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

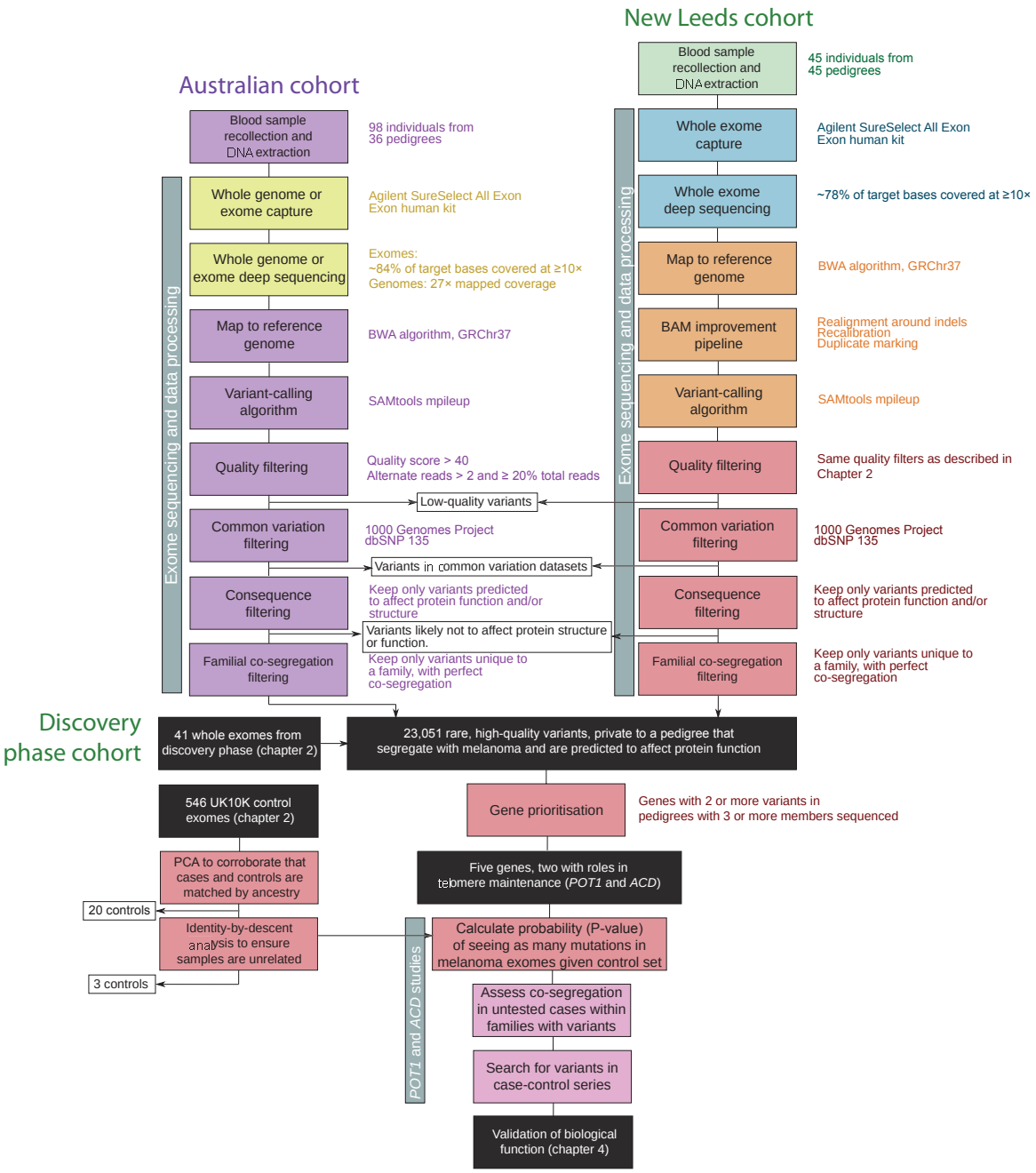

Figure 3.1: **Flowchart of analysis steps followed in the search for melanoma susceptibility genes, integrative phase.** Steps are colour-coded depending on the place where these were done, purple: QIMR Berghofer, yellow: Macrogen, Inc., green: Leeds or Leiden, blue: Beijing Genomics Institute, orange: 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