Identification of germline variants that predispose to familial melanoma

Aravind Sankar

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

St Edmund's College 2020

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

Declaration

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

Aravind Sankar
2020

Acknowledgements

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Abstract

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

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

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

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

Contents

Contents **i**

List of Figures **vii**

List of Tables **ix**

Nomenclature **xi**

1 Background **13**

1.1 Melanoma - A statistical overview 13

1.2 Melanoma through the ages 14

1.3 The biology of melanoma 16

1.3.1 An introduction to melanocytes 16

1.3.2 The progression of melanocytes to metastatic melanoma 18

1.3.3 The landscape of somatic variation in melanoma 21

1.3.3.1 The MAPK (Ras-Raf-MEK-ERK) Pathway 22

1.3.3.2 *BRAF* mutant melanoma 24

1.3.3.3 *NRAS* mutant melanoma 25

1.3.3.4 *NF1* mutant melanoma 26

1.3.3.5 Triple-Wild type melanoma 27

1.3.4 Germline familial melanoma genes and their clinical impact 27

1.3.4.1 The role of GWAS in melanoma research 27

1.3.4.2 Cyclin-dependent kinase inhibitor 2A *(CDKN2A)* 28

1.3.4.3 Cyclin dependent Kinase 4 *(CDK4)* 29

1.3.4.4 Telomere maintenance pathway 30

1.3.4.5 BRCA1-associated protein-1 *(BAP1)* 34

1.4 Classification of melanoma 35

1.4.1 Melanoma subtypes 35

1.4.1.1 Superficial spreading melanoma 36
1.4.1.2 Nodular melanoma 36
1.4.1.3 Lentigo maligna melanoma 37
1.4.1.4 Acral lentiginous melanoma 37
1.4.1.5 Rare melanoma subtypes 38
1.4.2 Cancer staging systems 38
1.4.2.1 A history of staging systems for cancer 38
1.5 Genetic testing and therapies for familial melanoma 40
1.6 Proposed approaches of sequence analysis undertaken for the identification of novel genes 41
1.7 Overarching aims of the project 44

2 Dataset description and methods used for the generation and analysis of the familial melanoma datasets 47
2.1 Introduction 47
2.2 Dataset description and assembly 48
2.2.1 An introduction to GenoMEL 48
2.2.2 Cohort description 48
2.2.3 Whole genome sequences - Sample selection and sequencing 52
2.2.3.1 Pilot whole genome dataset 52
2.2.3.2 Secondary Leiden whole genome dataset 53
2.2.4 Exome sequences - Sample selection and sequencing 53
2.2.4.1 Primary exome dataset 53
2.2.4.2 Secondary exome dataset 54
2.3 Alignment of DNA sequence data and variant calling 55
2.4 Exploration of population stratification bias within the dataset 57
2.5 Estimation of polygenic risk scores 60
2.5.1 Introduction 60
2.5.2 Methods 61
2.6 The determination of novel variants through association analysis 62
2.6.1 Selection of a control dataset 62
2.6.2 Initial filtering of variants 65
2.6.2.1 Control variants from gnomAD 65
2.6.2.2 Case variants 66
2.6.3 Joint processing of case and control variants 68

2.6.3.1 Annotation and filtering based on frequency of variants in gnomAD 68
2.6.3.2 Annotation and filtering based on coverage of samples in cases and controls 68
2.6.3.3 Annotation and filtering case variants based on alternate allele read depth and frequency 75
2.6.3.4 Annotation of cancer gene status 75
2.6.3.5 Calculation of total number of affected samples in genes . 76
2.6.4 Statistical testing to determine ranked list of genes 77
2.6.5 Limitations of an association analysis 78
2.7 Linkage analysis 80
2.7.1 An introduction to linkage analysis 80
2.7.2 The joint association and linkage approach - pVAAST 81
2.7.3 Methods 83
2.7.4 Limitations 85
2.8 The search for variants in known driver genes 86
2.9 Variants with high segregation within the cases 87
2.10 Pathogenic variants in ClinVar 87
2.11 Non-coding variants 90
2.11.1 Background 90
2.11.1.1 Introduction 90
2.11.1.2 Transcription factors and sequence logos 90
2.11.1.3 The role of non-coding variants that modify the function of transcription factors in cancer 93
2.11.2 Methods 94
2.12 Structural variants 95
2.12.1 Background 95
2.12.1.1 Introduction 95
2.12.1.2 Structural variants in genetic disorders 97
2.12.1.3 Determination of structural variants in next-generation sequencing data 99
2.12.2 Methods 101

3 Results from the analysis of the familial melanoma datasets 105
3.1 Introduction 105

3.2 Estimation of polygenic risk scores 105
3.3 The identification of novel variants through association analysis 108
3.4 The identification of novel variants using a joint association-linkage analysis 111
3.5 The search for variants in known driver genes 117
3.6 Variants with high segregation within the cases 120
3.6.1 Nonsense mutations 120
3.6.2 Missense mutations 121
3.7 Pathogenic variants in ClinVar 123
3.8 Analysis of non-coding variants affecting transcription factor binding motifs 126
3.9 Structural variant analysis 132

4 Discussion 135
4.1 A summary of the dissertation 135
4.2 Evaluating hypotheses and aims of the project 138
4.3 Major findings of the project 142
4.4 Future prospects and conclusion 144

References 149

A Supplementary Tables and Figures 171
A.1 Results from association analysis for all genes in the Cancer Gene Census . . 171
A.2 Results from association analysis for all protein coding genes as designated on Ensembl 171
A.3 Parameters used for the different steps involved in the execution of pVAAST for the joint association-linkage analysis 172
A.4 Results from pVAAST using the default background file from the 1000 genomes project 172
A.5 Results from pVAAST using INTERVAL exomes background file 172
A.6 Complete list of variants with high segregation in cases 173
A.7 Complete list of variants associated with cancer in ClinVAR that are present in the dataset 173
A.8 Results from the association analysis for the transcription factor binding motif variants for genes in the Cancer Gene Census 173
A.9 Parameters used for the generation of structural variants 173
A.10 Complete list of filtered and annotated structural variants 174
A.11 Supplementary Figure 1 176

A.12 Supplementary Figure 2 178

List of Figures

1.1 Clark's model of melanoma development. 15
1.2 The progression of melanoma from melanocytes to metastasis. 19
1.3 The MAPK (Ras-Raf-MEK-ERK) Pathway. 23
1.4 Alternative splicing of *CDKN2A*. 29
1.5 *TERT* promoter mutations in sporadic and familial melanoma. 32
1.6 Structure of the shelterin complex and the telomerase complex. 33
1.7 Germline truncating BAP1 mutations along with the protein domains of BAP1. 35
1.8 Germline *BAP1* mutations in sporadic melanoma. 36
1.9 The different common subtypes of melanoma. 37
1.10 Outline of PhD project. 43

2.1 An example of a pedigree sequenced as part of the study. 49
2.2 Workflow for sequencing of cases and variant-calling. 58
2.3 Principal component analysis to verify ethnicity. 59
2.4 Distribution of median genotype qualities. 67
2.5 Overview of the steps involved in the association analysis. 69
2.6 Distribution of allele frequencies for variants in cases. 70
2.7 Distribution of allele frequencies for variants in controls. 71
2.8 Distribution of affected family counts across all variants in cases. 72
2.9 Description of coverage generation, annotation and filtration of variants in cases and controls. 74
2.10 A graphical representation of the scoring pVAAST. 82
2.11 Depiction of chromatin looping to show enhancer-promotion interaction moderated through mediators, CTCFs and cohesion. 91
2.12 An example of a sequence logo for a transcription factor binding motif. . . . 92

2.13 Overview of steps involved in the analysis of variants within transcription factor binding motifs. 96
2.14 The different types of common structural variants within the human genome . 98
2.15 Different types of read errors used in the identification of structural variants. . 100
2.16 The relative locations of structural variants to the gene of interest 104

3.1 Distribution of polygenic risk scores for familial melanoma cases from the pilot study, sporadic melanoma cases and unaffected controls. 106
3.2 Original results from the joint association and linkage analysis. 113
3.3 Results from the joint association and linkage analysis. 115
3.4 Leiden pedigree with a segregating p.Y646Ffs frameshift variant in *BAP1*. . . 117
3.5 Sydney pedigree with the p.I49S missense variant in *CDKN2A*. 118
3.6 Leeds pedigree carrying the p.E312K missense variant in *MITF*. 119
3.7 Brisbane pedigree carrying the p.L890* stop-gain mutation in *ATR*. 121
3.8 Brisbane pedigree with the p.Q22Afs frameshift variant in *TP53AIP1*. 122
3.9 Leeds pedigree with the p.R344Afs frameshift variant in *EXO5*. 123
3.10 Sydney pedigree carrying the p.D233Y missense variant in *AMER1*. 124
3.11 Location of pathogenic OCA2 variants as identified in ClinVar. 125
3.12 Leiden pedigree carrying the p.N465D missense variant in *OCA2*. 125
3.13 Pedigrees with the p.V419I missense variant in *OCA2*. 127
3.14 Disruption of EGR1 binding motif in VAV1 as observed in a pedigree from Pennsylvania. 130
3.15 Disruption of GABP binding motif in *SKI* observed in 3 pedigrees. 131
3.16 Disruption of MSN2 binding motif in SRC as observed in a pedigree from Leeds. 132
3.17 Deletion upstream of *CDKN2A* observed in a Sydney pedigree with 11 sequenced members. 133

A.1 Original results from pVAAST using the INTERVAL exomes as the background.176
A.2 Results from pVAAST for all genes in the Cancer Gene Census excluding *MUC4* and *MUC16* using the INTERVAL exomes as the background. 178

List of Tables

2.1 The different collaborative institutions that provided samples for this project and their locations. 49
2.2 Distribution of samples by location, type of sequence and dataset. 51
2.3 Distribution of patients with multiple primary melanomas and early age of onset in each dataset. 52
2.4 Criteria for selection of whole genome samples in the pilot dataset. 53
2.5 Criteria for selection of exome samples in the primary exome dataset. 54
2.6 Criteria for selection of exome samples in the secondary exome dataset. . . . 55
2.7 List of VEP consequences retained for analysis. 56
2.8 Single Nucleotide Polymorphisms chosen for the polygenic risk score analysis. 62
2.9 Distribution of samples across different population groups in ExAC. 63
2.10 Distribution of samples across different population groups in gnomAD. . . . 64
2.11 Parameters used for running CrossMap to lift the aligned gnomAD sequences from GRCh37 to GRCh38. 65
2.12 List of parameters used for filtering variants based on their predicted consequence on protein function using VEP. 66
2.13 An example of a contingency table used in the association analysis. 77
2.14 Classification of clinical significance of variants in ClinVar. 88
2.15 Review status classification of supporting evidence for variants in ClinVar . . 89
2.16 Parameters used for the identification of structural variants using LUMPY. . . 101

3.1 Mean and median polygenic risk scores for different subgroups of samples. . 107
3.2 List of the top 10 genes associated with melanoma within the Cancer Gene Census. 109
3.3 List of the top 10 genes associated with melanoma within all protein-coding genes. 109

3.4 List of top 10 scoring genes from the joint association-linkage analysis using the 1000 genomes dataset as the background dataset. 111
3.5 List of top 10 scoring genes from the joint association-linkage analysis using the INTERVAL exomes dataset as the background dataset. 112
3.6 List of the top twenty genes associated with variants in transcription factor binding motifs within the Cancer Gene Census. 128

Nomenclature

Roman Symbols

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