Chapter 4

Discussion

4.1 A summary of the dissertation

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

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

The samples that were selected to be studied as part of this project were split into 4 different datasets: the pilot whole genome dataset, the secondary Leiden whole genome dataset, a primary exome dataset and a secondary exome dataset. The criteria for sample selection and the sequencing methodologies applied were different for each of these datasets. Once the samples had been sequenced, sequence alignment and joint variant calling were performed uniformly across all 4 datasets. A principal component analysis was performed on these samples with data from the 1000 genomes project being used as a control set; this was done to

avoid any potential bias in the sequencing due to population stratification[166]. No such bias was found. Additionally, there was an inherent possibility of the presence of melanoma in a few families due to an increased burden of common risk factors. This would indicate that the development of melanoma in such families was not due to the presence of a highly penetrant, low frequency variant but due to an additive burden of multiple low risk alleles. In such a case, these pedigrees would have to be removed from the dataset as they would not contribute to the identification of novel germline variants that predispose to melanoma. Polygenic risk scores were estimated for all samples and compared with a control dataset comprising of sporadic melanoma samples and unaffected individuals. While there was a significant difference between controls and affected individuals, no significant difference was found between the familial cases and the sporadic cases, implying that there was not enough evidence to indicate a higher burden of low risk variants across the cases. As a result, no families were removed from the dataset. Detailed descriptions of sample selection, dataset descriptions, sequencing and variant calling methodologies, population stratification analysis and estimation of polygenic risk scores.

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

While the joint association-linkage analysis helped in the identification of novel exonic variants, there were still a few edge cases that this did not account for. Multiple procedures were developed and applied on the exonic variants concurrent to the association-linkage analysis. Variants in known melanoma driver genes were examined to guarantee that the families in the dataset did not carry a nonsense variant in these genes. A mixture of previously known variants and novel variants were discerned in 8 families in genes including *BAP1*, *BRCA2*, *CDKN2A*, *POT1* and *MITF*. In a second method, the proportion of samples in a pedigree car-

rying a specific variant were estimated for all variants. This was carried out to find variants that segregated with the disease in all sequenced members of a given pedigree and to supplement the linkage analysis. Loss-of-function mutations in *ATR*, *TP53AIP1* and *EXO5* were found in 10 pedigrees using this procedure. Additionally, missense variants in *DMRTA1* and *AMER1* were also found in 3 other pedigrees through the same approach. All of these genes have previously been associated with cancer development and in the case of *TP53AIP1*, *EXO5* and *AMER1*, specifically to melanoma development. The third and final method for the secondary analysis of the exonic region variants focussed on the presence of known pathogenic variants within the cases. ClinVar, a curated database of variants, their estimated pathogenicity, and the associated disorders, was utilised for this purpose[188]. Pathogenic variants in genes associated with oculocutaneous albinism and hereditary cancer syndrome were observed.

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

on melanoma development.

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

4.2 Evaluating hypotheses and aims of the project

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

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

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

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

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

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

Variants were called across all samples using GATK haplotype caller. A multisample Variant Calling Format file with all mutations across all samples was generated. Variant Effect Predictor was used to identify the consequences of all variants, both in the coding region and the

non-coding region of the genome. A subset of variants predicted to be deleterious for protein function were identified using these annotations.

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

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

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

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

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

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

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

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

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

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

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

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

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

The complete set of variants from ClinVar were obtained and filtered to contain all pathogenic variants. The dataset of variants from the cases were analysed to determine the presence of

such pathogenic variants, including the ones associated with cancer development. Additionally, several variants in key genes associated with the onset of albinism were also discovered and reported.

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

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

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

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

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

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

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

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

- To identify novel structural variants disrupting known cancer genes.

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

4.3 Major findings of the project

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

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

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

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

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

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

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

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

4.4 Future prospects and conclusion

While the results of the dissertation helped provide further inroads to our understanding of familial melanoma, several questions remain unanswered. The outcomes of the different ap-

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

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

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

for each sample would considerably increase the accuracy of the burden estimation, which would in turn enhance the detection of candidate driver genes.

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

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

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

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

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