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.

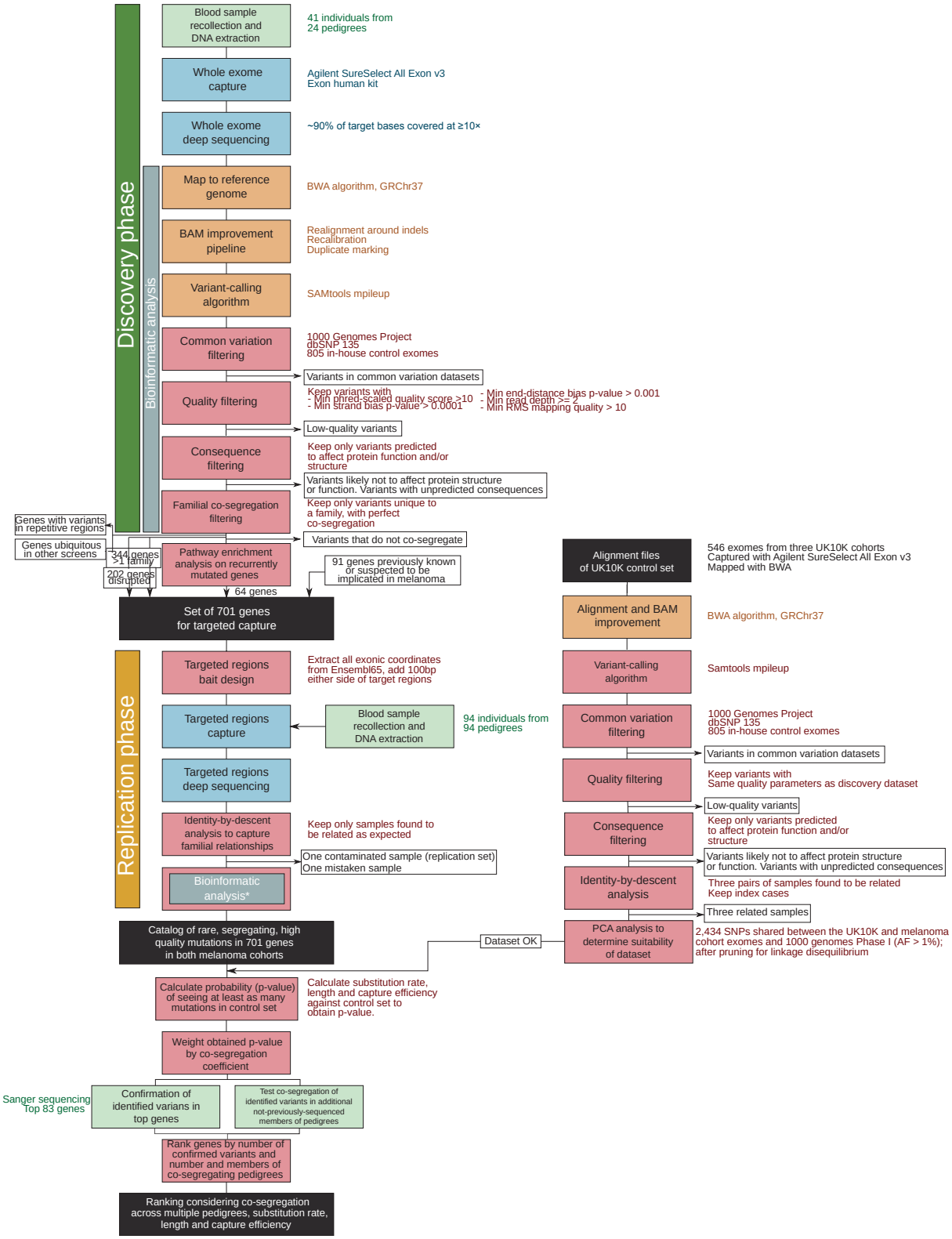

Figure 2.1: **Flowchart of analysis steps followed in the search for melanoma susceptibility genes, European phase.** Steps are colour-coded depending on the place where these were done, green: Leeds or Leiden, blue: Sanger Sequencing Facility, orange: Sanger Vertebrate Resequencing team, red: Sanger 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 ≥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 $-10log_{10}P$(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

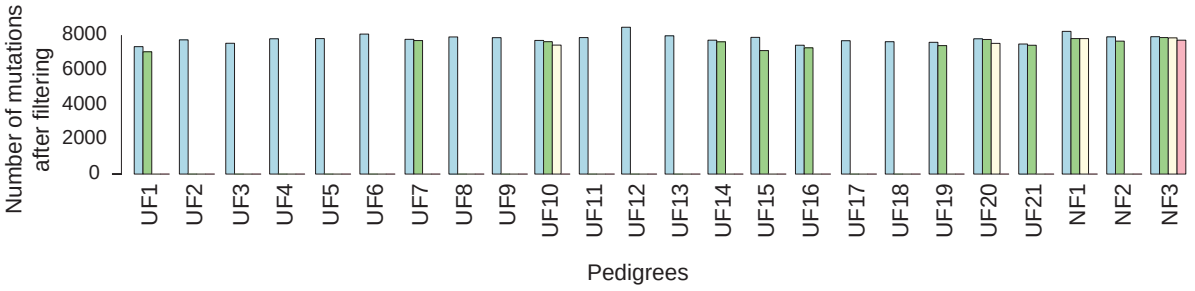

Figure 2.2: **Number of variants per exome remaining after filtering for common variation and quality, discovery phase.** Pedigree IDs are indicated below each set of bars, each bar represents one exome. Colours are used only to distinguish between different members within the same family. Similar numbers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

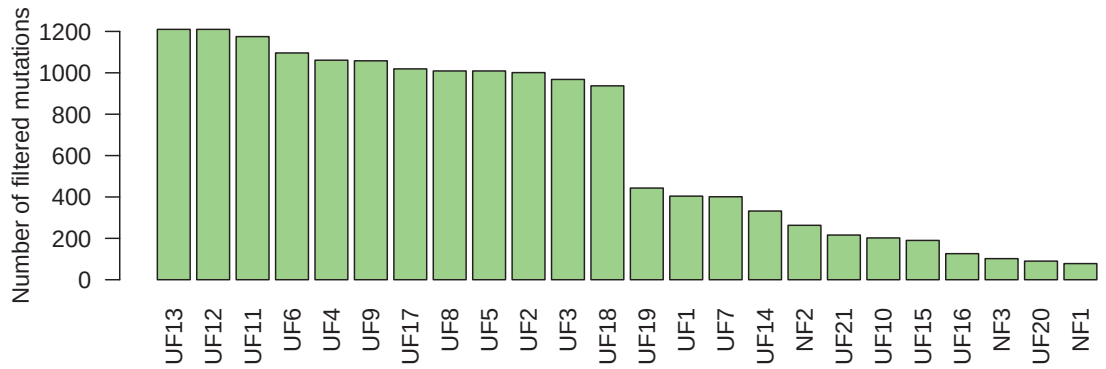

Figure 2.3: **Number of variants per pedigree remaining after filtering for co-segregation and protein-altering changes, discovery phase.** Pedigrees are ordered from the highest to the lowest number of variants passing all filters. In general, pedigrees with more members have less variants passing the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

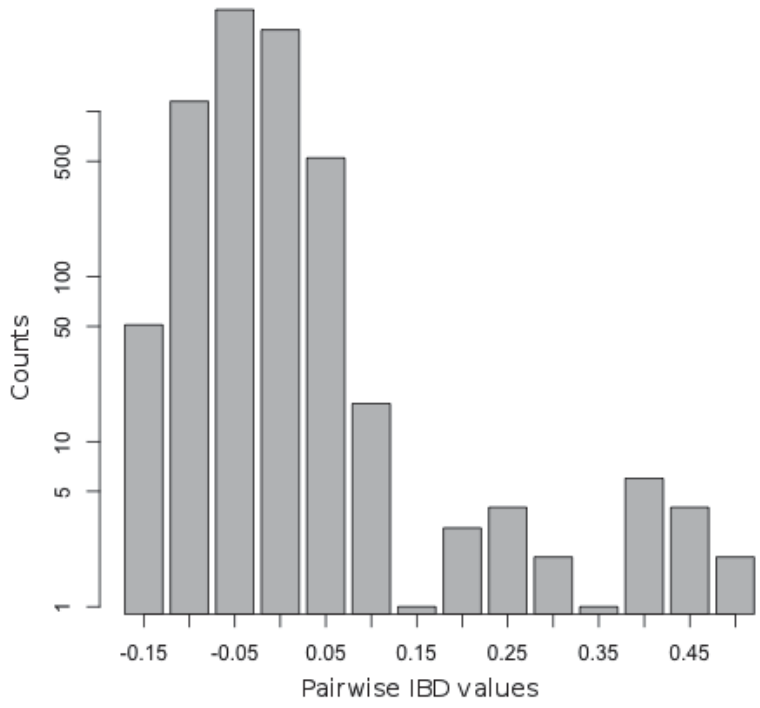

Figure 2.4: **Distribution of pairwise IBD values for samples in the discovery and replication sets.** The x-axis shows pairwise IBD values, whereas the y-axis shows counts of pairwise comparisons (out of $\binom{135}{2}$). Values are centered around 0.45 for first-degree relatives, around 0.25 for second-degree 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

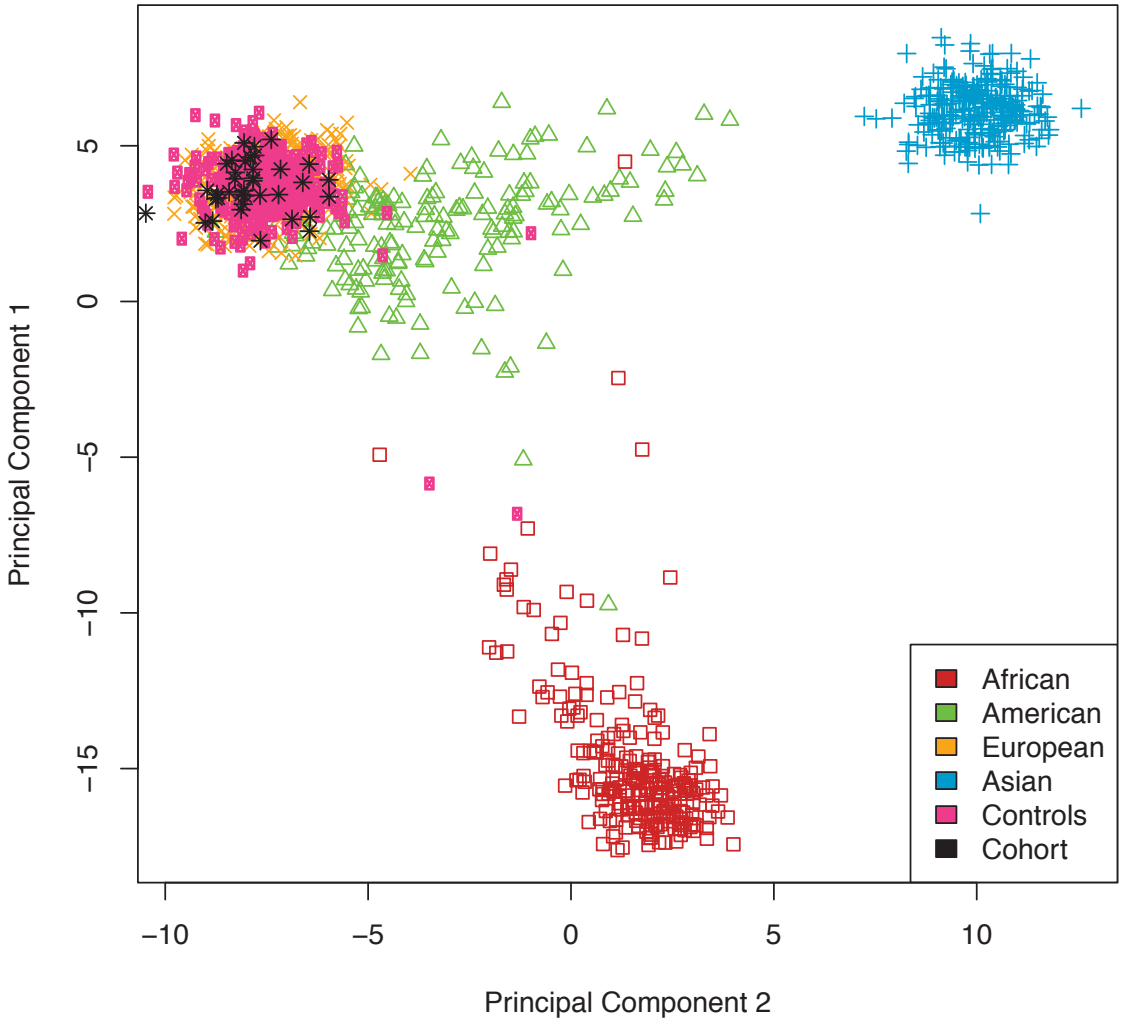

Figure 2.5: **Principal component analysis plot showing that cases and controls are matched by ancestry, European phase.** Plot showing the first and second principal components. Ancestry was estimated using the 1000 Genomes Project individuals and then projected onto the melanoma (black) and UK10K control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

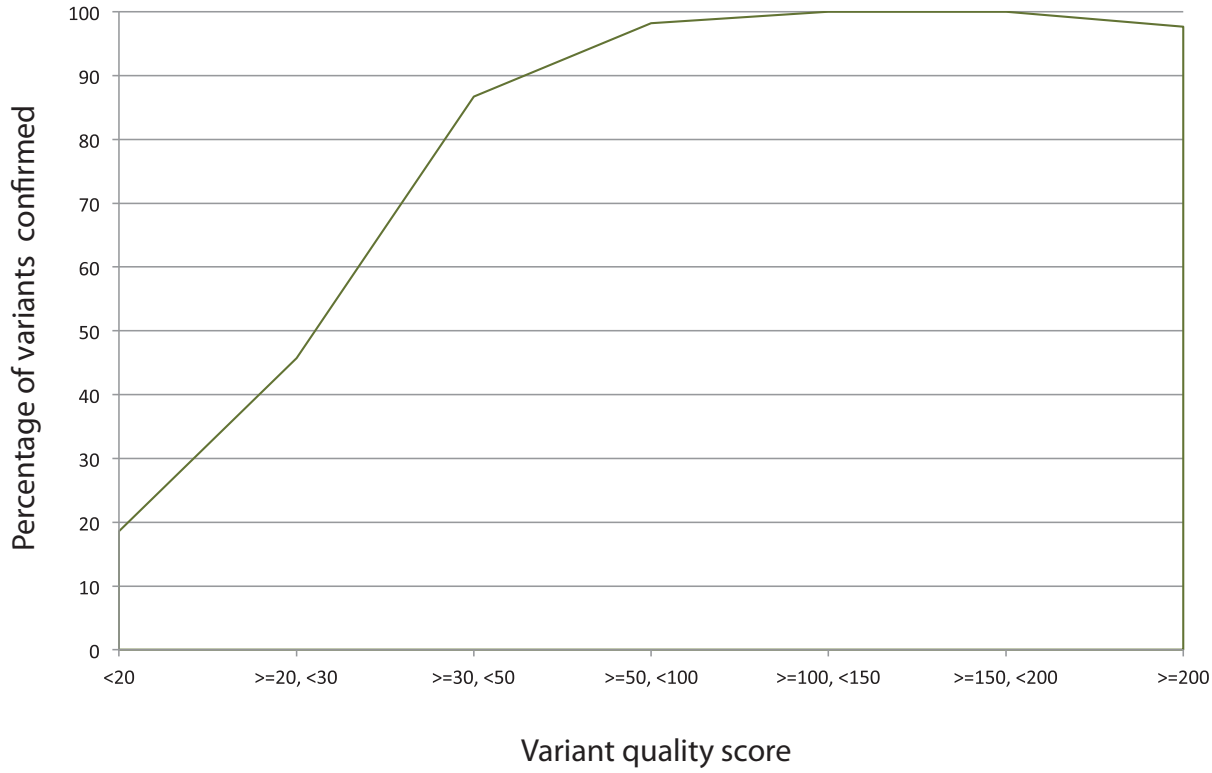

Figure 2.6: **Percentage of variants confirmed by PCR to be real grouped by quality score.** Overall, 608 variants were tested, and 522 were confirmed to be real.

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.

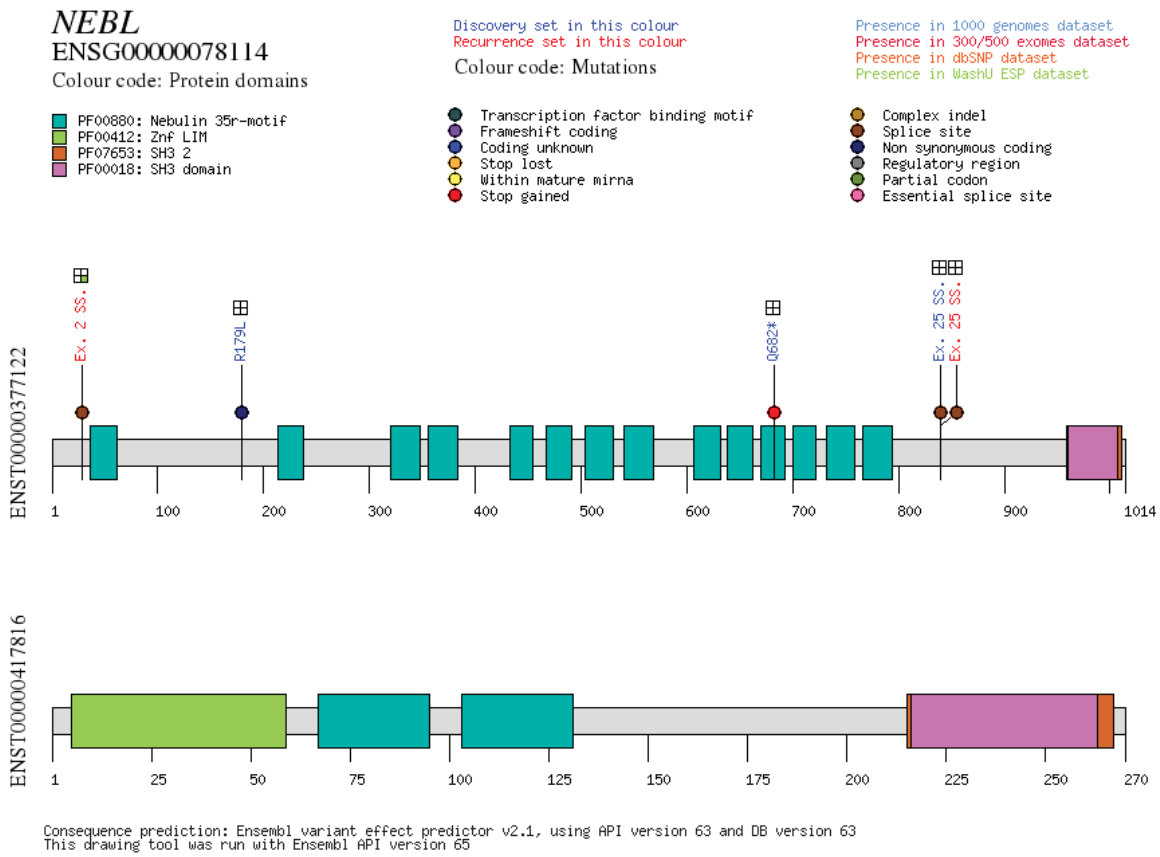

Figure 2.7: **Example of a protein plot showing NGS-detected variants.** The *NEBL* gene was chosen for illustrative purposes. For brevity, only two transcripts are shown, although the original plot depicts five. The gene name and associated Ensembl ID are shown at the left, followed by a list of domains (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).

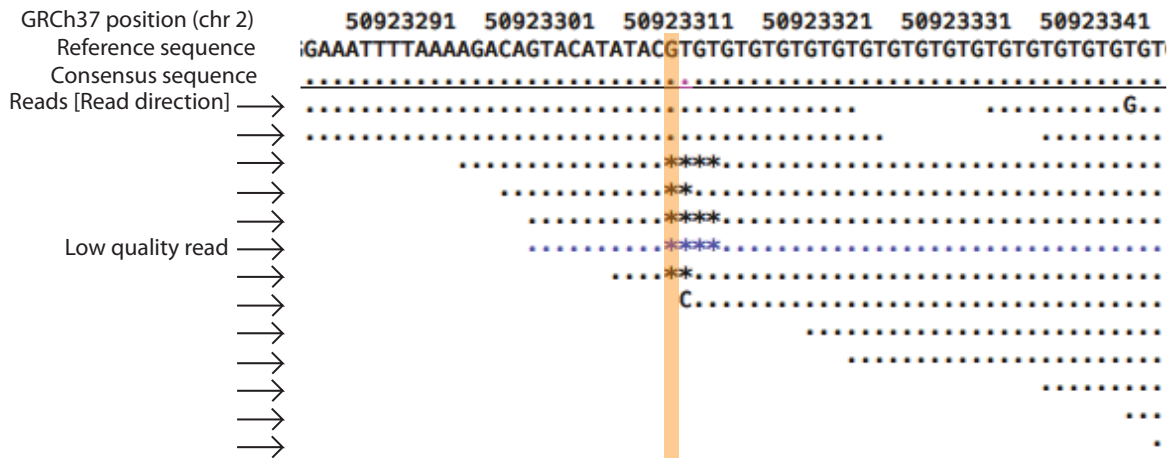

Figure 2.8: **Example of a putative founder variant found in melanoma pedigrees.** This variant is found co-segregating in 10 different melanoma pedigrees, the same or more than any other variant detected in this analysis. The position of the variant is indicated in orange. Dots indicate that the base 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.