23 did not. I explore the implications of this fact in Chapter 4 and in the Discussion.

Overall, we found that all nine *POT1* variant carrier individuals identified through whole-exome sequencing or by targeted PCR resequencing of additional family members (4 from UF20, 3 from AF1, one from UF23 and one from UF31) had developed melanoma, presenting from one primary (four cases) to eight melanomas at 21 to 80 years of age (Fig. 3.2). One variant carrier from these familial cases also developed breast cancer at 63 years of age, and another developed small cell lung cancer at 47 (pedigree UF20). Other malignancies in the untested first- or second-degree relatives of variant carriers included melanoma (pedigrees UF20 and UF31), endometrial cancer (pedigree UF20) and brain tumours (pedigrees UF20 and UF23). I cover follow-up experiments on these variants, as well as the role of *POT1* in familial cancer predisposition, in the next Chapter, as well as in the Discussion.

3.4.2 Variants in *ACD*

Because I used an updated version of dbSNP for the UK10K filtering step, I assessed whether the *ACD* variants found in this study passed this filter. The N249S variant is annotated in dbSNP 138, having been submitted by the NHLBI GO ESP [17], in which it is found at an overall allele frequency of 0.0002 (Fig. 3.4). The Q320X variant passed this filter. I then counted the number of individuals in the UK10K control set that carried rare variants in *ACD* and found a single, missense variant, resulting in an amino acid substitution (A72E) that does not occur in any functional domains. A Fisher test would not be informative in this case, as just one variant has been found in each cohort.

The QIMR Berghofer team, headed by Prof. Nicholas K. Hayward, led an investigation on the prevalence and haplotype of the *ACD* N249S variant. Briefly, they genotyped an additional 4 cases or obligate carriers of pedigree AF10, and found that all of them harboured the variant (7 out of 7 tested cases or obligate carriers) (individuals III-3, III-5, III-7, IV-3, IV-6, IV-7 and IV-8, Fig. 3.3b). Other carriers are individuals III-9, III-11, IV-1 and IV-9. Individual III-4 tested negative for the variant. They also studied the whole exomes of 7 families from Copenhagen, Denmark (recruited by the Danish project of hereditary malignant melanoma) and discovered one family harbouring the *ACD* N249S variant, segregating in 3 out of 4 members tested. This family was found to share the *ACD* haplotype with the Australian family. They also performed linkage analysis of both families and calculated a logarithm (base 10) of odds (LOD) score of 1.14.

They also searched for the N249S variant in Australian population-based case-control panel, but did not find it in 1,669 cases or 1,590 controls. The variant was also absent from 1,000 Danish diabetes cases and 1,000 metabolically healthy controls.

Overall, we found that *ACD* variant carriers presented with zero, one or two primaries, at 26 to 58 years of age (Fig. 3.3). Other malignancies, such as lung and breast cancer, are present in *ACD* variant carrier individuals, and brain cancer and chondrosarcoma are also present in untested first- or second-degree relatives of variant carriers (Fig. 3.3). I cover the implications of these results in the Discussion.

3.5 Summary and conclusion

In this phase of the study, I had access to exome sequences from a much larger set of familial melanoma pedigrees, recruited in Australia as part of the QFMP [416], totalling 98 exomes from 36 pedigrees. We also obtained whole exome sequences from an additional 45 Leeds patients, that were either familial cases (31 patients) or were single cases that presented with MPM or an early age of onset (14 patients) (Table 3.1). With this extended set of patients, an analysis of genes mutated in more than one pedigree with multiple members was more informative than when we performed a similar analysis in the European phase (shown in Table 2.3). We observed only five genes with co-segregating variants in more than one pedigree for which we had sequence for 3 or more members, and while all of them represent plausible candidates, we found it interesting that two of these genes seem to have very similar biological functions (telomere maintenance). These genes are *POT1* and *ACD*. Therefore, we decided to pursue these candidates and attempt to establish their role, if any, in melanoma predisposition.

Interestingly, *POT1* and *ACD* directly interact with each other, as ACD recruits POT1 to telomeres, where they belong to the six-protein complex shelterin [423, 424].

Shelterin has a paramount role in protecting telomere ends from the DNA damage response, as well as controlling telomerase access to telomeres [423]. The location of the mutations is also interesting, as missense variants fall within the OB folds of POT1, which allows it to bind to single-stranded (ss) DNA, or the POT1-interacting domain of ACD [423] (Fig. 3.4). The rest of the variants found in these genes are predicted to be disruptive, introducing a premature stop codon or affecting mRNA splicing.

A search for variants in these genes in a set of control exomes matched by ancestry to the melanoma cases suggested that germline mutations in *POT1* are rare, whereas results were inconclusive for variants in *ACD*. These hypotheses were further supported by genotyping of population-matched case control series for the particular variants found in this study. Therefore, we decided to further investigate the molecular mechanism by which these variants might increase melanoma risk in carriers, as well as their biological consequences.

I describe experiments investigating the function of these variants, as well as other genes detected in the European phase of this study, in the next Chapter. In the Discussion, I address the involvement of the protein complex to which POT1 and ACD belong, shelterin, in melanoma and generally in cancer predisposition.

Chapter 4

Biological validation of candidate melanoma susceptibility genes

Methods and results of this chapter have been published or accepted for publication (refs. [376] and [414]). Additionally, a review I wrote on sections covered by this chapter has also been accepted for publication (ref. [425]). Some parts of the text have been reproduced from these references; I confirm I have ownership of copyright for reproduction in this work.

In previous chapters, I have discussed the different methodologies we followed to pinpoint genes that might have a role in melanoma susceptibility. We found several genes with rare variants that co-segregate with melanoma in all affected members sequenced from each carrier family. These genes have roles in DNA repair upon UV irradiation or telomere protection, and the detected variants are predicted to alter protein structure or function. Therefore, these genes represent plausible candidates given their biological roles, and moreover, some of them are mutated at a higher-than-expected frequency in the melanoma cohort than in controls matched by ancestry.

However, *in silico* predictions of the consequences of these variants are insufficient to establish the causality of these genes in melanoma susceptibility, if any, and thus it is necessary to experimentally demonstrate their biological relevance. This Chapter explains the experimental procedures we performed to test the effect of these variants, and the consequences they may have in variant carriers.

4.1 *SMG1*, the top candidate gene from the European phase of the study

SMG1 was the highest-ranked candidate in the European phase of this study, and it also had a rare variant co-segregating with melanoma in multiple members of an Australian pedigree, which made this gene one of the top candidates of the integrative phase as well (Table 3.3). In order to investigate the role of *SMG1* in melanoma origin and progression, we attempted to assess whether *SMG1* had any influence on melanocyte transformation, to determine if changes of expression had any effect on cell cycle or proliferation.

4.1.1 Summary of variants found in *SMG1*

During the discovery phase of this study (Section 2.1), we had identified four variants in *SMG1* that segregated with the disease in four different families (Table 2.3). After PCR validation of these variants and testing of additional pedigree members, all of these with the exception of one were found to be real and co-segregate in all available affected members (Table 2.7). The variant that could not be confirmed due to poor sequencing traces, an insertion of one amino acid at position 18, was however detected by NGS in the two sequenced cases in that pedigree. Three variants in three different pedigrees were detected in the replication phase; however, two of these were found not to be real after PCR validation (see Table 2.7, these are the two variants that were successfully tested but that gave a negative result). During the integrative phase of this study, an additional rare variant in *SMG1* was found to co-segregate in all members of an Australian pedigree (Table 3.3). This variant was also confirmed by PCR. The PCR confirmation work and co-segregation assessment was performed by Dr. Mark Harland for Leeds samples and Lauren G. Aoude for Australian variants. All this information is summarised in Table 4.1.

We can appreciate, from visual inspection of the variants in a protein plot (generated by the tool discussed in Subsection 2.3.3), that all confirmed missense variants cluster in the C-terminus of the protein, outside any functional domains (Fig. 4.1). Two of these variants, S3047N and H3234Q, are found at a very low frequency in the set of exomes released by the NHLBI GO ESP [17] (1 out of 11,790 and 1 out of 12,146 alleles, respectively), but are found to co-segregate with melanoma in all three members of the pedigrees tested (Table 4.1). The two variants involving insertions or deletions, found

Table 4.1: **Summary of all variants found in *SMG1*.** The study phases in which each variant was identified is indicated. The number of individuals tested and that were found to carry the variant is shown in parentheses.

Study phase	Pedigree	Genomic position (GRCh37)	Protein	PCR	SIFT [417]	PolyPhen-2 [418]
Discovery	UF20 (4)	16:18840781, G/A	H3144Y	Real	Tolerated	Benign
Discovery	UF16 (3)	16:18841071, C/T	S3047N	Real	Tolerated	Benign
Discovery	UF17 (3)	16:18823187, C/A	V3602L, splice variant	Real	Tolerated	Probably damaging
Discovery	NF2 (2)	16:18937311, -/CGC	G18GG	Poor	-	-
Replication	RP1 (1)	16:18937327, CGT/-	S11-	Real	-	-
Replication	RP2	16:18880434, G/T	T942K	Not real	Tolerated	Probably damaging
Replication	RP3	16:18903640, G/A	S150L	Not real	Deleterious	Benign
Integrative	AF3 (3)	16:18839392, A/T	H3234Q	Real	Tolerated	Benign

in pedigrees NF2 and RP1, are in a repetitive region at the N-terminal portion of the protein that consists of two serines followed by six glycines.

4.1.2 Biological function of the SMG1 protein

SMG1 belongs to the PI3K-related kinase (PIKK) family of proteins, which includes ATM, ATR, MTOR and the DNA-dependent protein kinase, catalytic subunit (DNA-PK$_{cs}$, encoded by *PRKDC*). These proteins have important roles in the response to DNA damage, mitogenic signalling, and DNA break repair [426, 427]. Similarly to these proteins, SMG1 has been found to participate in DNA damage response, being activated upon UV or ionising radiation and participating in P53 phosphorylation and stabilisation [413]. It also participates in the nonsense-mediated mRNA decay (NMD) pathway, a process that monitors and destroys transcripts with premature termination codons (reviewed in [428]). In this pathway, it phosphorylates UPF1 regulator of nonsense transcripts homolog (UPF1), an essential helicase that is recruited to mRNA molecules

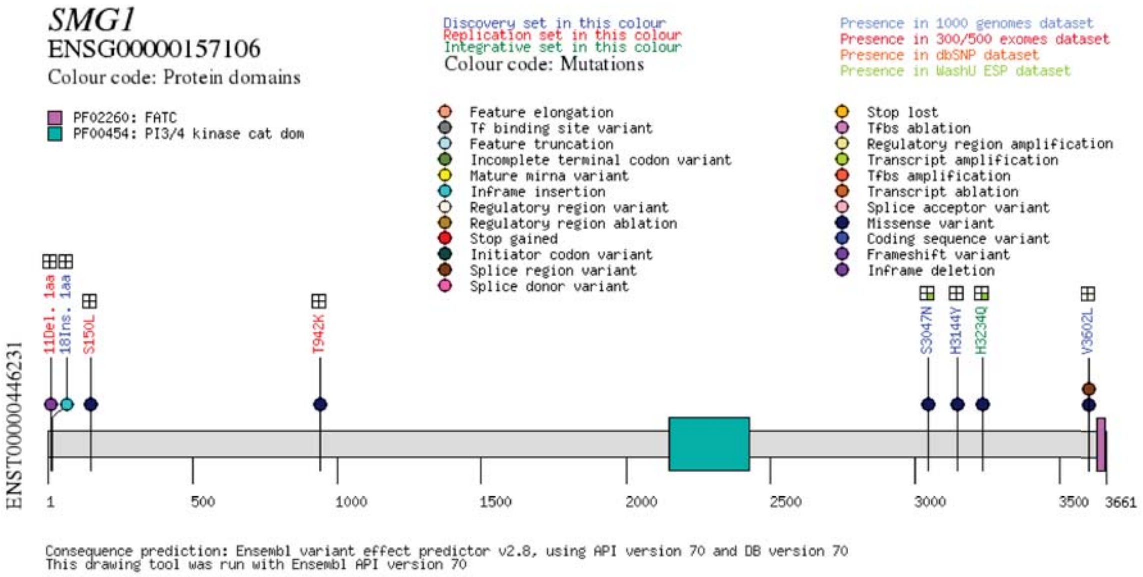
SMG1
ENSG00000157106
Colour code: Protein domains
PF02260: FATC
PF00454: PI3/4 kinase cat dom
Discovery set in this colour
Replication set in this colour
Integrative set in this colour
Colour code: Mutations
Feature elongation
Tf binding site variant
Feature truncation
Incomplete terminal codon variant
Mature mirna variant
Inframe insertion
Regulatory region variant
Regulatory region ablation
Stop gained
Initiator codon variant
Splice region variant
Splice donor variant
Presence in 1000 genomes dataset
Presence in 300/500 exomes dataset
Presence in dbSNP dataset
Presence in WashU ESP dataset
Stop lost
Tfbs ablation
Regulatory region amplification
Transcript amplification
Tfbs amplification
Transcript ablation
Splice acceptor variant
Missense variant
Coding sequence variant
Frameshift variant
Inframe deletion
ENST00000446231
11Del. 1aa
18Ins. 1aa
S150L
T942K
S3047N
H3144Y
H3234Q
V3602L
1
500
1000
1500
2000
2500
3000
3500
3661
Consequence prediction: Ensembl variant effect predictor v2.8, using API version 70 and DB version 70
This drawing tool was run with Ensembl API version 70

Figure 4.1: **Summary of variants detected in the SMG1 protein.** The variants we detected in familial melanoma pedigrees are shown in protein context, the transcript in the picture belongs to the consensus coding DNA sequence (CCDS45430). The four variants identified in the discovery part of the European phase are depicted in blue (Table 2.3), variants identified in the replication part are shown in red. The variant found co-segregating in an Australian pedigree is shown in green (Table 3.3). Subsequent PCR confirmation determined that two of the variants identified in the replication phase, S150L and T942K, were not real (see Table 2.7, these were the variants successfully tested but with a negative result). All other variants were confirmed as real with the exception of the 1 aminoacid insertion at position 18, as sequence in the trace was poor and could not be confirmed. However, this variant was found in two members of the same family by NGS. All variants co-segregate perfectly with melanoma in all members tested.

upon recognition of stop codons by the translation machinery (reviewed in [428]). This phosphorylation is central to the NMD pathway, as over-expression of a kinase-mutant form of SMG1 was found to suppress NMD, whereas that of wild-type SMG1 enhances it [429].

Additionally, SMG1 has been found to have roles in other biological processes. It has been implicated, for example, in the cellular response to low oxygen levels and in tumour necrosis factor (TNF)-α-induced apoptosis. In the first case, SMG1 was shown to be activated by hypoxia and its depletion augmented the activity of hypoxia inducible factor 1, alpha subunit (HIF1A), a master regulator of cellular responses to low oxygen levels [430]. In the second case, SMG1 depletion, but not that of other proteins part of the PIKK family, was shown to rapidly increase the rate of apoptosis induced by TNFα activity [431].

SMG1 was recently found to have a role in adipogenesis, possibly via its participation in another mRNA decay pathway, mediated by staufen double-stranded RNA binding protein 1 (STAU1) [432]. This pathway targets higher-order structures in the 3' untranslated region (3'UTR) of mRNAs, and is more active than NMD during adipogenesis. SMG1 down-regulation was found to delay adipogenesis possibly by phosphorylating UPF1, an enzyme that participates in both NMD and STAU1-mediated mRNA decay (SMD) [432]. Additionally, SMG1 might have a role in the regulation of brain phosphorylated α-synuclein (p-syn) levels, which are involved in neurodegenerative disorders such as Parkinson's disease with dementia. SMG1 knockdown was found to significantly increase the levels of p-syn levels in the brain, and its low expression correlated with higher levels of this phosphoprotein [433].

More intriguingly, SMG1 has also been linked to telomere maintenance, and independently, to cancer predisposition. Depletion of SMG1 was shown to induce loss of entire telomere tracts, as well as to induce chromosome and chromatid breaks [434]. It can localise to telomeres *in vivo* and antagonise the association of telomeric repeat–containing RNA (TERRA), which is the product of telomere transcription, with chromatin. By preventing TERRA association with telomeres, it might aid efficient telomere capping and cell cycle progression [435]. Although *Smg1* is homozygous lethal in mice, animals heterozygous for a knockout allele of *Smg1* were found to be predisposed to different cancers, such as lymphomas and lung adenocarcinomas, as well as having chronic inflammation [436]. These mice did not have any defects in the NMD machinery, but displayed tissue oxidative damage and low-level inflammation prior to the development of tumours. These conclusions are supported by human data, as *SMG1* has been found to be significantly mutated in lung adenocarcinoma [437], and mutations in this gene have been found in more than 5% of all assessed samples of cervix, stomach and intestine cancers in the Catalogue of Somatic Mutations in Cancer (COSMIC) database [438]. However, it may have a different role in myeloma and acute myeloid leukaemia, in which it could contribute positively to cancer progression [439, 440].

In conclusion, SMG1 has been found to play a role in many diverse biological pathways, namely NMD, SMD, hypoxia response, adipogenesis, TNFα-induced apoptosis, inflammation, DNA damage response upon UV and ionising radiation and telomere maintenance. Perhaps it is the last two, together with the observation that *Smg1*$^{+/-}$ mice are predisposed to cancer, those which are most relevant for the melanoma phenotype in these families with *SMG1* variants. In the next subsection, I describe the different experiments that were attempted in order to prove the role of *SMG1* and these variants in melanoma formation.

4.1.3 Description of biological assays

The experiments detailed in this section were performed by Drs. Jessamy C. Tiffen and James Hewinson, from the Experimental Cancer Genetics team at the Sanger, in the course of a year. The first experiment attempted was a colony formation assay in soft agar, using a melanocyte cell line, referred to as pmel*, that already carries the activating *BRAF* V600E mutation [259]. Cells that only have activated BRAF do not spontaneously transform in soft agar, possibly due to oncogene-induced senescence [441]. The rationale behind this experiment was that if *SMG1* functions as a tumour suppressor in melanocytes (as it may be assumed given the biological roles discussed above), then cells might transform and form colonies upon its knockdown. However, although a stable knockdown of *SMG1* was achieved via the use of lentiviral vectors, various technical problems prevented us from completing this experiment. Mainly, mycoplasma contamination before cells were shipped from the laboratory of origin meant that results obtained thus far had to be repeated, and spontaneous transformation of control cells (*i.e.*, those not treated with a *SMG1* knockdown vector) prevented us from deriving any conclusive results. Given these difficulties, a different strategy was attempted.

The second experiment that was tried was a knockdown or over-expression of *SMG1* in the melanoma cell lines A375 and Mel-ST [240, 442]. The first of these was derived from a 54-year-old female with malignant melanoma, and the second one is a human melanoma cell line that is immortalised but not transformed. The idea was to address whether *SMG1* expression changes had any effect on on apoptosis, cell cycle progression and proliferation upon UV or ionising radiation. Interestingly, Mel-ST cells that had been depleted of SMG1 via short hairpin (sh)-RNA interference showed an increase in the proportion of cells in the S and G2/M phases of the cell cycle upon UV irradiation when compared with irradiated controls (Fig. 4.2), in accordance with previous reports [413]. This result might indicate that *SMG1*-deficient cells have checkpoint signalling defects upon UV-induced DNA damage, but there might be another mechanism in place to arrest them at the S phase of the cell cycle (Fig. 4.2). This might indicate that cells depleted for *SMG1* need additional hits to progress towards malignancy. However, the behaviour of these cells could not be reproduced in A375 cells, as these did not show any changes in cell cycle profile upon UV irradiation.

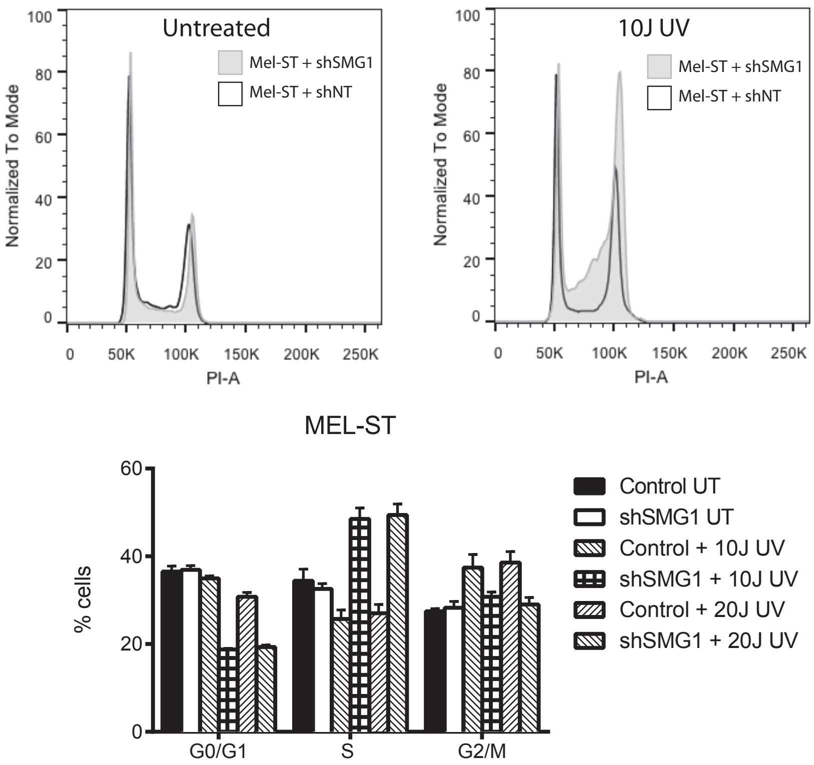

Figure 4.2: **Cell cycle profile for Mel-ST melanoma cells with a *SMG1* knockdown.** Top panel, cell cycle profiles. Gray line, Mel-ST with the anti-*SMG1* shRNA vector; black line, Mel-ST with a non-targeting vector. Two different UV intensities were assessed, note that only one is shown in the top panel. Bottom panel, bar chart showing the proportion of cells in each condition. PI: propidium iodide. I thank Drs. Jessamy C. Tiffen and Gabriel Balmus for this image.

Finally, site-directed mutagenesis is being attempted to assess whether the particular variants detected in melanoma pedigrees have any effect in SMG1 stability. So far, mutants have been generated in a cDNA expression vector (from ref. [429]), and have been confirmed to have the correct sequence. This experiment is on-going.

4.1.4 Conclusion

SMG1 is an attractive candidate for a melanoma susceptibility gene given the diverse biological roles it plays in the cell, among them, response to UV-induced DNA damage and P53 phosphorylation. We have attempted several different experimental methodologies to assess its role in melanocytes, namely *SMG1* knockdown, over-expression and site-directed mutagenesis in melanoma cell lines. However, studying *SMG1* has proven difficult for several reasons, among them, cell contamination and irreproducibility of results between cell lines. We hope that we can address these concerns with further experiments.

4.2 Candidate genes from the integrative phase

During the integrative phase of this analysis, we prioritised a list of five genes, because they were found to have co-segregating variants with melanoma in two or more pedigrees for which we had sequences available for three or more affected individuals (Table 3.3). Of these, *POT1* and *ACD* were of special interest because they participate in telomere maintenance, and furthermore, interact with each other as part of the six-protein complex shelterin [423]. The location of the detected variants is also noteworthy, as all of the missense variants detected fall within functional domains in the protein structures, within the ssDNA-binding OB folds in the case of *POT1* and in the POT1-binding domain in the case of *ACD* (Fig. 3.4). The rest of the detected variants are predicted to be disruptive, introducing premature stop codons or affecting mRNA splicing. Additionally, we found variants in *POT1* at a statistically higher frequency in the melanoma pedigrees when compared with controls, even when co-segregation with melanoma was not considered, and genotyping of a case-control series identified a case that carried one of the familial *POT1* variants. This makes *POT1* and *ACD* attractive candidates for being melanoma susceptibility genes. This Section details the experiments carried out to test the involvement of the detected variants in melanoma susceptibility.

4.2.1 The biological role of *POT1*, *ACD* and the shelterin complex

Telomeres are structures at the ends of linear chromosomes that consist of double-stranded DNA repeats followed by a short ssDNA protrusion. They play an essential role in regulating genomic stability by allowing the cell to distinguish between chromosome ends and double-stand DNA breaks, and as such, they need to be replicated each cell cycle and protected from DNA-processing enzymes [443, 444]. The shelterin complex, a large macromolecular structure to which POT1 and ACD belong, binds telomeres and has a paramount role in their protection [423]. Shelterin displays a wide range of functions that not only include telomere length maintenance and protection from DNA repair mechanisms, but also the regulation of diverse signalling cascades from telomeres

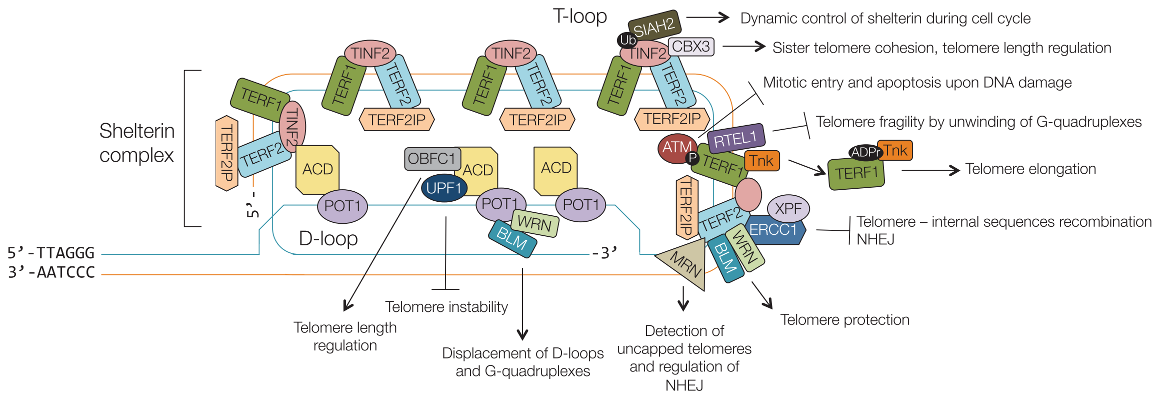
T-loop
SIAH2
Ub
TINF2
CBX3
Dynamic control of shelterin during cell cycle
Sister telomere cohesion, telomere length regulation
Mitotic entry and apoptosis upon DNA damage
Telomere fragility by unwinding of G-quadruplexes
ADPr
Tnk
TERF1
Telomere elongation
Shelterin complex
TERF1
TINF2
TERF2
TERF2IP
ACD
POT1
OBFC1
UPF1
WRN
BLM
D-loop
5'-
5'-TTAGGG
3'-AATCCC
-3'
ATM
P
RTEL1
XPF
ERCC1
MRN
Telomere – internal sequences recombination
NHEJ
Telomere protection
Detection of uncapped telomeres and regulation of NHEJ
Displacement of D-loops and G-quadruplexes
Telomere instability
Telomere length regulation

Figure 4.3: **The shelterin complex and its biological functions.** The shelterin complex, shown at the left of the figure, binds telomeres to protect them and regulate diverse signalling cascades. Chromosome ends can fold into T-loops and D-loops, which form when the 3' ssDNA protrusion invades the double-stranded telomere. Signalling cascades that shelterin regulates, and their biological effects, are shown at the right. SIAH2: Siah E3 ubiquitin protein ligase 2, CBX3: Chromobox homolog 3, RTEL1: Regulator of telomere elongation helicase 1, Tnk: Tankyrase, XPF: ERCC4, excision repair cross-complementation group 4, WRN: Werner syndrome, RecQ helicase-like, BLM: Bloom syndrome, RecQ helicase-like, MRN: MRE11-RAD50-NBN complex, OBFC1: oligonucleotide/oligosaccharide-binding fold containing 1, Ub: Ubiquitination, ADPr: ADP ribosylation, NHEJ: Non-homologous end joining.

[444] (Fig. 4.3).

The other components of shelterin are telomeric repeat-binding factors 1 and 2 (TERF1 and TERF2, also known as TRF1 and TRF2), TERF1-interacting protein 2 (TINF2, also known as TIN2) and TERF2-interacting protein 1 (TERF2IP, also known as RAP1) [423]. Although these proteins are fast-evolving, with the architecture of the complex being different in organisms such as ciliates and yeast when compared to mammals, the overall functionality of shelterin is highly conserved [445, 446].

TERF1 and TERF2 are double-stranded DNA-binding proteins that recognise telomeric repeats with high affinity upon homodimerisation, while POT1, the most evolutionarily conserved member of shelterin, can specifically recognise telomeric ssDNA [423, 447–449]. Therefore, the presence of several TTAGGG-binding domains in the complex gives shelterin its exquisite specificity for telomeric sequence. ACD binds to POT1, recruiting POT1 to telomeres and enhancing its recognition ability, whereas TERF2IP localises to telomeres via its interaction with TERF2 [423, 450]. TINF2 is able to bind TERF1, TERF2 and the ACD/POT1 sub-complex, therefore bringing all the shelterin components together [451, 452].

The importance of the shelterin complex is evidenced by the fact that null mutations in the majority of its components result in embryonic lethality in mice [453–456]. One of the most important functions of shelterin is to protect chromosome ends from DNA repair nucleases, which it achieves by inhibiting six DNA damage signalling pathways: ATM- and ATR-signalling, classical non-homologous end joining (NHEJ), alternative NHEJ, homologous recombination, and resection [457]. Shelterin also has an important role in regulating telomere structure and length. TERF1 and TERF2 have DNA remodelling activities, being able to bend DNA and contributing to T-loop formation, whereas POT1 is able to regulate telomere length by contributing to the nucleolytic processing of the telomeric 5' end and control telomerase access to the end of chromosomes [423, 447, 458, 459]. The absence of functional shelterin therefore leads to polyploidization, fragile telomeres and sister chromatid exchanges, among other chromosomal aberrations [457].

In particular, POT1 can function both as a negative and a positive regulator of telomere length. In support of the former role, its knockdown has been shown to elicit telomere lengthening in HTC75 cells [424], and so does a mutant form lacking the DNA-binding domain [460], or the ACD-binding domain [461]. Additionally, point mutations affecting the ability of POT1 to bind to ssDNA in CLL cells were shown to lead to telomere lengthening, as well as various types of chromosomal aberrations [462]. In support of the latter role, POT1 has been found to enhance the processivity of telomerase when found in a complex with ACD [463], and its exogenous expression in telomerase-positive HT1080 cells led to telomere elongation, albeit with some variability [464]. In order to explain these seemingly contradictory results, Feng Wang and colleagues proposed a three-state model for the role of POT1 at telomeres [463]: First, when POT1 is bound to the 3' overhang as part of the shelterin complex, it blocks access to telomerase by sequestering the telomere, thus acting as a negative regulator of telomere length. Second, the POT1-ACD complex is removed from chromosome ends, possibly by disruption of the shelterin complex. Third, the POT1-ACD complex is recruited to telomeres to serve as a telomerase processivity factor. The cycle is restarted as the telomere is elongated, as it can bind shelterin complexes and the 3' overhang can be re-bound by the POT1-ACD complex. POT1 depends on ACD for its recruitment to telomeres, as *ACD* knockdown has been shown to reduce POT1 localisation to the ends of chromosomes and to promote telomere dysfunction [445].

Additionally, knockdown of *POT1* in tumour cells has been shown to elicit a transient DNA damage response and to alter the sequence of chromosome termini, thus also determining telomere structure [459]. In this study, tumour cells depleted for POT1 displayed telomere dysfunction, yet they could proliferate as well as control cells without the *POT1* knockdown vector. In contrast, human primary fibroblasts responded to POT1 depletion by inducing senescence and reducing cell division, but proliferation was partially restored upon abrogation of the P53 and the INK4A/RB1 pathways. This observation is interesting, as it could mean that additional mutations are necessary in order for POT1-deficient cells to progress to tumourigenesis. Several studies have also shown that a single *POT1* allele that is unable to bind ssDNA is enough to exert a dominant-negative effect, indicating that abrogation of a single copy of the gene is sufficient to inhibit POT1 function [460, 462, 465]. This effect might be due to disrupted interactions at the telomere, as TERF2 has reduced association with its binding partners in the shelterin complex upon mutant POT1 expression [465].

In conclusion, POT1, ACD and the other shelterin components play an essential role in telomere length regulation and protection, and their malfunction might predispose cells to malignancy via telomere uncapping and length dysregulation, which then could lead to chromosomal abnormalities. In the next subsection, I describe the studies we performed to investigate the biological effect of the *POT1* and *ACD* variants detected in melanoma pedigrees.

4.2.2 The effect of *POT1* variants found in melanoma pedigrees

In total, we identified five pedigrees carrying different *POT1* variants, two predicted to affect mRNA splicing and three missense (Fig. 3.4, Table 3.4). We performed different analyses to study the function of these two types of mutations.

4.2.2.1 Investigation of the variants predicted to affect mRNA splicing

There were two variants predicted to affect *POT1* mRNA splicing detected in familial melanoma pedigrees, one falling in an intronic splice acceptor site and another one falling 6bp away from a splice donor site (Fig. 3.4, Table 3.4). In order to computationally predict whether these variants were likely to be deleterious, I used the MaxEntScan algorithm [466] to compare the sequences of the variant and wild-type splice sites. This algorithm was shown to be able to accurately discriminate real splice acceptor and donor sites from decoys, and works by supplying only a short sequence at the intron-exon boundary. It then assigns a score reflecting the likelihood that the sequence functions as a splice donor or acceptor. The scores for the region with the splice acceptor variant were 5.53 (wild-type) and -3.32 (with variant), whereas for the splice donor region these were 7.96 (wild-type) and 4.7 (with variant).

To put these scores in context, I retrieved 10,000 sequences at random from the human genome, using the Ensembl API release 70, and obtained their MaxEntScan scores. Subsequently, I retrieved 10,000 real splice acceptor sites and 10,000 real splice donors from human genes chosen at random, but always choosing the second exon. I then obtained a distribution of all of their MaxEntScan scores. The splice acceptor variant lowered the score of the wild-type sequence from the 9.2th to the 0.57th percentile when compared to the score distribution of real splice acceptors, while the splice donor variant lowered it from the 33.67th to the 12.31th percentile when compared to the score distribution of real splice donors (Fig. 4.4). Therefore, the splice acceptor variant is predicted to be highly deleterious, whereas the splice donor variant still falls well within the scores of real splice sites.

We then decided to test whether these variants had any effect in *POT1* mRNA splicing. To do this, RNA extracted from the whole blood of two carriers of the splice acceptor variant (pedigree AF1) was converted to cDNA using SuperScript III Reverse Transcriptase (Invitrogen). RT-PCR was then performed to confirm that this variant was indeed disruptive to splicing. M13-tagged forward and reverse primers were designed to flank the spliced region. The product was visualised on a 3% NuSieve Agarose gel, and the sequence was verified using standard Sanger sequencing methods. This work was done by Lauren G. Aoude, at QIMR Berghofer. This experiment showed that this splice acceptor variant affects transcript splicing, as it leads to a frameshift and the introduction of a premature stop codon 11 amino acids downstream of the intron-exon boundary (Fig. 4.5).

As we failed to see any effect of the splice donor variant on transcript splicing by PCR, we decided to perform whole blood RNA sequencing from the carrier to see whether we could find traces of any aberrant *POT1* transcripts. However, we could not detect anything abnormal (data not shown). The RNA extraction was performed by Marcela Sjöberg and the transcriptome analysis was done by Martin del Castillo, at Sanger.

These analyses support that the splice acceptor variant carried by pedigree AF1 is disruptive to the *POT1* transcript, causing a frameshift and introducing a premature stop codon. However, we did not find evidence to support that the splice donor variant is deleterious to the carrier.

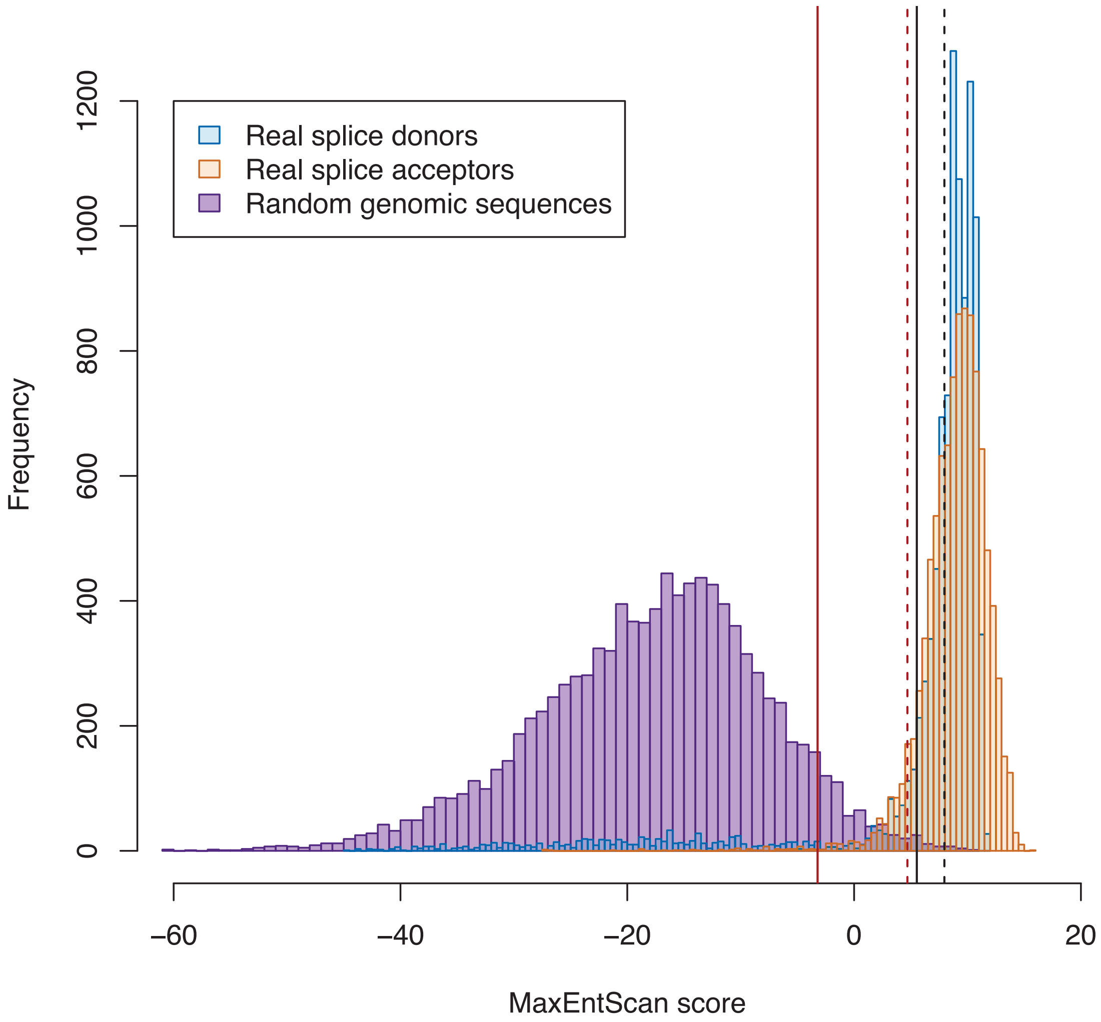

Figure 4.4: **MaxEntScan scores for the splice acceptor and splice donor variants detected in familial melanoma pedigrees.** The location of the scores for the wild-type (black) and variant (red) sequences for the splice acceptor (solid) and splice donor (dotted) sequences are shown against score distributions for real splice donors, real splice acceptors and random genomic sequences.

4.2.2.2 Investigation of the three missense variants detected in melanoma pedigrees

4.2.2.2.1 Evolutionary conservation of variant positions In order to assess the evolutionary conservation of the residues where we detected variants in POT1, I gathered amino acid sequences for the encoded protein in evolutionarily diverse species (*i.e.*, human, mouse, cow, armadillo, elephant, opossum, platypus, chicken, frog and zebrafish) from the National Centre for Biotechnology Information (NCBI) and aligned

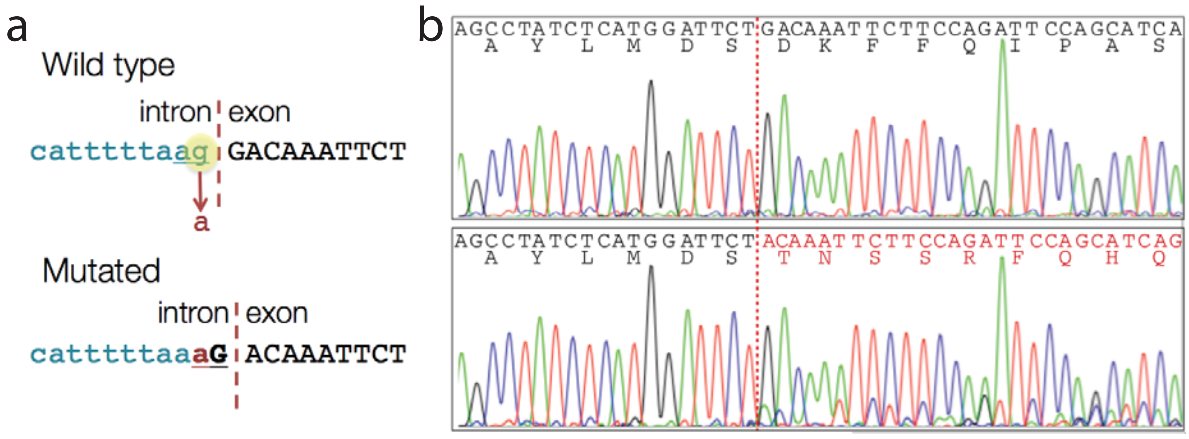

Figure 4.5: **The *POT1* splice acceptor variant affects transcript splicing.** a) Rationale for the effect of the splice acceptor variant. The splice acceptor signal sequence (AG) is underlined, and the variant is indicated. This variant is predicted to cause a one-base shift of the splice acceptor signal. b) Sequencing of the *POT1* product in a control with wild-type *POT1* (top) and a carrier of the splice acceptor variant (pedigree AF1, individual IV-1, Fig. 3.2) (bottom). The boundary between exons 17 and 18 is marked with a dotted red line. The wild-type sequence, in nucleotides and in amino acids, is indicated in black, and the sequence the variant results in is indicated in red. The mutant sequence leads to the introduction of a premature stop codon 11 amino acids downstream of the exon 17-exon 18 boundary.

them using Clustal Omega [467]. All of the missense variants we identified in *POT1* disrupt amino acids that are completely conserved through eutherians (Fig. 4.6).

To estimate the number of substitutions per site in this amino acid alignment, I used the ProtPars routine from the PHYLIP package of programmes [469]. This routine is able to infer an unrooted phylogeny from the sequences supplied by counting the number of mutations required to span all the amino acids observed at each site of the alignment, while being consistent with the genetic code. Synonymous variants are omitted from the final count, as it is assumed that they are not under selection. As such, it is capable of estimating the number of substitutions, at the DNA level, that occurred in the amino acid alignment shown in Fig. 4.6. This analysis showed higher conservation for the three altered amino acids (2, 2 and 0 substitutions at positions 89, 94 and 273, respectively) than the average for the OB folds (2.42 substitutions per site) and, in fact, the whole protein (3.49 substitutions per site) across ~450 million years of evolutionary history (since the divergence of the zebrafish and human lineages). If we only consider sequences from eutherian organisms, then no substitutions have occurred in any of these three residues, compared to 0.8 substitutions per site in the OB folds and 1.39 substitutions

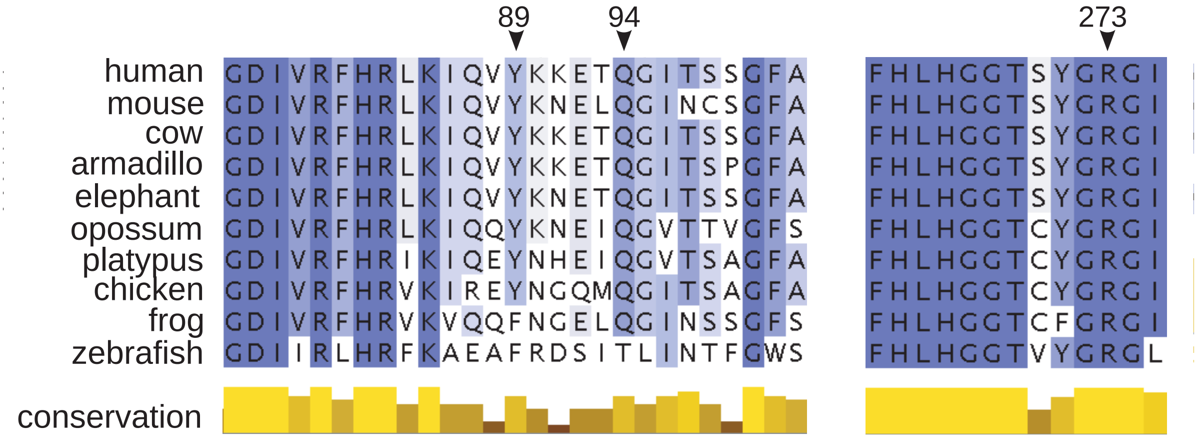
89
94
273
human GDIVRFHRLKIQVYKKETQGITSSGFA FHLHGGTSYGRGI
mouse GDIVRFHRLKIQVYKNELQGINCSGFA FHLHGGTSYGRGI
cow GDIVRFHRLKIQVYKKETQGITSSGFA FHLHGGTSYGRGI
armadillo GDIVRFHRLKIQVYKKETQGITSPGFA FHLHGGTSYGRGI
elephant GDIVRFHRLKIQVYKNETQGITSSGFA FHLHGGTSYGRGI
opossum GDIVRFHRLKIQQYKNEIQGVTTVGFS FHLHGGTCYGRGI
platypus GDIVRFHRIKIQEYNHEIQGVTSAGFA FHLHGGTCYGRGI
chicken GDIVRFHRVKIREYNGQMQGITSAGFA FHLHGGTCYGRGI
frog GDIVRFHRVKVQQFNGELQGINSSGFS FHLHGGTCFGRGI
zebrafish GDIIRLHRFKAEAFRDSITLINTFGWS FHLHGGTVYGRGL
conservation

Figure 4.6: **Highly conserved residues of POT1 are altered in familial melanoma.** Shown are the positions of the missense variants on a multi-species amino acid alignment. More conserved amino acids through evolution are shown in a darker colour. The alignment is displayed using Jalview v2.7 [468].

per site in the whole protein. I defined the OB fold regions as amino acids 8-299 in the human sequence, according to the annotation in SUPERFAMILY [419].

4.2.2.2.2 Structural modelling and characterisation of *POT1* variants Having corroborated that the positions where these variants lie are highly conserved, we then proceeded to structurally characterise the residues in which we find germline variants. To do this, we examined the structure of the POT1 protein bound to a telomere-like polynucleotide (dTUdAdGdGdGdTdTdAdG), obtained from the Protein Data Bank (PDB), id: 3KJP [470, 471]. According to this model, all three altered residues (Y89, Q94 and R273) were among 24 residues located in close proximity (<3.5 Ångströms) to the telomeric polynucleotide [462] (Fig. 4.7a). R273 interacts with the oxygen at position 2 of telomeric deoxythymidine 7, whereas Q94 and Y89 both interact with the G deoxynucleotide at position 4. Remarkably, the POT1 codon for Q94 has been found to be a target for recurrent somatic alteration (Q94R) in CLL, where ~5% of cases carry POT1 mutations that cluster in the sequences encoding the OB folds [462]. Therefore, the POT1 variants we identified are expected to weaken or abolish the interaction of POT1 with telomeres.

The list of 24 residues located in close proximity to the telomeric polynucleotide, based on this crystal structure, was defined by Andrew J. Ramsay *et al.* in a previous publication [462]. These residues are 31, 33, 36, 39–42, 48, 60, 62, 87, 89, 94, 159,

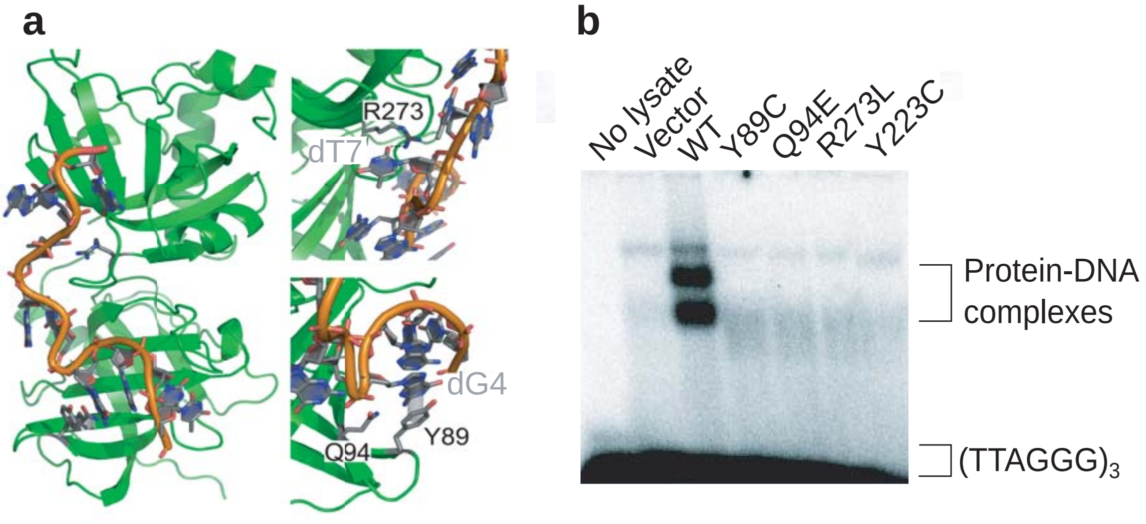

Figure 4.7: **Missense variants in *POT1* disrupt the interaction between POT1 and single-stranded DNA.** a) Shown are the locations of the POT1 Y89, Q94 and R273 residues in the two N-terminal OB folds (green). A telomere-like polynucleotide sequence is shown in orange. Interacting nucleotides in the telomeric sequence are labeled in gray. All three substitutions are predicted to disrupt the association of POT1 with telomeres. This image was generated by Víctor Quesada, at the University of Oviedo, Spain, and was rendered with PyMOL v0.99 [472]. b) Mutant Y89C, Q94E and R273L POT1 proteins are unable to bind telomeric $(TTAGGG)_3$ sequences as shown by an electrophoretic mobility shift assay. The Y223C POT1 mutant was used as a positive control representing a known disruptive alteration [462].

161, 223, 224, 243, 245, 266, 267, 270, 271 and 273. In order to assess the statistical significance of finding amino acid substitutions affecting these residues in the population, I searched 6,503 exomes released by the NHLBI GO ESP [17] for substitutions at any of the bases that would cause a change in these amino acids. The genomic positions that encode these 24 residues are shown in Table 4.2. In summary, a minimum of 6,498 exomes had all bases covered at a minimum average coverage of 59×. The variant encoding N224D was found at an overall allele frequency of 1 in 13,005. No other amino acid–changing variants were found. I then compared the number of variants found in these 24 OB domain residues in controls (1 in 6,498) to the number of variants found in all analysed pedigrees (3 in 105), obtaining a P-value of 1.54×10^{-5} using a two-tailed Fisher's exact test. This comparison indicates that our melanoma cohort is enriched for rare variants in DNA-interacting residues of POT1 when compared to the population.

The *in silico* analyses above point to these variants being deleterious, as not only are

Table 4.2: **Genomic location of bases encoding the 24 OB domain residues in close proximity to telomeres**

Amino acid	Genomic position (GRCh37)
31	g.124532351-124532353
33	g.124532345-124532347
36	g.124532336-124532338
39	g.124532327-124532329
40	g.124532324-124532326
41	g.124532321-124532323
42	g.124532320 and g.124511094-124511095
48	g.124511076-124511078
60	g.124511040-124511042
62	g.124511034-124511036
87	g.124503689-124503691
89	g.124503683-124503685
94	g.124503668-124503670
159	g.124503473-124503475
161	g.124503467-124503469
223	g.124499044-124499046
224	g.124499041-124499043
243	g.124493166-124493168
245	g.124493160-124493162
266	g.124493097-124493099
267	g.124493094-124493096
270	g.124493085-124493087
271	g.124493082-124493084
273	g.124493076-124493078

they predicted to be so by bioinformatic algorithms (Table 3.4), but variants in amino acids that participate in protein-DNA interactions seem to be extremely rare. The positions at which these variants are found are also highly conserved through evolution, and the variants themselves have the potential to impair POT1 binding to telomeres. Therefore, we decided to formally test whether the variants we identified disrupt POT1 function.

4.2.2.2.3 *In vitro* translation and G strand binding assays To test the effect of these variants on POT1 function, we assessed the ability of *in vitro*-translated POT1 proteins carrying the Y89C, Q94E and R273L variants to bind to (TTAGGG)$_3$ sequences. Human *POT1* in a T7 expression vector (Origene) was mutated by site-

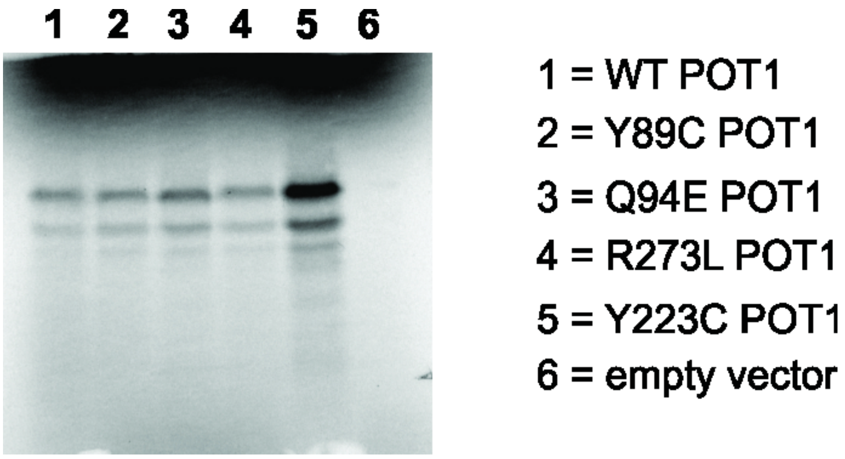

Figure 4.8: **^{35}S gel showing the *in vitro* translation products of *POT1* wild-type (WT) and OB domain mutants.** This gel confirms that each *in vitro* translation reaction successfully produced protein for the electrophoretic mobility shift assay shown in Fig. 4.7b. The Y223C variant was somatically acquired in CLL and has previously been shown to be unable to bind to telomeric DNA [462]. The DNA-protein complexes shown in 4.7b were visualised by [^{32}P]-labeling of (TTAGGG)$_3$ ssDNA.

directed mutagenesis to generate cDNAs encoding the POT1 Y89C, Q94E and R273L variants. Mutant and control T7 expression vectors were used in an *in vitro* translation reaction using the TNT coupled reticulocyte lysate kit (Promega) following the manufacturer's instructions. A 5-µl fraction of each reaction was analysed by SDS-PAGE; proteins were visualised and relative amounts were quantified using the FLA 7000 phosphorimager system (Fujifilm) (Fig. 4.8). DNA binding assays were performed as described previously with minor modifications [473]. Protein-DNA complexes were analysed by electrophoresis on a 6% polyacrylamide Tris-borate-EDTA gel run at 80 V for 3 h. Gels were visualised by exposure to a phosphorimager screen (Fig. 4.7b). DNA mutagenesis, *in vitro* translation reactions and DNA binding assays were performed by Andrew J. Ramsay, at the University of Oviedo, Spain.

Electrophoretic mobility shift assays showed a complete abolition of POT1-DNA complex formation with mutant POT1 (Fig. 4.7b). Notably, other similar variants that abolish the binding of POT1 with telomeres have been described to be somatically acquired in CLL [462]. The variants found in CLL, in particular Y36N and Y223C, promote uncapping of telomeres, telomere length extension and chromosomal aberrations and thereby promote tumourigenesis.

4.2.2.2.4 Analysis of telomere length in *POT1* variant carriers Given these observations, and the important role of POT1 in telomere length maintenance, we next asked whether melanoma cases from pedigrees with mutated *POT1* had telomere lengths that differed from those of non-carrier melanoma cases. Therefore, using exome sequence from the 41 cases that belong to the discovery phase (in which the three members of UF20 were sequenced), telomere length of each subject was estimated from NGS data. The algorithm used, called TelSeq and written by Zhihao Ding at the Sanger [474], estimates telomere length by counting TTAGGG sequences in unmapped reads in NGS data, and takes into account factors such as GC composition and read length in order to give an accurate estimation. The algorithm was shown to correlate with Southern blot measurements of telomere lengths, so it is a useful tool to compare telomere lengths in melanoma patients with and without *POT1* variants. Only the discovery phase cohort could be used because the exomes sequenced by BGI, to which the Q94 and R273 carriers belong, did not contain any unmapped reads and thus could not be assessed by TelSeq. The telomere length measurement was performed by Zhihao Ding, who was blinded to the *POT1* status of all samples.

After calculation of relative telomere length, I adjusted the 38 samples without germline *POT1* variants for age at blood draw and sex using a linear model (Fig. 4.9). I then estimated the corresponding values for *POT1* variant carriers on the basis of the same adjustment. I did these calculations with a custom R script, using the Hmisc library [475].

This analysis showed that all three members of pedigree UF20 had telomeres that were significantly longer than those in melanoma cases with wild-type *POT1*. I performed a Wilcoxon rank-sum test comparing the telomere lengths of the three Y89C cases to that for the 38 non-carrier controls (P-value <0.0002, Fig. 4.10a).

Because not all samples with missense variants could be included in the bioinformatic measurement, we decided to measure telomere lengths by PCR in all missense variant carriers, as well as other members of their families (both carriers and non carriers) and a panel of melanoma cases that did not carry *POT1* variants at these positions. These individuals had been genotyped for the detected variants before (see Subsection 3.4.1). Therefore, in this measurement, we included cases from the Leeds Melanoma case-control study, seven *POT1* missense variant carriers (pedigrees UF20, UF31 and UF23 and the carrier individual from the Leeds Melanoma cohort) and two non-carrier family controls (UF23, individual III-1 and UF20, individual III-1, Fig. 3.2). Relative mean telomere length was ascertained by SYBR Green RT-PCR using a version of the published

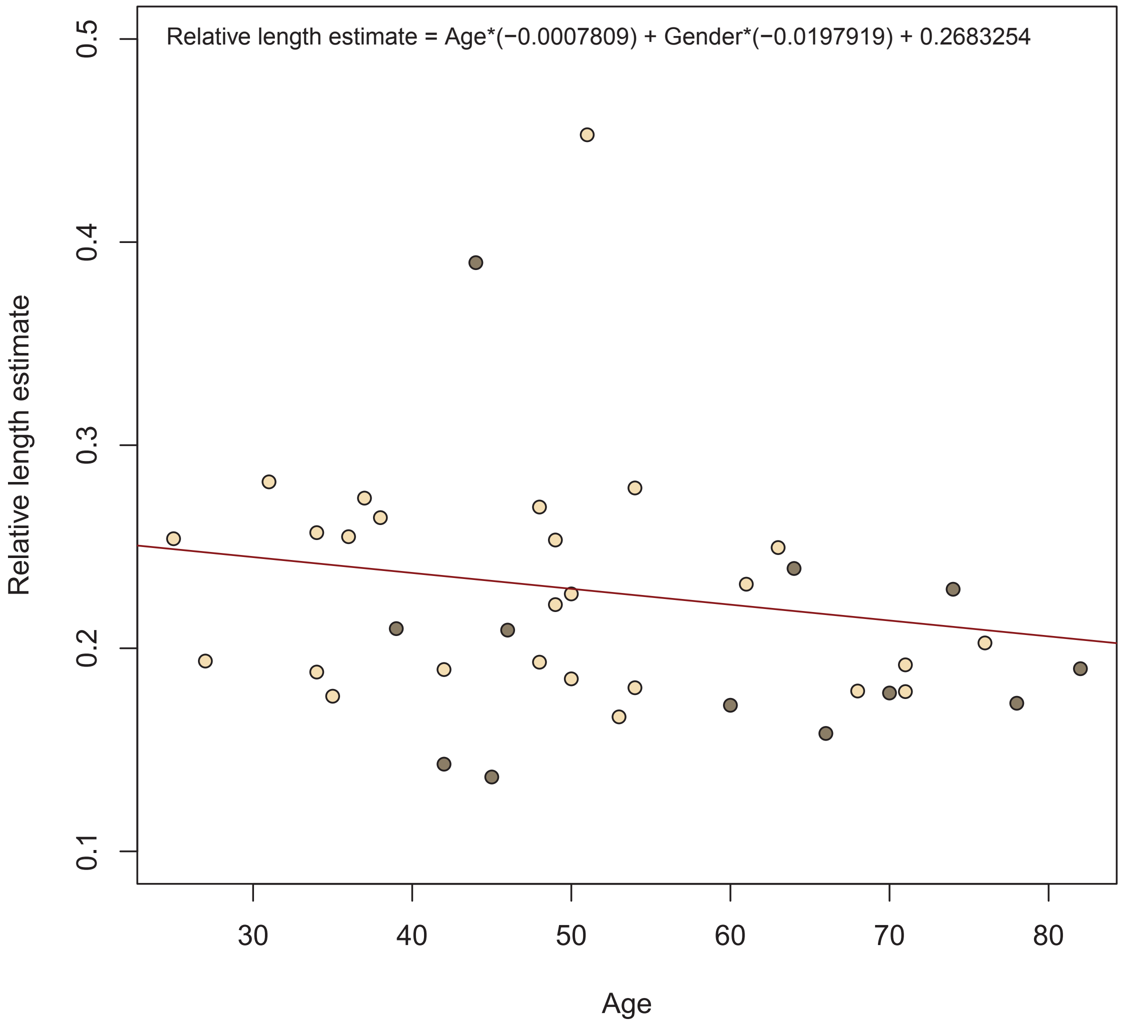

Figure 4.9: **Linear model used to adjust bioinformatically calculated telomere lengths for age at blood draw and sex.** Residuals were used as the adjusted relative telomere lengths. Dark circles represent male samples, and light circles represent females. The sex variable is coded as 0=female, 1=male. Note that only two dimensions (relative telomere length and age) are shown.

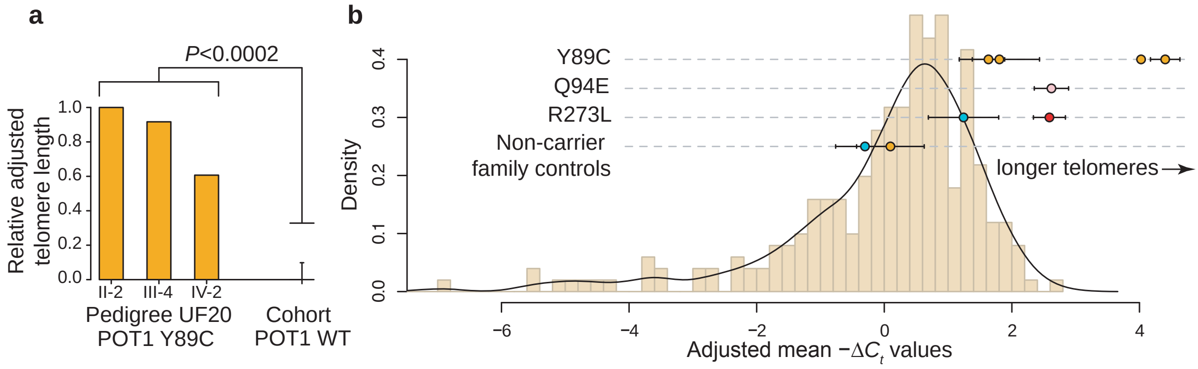

Figure 4.10: **Missense variants in *POT1* lead to elongated telomeres.** a) Calculation of telomere length from exome sequence data. Relative adjusted telomere lengths for the three sequenced members of pedigree UF20 are shown alongside the mean telomere length of 38 (all other) melanoma cases that were sequenced alongside them but were wild type for *POT1*. All values are shown relative to the largest sample measurement. Error bars, 1 s.d. b) PCR-based estimate of telomere length. Adjusted mean -ΔC_t values, which correlate positively with telomere length, for *POT1* missense variant carriers and non-carrier family controls are shown against a distribution of values from 252 melanoma cases recruited from the Leeds Melanoma cohort that are wild type at the above-mentioned positions. All measurements have been adjusted for age at blood draw and sex. The black line represents a Gaussian kernel density estimate for this set using Silverman's rule of thumb [476] for bandwidth smoothing. Orange dots, members of pedigree UF20; pink dots, members of pedigree UF31; blue dots, members of pedigree UF23; red dots, individual from the Leeds Melanoma case control study carrying the R273L variant. The number of biological replicates for each individual ranged from one to four, each with two technical replicates, for the *POT1* missense variant carriers and non-carrier family controls. Two technical replicates were performed for the 252 *POT1* non-carrier cases. Error bars, standard error of the mean.

quantitative PCR protocols [477, 478] that was modified as described previously [479]. In brief, genomic DNA was extracted from whole blood, and telomere length was ascertained by determining the ratio of detected fluorescence from the amplification of telomere repeat units (TEL) relative to fluorescence for a single-copy reference sequence from the *HBB* (β-globin) gene (CON). Telomere and control reactions were performed separately. For each assay, the PCR cycle at which each reaction crossed a predefined fluorescence threshold was determined (C_t value). The difference in the C_t values, $\Delta C_t = C_t\text{TEL} - C_t\text{CON}$, was the measure of telomere length used in the analysis, as in other published data generated using this assay [479, 480].

For the analysis, samples with $C_t\text{CON} < 18$, $C_t\text{CON} > 27$ or $C_t\text{CON} > 2$ s.d. away from the mean were removed and considered to represent failed reactions. This filtering left 252 samples from the Leeds Melanoma cohort for further analyses, with no missense variant carriers or non-carrier family controls removed. All samples had between two and eight technical replicates. The telomere length estimation by PCR was performed by Karen A. Pooley and Alison M. Dunning, at the University of Cambridge, UK, who were blinded to *POT1* status of all samples.

I then estimated mean ΔC_t values for each sample from all replicates. I then adjusted the estimated mean values of ΔC_t obtained from melanoma cases without germline *POT1* variants for age at blood draw and sex using a linear model (Fig. 4.11). The corresponding values for *POT1* variant carriers were estimated on the basis of the same adjustment. In Fig. 4.10b, adjusted mean ΔC_t values are plotted with the histogram showing the non-carrier melanoma cases compared to the missense variant carriers and the non-carrier family controls plotted above. I then performed a Wilcoxon rank-sum test comparing the adjusted mean ΔC_t values for the 252 non-carrier melanoma cases with those for the 7 missense variant carriers, yielding a P-value of 3.62×10^{-5}.

Both the bioinformatic estimation and the PCR measurements of telomere lengths agree that all *POT1* variant carriers have significantly longer telomeres than individuals without these variants, even when compared to members of the same family (Fig. 4.10b). Thus, missense variants in the OB domains of POT1 not only abolish telomere binding, but are also associated with increased telomere length. Interestingly, the melanoma case in pedigree UF23, who does not carry the R273L variant, has shorter telomeres than her carrier relative, with length comparable to non-carrier cases or the disease-free individual from pedigree UF20. I elaborate on this observation, as well as the potential tumourigenesis mechanisms of *POT1* variants, in the Discussion.

The difference in telomere lengths for *POT1* missense variant carriers when compared to controls made us question whether there would be any effect in telomere length for the splice variant carriers. Unfortunately, both the bioinformatic algorithm and the PCR measurements are extremely sensitive to variations in DNA extraction methods, sequencing methodology and the plate where the sample resides (see Subsection A.1.10). These facts did not allow us to assess samples from the Australian cohort in this analysis. However, the sample carrying the *POT1* splice donor variant showed longer telomeres when compared to samples from the Leeds case-control study when analysed by PCR, suggesting that that this variant might have some effect on telomere regulation (Fig. 4.12).

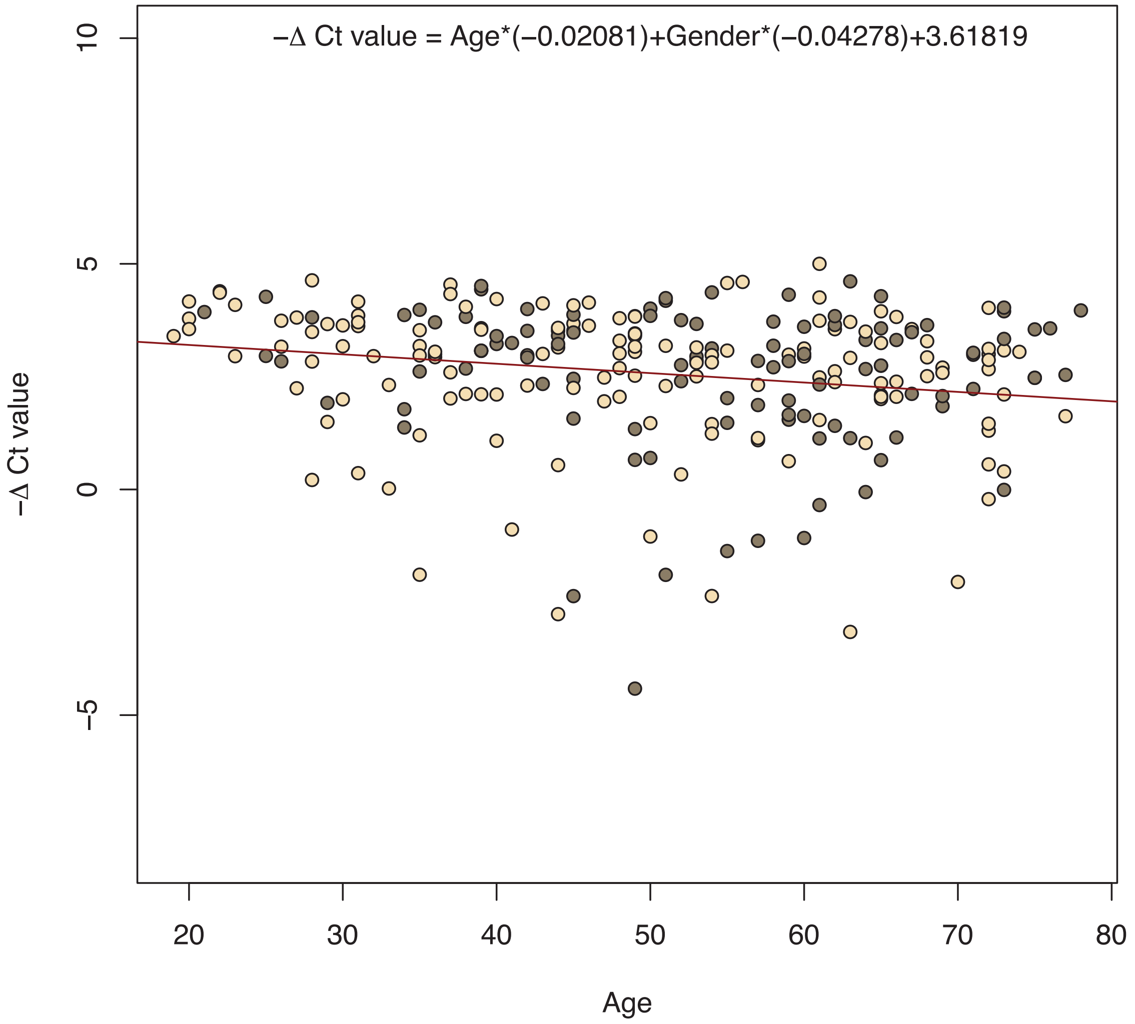

Figure 4.11: **Linear model used to adjust PCR mean ΔC_t values for age at blood draw and sex.** Residuals were used as the adjusted mean ΔC_t values. Dark circles represent male samples, and light circles represent females. The sex variable is coded as 0=female, 1=male. Note that only two dimensions (mean ΔC_t value and age) are shown.

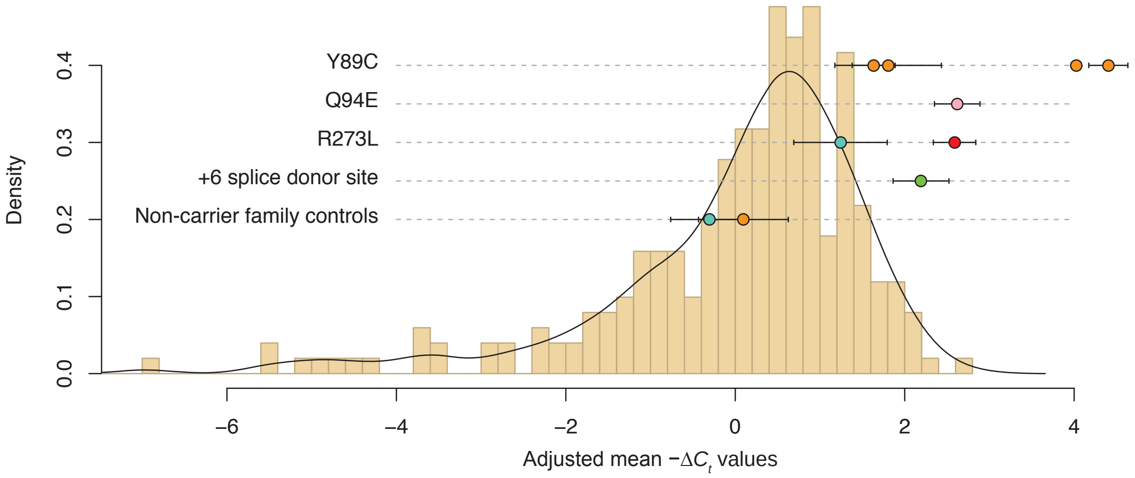

Figure 4.12: **PCR-based estimate of telomere length showing the carrier of a *POT1* intronic splice donor variant.** The carrier of the intronic splice donor variant is shown in the fourth row, in green. The rest of the figure is identical to Fig. 4.10b.

4.2.2.2.5 Investigating the role of *POT1* mutations in other cancers The identification of *POT1* mutations in CLL and the probable susceptibility of our *POT1*-mutated familial melanoma pedigrees to other tumour types (Fig. 3.2) suggest that *POT1* might have a more general role in tumourigenesis. To investigate this possibility, we examined pan-cancer data from the COSMIC [438] and IntOGen [481] databases (data from The Cancer Genome Atlas [TCGA] and the International Cancer Genome Consortium [ICGC]).

To statistically assess the mutational patterns affecting *POT1* in cancer, the COSMIC database, version 66, was mined for confirmed somatic mutations absent from the 1000 Genomes Project affecting the ORF of *POT1* across 14 cancer types (breast, central nervous system, endometrium, hematopoietic and lymphoid tissue, kidney, large intestine, liver, lung, ovary, parathyroid, prostate, skin, urinary tract and not specified). This analysis yielded 35 somatic mutations, including 4 that were silent. The total frequency of each reference/mutated base pair in the same COSMIC database was compiled. Finally, a Monte Carlo simulation with 100,000 groups of 35 mutations at random locations in the *POT1* ORF was performed. The probability of a given mutation from a reference base (*e.g.*, A to G) was forced to equal the frequency for that pair in the whole COSMIC database.

Of the 100,000 simulations performed with this method, only 2,971 contained 4 or fewer silent mutations. Therefore, the COSMIC database contains fewer silent mutations affecting *POT1* than expected by chance (P-value<0.03). To assess the clustering of mutations at sites encoding DNA-interacting residues (Table 4.2), only missense mutations were considered, as no selection would be expected for nonsense mutations. In the COSMIC database, 27 *POT1* missense mutations were found, 4 of which affected telomere-binding residues. In the Monte Carlo experiment, 9,244 simulations had exactly 27 missense mutations. In only 176 of these simulations were 4 or more residues identified that were classified as disrupting telomere binding. This result suggests that *POT1* missense mutations affect DNA-binding residues at a higher than expected rate in the COSMIC database (P-value<0.02). The Monte Carlo simulations were performed by Víctor Quesada, at the University of Oviedo, Spain.

Finally, to assess the functional impact bias of somatic mutations in *POT1*, I also considered all mutations in *POT1* that are present in the IntOGen database [481]. We chose this database because it integrates only samples that have been whole-exome sequenced and thus can provide a valid, non-biased estimate of the functional impact of mutations in *POT1* when they are compared with mutations in the rest of the exome. The frequency of *POT1* mutations in this data set is ~0.01 across 9 cancer sites (all of those available in the database, contained in the 14 sites listed above). P-values for the three studies for which the gene passed set thresholds defined in the database [481], calculated with Oncodrive-fm [482], were combined to yield a P-value of 0.021, indicating that this gene is biased toward the accumulation of functional mutations.

Therefore, although mutations in *POT1* have not been found at a high frequency in the cancer studies deposited in COSMIC and IntOGen (which integrates only whole-exome data from ICGC and TCGA as well as other studies), the mutations that have been reported show a tendency to be missense (P-value<0.03), to alter residues in close proximity to DNA (P-value<0.02), and to have a higher functional impact bias (P-value<0.03) than expected by chance. These results suggest that, although rare, somatic *POT1* mutations may drive tumourigenesis across multiple histologies.

4.2.2.3 Conclusion

We identified five different variants in *POT1* in melanoma pedigrees, three that alter amino acid residues, one predicted to affect splicing and one of uncertain significance. In this part of the study, we investigated the mechanisms by which *POT1* variant carriers might be predisposed to the development of melanoma, and possibly other cancers (Fig. 3.2).

The missense variants are found in highly evolutionarily conserved positions, in the

OB folds of *POT1* (Figs. 3.4 and 4.6). We show here that these not only abolish the interaction of POT1 with telomeres *in vitro* (Fig. 4.7b) but also that variant carriers have longer telomeres than melanoma cases without *POT1* variants, and indeed that members of their own families without *POT1* variants (Fig. 4.10). These differences persist when measurements are adjusted by age at blood draw and sex. Additionally, *POT1* variants falling in amino acids in close proximity to DNA seem to be extremely rare, as only one variant allele was found when bases coding for these 24 amino acids were examined in more than 6,500 control exomes. Accordingly, somatically acquired variants in this gene in cancer databases show a tendency to be missense rather than silent, to affect amino acids in close proximity to DNA, and to have a higher functional impact bias than expected by chance.

We also show that one of the two splice variants affects *POT1* mRNA splicing, introducing a frameshift that leads to a premature stop codon (Fig. 4.5). Although we could not show this effect for the other variant, some impact on telomere length was suggested when we examined it by PCR (Fig. 4.12).

Given the essential role of POT1 in telomere protection and length regulation, the effect that these variants have on carriers, and the dominant-negative effect that has been described for similar mutations, it is not surprising that individuals with one non-functional allele are predisposed to malignancy. I cover the biological mechanisms by which this might happen in the Discussion.

4.2.3 The effect of *ACD* variants found in melanoma pedigrees

We identified two pedigrees with variants in *ACD*, one introducing a premature stop codon and another affecting an amino acid in the POT1-binding domain (Fig. 3.4 and Table 3.4). The QIMR Berghofer team led an investigation into the effect of these variants in protein associations within the shelterin complex. Because ACD does not bind DNA, an electrophoretic mobility shift assay could not be used. Therefore, in order to analyse protein-protein interactions, our collaborators used Octet RED biolayer interferometry technology, which is able to detect changes in biomolecular interactions between an immobilised receptor attached to a biosensor surface and analytes in solution [483]. Therefore, this method allows the comparison of binding wild-type and mutated proteins to an immobilised binding partner.

Their preliminary data has shown that the ACD Q320X variant results in altered binding to POT1, compared to ACD wild-type protein (Fig. 4.13). They are in the process of generating the ACD N249S protein for testing in a similar manner. This work is on-going and the kinetics of the interactions, assessed by the same technology, will be performed shortly. The protein-protein interaction assays were performed by Dr. Antonia L. Pritchard, at QIMR Berghofer, whom I thank for supplying methods and figure.

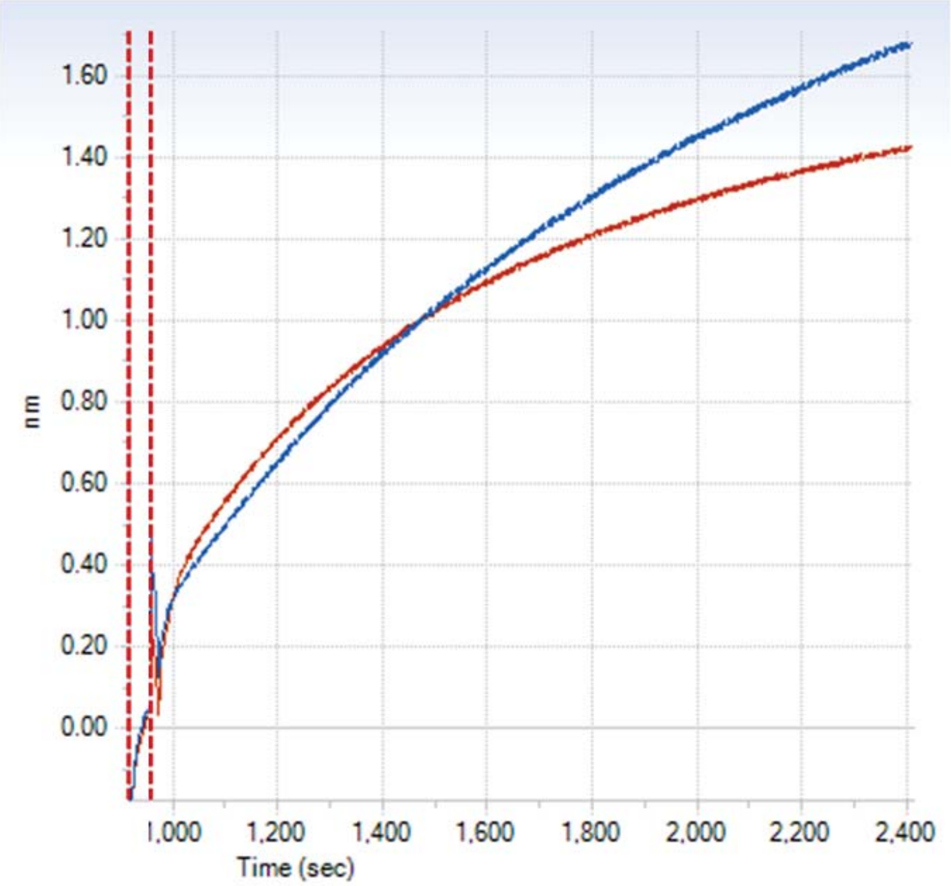

Figure 4.13: **Biolayer interferometry investigation of wild-type and mutated ACD interactions with POT1.** The POT1-ACD binding curve for the ACD wild-type protein is shown in red, whereas the one for the ACD Q320X protein is shown in blue. A lower end point indicates a lower total association, a similar slope and end point indicates a similar association between wild-type and mutated proteins, a higher end point indicates a potentially higher degree of association, while a steeper slope indicates a faster association rate. I thank Dr. Antonia L. Pritchard for this image.

Measurements for wild-type and mutant ACD when binding POT1, although they vary, were not significantly different. Although a cumulative binding effect over a lifetime cannot be discarded, we were unable to draw any conclusions from this experiment. Telomere length measurements were also performed by PCR on the ACD variant carriers, along with additional variants identified in other members in the shelterin complex (see

Subsection 4.2.4), but telomere lengths were not found to differ significantly (data not shown).

4.2.3.1 Conclusion

We identified two variants in *ACD* co-segregating with melanoma in two different pedigrees, one missense and another introducing a premature stop codon. The QIMR Berghofer team performed protein binding assays and telomere length measurements, but could not find any functional evidence for the mechanism by which these variants might be predisposing to melanoma, if indeed they are. However, it is likely that the stop codon is deleterious, as there are no stop codons found throughout the length of the protein in dbSNP release 138 [381] or the 6,500 exomes released by NHLBI GO ESP [17], but there are examples of somatically acquired stop codons in *ACD* in the COSMIC database [438]. This observation evidences that these mutational events are extremely rare in the germline, and, added to the perfect co-segregation with melanoma found in pedigree AF19, might indicate that this variant plays a role in the melanoma predisposition seen in this pedigree. However, more biological studies should be carried out to address the mechanism definitively, and to determine whether these *ACD* variants have any effect on telomere protection.

4.2.4 Other genes in the shelterin complex

Given their biological role in telomere length and protection, the rest of the members in the shelterin complex represent attractive candidate genes in melanoma susceptibility. The QIMR Berghofer team screened 601 individuals belonging to 510 families, from Australia, UK, The Netherlands, Denmark and Sweden, for germline variants in any of the members of the shelterin complex apart from *POT1*. The families tested did not have variants in *CDKN2A*, *BAP1*, *POT1*, *BRCA2*, *CDK4* and the promoter of *TERT*. We found, in addition to the two variants in *ACD* described above, a nonsense variant in *TERF2IP* co-segregating perfectly with melanoma in three members of pedigree UF3, of which one member had been sequenced as part of the discovery phase (Fig. A.1.1 and Table A.1.8).

Other variants they found co-segregating perfectly with melanoma in members of the shelterin complex are a V272M substitution in *ACD* in two members of one pedigree, a Q191R variant in *TERF2IP* in two members of another pedigree, and a M5I substitution in the same gene in one evaluated member of a pedigree, all in Australian families.

Also, they found other variants not co-segregating perfectly in these genes, as well as other genes in the shelterin complex, but these were not convincingly associated with melanoma.

4.3 Summary and conclusion

In this Chapter, I described the studies we performed to assess the biological function by which the variants detected in prioritised genes might be predisposing carriers to melanomagenesis. We were able to establish a role for variants in the shelterin complex member *POT1* in telomere length maintenance, with carriers of the variants having longer telomeres than both other non-carrier members in their families and other melanoma patients without these variants. I discuss the mechanism by which longer, and potentially unprotected, telomeres predispose to malignancy in the next Chapter.

We could not conclusively define the biological mechanism, if any, by which *SMG1* and *ACD* variants might be contributing to tumourigenesis. In the case of *SMG1*, difficulties such as cell culture contamination or assay irreproducibility prevented us from drawing any conclusions from the experiments we performed. In the case of *ACD*, although we could not pinpoint the function that might be altered in variant carriers, its potential involvement in melanoma predisposition is supported by its biological role and the absence of stop codon variants in common genomic variation datasets from thousands of individuals. However, biological assays focusing in other aspects relevant to melanoma, such as the efficiency of DNA repair upon UV damage, the rate of acquisition of chromosomal aberrations, or cell replicative ability should be performed in order to investigate the potential role of these variants.

In the next and final Chapter, I discuss the relevance of the findings presented in this dissertation and I cover the biological mechanisms that might be behind the variants identified in familial melanoma pedigrees.

Chapter 5

Discussion

5.1 Summary of the work described in this dissertation

In this dissertation, I have described the methodology we followed to search for genes that predispose, with high penetrance, to the development of melanoma. This search started with whole-exome sequencing from several individuals from families with a high clustering of melanoma cases. We started an analysis of UK and Dutch families (European phase), but during the course of that study, a large collection of Australian pedigrees became available for analysis (integrative phase). For both phases, we implemented different gene prioritisation strategies, and then attempted to experimentally validate our findings.

The European phase was further divided into two parts, the discovery and replication phases. The discovery phase involved sequencing 24 pedigrees from the UK and The Netherlands and implementing a gene prioritisation strategy to define candidate melanoma susceptibility genes. This strategy took into account diverse criteria such as the number, position and consequence of gene variants at the protein level, the probability that these are shared among members of a given pedigree, their allelic frequency in catalogues of common human genetic variation, the likelihood of finding variants in these genes in a population matched by ancestry, and the biological function of the gene. We reasoned that a high score in these attributes would increase the likelihood that these are involved in melanoma predisposition.

Having a list of candidate melanoma susceptibility genes, we then sequenced these in an additional 92 pedigrees (replication phase). We implemented a novel prioritisation strategy taking into account results from both phases of the study, and validated by PCR variants in the highest-ranked genes. From this part of the study, detailed in Chapter

2, the highest-ranked candidate gene was *SMG1*.

For the integrative phase, we extended our dataset to include a large collection of Australian melanoma pedigrees, bringing the total number of pedigrees that had been whole exome- or genome-sequenced to 105. As this set included a large proportion of families with multiple affected members sequenced, we performed a new prioritisation strategy relying mainly on co-segregation of rare, potentially deleterious variants with melanoma. From this strategy, we prioritised two genes that contribute to telomere maintenance, *POT1* and *ACD*. I cover the analyses we performed in this phase in Chapter 3.

Although the experimental evidence we could gather for the involvement of *SMG1* and *ACD* in melanoma predisposition was inconclusive, we were able to establish a role for the variants within *POT1* in the aetiology of this disease. We found that *POT1* variants in these families either interfered with mRNA splicing or rendered the protein unable to bind to ssDNA, and therefore, affected its ability to mediate telomere protection and regulate telomere length. Accordingly, carrier individuals had telomere lengths that were significantly longer than *POT1* wild-type melanoma cases and non-carrier members of their own families. Variants affecting the ability of *POT1* to bind to telomeres seem to be extremely rare in the populations examined. I explain these analyses and biological assays in Chapter 4.

In summary, this work describes the methodology we followed to identify candidate melanoma susceptibility genes starting with exome sequencing data. It discusses the difficulties that are faced by large-scale sequencing projects investigating genetic diseases, namely, the diverse attributes that may or may not be indicative of the involvement of a gene in disease risk, problems with sequencing errors, the need for availability of control datasets and the suitability of biological assays to test genetic hypotheses. It also touches on complications arising from the genetic heterogeneity of the disease and the occurrence of phenocopies in familial studies. However, this dissertation also shows a successful example of the use of exome sequencing to pinpoint causal variants in a gene, and moreover, directly implicate a biological pathway in melanoma risk.

5.2 The identification of telomere dysregulation as an important contributor to familial melanoma

In this dissertation, I describe the identification of germline variants in *POT1* in more than 4% of familial melanoma pedigrees that do not carry variants in *CDKN2A* and *CDK4*, and in 2 out of 34 pedigrees (almost 6%) with 5 or more cases, making *POT1* the second most frequently mutated high-penetrance melanoma susceptibility gene reported thus far. At the same time we reported our results, another group based at the National Cancer Institute in the US independently reported French, US and Italian melanoma-prone families with rare germline variants in *POT1* [484]. They found that in families from Romagna, Italy, the frequency of rare *POT1* variants is comparable to that of *CDKN2A* variants, and identified a founder mutation occurring in five unrelated melanoma-prone families from this region. Carriers of this variant not only had longer telomeres than controls, but also a significant increase in the number of fragile telomeres, which indicates that these disruptive *POT1* variants alter not only telomere length maintenance but also telomere integrity. In accordance with these observations, susceptibility to cutaneous malignant melanoma due to disruptive variants in *POT1* has recently been included in the Mendelian Inheritance in Man catalogue of human genetic diseases (MIM #615848).

These results implicate telomere dysregulation as an important factor in the aetiology of familial melanoma. It had been recognised by previous studies that both short and long telomeres can be risk factors for cancer development. Some examples of telomere shortening syndromes due to deleterious mutations in telomere-associated proteins are dyskeratosis congenita, ataxia telangiectasia and the Bloom and Werner syndromes (reviewed in [200]). Individuals with these conditions have higher frequencies of cancer than the general population [485–487], possibly due to chromosomal breakage and telomere fusions.

Nonetheless, individuals with long telomeres are also cancer-prone. For example, it has been observed that both breast cancer cases and women at high genetic risk for developing the disease have longer telomeres than controls, with telomere length displaying a positive correlation with risk [488, 489]. In fact, *BRCA1* and *BRCA2* mutation carriers have recently been found to have longer telomeres than controls [490], and longer telomeres have also been associated with a worse prognosis in a subset of breast cancer patients with advanced disease [489]. Other cancers in which longer telomeres have been associated with increased risk are non-Hodgkin lymphoma [491], and lung cancer in both smokers [492] and non-smokers [493]. Interestingly, a higher risk of developing melanoma, but not squamous or basal cell carcinomas, has also been associated with longer telomeres [494]. This risk is not only seen in sporadic cases, but in *CDKN2A*-negative familial cases as well [495], and in both studies, longer telomeres were associated with a higher naevus count. In accordance with these results, shorter telomeres have also been found to be associated with decreased melanoma risk [496].

The studies cited above report comparisons of telomere lengths, usually measured in whole blood or leukocytes, between groups of cancer cases and controls. Here I have described what is, to the best of our knowledge, the first hereditary mechanism underlying telomere lengthening in humans. It is known that cancer cells activate telomerase or the alternative lengthening of telomeres (ALT) pathway to bypass replicative senescence (reviewed in [497]), and thus cells from individuals with inactivating *POT1* variants might be behaving in a similar way: If *POT1* cannot inhibit telomerase, then cells might have a longer lifespan, allowing the accumulation of somatic mutations. A non-functional *POT1* might lead to progressive lengthening of telomeres, despite the telomere erosion that ensues every cell division.

In accordance with this idea, previous experiments have shown that in *POT1*-deficient cells, telomeres get progressively longer with each cell division [460, 462]. In fact, although we have limited evidence, it is tempting to speculate that families with *POT1* variants in both our study and that one from the US might be showing genetic anticipation, in the form of a higher number of primaries or an earlier age of onset with each successive generation. It has long been known that families with progressive telomere shortening show this effect due to shortening of telomeres in germ cells (reviewed in [498]), and it might be the case that *POT1* families display the same effect. However, assessment of a higher number of families and molecular studies will be necessary in order to address this question.

Moreover, telomere lengthening might not be the only mechanism by which *POT1* loss-of-function variants might contribute to tumourigenesis, but telomere dysfunction is likely to play an important role as well. As noted by previous publications [462, 484, 499], cells with defects in POT1 display loss of telomeric overhangs, chromosomal fusions and breaks, multitelomeric signals and fragile telomeres. Even so, this effect might not be severe enough to arrest cell division or lead to cell death, thus rendering *POT1* loss-of-function variants compatible with life. In accordance with this hypothesis, it has been shown that *POT1* knockdown in primary fibroblasts reduced their proliferative potential, but they nonetheless could continue dividing [459]. Interestingly, this effect could be rescued by the abrogation of the P53 or INK4A/RB1 pathways, which could offer some clues as to the additional molecular events that are necessary for *POT1*-deficient cells to progress to malignancy. The effect of a *POT1* knockdown is better tolerated than that of *TERF2* (which leads to severe telomere fusions [459]), which might also explain why we did not find any deleterious variants in *TERF1* and *TERF2* when we examined 510 melanoma-prone families from around the world (discussed in Chapter 4).

An interesting observation arises from the examination of pedigree UF23 (Fig. 3.2d). Whereas the individual that was sequenced, III-3, is a carrier of the POT1 R273L variant, her half sister (III-1), who also developed melanoma, is not. However, these two individuals presented with very different clinical characteristics: The carrier of R273L presented with three primaries, the first one at 45 years of age, whereas the non-carrier presented with a single melanoma *in situ* at 62 years of age. Additionally, the POT1 R273L variant was found in one case part of the Leeds Melanoma Case-Control study, who presented with MPM and an early age of onset, similar to the phenotype of the familial cases. This fact could point to other cancer-predisposing genetic or non-genetic aspects being shared by the two individuals of pedigree UF23, or just to the occurrence of phenocopies within the same family. Notably, individual III-1 did not seem to have longer telomeres when compared to *POT1* wild-type melanoma cases (Fig. 4.10b), further implicating the involvement of telomere dysregulation in the development of a more severe phenotype.

It is still not completely clear whether the biological mechanism by which these *POT1* variants predispose to familial melanoma is the same as that one by which mutations in the promoter of *TERT* do so, but it is unlikely. Telomere length measurements for germline *TERT* variant carriers have not been reported in the literature, but our own bioinformatic measurements in the Leeds family that was found to harbour the *TERT* promoter variant did not indicate that they had telomeres that differed significantly from controls (data not shown). Although similar somatically-acquired variants in *TERT* do lead to its increased expression [500], *TERT* has many other roles outside telomere maintenance that could be contributing to the elevated cancer risk. For example, it can act as a transcriptional modulator of the Wnt-β-catenin signalling pathway and physically occupy Wnt-dependent promoters [501], and bind the RNA component of mitochondrial RNA processing endoribonuclease (*RMRP*), [502] possibly to regulate gene expression. In addition to this, individuals with activating variants in *TERT* would not, in principle, show the fragile telomere syndrome displayed by *POT1*-deficient cells. More experiments will be necessary to address how *TERT* over-expression affects telomere maintenance, if indeed it does.

But why would variants in *POT1* lead to the development of melanoma? What is so special about the relationship between melanocytes and telomeres? This is indeed something we do not know at the moment, not unlike the question of why germline defects in the major tumour suppressor *CDKN2A* lead primarily to the development of melanoma, with other cancers manifesting at lower frequencies in carriers. It may be the case that defects in *POT1* also predispose to other cancers, as we can see in the carrier families (Fig. 3.2), and pedigrees with other cancers and germline variants in *POT1* will be described in the future. However, as case-control studies analysing melanoma risk factors have shown [494–496], longer telomere length seems to be associated with melanoma, but not squamous or basal cell carcinomas. It can also be that longer telomere length is indicative of a defect in the telomere maintenance machinery that also leads to fragile telomeres and a higher tendency to suffer telomere breaks and fusions, just as has been shown in *POT1* variant carriers. This indeed points to the idea that there is a special relationship between melanocytes and their telomeres.

Some clues as to the above question might come from experiments that have shown that DNA-damaging agents, including UV irradiation, cause damage throughout the genome but cause proportionately more damage in telomeric sequences [503]. Telomeres have been shown to be hypersensitive to UV-induced DNA damage, and to be refractory to DNA repair [504, 505]. Perhaps this effect could be exacerbated by the longer and unprotected telomeres in *POT1* variant carriers. This hypothesis would point to the involvement of other genes that contribute to UV-induced melanomagenesis, such as *TERT* and *TP53*, in melanoma tumours initiated by telomere dysfunction. However, future experiments should be performed to address the biological pathway, or pathways, by which these cells progress toward malignancy.

In conclusion, our study of the genomes of melanoma-prone pedigrees has led us to describe what is, to the best of our knowledge, the first mechanism underlying hereditary telomere lengthening in humans. This telomere lengthening is associated with high melanoma risk, and possibly other cancers. Many questions are posed by this discovery, for example, whether families with this defect display genetic anticipation, what other genetic hits are necessary for melanoma progression, and whether there is a reason why melanocytes would be more susceptible to telomere dysfunction or this is just an effect of sample ascertainment bias. The investigation of these questions might have the potential to facilitate better clinical management of families with *POT1*, and potentially variants of a similar consequence in other genes, in the future.

5.3 Future directions

As mentioned above, there are many questions that need to be addressed relating to the manner by which individuals with germline variants leading to telomere dysfunction develop cancer. Initially, to help translate this discovery to clinical practice, we need to establish the true penetrance of the variants described in this study and the one from the US. In the work presented here, we did not find any *POT1* variant carrier that did not develop melanoma, whereas the US group determined that the founder variant they identified showed dominant inheritance with incomplete penetrance [484]. Whether this effect is related to the positions of the variants themselves (*i.e.*, the variants identified in our study all affect DNA-binding residues whereas the founder mutation identified by the US group does not) or whether we need a larger sample size to see the same effect remains to be determined. We also need to determine the penetrance and biological mode of action of the splice acceptor variant. Although we only have one family with this type of variant (Fig. 3.2b), its affected members seem to have presented with melanoma later in life than the missense carriers. This observation might arise from the small sample size, but it can also mean that a variant affecting splicing does not generate a protein able to compete with wild-type POT1 and perturb protein interactions at the telomere [465]. This could imply that the mechanism by which this variant predisposes to melanoma is not dominant-negative, but haploinsufficiency. We will be able to answer these questions, hopefully, as we gather additional data from families with these types of variants. We are in the process of doing so as part of the Melanoma Genetics Consortium (GenoMEL, `http://www.genomel.org/`).

A mouse model might be useful in elucidating the additional somatic variants that must occur in *POT1* variant carriers to develop melanoma. Within our group, Chi Wong has successfully created, via CRISPR/Cas9 technology [506], mice carrying variants orthologous to the human *POT1* Y89C and Q94E variants. Analyses of their consequences might be complicated by the fact that mice have much longer telomeres than humans [507] and two *POT1* orthologues, *Pot1a* and *Pot1b* [454]. Nonetheless, we hope to be able to recapitulate the human cancer-prone phenotype, just as it was possible with shortening telomere syndromes due to variants in *TERC* [508]. We can then seek to answer questions such as whether telomeres get progressively longer throughout life, the spectrum of cancer types that these mice are predisposed to and the contribution of UV irradiation toward melanoma development.

However, mouse studies should be compared to and supported by studies of human tissue whenever possible, as it has been shown that mouse models not always recapitulate human disease perfectly (reviewed in [509]). Undoubtedly, the study of tumours from patients with *POT1* variants will prove invaluable as we endeavour to establish the genomic events leading to malignancy. We can then assess the contribution of UV-induced damage in tumour DNA and find recurrently mutated oncogenes or tumour suppressors, for example. As more patients with rare *POT1* variants are being identified by dermatologists part of GenoMEL, we hope to be able to obtain this important biological material in the coming months.

Additionally, a more in-depth investigation into the biological consequences of rare variants in *ACD* and *TERF2IP* should hopefully help us complete the picture of how telomere dysfunction predisposes to the development of melanoma. Presumably, they function as haploinsufficient alleles and lead to a malformed shelterin complex, as the variants that segregate with melanoma in these pedigrees introduce premature termination codons into the affected proteins. We hope to be able to address these questions by performing protein-protein interaction assays and introducing the detected variants into cells *in vitro*, thus being able to assess their contribution to biological processes such as cell cycle progression and telomere maintenance.

Finally, although we have identified a novel biological pathway relevant to melanoma predisposition, we cannot disregard the bigger picture: We still have 100 families in the Leeds and Leiden collections for which we have not identified any genetic predisposition loci. It could be that some of these have high-penetrance variants in genes that we have sequenced, such as *SMG1*, and we just need to identify a suitable biological assay to test them. One of the genes prioritised in the integrative phase, NIMA-related kinase 10 (*NEK10*), which plays a role in G2/M cell cycle arrest upon UV irradiation [510], is currently being investigated by our Dutch colleagues.

Variants in genes with roles in biological processes previously found to be important in melanoma development could also represent interesting candidates. In this category, we have variants segregating with melanoma within a single family in discs, large homolog 1 (*DLG1*), proteasome 26S subunit, non-ATPase, 3 (*PSMD3*), *PSMD12*, *PSMD13*, *AKT1*, minichromosome maintenance complex component 5 (*MCM5*) and nuclear protein, ataxia-telangiectasia locus (*NPAT*). All these genes encode proteins that play a role in the G1/S checkpoint of the mitotic cell cycle, the same pathway in which CDKN2A and CDK4 participate. Genes encoding proteins that participate in other cell cycle regulatory tasks or cell differentiation could represent interesting candidates, such as centrosomal protein 250kDa (*CEP250*), *CEP290*, eukaryotic translation initiation factor 4E binding protein 2 (*EIF4EBP2*) or *SOX3*. All these genes have variants segregating with melanoma in the families we studied, and could represent plausible melanoma susceptibility candidates.

Additionally, it could also be that these families have high-penetrance predisposition variants in non-exonic regions, in which case we would not have the data as we did not sequence the non-coding genome. A good example of this type of variant is the activating mutations in the promoter of *TERT* [297, 298, 500], which have been found subsequently in a large number of cancer types. In order to address this question, we are in the process of whole genome-sequencing 29 familial melanoma pedigrees from the Dutch cohort. Other possibilities that we need to consider relate to the occurrence of several low-penetrance alleles within families or the contribution of non-genetic or epigenetic effects.

In conclusion, there are many questions we are looking forward to answer relating to the role of telomere dysregulation in melanoma susceptibility, and possibly other cancers. We expect to be able to investigate these as we gather more families with rare variants in *POT1* and related genes, and with the generation of mouse models and the study of human tumours from carriers. Additionally, we need to investigate the involvement of other biological processes, and the contribution of non-genic effects, in the remaining families for which have not been able to identify any predisposition loci thus far. Hopefully, as we gather more data and exploit alternative technologies such as whole-genome or bisulphite sequencing, we will be able to help complete the description of the processes that influence genetic susceptibility to familial melanoma.

References

[1] Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines, Abhandlungern, Brünn* **4**, 3–47 (1866).

[2] Paweletz, N. Walther Flemming: pioneer of mitosis research. *Nat Rev Mol Cell Biol* **2**, 72–5 (2001).

[3] Allen, G. E. Mendel and modern genetics: the legacy for today. *Endeavour* **27**, 63–8 (2003).

[4] Sutton, W. S. On the Morphology of the Chromosome group in *Brachystola Magna*. *Biol Bull* **4**, 24–39 (1902).

[5] Boveri, T. *Ergebnisse Über Die Konstitution Der Chromatischen Substanz Des Zellkerns* (G. Fischer, Jena, 1904).

[6] Dahm, R. From discovering to understanding. Friedrich Miescher's attempts to uncover the function of DNA. *EMBO Rep* **11**, 153–60 (2010).

[7] Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–8 (1953).

[8] Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**, 729–73 (1982).

[9] Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–95 (1977).

[10] Fiers, W. *et al.* Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113–20 (1978).

[11] Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–65 (1981).

[12] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[13] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–45 (2004).

[14] Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).

[15] The 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

[16] Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur J Hum Genet* (2014).

[17] Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–20 (2013).

[18] Encode Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[19] Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045–8 (2010).

[20] Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* **30**, 224–6 (2012).

[21] Ford, C. E. & Hamerton, J. L. The chromosomes of man. *Nature* **178**, 1020–3 (1956).

[22] Tjio, J. H. & Levan, A. The Chromosome Number of Man. *Hereditas* **42**, 1–6 (1956).

[23] Gersen, S. & Keagle, M. *The Principles of Clinical Cytogenetics* (Humana Press, 2008).

[24] Morton, N. E. Parameters of the human genome. *Proc Natl Acad Sci U S A* **88**, 7474–6 (1991).

[25] Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nat Genet* **7**, 345–6 (1994).

[26] Hillier, L. W. *et al.* Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res* **15**, 1651–60 (2005).

[27] Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* **315**, 207–12 (2007).

[28] Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–5 (2009).

[29] Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–10 (2012).

[30] Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–5 (2008).

[31] Stumpf, M. P. *et al.* Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* **105**, 6959–64 (2008).

[32] Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24 (2008).

[33] Jensen, O. N. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* **7**, 391–403 (2006).

[34] Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100–4 (2013).

[35] Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–51 (2003).

[36] Mattick, J. S. RNA regulation: a new genetics? *Nat Rev Genet* **5**, 316–23 (2004).

[37] Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930–9 (2003).

[38] Encode Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–40 (2004).

[39] Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–8 (2012).

[40] The 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).

[41] Barbujani, G., Ghirotto, S. & Tassi, F. Nine things to remember about human genome diversity. *Tissue Antigens* **82**, 155–64 (2013).

[42] Macconaill, L. E. & Garraway, L. A. Clinical implications of the cancer genome. *J Clin Oncol* **28**, 5219–28 (2010).

[43] von Hansemann, D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchows Arch Path Anat* **119**, 299 (1890).

[44] Mukherjee, S. *The emperor of all maladies : a biography of cancer* (Scribner, New York, 2011), 1st Scribner trade paperback edn.

[45] Balmain, A. Cancer genetics: from Boveri and Mendel to microarrays. *Nat Rev Cancer* **1**, 77–82 (2001).

[46] Boveri, T. Uber mehrpolige mitosen als mittel zur analyse des zellkerns. *Verh D Phys Med Ges Wurzberg N F* **35**, 67–70 (1902).

[47] Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* **68**, 6863–72 (2008).

[48] Yamagiwa, K. & Ichikawa, K. Experimental Study of the Pathogenesis of Carcinoma. *J Cancer Res* **3**, 1–29 (1918).

[49] Kennaway, E. L. Further experiments on cancer-producing substances. *Biochem J* **24**, 497–504 (1930).

[50] Balmain, A. & Pragnell, I. B. Mouse skin carcinomas induced in vivo by chemical carcinogens have a transforming Harvey-*ras* oncogene. *Nature* **303**, 72–4 (1983).

[51] Carrell, C. J., Carrell, T. G., Carrell, H. L., Prout, K. & Glusker, J. P. Benzo[*a*]pyrene and its analogues: structural studies of molecular strain. *Carcinogenesis* **18**, 415–22 (1997).

[52] Croy, R. G., Essigmann, J. M., Reinhold, V. N. & Wogan, G. N. Identification of the principal aflatoxin B_1-DNA adduct formed *in vivo* in rat liver. *Proc Natl Acad Sci U S A* **75**, 1745–9 (1978).

[53] Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149–52 (1982).

[54] Brown, J. R. & Thornton, J. L. Percivall Pott (1714-1788) and chimney sweepers' cancer of the scrotum. *Br J Ind Med* **14**, 68–70 (1957).

[55] American Cancer Society. Occupation and Cancer. URL http://www.cancer.org/acs/groups/content/@nho/documents/document/occupationandcancerpdf.pdf.

[56] Boffetta, P. & Nyberg, F. Contribution of environmental factors to cancer risk. *Br Med Bull* **68**, 71–94 (2003).

[57] Cogliano, V. J. *et al.* Preventable exposures associated with human cancers. *J Natl Cancer Inst* **103**, 1827–39 (2011).

[58] Hernandez, L. G., van Steeg, H., Luijten, M. & van Benthem, J. Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach. *Mutat Res* **682**, 94–109 (2009).

[59] Liu, G., Cheresh, P. & Kamp, D. W. Molecular basis of asbestos-induced lung disease. *Annu Rev Pathol* **8**, 161–87 (2013).

[60] Tweedale, G. & Hansen, P. Protecting the workers: the medical board and the asbestos industry, 1930s-1960s. *Med Hist* **42**, 439–57 (1998).

[61] Luus, K. Asbestos: mining exposure, health effects and policy implications. *Mcgill J Med* **10**, 121–6 (2007).

[62] Committee on Asbestos: Selected Health Effects, Board on Population Health and Public Health Practice, Institute of Medicine. *Asbestos:: Selected Cancers* (National Academies Press, 2006).

[63] Kamp, D. W. & Weitzman, S. A. The molecular basis of asbestos induced lung injury. *Thorax* **54**, 638–52 (1999).

[64] Boffetta, P. Epidemiology of environmental and occupational cancer. *Oncogene* **23**, 6392–403 (2004).

[65] Newhouse, M. L., Berry, G. & Wagner, J. C. Mortality of factory workers in east London 1933-80. *Br J Ind Med* **42**, 4–11 (1985).

[66] Case, B. W., Abraham, J. L., Meeker, G., Pooley, F. D. & Pinkerton, K. E. Applying definitions of "asbestos" to environmental and "low-dose" exposure levels and health effects, particularly malignant mesothelioma. *J Toxicol Environ Health B Crit Rev* **14**, 3–39 (2011).

[67] Cordier, S. *et al.* Epidemiologic investigation of respiratory effects related to environmental exposure to asbestos inside insulated buildings. *Arch Environ Health* **42**, 303–9 (1987).

[68] Substance Abuse and Mental Health Services Administration (SAMHSA). National Survey on Drug Use and Health (NSDUH). URL http://www.samhsa.gov/data/NSDUH/2012SummNatFindDetTables/DetTabs/NSDUH-DetTabsSect2peTabs43to84-2012.htm#Tab2.71B.

[69] Office for National Statistics. 5 interesting facts about alcohol consumption in Great Britain. URL http://www.ons.gov.uk/ons/rel/ghs/opinions-and-lifestyle-survey/drinking-habits-amongst-adults--2012/sty-alcohol-consumption.html.

[70] Lamu, L. Etude de statistique clinique de 131 cas de cancer de l'oesophage et du cardia. *Archives des Maladies Digestifs et de Malnutrition* **4**, 451–456 (1910).

[71] Poschl, G. & Seitz, H. K. Alcohol and cancer. *Alcohol Alcohol* **39**, 155–65 (2004).

[72] International Agency for Research on Cancer. Alcohol Drinking. Tech. Rep. 44, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (1988). URL http://monographs.iarc.fr/ENG/Monographs/vol44/volume44.pdf.

[73] Seitz, H. K. & Stickel, F. Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat Rev Cancer* **7**, 599–612 (2007).

[74] Brooks, P. J. & Theruvathu, J. A. DNA adducts from acetaldehyde: implications for alcohol-related carcinogenesis. *Alcohol* **35**, 187–93 (2005).

[75] Travis, R. C. & Key, T. J. Oestrogen exposure and breast cancer risk. *Breast Cancer Res* **5**, 239–47 (2003).

[76] Lopes, C. F. *et al.* Concomitant consumption of marijuana, alcohol and tobacco in oral squamous cell carcinoma development and progression: recent advances and challenges. *Arch Oral Biol* **57**, 1026–33 (2012).

[77] Eichner, E. R. & Hillman, R. S. Effect of alcohol on serum folate level. *J Clin Invest* **52**, 584–91 (1973).

[78] Boffetta, P., Hashibe, M., La Vecchia, C., Zatonski, W. & Rehm, J. The burden of cancer attributable to alcohol drinking. *Int J Cancer* **119**, 884–7 (2006).

[79] Nelson, D. E. *et al.* Alcohol-attributable cancer deaths and years of potential life lost in the United States. *Am J Public Health* **103**, 641–8 (2013).

[80] International Agency for Research on Cancer. A Review of Human Carcinogens: Chemical Agents and Related Occupations. Tech. Rep. 100F, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (2012). URL http://monographs.iarc.fr/ENG/Monographs/vol100F/mono100F-21.pdf.

[81] Butlin, H. T. Cancer of the scrotum in chimney sweeps and others. II. Why foreign sweeps do not suffer from scrotal cancer. *British Medical Journal* **2**, 1–6 (1892).

[82] Poirier, M. C. Chemical-induced DNA damage and human cancer risk. *Nat Rev Cancer* **4**, 630–7 (2004).

[83] Cook, J. & Kennaway, E. L. Chemical Compounds as Carcinogenic Agents: First Supplementary Report: Literature of 1937. *Cancer Res* **33**, 50–97 (1938).

[84] Levin, W. *et al.* Carcinogenicity of benzo[*a*]pyrene 4,5-, 7,8-, and 9,10-oxides on mouse skin. *Proc Natl Acad Sci U S A* **73**, 243–7 (1976).

[85] Kim, J. H. *et al.* Metabolism of benzo[*a*]pyrene and benzo[*a*]pyrene-7,8-diol by human cytochrome P450 1B1. *Carcinogenesis* **19**, 1847–53 (1998).

[86] Volk, D. E. *et al.* Solution structure of a cis-opened (10R)-N6-deoxyadenosine adduct of (9S,10R)-9,10-epoxy-7,8,9,10-tetrahydrobenzo[*a*]pyrene in a DNA duplex. *Biochemistry* **42**, 1410–20 (2003).

[87] Mao, B. *et al.* Solution structure of the (+)-cis-anti-benzo[*a*]pyrene-dA ([BP]dA) adduct opposite dT in a DNA duplex. *Biochemistry* **38**, 10831–42 (1999).

[88] Doll, R. & Hill, A. B. Smoking and carcinoma of the lung; preliminary report. *Br Med J* **2**, 739–48 (1950).

[89] Witschi, H. A short history of lung cancer. *Toxicol Sci* **64**, 4–6 (2001).

[90] Wynder, E. L. & Graham, E. A. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *J Am Med Assoc* **143**, 329–36 (1950).

[91] International Agency for Research on Cancer. Tobacco Smoke and Involuntary Smoking. Tech. Rep. 83, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (2004). URL http://monographs.iarc.fr/ENG/Monographs/vol83/mono83-6C.pdf.

[92] Centers for Disease Control and Prevention (US); National Center for Chronic Disease Prevention and Health Promotion (US); Office on Smoking and Health (US). *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Chapter 5: Cancer* (Centers for Disease Control and Prevention (US), 2010). URL http://www.ncbi.nlm.nih.gov/books/NBK53010/.

[93] Feng, Z. *et al.* Preferential DNA damage and poor repair determine *ras* gene mutational hotspot in human cancer. *J Natl Cancer Inst* **94**, 1527–36 (2002).

[94] Tang, M. S., Zheng, J. B., Denissenko, M. F., Pfeifer, G. P. & Zheng, Y. Use of UvrABC nuclease to quantify benzo[*a*]pyrene diol epoxide-DNA adduct formation at methylated versus unmethylated CpG sites in the p53 gene. *Carcinogenesis* **20**, 1085–9 (1999).

[95] Sasco, A. J., Secretan, M. B. & Straif, K. Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer* **45 Suppl 2**, S3–9 (2004).

[96] Goel, S., Ravindra, K., Singh, R. J. & Sharma, D. Effective smoke-free policies in achieving a high level of compliance with smoke-free law: experiences from a district of North India. *Tob Control* (2013).

[97] Kostova, D. *et al.* Cigarette prices and smoking prevalence after a tobacco tax increase - Turkey, 2008 and 2012. *MMWR Morb Mortal Wkly Rep* **63**, 457–61 (2014).

[98] Abascal, W. *et al.* Tobacco control campaign in Uruguay: a population-based trend analysis. *Lancet* **380**, 1575–82 (2012).

[99] McAfee, T., Davis, K. C., Alexander, J., R. L., Pechacek, T. F. & Bunnell, R. Effect of the first federally funded US antismoking national media campaign. *Lancet* **382**, 2003–11 (2013).

[100] de Gruijl, F. R. Skin cancer and solar UV radiation. *Eur J Cancer* **35**, 2003–9 (1999).

[101] MacKie, R. M. *Skin cancer : an illustrated guide to the aetiology, clinical features, pathology and management of benign and malignant cutaneous tumours*. Focal points in dermatology (M. Dunitz ; Year Book Medical Publishers, London Chicago, 1989).

[102] International Agency for Research on Cancer. Solar and Ultraviolet Radiation. Tech. Rep. 55, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (1992). URL http://monographs.iarc.fr/ENG/Monographs/vol100F/mono100F-21.pdf.

[103] Young, C. Solar ultraviolet radiation and skin cancer. *Occup Med (Lond)* **59**, 82–8 (2009).

[104] Taylor, J. S. Unraveling the Molecular Pathway from Sunlight to Skin Cancer. *Accounts of Chemical Research* **27**, 76–82 (1994).

[105] Devary, Y., Rosette, C., DiDonato, J. A. & Karin, M. NF-kappa B activation by ultraviolet light not dependent on a nuclear signal. *Science* **261**, 1442–5 (1993).

[106] Chen, A. C., Halliday, G. M. & Damian, D. L. Non-melanoma skin cancer: carcinogenesis and chemoprevention. *Pathology* **45**, 331–41 (2013).

[107] Brash, D. E. Sunlight and the onset of skin cancer. *Trends Genet* **13**, 410–4 (1997).

[108] Ikehata, H. & Ono, T. The mechanisms of UV mutagenesis. *J Radiat Res* **52**, 115–25 (2011).

[109] Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry* (W.H. Freeman and Company, 2013), 6th edn.

[110] Parkin, D. M., Mesher, D. & Sasieni, P. 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010. *Br J Cancer* **105 Suppl 2**, S66–9 (2011).

[111] de Vries, E. & Coebergh, J. W. Cutaneous malignant melanoma in Europe. *Eur J Cancer* **40**, 2355–66 (2004).

[112] Diffey, B. L. & Norridge, Z. Reported sun exposure, attitudes to sun protection and perceptions of skin cancer risk: a survey of visitors to Cancer Research UK's SunSmart campaign website. *Br J Dermatol* **160**, 1292–8 (2009).

[113] Makin, J. K., Warne, C. D., Dobbinson, S. J., Wakefield, M. A. & Hill, D. J. Population and age-group trends in weekend sun protection and sunburn over two decades of the SunSmart programme in Melbourne, Australia. *Br J Dermatol* **168**, 154–61 (2013).

[114] Reeder, A. I., Jopson, J. A. & Gray, A. Baseline survey of sun protection policies and practices in primary school settings in New Zealand. *Health Educ Res* **24**, 778–87 (2009).

[115] Devesa, S. S., Blot, W. J. & Fraumeni, J., J. F. Declining lung cancer rates among young men and women in the United States: a cohort analysis. *J Natl Cancer Inst* **81**, 1568–71 (1989).

[116] Polednak, A. P. Tobacco control indicators and lung cancer rates in young adults by state in the United States. *Tob Control* **17**, 66–9 (2008).

[117] Klauber-DeMore, N. *Breast cancer in young women* (IOS Press, Amsterdam ; Washington, DC, 2006), Breast disease book edn.

[118] Warthin, A. Heredity with reference to carcinoma. *Arch Intern Med (Chic)* **XII(5)**, 546–555 (1913).

[119] Douglas, J. A. *et al.* History and molecular genetics of Lynch syndrome in family G: a century later. *JAMA* **294**, 2195–202 (2005).

[120] Lynch, H. T. Classics in oncology. Aldred Scott Warthin, M.D., Ph.D. (1866-1931). *CA Cancer J Clin* **35**, 345–7 (1985).

[121] Hansen, M. F. & Cavenee, W. K. Genetics of cancer predisposition. *Cancer Res* **47**, 5518–27 (1987).

[122] Fletcher, O. & Houlston, R. S. Architecture of inherited susceptibility to common cancer. *Nat Rev Cancer* **10**, 353–61 (2010).

[123] Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nat Genet* **33 Suppl**, 238–44 (2003).

[124] Schimke, R. N. Genetic aspects of multiple endocrine neoplasia. *Annu Rev Med* **35**, 25–31 (1984).

[125] Sparkes, R. S. *et al.* Gene for hereditary retinoblastoma assigned to human chromosome 13 by linkage to esterase D. *Science* **219**, 971–3 (1983).

[126] Strong, L. C., Riccardi, V. M., Ferrell, R. E. & Sparkes, R. S. Familial retinoblastoma and chromosome 13 deletion transmitted via an insertional translocation. *Science* **213**, 1501–3 (1981).

[127] Knudson, J., A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820–3 (1971).

[128] Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–6 (1986).

[129] Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–8 (2014).

[130] Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–9 (1990).

[131] Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).

[132] Nagy, R., Sweet, K. & Eng, C. Highly penetrant hereditary cancer syndromes. *Oncogene* **23**, 6445–70 (2004).

[133] Fearon, E. R. Human cancer syndromes: clues to the origin and nature of cancer. *Science* **278**, 1043–50 (1997).

[134] Chang, C. Q. *et al.* A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur J Hum Genet* **22**, 402–8 (2014).

[135] Frank, S. A. Genetic predisposition to cancer - insights from population genetics. *Nat Rev Genet* **5**, 764–72 (2004).

[136] Marsh, D. & Zori, R. Genetic insights into familial cancers– update and recent discoveries. *Cancer Lett* **181**, 125–64 (2002).

[137] Vogel, F. Genetics of retinoblastoma. *Hum Genet* **52**, 1–54 (1979).

[138] Classon, M. & Harlow, E. The retinoblastoma tumour suppressor in development and cancer. *Nat Rev Cancer* **2**, 910–7 (2002).

[139] DeCaprio, J. A. *et al.* SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. *Cell* **54**, 275–83 (1988).

[140] Whyte, P. *et al.* Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* **334**, 124–9 (1988).

[141] Henley, S. A. & Dick, F. A. The retinoblastoma family of proteins and their regulatory functions in the mammalian cell division cycle. *Cell Div* **7**, 10 (2012).

[142] Broaddus, E., Topham, A. & Singh, A. D. Incidence of retinoblastoma in the USA: 1975-2004. *Br J Ophthalmol* **93**, 21–3 (2009).

[143] Valverde, J. R., Alonso, J., Palacios, I. & Pestaña, A. *RB1* gene mutation update, a meta-analysis based on 932 reported mutations available in a searchable database. *BMC Genet* **6**, 53 (2005).

[144] Dommering, C. J. *et al.* *RB1* mutation spectrum in a comprehensive nationwide cohort of retinoblastoma patients. *J Med Genet* **51**, 366–74 (2014).

[145] Szijan, I., Lohmann, D. R., Parma, D. L., Brandt, B. & Horsthemke, B. Identification of RB1 germline mutations in Argentinian families with sporadic bilateral retinoblastoma. *J Med Genet* **32**, 475–9 (1995).

[146] Harding, F. *Breast cancer: cause, prevention, cure* (Tekline Pub., Aylesbury, 2006).

[147] Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J Cancer Sci Ther* **1**, 1–4 (2009).

[148] Bray, F., Ren, J.-S., Masuyer, E. & Ferlay, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer* **132**, 1133–45 (2013).

[149] Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–92 (1995).

[150] Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **62**, 676–89 (1998).

[151] Chen, S. & Parmigiani, G. Meta-analysis of *BRCA1* and *BRCA2* penetrance. *J Clin Oncol* **25**, 1329–33 (2007).

[152] Venkitaraman, A. R. Functions of BRCA1 and BRCA2 in the biological response to DNA damage. *J Cell Sci* **114**, 3591–8 (2001).

[153] Yoshida, K. & Miki, Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci* **95**, 866–71 (2004).

[154] Gardini, A., Baillat, D., Cesaroni, M. & Shiekhattar, R. Genome-wide analysis reveals a role for BRCA1 and PALB2 in transcriptional co-activation. *EMBO J* **33**, 890–905 (2014).

[155] Konishi, H. *et al.* Mutation of a single allele of the cancer susceptibility gene *BRCA1* leads to genomic instability in human breast epithelial cells. *Proc Natl Acad Sci U S A* **108**, 17773–8 (2011).

[156] Yata, K. *et al.* BRCA2 Coordinates the Activities of Cell-Cycle Kinases to Promote Genome Stability. *Cell Rep* (2014).

[157] Connor, F. *et al.* Tumorigenesis and a DNA repair defect in mice with a truncating *Brca2* mutation. *Nat Genet* **17**, 423–30 (1997).

[158] Collins, N. *et al.* Consistent loss of the wild type allele in breast cancers from a family linked to the *BRCA2* gene on chromosome 13q12-13. *Oncogene* **10**, 1673–5 (1995).

[159] Staff, S., Nupponen, N. N., Borg, A., Isola, J. J. & Tanner, M. M. Multiple copies of mutant *BRCA1* and *BRCA2* alleles in breast tumors from germ-line mutation carriers. *Genes Chromosomes Cancer* **28**, 432–42 (2000).

[160] Thompson, E. R. *et al.* Exome sequencing identifies rare deleterious mutations in DNA repair genes *FANCC* and *BLM* as potential breast cancer susceptibility alleles. *PLoS Genet* **8**, e1002894 (2012).

[161] Easton, D. F. & Eeles, R. A. Genome-wide association studies in cancer. *Hum Mol Genet* **17**, R109–15 (2008).

[162] Rodriguez-Bigas, M. A. *Hereditary colorectal cancer.* M.D. Anderson solid tumor oncology series (Springer, New York, 2010).

[163] Yan, H. *et al.* Conversion of diploidy to haploidy. *Nature* **403**, 723–4 (2000).

[164] Silva, F. C. C. d., Valentin, M. D., Ferreira, F. d. O., Carraro, D. M. & Rossi, B. M. Mismatch repair genes in Lynch syndrome: a review. *Sao Paulo Med J* **127**, 46–51 (2009).

[165] Nagasaka, T. *et al.* Somatic hypermethylation of *MSH2* is a frequent event in Lynch Syndrome colorectal cancers. *Cancer Res* **70**, 3098–108 (2010).

[166] Bodmer, W. F. *et al.* Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* **328**, 614–6 (1987).

[167] Powell, S. M. *et al.* *APC* mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–7 (1992).

[168] Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–70 (1996).

[169] Fearon, E. R. Molecular genetics of colorectal cancer. *Annu Rev Pathol* **6**, 479–507 (2011).

[170] Li, F. P. & Fraumeni, J. F., Jr. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Ann Intern Med* **71**, 747–52 (1969).

[171] Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233–8 (1990).

[172] Srivastava, S., Zou, Z. Q., Pirollo, K., Blattner, W. & Chang, E. H. Germ-line transmission of a mutated *p53* gene in a cancer-prone family with Li-Fraumeni syndrome. *Nature* **348**, 747–9 (1990).

[173] Malkin, D. Li-fraumeni syndrome. *Genes Cancer* **2**, 475–84 (2011).

[174] Sedlacek, Z. *et al.* Two Li-Fraumeni syndrome families with novel germline *p53* mutations: loss of the wild-type *p53* allele in only 50% of tumours. *Br J Cancer* **77**, 1034–9 (1998).

[175] Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–10 (2000).

[176] Efeyan, A. & Serrano, M. p53: guardian of the genome and policeman of the oncogenes. *Cell Cycle* **6**, 1006–10 (2007).

[177] Walls, G. V. Multiple endocrine neoplasia (MEN) syndromes. *Semin Pediatr Surg* **23**, 96–101 (2014).

[178] Wells, S. A., Jr, Pacini, F., Robinson, B. G. & Santoro, M. Multiple endocrine neoplasia type 2 and familial medullary thyroid carcinoma: an update. *J Clin Endocrinol Metab* **98**, 3149–64 (2013).

[179] Donis-Keller, H. *et al.* Mutations in the RET proto-oncogene are associated with MEN 2A and FMTC. *Hum Mol Genet* **2**, 851–6 (1993).

[180] Mulligan, L. M. *et al.* Germ-line mutations of the *RET* proto-oncogene in multiple endocrine neoplasia type 2A. *Nature* **363**, 458–60 (1993).

[181] Hofstra, R. M. *et al.* A mutation in the *RET* proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. *Nature* **367**, 375–6 (1994).

[182] Mulligan, L. M. RET revisited: expanding the oncogenic portfolio. *Nat Rev Cancer* **14**, 173–86 (2014).

[183] Besset, V., Scott, R. P. & Ibáñez, C. F. Signaling complexes and protein-protein interactions involved in the activation of the Ras and phosphatidylinositol 3-kinase pathways by the c-Ret receptor tyrosine kinase. *J Biol Chem* **275**, 39159–66 (2000).

[184] Davenport, M. P., Ward, R. L. & Hawkins, N. J. The null oncogene hypothesis and protection from cancer. *J Med Genet* **39**, 12–4 (2002).

[185] Chin, L. The genetics of malignant melanoma: lessons from mouse and man. *Nat Rev Cancer* **3**, 559–70 (2003).

[186] Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264**, 436–40 (1994).

[187] Spruck, C. H., 3rd *et al.* p16 gene in uncultured tumours. *Nature* **370**, 183–4 (1994).

[188] Cannon-Albright, L. A. *et al.* Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* **258**, 1148–52 (1992).

[189] Cairns, P. *et al.* Frequency of homozygous deletion at *p16/CDKN2* in primary human tumours. *Nat Genet* **11**, 210–2 (1995).

[190] Gruis, N. A. *et al.* Homozygotes for *CDKN2* (p16) germline mutation in Dutch familial melanoma kindreds. *Nat Genet* **10**, 351–3 (1995).

[191] Liggett, W. H., Jr & Sidransky, D. Role of the p16 tumor suppressor gene in cancer. *J Clin Oncol* **16**, 1197–206 (1998).

[192] Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nat Genet* **8**, 15–21 (1994).

[193] Ranade, K. *et al.* Mutations associated with familial melanoma impair p16INK4 function. *Nat Genet* **10**, 114–6 (1995).

[194] Hayward, N. K. Genetics of melanoma predisposition. *Oncogene* **22**, 3053–62 (2003).

[195] de Snoo, F. A. & Hayward, N. K. Cutaneous melanoma susceptibility and progression genes. *Cancer Lett* **230**, 153–86 (2005).

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

[197] Rizos, H. *et al.* A melanoma-associated germline mutation in exon 1beta inactivates p14ARF. *Oncogene* **20**, 5543–7 (2001).

[198] Flores, J. F. *et al.* Analysis of the *CDKN2A*, *CDKN2B* and *CDK4* genes in 48 Australian melanoma kindreds. *Oncogene* **15**, 2999–3005 (1997).

[199] Zuo, L. *et al.* Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma. *Nat Genet* **12**, 97–9 (1996).

[200] Kong, C. M., Lee, X. W. & Wang, X. Telomere shortening in human diseases. *FEBS J* **280**, 3180–93 (2013).

[201] Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33–6 (2002).

[202] American Cancer Society. Cancer facts & figures (2014). URL http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf.

[203] Cancer Research UK. Skin cancer incidence statistics. URL http://www.cancerresearchuk.org/cancer-info/cancerstats/types/skin/incidence/uk-skin-cancer-incidence-statistics.

[204] Rahib, L. *et al.* Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res* **74**, 2913–21 (2014).

[205] Cancer Research UK. Cancer incidence for common cancers (2014). URL http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/commoncancers/uk-cancer-incidence-statistics-for-common-cancers.

[206] Iannacone, M. R., Youlden, D. R., Baade, P. D., Aitken, J. F. & Green, A. C. Melanoma incidence trends and survival in adolescents and young adults in Queensland, Australia. *Int J Cancer* (2014).

[207] Surveillance, Epidemiology, and End Results (SEER) Program. Cancer Statistics Review 1975–2010. 5-Year relative survival for the top 5 cancer sites by age, all races, both sexes. URL http://seer.cancer.gov/csr/1975_2010/browse_csr.php?sectionSEL=32&pageSEL=sect_32_table.20.html.

[208] Coates, A. S. Systemic chemotherapy for malignant melanoma. *World J Surg* **16**, 277–81 (1992).

[209] Tsao, H., Chin, L., Garraway, L. A. & Fisher, D. E. Melanoma: from mutations to medicine. *Genes Dev* **26**, 1131–55 (2012).

[210] Sauka-Spengler, T. & Bronner-Fraser, M. A gene regulatory network orchestrates neural crest formation. *Nat Rev Mol Cell Biol* **9**, 557–68 (2008).

[211] Colombo, S., Berlin, I., Delmas, V. & Larue, L. *Classical and Nonclassical Melanocytes in Vertebrates*, chap. 2, 21–61. Melanins and Melanosomes (Wiley-VCH Verlag GmbH & Co. KGaA, 2011).

[212] Uong, A. & Zon, L. I. Melanocytes in development and cancer. *J Cell Physiol* **222**, 38–41 (2010).

[213] Lemke, G. *Developmental neurobiology* (Elsevier, Academic Press, London, 2009).

[214] Thomas, A. J. & Erickson, C. A. The making of a melanocyte: the specification of melanoblasts from the neural crest. *Pigment Cell Melanoma Res* **21**, 598–610 (2008).

[215] Bertolotto, C. Melanoma: from melanocyte to genetic alterations and clinical options. *Scientifica (Cairo)* **2013**, 635203 (2013).

[216] Gammill, L. S. & Bronner-Fraser, M. Neural crest specification: migrating into genomics. *Nat Rev Neurosci* **4**, 795–805 (2003).

[217] Bondurand, N. *et al.* Interaction among *SOX10*, *PAX3* and *MITF*, three genes altered in Waardenburg syndrome. *Hum Mol Genet* **9**, 1907–17 (2000).

[218] Hou, L., Panthier, J. J. & Arnheiter, H. Signaling and transcriptional regulation in the neural crest-derived melanocyte lineage: interactions between KIT and MITF. *Development* **127**, 5379–89 (2000).

[219] Lin, J. Y. & Fisher, D. E. Melanocyte biology and skin pigmentation. *Nature* **445**, 843–50 (2007).

[220] Slominski, A. Neuroendocrine activity of the melanocyte. *Exp Dermatol* **18**, 760–3 (2009).

[221] Le Poole, I. C. *et al.* A novel, antigen-presenting function of melanocytes and its possible relationship to hypopigmentary disorders. *J Immunol* **151**, 7284–92 (1993).

[222] Costin, G.-E. & Hearing, V. J. Human skin pigmentation: melanocytes modulate skin color in response to stress. *FASEB J* **21**, 976–94 (2007).

[223] Nestle, F. O., Di Meglio, P., Qin, J.-Z. & Nickoloff, B. J. Skin immune sentinels in health and disease. *Nat Rev Immunol* **9**, 679–91 (2009).

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

[225] Cichorek, M., Wachulska, M., Stasiewicz, A. & Tymińska, A. Skin melanocytes: biology and development. *Postepy Dermatol Alergol* **30**, 30–41 (2013).

[226] Tolleson, W. H. Human melanocyte biology, toxicology, and pathology. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **23**, 105–61 (2005).

[227] Mjaatvedt, C. H., Kern, C. B., Norris, R. A., Fairey, S. & Cave, C. L. Normal distribution of melanocytes in the mouse heart. *Anat Rec A Discov Mol Cell Evol Biol* **285**, 748–57 (2005).

[228] Yajima, I. & Larue, L. The location of heart melanocytes is specified and the level of pigmentation in the heart may correlate with coat color. *Pigment Cell Melanoma Res* **21**, 471–6 (2008).

[229] Goldgeier, M. H., Klein, L. E., Klein-Angerer, S., Moellmann, G. & Nordlund, J. J. The distribution of melanocytes in the leptomeninges of the human brain. *J Invest Dermatol* **82**, 235–8 (1984).

[230] Shosuke, I., Wakamatsu, K., d'Ischia, M., Napolitano, A. & Pezzella, A. *Structure of Melanins*, chap. 2, 167–185. Melanins and Melanosomes (Wiley-VCH Verlag GmbH I& Co. KGaA, 2011).

[231] Wenczl, E. *et al.* (Pheo)melanin photosensitizes UVA-induced DNA damage in cultured human melanocytes. *J Invest Dermatol* **111**, 678–82 (1998).

[232] Fedorow, H. *et al.* Neuromelanin in human dopamine neurons: comparison with peripheral melanins and relevance to Parkinson's disease. *Prog Neurobiol* **75**, 109–24 (2005).

[233] Delevoye, C., Giordano, F., Marks, M. & Raposo, G. *Biogenesis of Melanosomes*, chap. 9, 247–294. Melanins and Melanosomes (Wiley-VCH Verlag GmbH & Co. KGaA, 2011).

[234] Hearing, V. J. Biogenesis of pigment granules: a sensitive way to regulate melanocyte function. *J Dermatol Sci* **37**, 3–14 (2005).

[235] Marks, M. S. & Seabra, M. C. The melanosome: membrane dynamics in black and white. *Nat Rev Mol Cell Biol* **2**, 738–48 (2001).

[236] García-Borrón, J. & Olivares Sánchez, M. *Biosynthesis of Melanins*, chap. 4, 87–116. Melanins and Melanosomes (Wiley-VCH Verlag GmbH & Co. KGaA, 2011).

[237] Chakraborty, A. K. *et al.* Production and release of proopiomelanocortin (POMC) derived peptides by human melanocytes and keratinocytes in culture: regulation by ultraviolet B. *Biochim Biophys Acta* **1313**, 130–8 (1996).

[238] Kippenberger, S. *et al.* Melanocytes respond to mechanical stretch by activation of mitogen-activated protein kinases (MAPK). *Pigment Cell Res* **13**, 278–80 (2000).

[239] Berridge, M. J. Cell Signaling Biology (2012). URL http://www.biochemj.org/csb/007/csb007.pdf.

[240] Gupta, P. B. *et al.* The melanocyte differentiation program predisposes to metastasis after neoplastic transformation. *Nat Genet* **37**, 1047–54 (2005).

[241] Peinado, H., Olmeda, D. & Cano, A. Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nat Rev Cancer* **7**, 415–28 (2007).

[242] Bailey, C. M., Morrison, J. A. & Kulesa, P. M. Melanoma revives an embryonic migration program to promote plasticity and invasion. *Pigment Cell Melanoma Res* **25**, 573–83 (2012).

[243] Seong, I. *et al.* Sox10 controls migration of B16F10 melanoma cells through multiple regulatory target genes. *PLoS One* **7**, e31477 (2012).

[244] Serrone, L., Zeuli, M., Sega, F. M. & Cognetti, F. Dacarbazine-based chemotherapy for metastatic melanoma: thirty-year experience overview. *J Exp Clin Cancer Res* **19**, 21–34 (2000).

[245] Quirbt, I. *et al.* Temozolomide for the treatment of metastatic melanoma. *Curr Oncol* **14**, 27–33 (2007).

[246] Legha, S. S. *et al.* Treatment of metastatic melanoma with combined chemotherapy containing cisplatin, vinblastine and dacarbazine (CVD) and biotherapy using interleukin-2 and interferon-alpha. *Ann Oncol* **7**, 827–35 (1996).

[247] Marchesi, F. *et al.* Triazene compounds: mechanism of action and related DNA repair systems. *Pharmacol Res* **56**, 275–87 (2007).

[248] Swift, L. H. & Golsteyn, R. M. Genotoxic anti-cancer agents and their relationship to DNA damage, mitosis, and checkpoint adaptation in proliferating cancer cells. *Int J Mol Sci* **15**, 3403–31 (2014).

[249] Soengas, M. S. & Lowe, S. W. Apoptosis and melanoma chemoresistance. *Oncogene* **22**, 3138–51 (2003).

[250] Castillo Arias, J. & Galvonas Jasiulionis, M. *Melanoma: Treatments and Resistance*, chap. 16, 439–473. Melanoma - From Early Detection to Treatment (InTech, 2013).

[251] Bhatia, S., Tykodi, S. S. & Thompson, J. A. Treatment of metastatic melanoma: an overview. *Oncology (Williston Park)* **23**, 488–96 (2009).

[252] Longo, C. *et al.* De novo melanoma and melanoma arising from pre-existing nevus: in vivo morphologic differences as evaluated by confocal microscopy. *J Am Acad Dermatol* **65**, 604–14 (2011).

[253] Clark, W. H., Jr *et al.* A study of tumor progression: the precursor lesions of superficial spreading and nodular melanoma. *Hum Pathol* **15**, 1147–65 (1984).

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

[255] Poynter, J. N. *et al.* *BRAF* and *NRAS* mutations in melanoma and melanocytic nevi. *Melanoma Res* **16**, 267–73 (2006).

[256] Michaloglou, C. *et al.* BRAF$^{\mathrm{E600}}$-associated senescence-like cell cycle arrest of human naevi. *Nature* **436**, 720–4 (2005).

[257] Tsao, H., Mihm, M. C., Jr & Sheehan, C. PTEN expression in normal skin, acquired melanocytic nevi, and cutaneous melanoma. *J Am Acad Dermatol* **49**, 865–72 (2003).

[258] Miller, A. J. & Mihm, M. C., Jr. Melanoma. *N Engl J Med* **355**, 51–65 (2006).

[259] Garraway, L. A. *et al.* Integrative genomic analyses identify *MITF* as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–22 (2005).

[260] Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol Med* **12**, 406–14 (2006).

[261] Ramirez, R. D. *et al.* Progressive increase in telomerase activity from benign melanocytic conditions to malignant melanoma. *Neoplasia* **1**, 42–9 (1999).

[262] Guo, H., Carlson, J. A. & Slominski, A. Role of TRPM in melanocytes and melanoma. *Exp Dermatol* **21**, 650–4 (2012).

[263] Takata, M., Murata, H. & Saida, T. Molecular pathogenesis of malignant melanoma: a different perspective from the studies of melanocytic nevus and acral melanoma. *Pigment Cell Melanoma Res* **23**, 64–71 (2010).

[264] Kumasaka, M. Y. *et al.* A novel mouse model for *de novo* melanoma. *Cancer Res* **70**, 24–9 (2010).

[265] Soufir, N. *et al.* Association between endothelin receptor B nonsynonymous variants and melanoma risk. *J Natl Cancer Inst* **97**, 1297–301 (2005).

[266] Clark, W. H., Goldman, L. I. & Mastrangelo, M. J. *Human malignant melanoma* (Grune & Stratton, New York, 1979).

[267] Bandarchi, B., Ma, L., Navab, R., Seth, A. & Rasty, G. From melanocyte to metastatic malignant melanoma. *Dermatol Res Pract* **2010** (2010).

[268] Melanoma Know More. Types of melanoma. URL http://melanomaknowmore.com/types-of-melanoma/.

[269] Menzies, S. W. *Superficial spreading melanoma*, chap. 9a. An Atlas of Dermoscopy (CRC Press, 2004).

[270] Cohen, L. M. Lentigo maligna and lentigo maligna melanoma. *J Am Acad Dermatol* **33**, 923–36; quiz 937–40 (1995).

[271] Erkurt, M. A., Aydogdu, I., Kuku, I., Kaya, E. & Basaran, Y. Nodular melanoma presenting with rapid progression and widespread metastases: a case report. *J Med Case Rep* **3**, 50 (2009).

[272] Chamberlain, A. J., Fritschi, L. & Kelly, J. W. Nodular melanoma: patients' perceptions of presenting features and implications for earlier detection. *J Am Acad Dermatol* **48**, 694–701 (2003).

[273] Harmelin, E. S., Holcombe, R. N., Goggin, J. P., Carbonell, J. & Wellens, T. Acral lentiginous melanoma. *J Foot Ankle Surg* **37**, 540–5 (1998).

[274] Coleman, W. P., 3rd, Loria, P. R., Reed, R. J. & Krementz, E. T. Acral lentiginous melanoma. *Arch Dermatol* **116**, 773–6 (1980).

[275] World Health Organisation Classification of Tumours. *Pathology & Genetics: Skin tumours*, chap. 2 (IARC Press, 2006).

[276] Curtin, J. A. *et al.* Distinct sets of genetic alterations in melanoma. *N Engl J Med* **353**, 2135–47 (2005).

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

[278] Balch, C. M. *et al.* Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. *J Clin Oncol* **19**, 3635–48 (2001).

[279] Balch, C. M. *et al.* Final version of 2009 AJCC melanoma staging and classification. *J Clin Oncol* **27**, 6199–206 (2009).

[280] Breslow, A. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Ann Surg* **172**, 902–8 (1970).

[281] Friedman, R. J., Rigel, D. S. & Kopf, A. W. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA Cancer J Clin* **35**, 130–51 (1985).

[282] Abbasi, N. R. *et al.* Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA* **292**, 2771–6 (2004).

[283] Nachbar, F. *et al.* The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol* **30**, 551–9 (1994).

[284] Thomas, L. *et al.* Semiological value of ABCDE criteria in the diagnosis of cutaneous pigmented tumors. *Dermatology* **197**, 11–7 (1998).

[285] Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–54 (2002).

[286] Ball, N. J. *et al.* *RAS* mutations in human melanoma: a marker of malignant progression. *J Invest Dermatol* **102**, 285–90 (1994).

[287] Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–63 (2012).

[288] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).

[289] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–8 (2013).

[290] Prickett, T. D. *et al.* Exon capture analysis of G protein-coupled receptors identifies activating mutations in *GRM3* in melanoma. *Nat Genet* **43**, 1119–26 (2011).

[291] Nikolaev, S. I. *et al.* Exome sequencing identifies recurrent somatic *MAP2K1* and *MAP2K2* mutations in melanoma. *Nat Genet* **44**, 133–9 (2012).

[292] Prickett, T. D. *et al.* Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in *ERBB4*. *Nat Genet* **41**, 1127–32 (2009).

[293] Stark, M. S. *et al.* Frequent somatic mutations in *MAP3K5* and *MAP3K9* in metastatic melanoma identified by exome sequencing. *Nat Genet* **44**, 165–9 (2012).

[294] Kwong, L. N. & Davies, M. A. Navigating the therapeutic complexity of PI3K pathway inhibition in melanoma. *Clin Cancer Res* **19**, 5310–9 (2013).

[295] Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–6 (2012).

[296] Curtin, J. A., Busam, K., Pinkel, D. & Bastian, B. C. Somatic activation of KIT in distinct subtypes of melanoma. *J Clin Oncol* **24**, 4340–6 (2006).

[297] Horn, S. *et al.* *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–61 (2013).

[298] Huang, F. W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–9 (2013).

[299] Gerami, P. *et al.* Copy number gains in 11q13 and 8q24 [corrected] are highly linked to prognosis in cutaneous malignant melanoma. *J Mol Diagn* **13**, 352–8 (2011).

[300] Puig-Butillé, J. A. *et al.* Genetic alterations in RAS-regulated pathway in acral lentiginous melanoma. *Exp Dermatol* **22**, 148–50 (2013).

[301] Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic *RAC1* mutations in melanoma. *Nat Genet* **44**, 1006–14 (2012).

[302] Stefansson, B. & Brautigan, D. L. Protein phosphatase PP6 N terminal domain restricts G1 to S phase progression in human cancer cells. *Cell Cycle* **6**, 1386–92 (2007).

[303] Muthusamy, V. *et al.* Amplification of *CDK4* and *MDM2* in malignant melanoma. *Genes Chromosomes Cancer* **45**, 447–54 (2006).

[304] Hocker, T. & Tsao, H. Ultraviolet radiation and melanoma: a systematic review and analysis of reported sequence variants. *Hum Mutat* **28**, 578–88 (2007).

[305] Gartner, J. J. *et al.* Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A* **110**, 13481–6 (2013).

[306] Cronin, J. C. *et al.* Frequent mutations in the MITF pathway in melanoma. *Pigment Cell Melanoma Res* **22**, 435–44 (2009).

[307] Wei, X. *et al.* Exome sequencing identifies *GRIN2A* as frequently mutated in melanoma. *Nat Genet* **43**, 442–6 (2011).

[308] Takano, T. *et al.* Glutamate release promotes growth of malignant gliomas. *Nat Med* **7**, 1010–5 (2001).

[309] Gandini, S. *et al.* Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure. *Eur J Cancer* **41**, 45–60 (2005).

[310] Boniol, M., Autier, P., Boyle, P. & Gandini, S. Cutaneous melanoma attributable to sunbed use: systematic review and meta-analysis. *BMJ* **345**, e4757 (2012).

[311] International Agency for Research on Cancer Working Group on artificial ultraviolet (UV) light and skin cancer. The association of use of sunbeds with cutaneous malignant melanoma and other skin cancers: A systematic review. *Int J Cancer* **120**, 1116–22 (2007).

[312] El Ghissassi, F. *et al.* A review of human carcinogens–part D: radiation. *Lancet Oncol* **10**, 751–2 (2009).

[313] Sunbeds (Regulation) Act 2010 (2010).

[314] Lim, H. W. *et al.* Adverse effects of ultraviolet radiation from the use of indoor tanning equipment: time to ban the tan. *J Am Acad Dermatol* **64**, e51–60 (2011).

[315] Beane Freeman, L. E., Dennis, L. K., Lynch, C. F., Thorne, P. S. & Just, C. L. Toenail arsenic content and cutaneous melanoma in Iowa. *Am J Epidemiol* **160**, 679–87 (2004).

[316] Nelemans, P. J. *et al.* Swimming and the risk of cutaneous melanoma. *Melanoma Res* **4**, 281–6 (1994).

[317] Tynes, T., Klaeboe, L. & Haldorsen, T. Residential and occupational exposure to 50 Hz magnetic fields and malignant melanoma: a population based study. *Occup Environ Med* **60**, 343–7 (2003).

[318] Nelemans, P. J. *et al.* Melanoma and occupation: results of a case-control study in The Netherlands. *Br J Ind Med* **50**, 642–6 (1993).

[319] Rees, J. L. Genetics of hair and skin color. *Annu Rev Genet* **37**, 67–90 (2003).

[320] Schiöth, H. B. *et al.* Loss of function mutations of the human melanocortin 1 receptor are common and are associated with red hair. *Biochem Biophys Res Commun* **260**, 488–91 (1999).

[321] Mitra, D. *et al.* An ultraviolet-radiation-independent pathway to melanoma carcinogenesis in the red hair/fair skin background. *Nature* **491**, 449–53 (2012).

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

[323] Bastiaens, M. *et al.* The melanocortin-1-receptor gene is the major freckle gene. *Hum Mol Genet* **10**, 1701–8 (2001).

[324] Youl, P. *et al.* Melanoma in adolescents: a case-control study of risk factors in Queensland, Australia. *Int J Cancer* **98**, 92–8 (2002).

[325] Bliss, J. M. *et al.* Risk of cutaneous melanoma associated with pigmentation characteristics and freckling: systematic overview of 10 case-control studies. The International Melanoma Analysis Group (IMAGE). *Int J Cancer* **62**, 367–76 (1995).

[326] Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A. & Kingsley, D. M. A molecular basis for classic blond hair color in Europeans. *Nat Genet* **46**, 748–52 (2014).

[327] Grichnik, J. M., Burch, J. A., Burchette, J. & Shea, C. R. The SCF/KIT pathway plays a critical role in the control of normal human melanocyte homeostasis. *J Invest Dermatol* **111**, 233–8 (1998).

[328] White, D. & Rabago-Smith, M. Genotype-phenotype associations and human eye color. *J Hum Genet* **56**, 5–7 (2011).

[329] Ibarrola-Villava, M. *et al.* Genetic analysis of three important genes in pigmentation and melanoma susceptibility: *CDKN2A*, *MC1R* and *HERC2/OCA2*. *Exp Dermatol* **19**, 836–44 (2010).

[330] Jannot, A.-S. *et al.* Allele variations in the *OCA2* gene (pink-eyed-dilution locus) are associated with genetic susceptibility to melanoma. *Eur J Hum Genet* **13**, 913–20 (2005).

[331] Gandini, S. *et al.* Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi. *Eur J Cancer* **41**, 28–44 (2005).

[332] Duffy, K. & Grossman, D. The dysplastic nevus: from historical perspective to management in the modern era: part I. Historical, histologic, and clinical aspects. *J Am Acad Dermatol* **67**, 1.e1–16; quiz 17–8 (2012).

[333] Bishop, J. A. *et al.* Genotype/phenotype and penetrance studies in melanoma families with germline CDKN2A mutations. *J Invest Dermatol* **114**, 28–33 (2000).

[334] Tucker, M. A. *et al.* Risk of melanoma and other cancers in melanoma-prone families. *J Invest Dermatol* **100**, 350S–355S (1993).

[335] Cancer Research UK. Melanoma risks and causes (2014). URL http://www.cancerresearchuk.org/about-cancer/type/melanoma/about/melanoma-risks-and-causes.

[336] Jensen, P. *et al.* Skin cancer in kidney and heart transplant recipients and different long-term immunosuppressive therapy regimens. *J Am Acad Dermatol* **40**, 177–86 (1999).

[337] Kubica, A. W. & Brewer, J. D. Melanoma in immunosuppressed patients. *Mayo Clin Proc* **87**, 991–1003 (2012).

[338] Bertoni, J. M. *et al.* Increased melanoma risk in Parkinson disease: a prospective clinicopathological study. *Arch Neurol* **67**, 347–52 (2010).

[339] Lens, M. B. & Newton-Bishop, J. A. An association between cutaneous melanoma and non-Hodgkin's lymphoma: pooled analysis of published data with a review. *Ann Oncol* **16**, 460–5 (2005).

[340] Travis, L. B., Curtis, R. E., Hankey, B. F. & Fraumeni, J. F., Jr. Second cancers in patients with chronic lymphocytic leukemia. *J Natl Cancer Inst* **84**, 1422–7 (1992).

[341] Beisland, C., Talleraas, O., Bakke, A. & Norstein, J. Multiple primary malignancies in patients with renal cell carcinoma: a national population-based cohort study. *BJU Int* **97**, 698–702 (2006).

[342] Washington, K. & McDonagh, D. Secondary tumors of the gastrointestinal tract: surgical pathologic findings and comparison with autopsy survey. *Mod Pathol* **8**, 427–33 (1995).

[343] Draper, G. J., Sanders, B. M. & Kingston, J. E. Second primary neoplasms in patients with retinoblastoma. *Br J Cancer* **53**, 661–71 (1986).

[344] Moll, A. C., Imhof, S. M., Bouter, L. M. & Tan, K. E. Second primary tumors in patients with retinoblastoma. A review of the literature. *Ophthalmic Genet* **18**, 27–34 (1997).

[345] Gandini, S. *et al.* Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors. *Eur J Cancer* **41**, 2040–59 (2005).

[346] Ford, D. *et al.* Risk of cutaneous melanoma associated with a family history of the disease. The International Melanoma Analysis Group (IMAGE). *Int J Cancer* **62**, 377–81 (1995).

[347] Breast Cancer Linkage Consortium. Cancer risks in *BRCA2* mutation carriers. *J Natl Cancer Inst* **91**, 1310–6 (1999).

[348] Risch, H. A. *et al.* Population BRCA1 and BRCA2 mutation frequencies and cancer penetrances: a kin-cohort study in Ontario, Canada. *J Natl Cancer Inst* **98**, 1694–706 (2006).

[349] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).

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

[351] Ward, K. A., Lazovich, D. & Hordinsky, M. K. Germline melanoma susceptibility and prognostic genes: a review of the literature. *J Am Acad Dermatol* **67**, 1055–67 (2012).

[352] Bishop, D. T. *et al.* Geographical variation in the penetrance of CDKN2A mutations for melanoma. *J Natl Cancer Inst* **94**, 894–903 (2002).

[353] Begg, C. B. *et al.* Lifetime risk of melanoma in CDKN2A mutation carriers in a population-based sample. *J Natl Cancer Inst* **97**, 1507–15 (2005).

[354] Wiesner, T. *et al.* Germline mutations in *BAP1* predispose to melanocytic tumors. *Nat Genet* **43**, 1018–21 (2011).

[355] Vinagre, J. *et al.* Frequency of *TERT* promoter mutations in human cancers. *Nat Commun* **4**, 2185 (2013).

[356] Raimondi, S. *et al.* *MC1R* variants, melanoma and red hair color phenotype: a meta-analysis. *Int J Cancer* **122**, 2753–60 (2008).

[357] Bertolotto, C. *et al.* A SUMOylation-defective MITF germline mutation predisposes to melanoma and renal carcinoma. *Nature* **480**, 94–8 (2011).

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

[359] Gudbjartsson, D. F. *et al.* *ASIP* and *TYR* pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nat Genet* **40**, 886–91 (2008).

[360] Kvaskoff, M. *et al.* Polymorphisms in nevus-associated genes *MTAP*, *PLA2G6*, and *IRF4* and the risk of invasive cutaneous melanoma. *Twin Res Hum Genet* **14**, 422–32 (2011).

[361] Park, J. Y. *et al.* Gene variants in angiogenesis and lymphangiogenesis and cutaneous melanoma progression. *Cancer Epidemiol Biomarkers Prev* **22**, 827–34 (2013).

[362] Shahbazi, M. *et al.* Association between functional polymorphism in *EGF* gene and malignant melanoma. *Lancet* **359**, 397–401 (2002).

[363] Norris, W. A case of fungoid disease. *Edinb. Med. Surg.* **16**, 562–565 (1820).

[364] Lynch, H. T., Frichot, B. C., 3rd & Lynch, J. F. Familial atypical multiple mole-melanoma syndrome. *J Med Genet* **15**, 352–6 (1978).

[365] Mize, D. E., Bishop, M., Resse, E. & Sluzevich, J. Familial atypical multiple mole melanoma syndrome. *Bethesda, MD: National Center for Biotechnology Information* (2009).

[366] Vasen, H. F. *et al.* Risk of developing pancreatic cancer in families with familial atypical multiple mole melanoma associated with a specific 19 deletion of p16 (p16-Leiden). *Int J Cancer* **87**, 809–11 (2000).

[367] Shekar, S. N. *et al.* A population-based study of Australian twins with melanoma suggests a strong genetic contribution to liability. *J Invest Dermatol* **129**, 2211–9 (2009).

[368] Law, M. H., Macgregor, S. & Hayward, N. K. Melanoma genetics: recent findings take us beyond well-traveled pathways. *J Invest Dermatol* **132**, 1763–74 (2012).

[369] Coory, M. *et al.* Trends for *in situ* and invasive melanoma in Queensland, Australia, 1982-2002. *Cancer Causes Control* **17**, 21–7 (2006).

[370] Leachman, S. A. *et al.* Selection criteria for genetic assessment of patients with familial melanoma. *J Am Acad Dermatol* **61**, 677.e1–14 (2009).

[371] Lesueur, F. *et al.* The contribution of large genomic deletions at the cdkn2a locus to the burden of familial melanoma. *Br J Cancer* **99**, 364–70 (2008).

[372] Eletr, Z. M. & Wilkinson, K. D. An emerging model for BAP1's role in regulating cell cycle progression. *Cell Biochem Biophys* **60**, 3–11 (2011).

[373] Marcand, S., Brevet, V., Mann, C. & Gilson, E. Cell cycle restriction of telomere elongation. *Curr Biol* **10**, 487–90 (2000).

[374] Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* **19**, R145–51 (2010).

[375] Singleton, A. B. Exome sequencing: a transformative technology. *Lancet Neurol* **10**, 942–6 (2011).

[376] Robles-Espinoza, C. D. *et al.* *POT1* loss-of-function variants predispose to familial melanoma. *Nat Genet* **46**, 478–81 (2014).

[377] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

[378] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–8 (2011).

[379] McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–303 (2010).

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

[381] The Single Nucleotide Polymorphism Database (dbSNP). URL http://www.ncbi.nlm.nih.gov/projects/SNP/.

[382] Ledergerber, C. & Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief Bioinform* **12**, 489–97 (2011).

[383] Durtschi, J., Margraf, R. L., Coonrod, E. M., Mallempati, K. C. & Voelkerding, K. V. VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC Bioinformatics* **14 Suppl 13**, S2 (2013).

[384] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–93 (2011).

[385] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

[386] McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–70 (2010).

[387] Rashid, M., Robles-Espinoza, C. D., Rust, A. G. & Adams, D. J. Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics* **29**, 2208–10 (2013).

[388] Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).

[389] Ensembl Variation - Predicted data. URL http://www.ensembl.org/info/genome/variation/predicted_data.html.

[390] The Sequence Ontology project. URL http://www.sequenceontology.org/.

[391] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–50 (2005).

[392] Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

[393] Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691–7 (2011).

[394] BioCarta. URL http://www.biocarta.com/genes/index.asp.

[395] Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84–90 (2012).

[396] Wang, X., Terfve, C., Rose, J. C. & Markowetz, F. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics* **27**, 879–80 (2011).

[397] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300 (1995).

[398] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).

[399] Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med* **5**, 100 (2013).

[400] Tian, C., Gregersen, P. K. & Seldin, M. F. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* **17**, R143–50 (2008).

[401] Stein, L. D. Graphic Design (GD) module for Perl. URL https://github.com/lstein/Perl-GD.

[402] Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **38**, D211–22 (2010).

[403] Ferla, R. *et al.* Founder mutations in *BRCA1* and *BRCA2* genes. *Ann Oncol* **18 Suppl 6**, vi93–8 (2007).

[404] Liu, Y., Wang, L. & Zheng, P. X-linked tumor suppressors: perplexing inheritance, a unique therapeutic opportunity. *Trends Genet* **26**, 260–5 (2010).

[405] HUGO Gene Nomenclature Committee: Symbol report: WASH6P. URL http://www.genenames.org/cgi-bin/gene_symbol_report?q=data/hgnc_data.php&hgnc_id=31685.

[406] Caputo, S. *et al.* Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. *Nucleic Acids Res* **40**, D992–1002 (2012).

[407] Vallée, M. P. *et al.* Classification of missense substitutions in the BRCA genes: a database dedicated to Ex-UVs. *Hum Mutat* **33**, 22–8 (2012).

[408] Fargnoli, M. C. *et al.* Contribution of melanocortin-1 receptor gene variants to sporadic cutaneous melanoma risk in a population in central Italy: a case-control study. *Melanoma Res* **16**, 175–82 (2006).

[409] Galore-Haskel, G. *et al.* *MC1R* variant alleles and malignant melanoma risk in Israel. *Eur J Cancer* **45**, 2015–22 (2009).

[410] Landi, M. T. *et al.* MC1R, ASIP, and DNA repair in sporadic and familial melanoma in a Mediterranean population. *J Natl Cancer Inst* **97**, 998–1007 (2005).

[411] Cust, A. E. *et al.* *MC1R* genotypes and risk of melanoma before age 40 years: a population-based case-control-family study. *Int J Cancer* **131**, E269–81 (2012).

[412] Fernandez, L. *et al.* *MC1R*: three novel variants identified in a malignant melanoma association study in the Spanish population. *Carcinogenesis* **28**, 1659–64 (2007).

[413] Brumbaugh, K. M. *et al.* The mRNA surveillance protein hSMG-1 functions in genotoxic stress response pathways in mammalian cells. *Mol Cell* **14**, 585–98 (2004).

[414] Aoude, L. G. *et al.* Nonsense mutations in the shelterin complex genes *ACD* and *TERF2IP* in familial melanoma. *J Natl Cancer Inst* (2014).

[415] Bataille, V. *et al.* The association between naevi and melanoma in populations with different levels of sun exposure: a joint case-control study of melanoma in the UK and Australia. *Br J Cancer* **77**, 505–10 (1998).

[416] Aitken, J. F., Green, A. C., MacLennan, R., Youl, P. & Martin, N. G. The Queensland Familial Melanoma Project: study design and characteristics of participants. *Melanoma Res* **6**, 155–65 (1996).

[417] Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–4 (2003).

[418] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–9 (2010).

[419] Wilson, D. *et al.* SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* **37**, D380–6 (2009).

[420] Pfam. Family: *TPP1* (PF10341) (2014). URL http://pfam.xfam.org/family/PF10341.

[421] Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–8 (2012).

[422] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007).

[423] de Lange, T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* **19**, 2100–10 (2005).

[424] Ye, J. Z.-S. *et al.* POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. *Genes Dev* **18**, 1649–54 (2004).

[425] Robles-Espinoza, C. D., del Castillo Velasco-Herrera, M., Hayward, N. K. & Adams, D. J. Telomere-regulating genes and the telomere interactome in familial cancers. *Mol Cancer Res* (2014).

[426] Denning, G., Jamieson, L., Maquat, L. E., Thompson, E. A. & Fields, A. P. Cloning of a novel phosphatidylinositol kinase-related kinase: characterization of the human SMG-1 RNA surveillance protein. *J Biol Chem* **276**, 22709–14 (2001).

[427] McIlwain, D. R. *et al.* Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc Natl Acad Sci U S A* **107**, 12186–91 (2010).

[428] Chang, Y.-F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51–74 (2007).

[429] Yamashita, A., Ohnishi, T., Kashima, I., Taya, Y. & Ohno, S. Human SMG-1, a novel phosphatidylinositol 3-kinase-related protein kinase, associates with components of the mRNA surveillance complex and is involved in the regulation of nonsense-mediated mRNA decay. *Genes Dev* **15**, 2215–28 (2001).

[430] Chen, R.-Q. *et al.* Kinome siRNA screen identifies SMG-1 as a negative regulator of hypoxia-inducible factor-1alpha in hypoxia. *J Biol Chem* **284**, 16752–8 (2009).

[431] Oliveira, V. *et al.* A protective role for the human SMG-1 kinase against tumor necrosis factor-alpha-induced apoptosis. *J Biol Chem* **283**, 13174–84 (2008).

[432] Cho, H., Han, S., Park, O. H. & Kim, Y. K. SMG1 regulates adipogenesis via targeting of staufen1-mediated mRNA decay. *Biochim Biophys Acta* **1829**, 1276–87 (2013).

[433] Henderson-Smith, A. *et al.* SMG1 identified as a regulator of Parkinson's disease-associated alpha-synuclein through siRNA screening. *PLoS One* **8**, e77711 (2013).

[434] Azzalin, C. M., Reichenbach, P., Khoriauli, L., Giulotto, E. & Lingner, J. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318**, 798–801 (2007).

[435] Le, P. N., Maranon, D. G., Altina, N. H., Battaglia, C. L. R. & Bailey, S. M. TERRA, hnRNP A1, and DNA-PKcs Interactions at Human Telomeres. *Front Oncol* **3**, 91 (2013).

[436] Roberts, T. L. *et al.* Smg1 haploinsufficiency predisposes to tumor formation and inflammation. *Proc Natl Acad Sci U S A* **110**, E285–94 (2013).

[437] Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–75 (2008).

[438] Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945–50 (2011).

[439] Neben, K. *et al.* Distinct gene expression patterns associated with FLT3- and NRAS-activating mutations in acute myeloid leukemia with normal karyotype. *Oncogene* **24**, 1580–8 (2005).

[440] Tiedemann, R. E. *et al.* Kinome-wide RNAi studies in human multiple myeloma identify vulnerable kinase targets, including a lymphoid-restricted kinase, GRK6. *Blood* **115**, 1594–604 (2010).

[441] Mishra, P. J. *et al.* Dissection of RAS downstream pathways in melanomagenesis: a role for Ral in transformation. *Oncogene* **29**, 2449–56 (2010).

[442] ATCC:A-375 (ATCC® CRL-1619™) (2014). URL http://www.atcc.org/products/all/CRL-1619.aspx.

[443] Blackburn, E. H. & Gall, J. G. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena*. *J Mol Biol* **120**, 33–53 (1978).

[444] O'Sullivan, R. J. & Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat Rev Mol Cell Biol* **11**, 171–81 (2010).

[445] Xin, H. *et al.* TPP1 is a homologue of ciliate TEBP-beta and interacts with POT1 to recruit telomerase. *Nature* **445**, 559–62 (2007).

[446] Linger, B. R. & Price, C. M. Conservation of telomere protein complexes: shuffling through evolution. *Crit Rev Biochem Mol Biol* **44**, 434–46 (2009).

[447] Bianchi, A., Smith, S., Chong, L., Elias, P. & de Lange, T. TRF1 is a dimer and bends telomeric DNA. *EMBO J* **16**, 1785–94 (1997).

[448] Court, R., Chapman, L., Fairall, L. & Rhodes, D. How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. *EMBO Rep* **6**, 39–45 (2005).

[449] Loayza, D., Parsons, H., Donigian, J., Hoke, K. & de Lange, T. DNA binding features of human POT1: a nonamer 5'-TAGGGTTAG-3' minimal binding site, sequence specificity, and internal binding to multimeric sites. *J Biol Chem* **279**, 13241–8 (2004).

[450] Hockemeyer, D. *et al.* Telomere protection by mammalian Pot1 requires interaction with Tpp1. *Nat Struct Mol Biol* **14**, 754–61 (2007).

[451] Nandakumar, J. & Cech, T. R. Finding the end: recruitment of telomerase to telomeres. *Nat Rev Mol Cell Biol* **14**, 69–82 (2013).

[452] Ye, J. Z.-S. *et al.* TIN2 binds TRF1 and TRF2 simultaneously and stabilizes the TRF2 complex on telomeres. *J Biol Chem* **279**, 47264–71 (2004).

[453] Chiang, Y. J., Kim, S.-H., Tessarollo, L., Campisi, J. & Hodes, R. J. Telomere-associated protein TIN2 is essential for early embryonic development through a telomerase-independent pathway. *Mol Cell Biol* **24**, 6631–4 (2004).

[454] Hockemeyer, D., Daniels, J.-P., Takai, H. & de Lange, T. Recent expansion of the telomeric complex in rodents: Two distinct POT1 proteins protect mouse telomeres. *Cell* **126**, 63–77 (2006).

[455] Karlseder, J. *et al.* Targeted deletion reveals an essential function for the telomere length regulator Trf1. *Mol Cell Biol* **23**, 6533–41 (2003).

[456] Celli, G. B. & de Lange, T. DNA processing is not required for ATM-mediated telomere damage response after *TRF2* deletion. *Nat Cell Biol* **7**, 712–8 (2005).

[457] Sfeir, A. & de Lange, T. Removal of shelterin reveals the telomere end-protection problem. *Science* **336**, 593–7 (2012).

[458] Stansel, R. M., de Lange, T. & Griffith, J. D. T-loop assembly in vitro involves binding of TRF2 near the 3' telomeric overhang. *EMBO J* **20**, 5532–40 (2001).

[459] Hockemeyer, D., Sfeir, A. J., Shay, J. W., Wright, W. E. & de Lange, T. POT1 protects telomeres from a transient DNA damage response and determines how human chromosomes end. *EMBO J* **24**, 2667–78 (2005).

[460] Loayza, D. & De Lange, T. POT1 as a terminal transducer of TRF1 telomere length control. *Nature* **423**, 1013–8 (2003).

[461] Liu, D. *et al.* PTOP interacts with POT1 and regulates its localization to telomeres. *Nat Cell Biol* **6**, 673–80 (2004).

[462] Ramsay, A. J. *et al.* *POT1* mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet* **45**, 526–30 (2013).

[463] Wang, F. *et al.* The POT1-TPP1 telomere complex is a telomerase processivity factor. *Nature* **445**, 506–10 (2007).

[464] Colgin, L. M., Baran, K., Baumann, P., Cech, T. R. & Reddel, R. R. Human POT1 facilitates telomere elongation by telomerase. *Curr Biol* **13**, 942–6 (2003).

[465] Kendellen, M. F., Barrientos, K. S. & Counter, C. M. POT1 association with TRF2 regulates telomere length. *Mol Cell Biol* **29**, 5611–9 (2009).

[466] Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377–94 (2004).

[467] Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).

[468] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–91 (2009).

[469] Felsenstein, J. PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).

[470] Nandakumar, J., Podell, E. R. & Cech, T. R. How telomeric protein POT1 avoids RNA to achieve specificity for single-stranded DNA. *Proc Natl Acad Sci U S A* **107**, 651–6 (2010).

[471] Lei, M., Podell, E. R. & Cech, T. R. Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat Struct Mol Biol* **11**, 1223–9 (2004).

[472] PyMOL Molecular Graphics System, Version 0.99. URL http://www.pymol.org/.

[473] Baumann, P., Podell, E. & Cech, T. R. Human Pot1 (protection of telomeres) protein: cytolocalization, gene structure, and alternative splicing. *Mol Cell Biol* **22**, 8079–87 (2002).

[474] Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Res* **42**, e75 (2014).

[475] Harrell, F. E. Hmisc S function library (2004). URL http://biostat.mc.vanderbilt.edu/s/Hmisc.

[476] Silverman, B. *Density estimation for statistics and data analysis* (Chapman and Hall, 1986).

[477] McGrath, M., Wong, J. Y. Y., Michaud, D., Hunter, D. J. & De Vivo, I. Telomere length, cigarette smoking, and bladder cancer risk in men and women. *Cancer Epidemiol Biomarkers Prev* **16**, 815–9 (2007).

[478] Cawthon, R. M. Telomere measurement by quantitative PCR. *Nucleic Acids Res* **30**, e47 (2002).

[479] Pooley, K. A. *et al.* Telomere length in prospective and retrospective cancer case-control studies. *Cancer Res* **70**, 3170–6 (2010).

[480] Bojesen, S. E. *et al.* Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* **45**, 371–84, 384e1–2 (2013).

[481] Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **10**, 1081–2 (2013).

[482] Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169 (2012).

[483] Cooper, M. A. Optical biosensors in drug discovery. *Nat Rev Drug Discov* **1**, 515–28 (2002).

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

[485] Alter, B. P., Giri, N., Savage, S. A. & Rosenberg, P. S. Cancer in dyskeratosis congenita. *Blood* **113**, 6549–57 (2009).

[486] Shiloh, Y. Ataxia-telangiectasia and the Nijmegen breakage syndrome: related disorders but genes apart. *Annu Rev Genet* **31**, 635–62 (1997).

[487] Arora, H. *et al.* Bloom syndrome. *Int J Dermatol* **53**, 798–802 (2014).

[488] Gramatges, M. M., Telli, M. L., Balise, R. & Ford, J. M. Longer relative telomere length in blood from women with sporadic and familial breast cancer compared with healthy controls. *Cancer Epidemiol Biomarkers Prev* **19**, 605–13 (2010).

[489] Svenson, U. *et al.* Breast cancer survival is associated with telomere length in peripheral blood cells. *Cancer Res* **68**, 3618–23 (2008).

[490] Pooley, K. A. *et al.* Lymphocyte telomere length is long in *BRCA1* and *BRCA2* mutation carriers regardless of cancer-affected status. *Cancer Epidemiol Biomarkers Prev* **23**, 1018–24 (2014).

[491] Lan, Q. *et al.* A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of non-Hodgkin lymphoma. *Clin Cancer Res* **15**, 7429–33 (2009).

[492] Shen, M. *et al.* A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of lung cancer. *Lung Cancer* **73**, 133–7 (2011).

[493] Lan, Q. *et al.* Longer telomere length in peripheral white blood cells is associated with risk of lung cancer and the rs2736100 (*CLPTM1L-TERT*) polymorphism in a prospective cohort study among women in China. *PLoS One* **8**, e59230 (2013).

[494] Anic, G. M. *et al.* Telomere length and risk of melanoma, squamous cell carcinoma, and basal cell carcinoma. *Cancer Epidemiol* **37**, 434–9 (2013).

[495] Burke, L. S. *et al.* Telomere length and the risk of cutaneous malignant melanoma in melanoma-prone families with and without *CDKN2A* mutations. *PLoS One* **8**, e71121 (2013).

[496] Nan, H. *et al.* Shorter telomeres associate with a reduced risk of melanoma development. *Cancer Res* **71**, 6758–63 (2011).

[497] Kuilman, T., Michaloglou, C., Mooi, W. J. & Peeper, D. S. The essence of senescence. *Genes Dev* **24**, 2463–79 (2010).

[498] Armanios, M. & Blackburn, E. H. The telomere syndromes. *Nat Rev Genet* **13**, 693–704 (2012).

[499] Yang, Q., Zheng, Y.-L. & Harris, C. C. POT1 and TRF2 cooperate to maintain telomeric integrity. *Mol Cell Biol* **25**, 1070–80 (2005).

[500] Heidenreich, B. *et al.* Telomerase reverse transcriptase promoter mutations in primary cutaneous melanoma. *Nat Commun* **5**, 3401 (2014).

[501] Park, J.-I. *et al.* Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature* **460**, 66–72 (2009).

[502] Maida, Y. *et al.* An RNA-dependent RNA polymerase formed by TERT and the *RMRP* RNA. *Nature* **461**, 230–5 (2009).

[503] Gilchrest, B. A., Eller, M. S. & Yaar, M. Telomere-mediated effects on melanogenesis and skin aging. *J Investig Dermatol Symp Proc* **14**, 25–31 (2009).

[504] Kruk, P. A., Rampino, N. J. & Bohr, V. A. DNA damage and repair in telomeres: relation to aging. *Proc Natl Acad Sci U S A* **92**, 258–62 (1995).

[505] Rochette, P. J. & Brash, D. E. Human telomeres are hypersensitive to UV-induced DNA Damage and refractory to repair. *PLoS Genet* **6**, e1000926 (2010).

[506] Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–8 (2013).

[507] Kipling, D. & Cooke, H. J. Hypervariable ultra-long telomeres in mice. *Nature* **347**, 400–2 (1990).

[508] Blasco, M. A. *et al.* Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. *Cell* **91**, 25–34 (1997).

[509] Robles-Espinoza, C. D. & Adams, D. J. Cross-species analysis of mouse and human cancer genomes. *Cold Spring Harb Protoc* **2014**, 350–8 (2014).

[510] Moniz, L. S. & Stambolic, V. Nek10 mediates G2/M cell cycle arrest and MEK autoactivation in response to UV irradiation. *Mol Cell Biol* **31**, 30–42 (2011).

[511] Dreszer, T. R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* **40**, D918–23 (2012).

Appendix A

Electronic files and supplementary information

A.1 Description of electronic files

I have placed these files in the accompanying CD, as they are large files that I do not think need to be printed. In this Appendix, I have written a short description detailing each of these, as well as the file name with which they can be located.

A.1.1 Pedigree structures for 24 melanoma-prone families sequenced as part of the discovery phase

File name: `Figure_A_1_1.pdf`

This PDF file contains pedigree structures for the 24 pedigrees that were sequenced as part of the discovery phase, one per page. Squares represent males, circles represent females, diamonds represent individuals of undisclosed sex. Individuals that had their whole exome sequenced are shown with a red outline. Types of cancer are indicated in the legend. Note that pedigrees have been adjusted to protect the identity of the families without loss of scientific integrity.

A.1.2 Detailed information about recurrently mutated genes

File name: `Table_A_1_2.xlsx`

This table contains the list of 344 genes that had mutations in two or more pedigrees. The number of families with mutations, the number of members in each family and the coding length of the gene are indicated.

A.1.3 List of biological pathways ordered by *P*-value after hypergeometric tests on list of recurrently mutated genes

File name: `Table_A_1_3.xlsx`

This table contains the results of the analysis to identify overrepresented biological pathways in our set of 344 recurrently mutated genes. All pathways with a P-value<0.5 are in this table, pathways in green were the ones included in the list for targeted sequencing. Note that the acute myocardial infarction pathway from Biocarta has an adjusted P-value < 0.05 and was not included in the set for targeted sequencing; the reason is that the original analysis was performed with the R package HTSanalyzeR version 2.8.0, which had an error in the code for the hypergeometric function. This error was corrected for the next version (2.8.3), which is the one I used to generate the table presented here.

A.1.4 All genes included for capture in the replication phase

File name: `Table_A_1_4.xlsx`

All 701 genes in that were targeted for sequencing in an additional set of 94 melanoma cases. The Ensembl identifier, along with the coding length and the reason for inclusion are indicated. “Direct evidence” means the gene was mutated in two or more pedigrees, “ABC transporter”, “Pantothenate and CoA biosynthesis pathway” and “Linkage to MAPK signalling for integrins” indicate that the gene was included because it is part of a pathway that was found to be overrepresented in the recurrently mutated genes. “Disruptive consequence” means that the gene was included because it had a premature stop codon, a frameshift variant or a splice acceptor or donor variant in any one of the familial melanoma pedigrees. Genes marked as “previous evidence for involvement in melanoma/cancer” were included because they have been found linked to processes relevant to melanoma.

A.1.5 List of ranked genes after prioritisation stage

File name: `Table_A_1_5.xlsx`

This table includes the genes captured in both the discovery and replication phases, with their respective values for the variables used in the prioritisation method. The genes highlighted in red are regarded as uninformative, as they have either a single variant detected in the melanoma pedigrees and no variants in the control exomes, or were found to have a coding length of 0.

A.1.6 List of variants in the analysis of founder mutations and those in single pedigrees

File name: `Table_A_1_6.xlsx`

This table includes the list of variants predicted to affect protein function that were present in more than one pedigree for which co-segregation information was available. The position of the variants, the number of pedigrees with their respective number of members and the consequences of the variants are indicated.

A.1.7 Genes found with only one variant in multi-case pedigrees

File name: `Table_A_1_7.xlsx`

List of the genes that were found with variants segregating with melanoma in only one pedigree, and that were not present in the prioritisation stage.

A.1.8 Pedigrees sequenced as part of the integrative phase

File name: `Table_A_1_8.xlsx`

The number of cases in each pedigree is indicated, as well as the number of cases sequenced (exomes or whole genomes). Whole genomes were only sequenced as part of the QFMP cohort.

A.1.9 Genes with co-segregating variants from the 28 pedigrees for which we had sequence data for 3 or more members and their GO terms

File name: `Table_A_1_9.xlsx`

Genes that had variants co-segregating with melanoma from the 28 pedigrees for which we had sequence data for 3 or more members and their GO terms.

A.1.10 Wide variation in telomere measurements when samples have not been processed in the same manner

File name: `Figure_A_1_10.pdf`

This graph shows the telomere measurements for the 41 samples that belong to the discovery phase cohort and that were sequenced at the Sanger Institute alongside telomere length estimates for samples part of the UK10K, sequenced at the same institute (all shown in a white background), and samples processed at QFMP (in a blue background, sample origin is indicated at the bottom). Samples with *POT1* variants are indicated with red arrows, two whole genomes part of the QFMP cohort are indicated with green arrows. Samples within the QFMP cohort showed much more variability in telomere length measurement and thus could not be assessed with the bioinformatic method.

A.2 Supplementary tables, figures and notes

A.2.1 Removal of genes likely to be false positives after filtering

I manually scanned the list of candidates after filtering and decided to remove four that are likely to be false positives (titin [*TTN*], obscurin [*OBSCN*], dystrophin [*DMD*] and maestro heat-like repeat family member 2A [*MROH2A*]) due to their length and/or their ubiquity in other cancer screens (based on analyses performed by Vertebrate Resequencing Informatics at the Sanger). I also inspected the variants for their presence in repeat regions, and a further five genes were excluded given that the mutations we detected on these overlapped with the RepeatMasker track from the University of California, Santa Cruz (UCSC) Genome Browser [511]: (lysine-specific demethylase 6B [*KDM6B*], WD repeat domain 87 [*WDR87*], Zinc finger protein 589 [*ZNF589*], choline kinase alpha [*CHKA*] and abnormal spindle homolog, microcephaly associated [*ASPM*]). This left 344 genes for further consideration.

Table A.1: Tools and parameters used for read alignment and variant calling in the discovery and replication phases

Step	Reference dataset	Tool	Version	Parameters
Discovery phase				
Read alignment to reference genome	GRCh37	BWA [377]	0.5.8c, 0.5.9	`-q 15 -t 6`
Alignment improvement				
Duplicate marking	-	Picard MarkDuplicates [378]	1.47	-
Indel realignment	dbSNP 129	GATK IndelRealigner [379]	1.1-5	`-LOD 0.4 -model KNOWNS_ONLY -entropy 0.15`
Quality score recalibration		GATK TableRecalibration [379]	1.1-5	-
Variant calling	-	SAMtools mpileup [380]	0.1.17	`-DRS -d 10000 -C50 -m2 -F0.0005 -aug -P ILLUMINA`
Replication phase				
Read alignment to reference genome	GRCh37d5	BWA [377]	0.5.9	`-q 15 -t 6`
Alignment improvement				
Duplicate marking	-	Picard MarkDuplicates [378]	1.5-9	-
Indel realignment	1000 Genomes Phase 2 Indels [15]	GATK IndelRealigner GATK CountCovariates [379]	1.5-9	`-LOD 0.4 -model KNOWNS_ONLY`
Quality score recalibration	-	GATK TableRecalibration [379]	1.5-9	-
Variant calling	-	SAMtools mpileup / Bcftools [380]	SAMTools: 0.1.18 Bcftools: 0.1.17-dev	SAMtools: `-EDS -d 10000 -C50 -m2 -F0.0005` Bcftools: `view -p 0.99 -vcgN`

Appendix B

Articles published during my PhD

During the course of my PhD, I was part of several publications, that were either directly associated with the work described in this dissertation or represented other work I carried out. In this Appendix, I list these publications and provide a short explanation of their main findings and my contribution to each of them. The physical articles can be found at the end of the dissertation. An asterisk denotes that those authors contributed equally.

B.1 Articles directly associated with this dissertation

- Robles-Espinoza CD*, Harland M*, Ramsay AJ*, Aoude LG*, Quesada V, Ding Z, Pooley KA, Pritchard AL, Tiffen JC, Petljak M, Palmer JM, Symmons J, Johansson P, Stark MS, Gartside MG, Snowden H, Montgomery GW, Martin NG, Liu JZ, Choi J, Makowski M, Brown KM, Dunning AM, Keane TM, López-Otín C, Gruis NA, Hayward NK, Bishop DT, Newton-Bishop JA and Adams DJ (2014). *POT1* loss-of-function variants predispose to familial melanoma. *Nature Genetics* **45**(5): 478-81. doi: 10.1038/ng.2947.

This is the main publication explaining the results from my dissertation. This letter explains the finding of familial melanoma pedigrees with rare variants in *POT1* and their consequences in carriers. I performed most of the bioinformatic analyses, including data processing and filtering, analysis and comparison with control exomes, computational assessment of variant conservation and pathogenicity, and analysis of telomere length data.

- Aoude LG*, Pritchard AL*, Robles-Espinoza CD*, Wadt K*, Harland M*, Choi J, Gartside M, Quesada V, Johansson P, Palmer JM, Ramsay AJ, Zhang X, Jones K, Symmons J, Holland EA, Schmid H, Bonazzi V, Woods S, Dutton-Regester K, Stark MS, Snowden H, van Doorn R, Montgomery GW, Martin NG, Keane TM, López-Otín C, Gerdes AM, Olsson H, Ingvar C, Borg A, Gruis NA, Trent JM, Jonsson G, Bishop DT, Mann GJ, Newton-Bishop JA, Brown KM, Adams DJ and Hayward NK (2014). Nonsense mutations in the shelterin complex genes *ACD* and *TERF2IP* in familial melanoma. *Journal of the National Cancer Institute* (Accepted for publication).

This publication explains an extended search for variants in members of the shelterin complex in 510 melanoma-prone pedigrees from Australia, UK, The Netherlands, Denmark and Sweden. We found additional nonsense variants that co-segregate with the disease in *ACD* and *TERF2IP*, providing further support for telomere dysregulation as an important contributor to this phenotype. I did the data analysis for samples from the UK and The Netherlands, as well as comparisons with control exomes.

- Robles-Espinoza CD, del Castillo Velasco-Herrera M, Hayward NK and Adams DJ (2014). Telomere-regulating genes and the telomere interactome in familial cancers. *Molecular Cancer Research* (Accepted for publication).

This review explores the role that telomere-regulating proteins, including members of the shelterin complex, have in cancer predisposition. The structures and functions of telomerase, shelterin and the telomere interactome are discussed. I wrote most of this review.

- Rashid M, Robles-Espinoza CD, Rust AG and Adams DJ (2013). Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics* **29**(17): 2208-10. doi: 10.1093/bioinformatics/btt371.

This application note provides a tool for somatic variant discovery from NGS data in cancer genomes. It combines other software tools previously developed and provides a strategy to optimise the sensitivity and accuracy of candidate variant calls when compared to any of these tools alone. I wrote some of the functions in this piece of software, which were adapted mainly from the analyses I did during the discovery phase of this study.

B.2 Other articles

- Robles-Espinoza CD and Adams DJ (2013). Cross-species analysis of mouse and human cancer genomes. *Cold Spring Harbor Protocols* **2014**(4).

  doi: 10.1101/pdb.top078824

- van der Weyden L, Rust AG, McIntyre RE, Robles-Espinoza CD, del Castillo Velasco-Herrera M, Strogantsev R, Ferguson-Smith AC, McCarthy S, Keane TM, Arends MJ and Adams DJ (2013). *Jdp2* downregulates *Trp53* transcription to promote leukaemogenesis in the context of *Trp53* heterozygosity. *Oncogene* **32**(3): 397-402. doi: 10.1038/onc.2012.56