Chapter 3

Results from the analysis of the familial melanoma datasets

3.1 Introduction

This chapter includes the results of all the analysis described in Chapter 2. The first section deals with the description and the implications of the polygenic risk score analysis on the dataset composition. Sections 3.3 and 3.4 describe the novel candidates identified in the association analysis and the joint association-linkage analysis on the coding region variants respectively. The results from the secondary, complementary methods used for discerning novel disruptive variants in addition to the association analyses are discussed in Sections 3.5,3.6 and 3.7. The investigation of non-coding variants and structural variants from the whole genome sequences resulted in novel findings, described in Sections 3.8 and 3.9 respectively. The implications of these results are discussed in Chapter 4.

3.2 Estimation of polygenic risk scores

The distribution of polygenic risk scores from the 123 familial melanoma cases in the pilot study, the sporadic cases and the unaffected controls are shown in Figure 3.1.

From the figure, it can be observed that the risk scores are distributed between -0.6 and 1.2 with the majority lying between -0.5 to 0.6. No discernible differences can be observed between the three groups directly. In addition to the genotype information obtained for the dataset from Leeds, information on the presence of multiple primary melanomas within the samples, early age of onset and the number of members in the family who suffered from

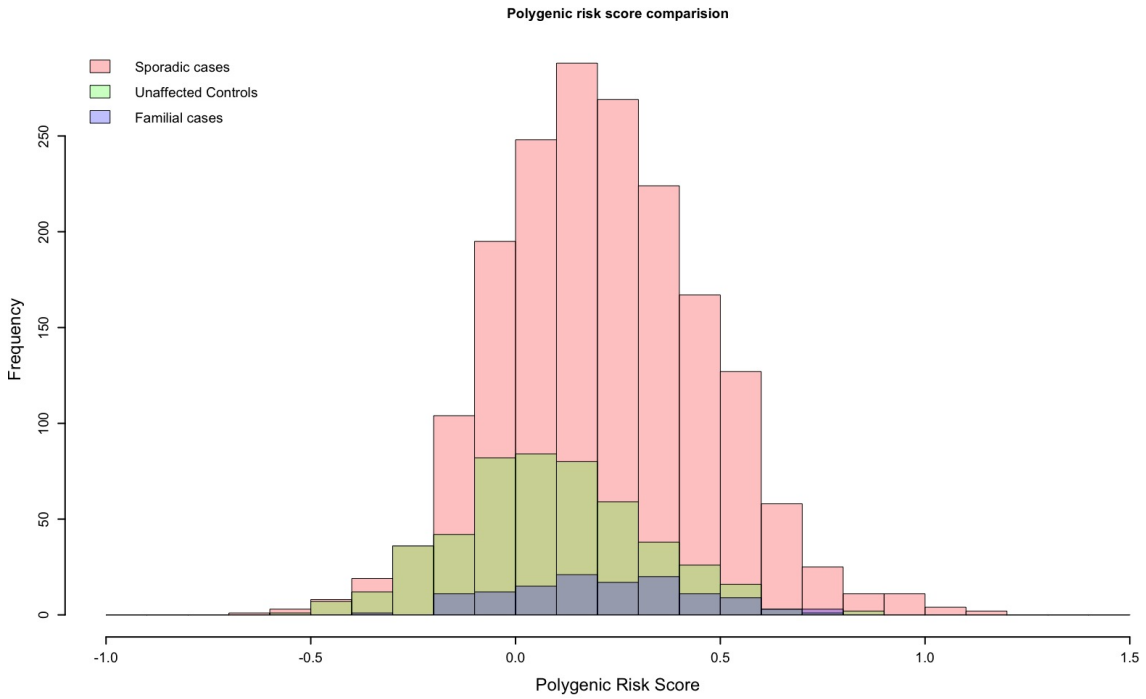


Figure 3.1: Distribution of polygenic risk scores for familial melanoma cases, sporadic melanoma cases and unaffected controls.

melanoma were also provided. The mean and median polygenic risk scores for each of these categories are given in Table 3.1.

The distribution of the risk scores and the mean and median risk scores in each category imply the following:

1. The risk scores are a representation of the burden of risk factors arising from common variants for the development of melanoma within the samples. As the controls do not have melanoma, they would be expected to have lower risk scores as compared to the cases, and this is indeed evident from the histograms, as the distribution of the risk scores for the controls samples lies to the "left" of the distribution of both sporadic and familial cases. This is also borne out by comparing mean and median risk scores between distributions. The controls have a mean risk score of .086 with a median of .082, which is significantly lower than the mean and median risk scores of the Leeds familial cases, the pilot study familial cases and the Leeds sporadic cases. The risk scores of the controls, sporadic values and the pilot study familial values were used as inputs for pairwise t-tests to identify if both the sets in consideration could be obtained from the same distribution. The null hypothesis is that there is no significant difference be-

Group	Mean	Median
Familial melanoma cases from the pilot study	0.2194	0.2203
Sporadic cases	0.2118	0.1993
Leeds familial cases	0.2254	0.1980
Unaffected controls	0.0868	0.0825
Early onset cases (inclusive of sporadic cases and the Leeds familial cases)	0.2249	0.2068
Cases with multiple primary melanomas (inclusive of sporadic cases and the Leeds familial cases)	0.2253	0.2210
Sporadic cases with multiple primary melanomas	0.2212	0.2216
Sporadic early onset cases	0.2125	0.1924
Familial cases from Leeds with multiple primary melanomas	0.2536	0.1277
Familial early onset cases from Leeds	0.3263	0.3579
Early onset sporadic cases with multiple primaries	0.1566	0.1650
Early onset familial cases from Leeds with multiple primaries	0.4039	0.3960

Table 3.1: Mean and median polygenic risk scores for different subgroups of samples.

tween the two sets while the alternate hypothesis is that there is a significant difference. The p-value of the t-test comparing the unaffected controls and the sporadic cases was $2.2*10^{-16}$while the p-value comparing the controls and pilot study familial melanoma cases was $3.988*10^{-8}$. As the p-values are less than .05, the null hypothesis is rejected, indicating that the controls are significantly different based on their risk scores compared to the sporadic and pilot study familial cases.

2. The distribution of the risk scores of the samples from the pilot study lie within the distribution of sporadic cases from Leeds. Although there is a large difference in the number of sporadic cases as compared to the familial cases, the difference in their mean and median values are not significantly different. The p-value obtained when tested for significant difference between the two groups was 0.7235, indicating that both the categories could be from the same distribution. The risk scores of the pilot study cases are also comparable with the different subcategories of cases from Leeds (inclusive the sporadic and the familial cases) involving early onset and the presence of multiple primary melanomas.

3. Interestingly, the risk scores for early onset familial cases from Leeds, both with and without multiple primary melanomas, are quite high compared to the sporadic risk scores, unaffected control risk scores and the pilot study risk scores. This indicates that these cases contain a high burden of risk factors which predisposes them to the early development of melanoma.

In conclusion, the familial melanoma cases have a higher polygenic risk score on average compared to unaffected controls and a similar risk score to sporadic cases. While this indicates a higher burden of common risk alleles compared to the controls, it does not rule out the presence of a high penetrant allele. None of the familial melanoma cases have an abnormally high risk score to merit their exclusion from the dataset. All samples from the dataset were therefore retained for further analysis.

3.3 The identification of novel variants through association analysis

The complete table of genes and their corresponding p-values both the genes in the Cancer Gene Census and for all protein-coding genes are attached as Supplementary Tables 1 and 2 respectively. The top 10 ranked genes from the Cancer Gene Census is shown in Table 3.2

while the top 10 ranked genes from all protein-coding genes is shown in Table 3.3. Variants from Table 3.2were investigated in detail to determine candidate driver mutations as they had the highest likelihood of playing an important role in familial melanoma development.

Ensembl ID	Gene name	Maximum segregation in families (%)	Fisher's Test p-value	Corrected P-value
ENSG00000145113	*MUC4*	100	1.25E-09	8.50E-07
ENSG00000178104	*PDE4DIP*	100	6.95E-06	0.00196
ENSG00000104517	*UBR5*	100	8.66E-06	0.00196
ENSG00000135333	*EPHA7*	50	4.35E-05	0.00739
ENSG00000163930	*BAP1*	100	7.28E-05	0.00989
ENSG00000087460	*GNAS*	66	0.00018	0.02111
ENSG00000046889	*PREX2*	50	0.00032	0.03090
ENSG00000138448	*ITGAV*	100	0.00036	0.03090
ENSG00000156650	*KAT6B*	54.5	0.00064	0.04645
ENSG00000204713	*TRIM27*	50	0.00071	0.04645

Table 3.2: List of the top 10 genes associated with melanoma within the Cancer Gene Census.

Ensembl ID	Gene name	Maximum segregation in families (%)	Fisher's Test p-value	Corrected P-value
ENSG00000204172	*AGAP10*	75	5.61E-22	1.06E-17
ENSG00000175820	*CCDC168*	75	5.45E-17	5.13E-13
ENSG00000185926	*OR4C46*	75	1.24E-16	7.80E-13
ENSG00000188649	*CC2D2B*	75	1.03E-13	4.83E-10
ENSG00000216937	*CCDC7*	66.6	4.49E-13	1.69E-09
ENSG00000112592	*TBP*	50	1.01E-11	3.15E-08
ENSG00000155495	*MAGEC1*	66.6	4.24E-11	1.14E-07
ENSG00000266714	*MYO15B*	66.6	1.09E-10	2.56E-07
ENSG00000177182	*CLVS1*	100	2.66E-10	5.57E-07
ENSG00000213401	*MAGEA12*	100	9.50E-10	1.74E-06

Table 3.3: List of the top 10 genes associated with melanoma within all protein coding genes.

From Table 3.2, it can be observed that *BAP1,* a gene previously described as a familial melanoma driver gene in Section 1.3.4.5, is ranked fifth on the ordered list of genes. This suggests that the approach in use has the statistical power and capacity to detect other potential familial melanoma driver genes as *BAP1* is a known driver gene. A family from Leiden, the Netherlands carried a *BAP1* variant that completely segregated with this disease. The importance of this variant is further discussed in Section 2.8.

Mucin 4 (MUC4) was the highest ranking gene from the Cancer Gene Census in the association analysis. It a member of the Mucin family, a set of high molecular weight glycoproteins present in the epithelial cells which are responsible for controlling the activity of inflammatory responses. Mucins are classified into two categories: secreted mucins and membrane bound mucins. *MUC4* is a membrane bound mucin responsible for several functions including the activation of an oncoprotein called ERBB2[223]. Two variants segregated completely with the disease in *MUC4 :* A variant at p.R906W (GRCh38 reference build, Chromosome 3, genomic position 195788864, c.2716G>A) in a three member pedigree from Leeds and a p.R468K variant (GRCh38 reference build, Chromosome 3, genomic position 195790177, c.1403C>T) in a three member pedigree from Stockholm. Two other variants also had high but not complete co-occurence of the variant with the disease, with three out of four sequenced members in two pedigrees carrying additional variants. Overexpression of *MUC4* has been observed in multiple cancers including pancreas, gall bladder, ovary, breast and lung carcinomas[224]. It is therefore seen as a potential therapeutic target.

Ubiquitin protein ligase E3 component n-recognin 5 (UBR5) was ranked third on the list of genes. It is an important constituent of the Ubiquitin-Proteasome System (UPS)which is an essential regulator of the DNA damage repair pathway. *UBR5* plays a key role in the development of several forms of cancer, as reviewed in 2015 by Shearer et al*[225].* Two variants segregated with the disease in *UBR5 :* A variant at p.S552I (GRCh38 reference build, Chromosome 8, genomic position 102323440, c.1655C>A) in a two member pedigree from Barcelona and a p.T1721P variant (GRCh38 reference build, Chromosome 8, genomic position 102286414, c.5161T>G) in a two member pedigree from Leeds. However, *UBR5* has not been previously implicated in either sporadic or familial melanoma. The mechanism of activation of melanoma development through *UBR5* disruption is therefore as yet undetermined.

Integrin Subunit Alpha V (*ITGAV*) ranked eighth in the association analysis. *ITGAV* encodes for an integrin membrane protein that regulates angiogenesis and cancer progression. A variant in ITGAV at p.R573Q mutation(GRCh38 reference build, Chromosome 2, genomic position 186656400, c.1718G>A) segregating with the disease was observed in a two member pedigree from Leeds, United Kingdom. Additional variants were observed in *ITGAV*; none

Gene	CLRT_score	P-value	LOD_score
MUC4	656.51	0.000999	5.22
MUC16	221.15	0.000999	1.99
FAM47C	156.23	0.000999	3.542
PDE4DIP	136.42	0.000999	3.67
RNF213	123.8	0.00699	3.22
MLL3	123.68	0.024	0.32
MLL2	117.9	0.0569	2.62
HLA-A_DUP_07	115.23	0.000999	0.49
NOTCH1	106.95	0.000999	0.51
FAT1	106.53	0.138	2.64

Table 3.4: List of top 10 scoring genes from the joint association-linkage analysis using the 1000 genomes dataset as the background dataset.

of which segregated with the disease in any of the pedigrees. Increased expression of *ITGAV* has previously been associated with increased invasion in colorectal cancer tumours[226] and it also facilitates prostate cancer metastasis[227]. Similar to *UBR5*, the role of *ITGAV* in skin disorders, particularly in cutaneous melanoma is currently unknown and warrants further investigation.

Whilst *EPH receptor A7* (*EPHA7*), *GNAS complex locus* (*GNAS*), *phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2* (*PREX2*), *lysine acetyltransferase 6B* (*KAT6B*) and *tripartite motif containing 27* (*TRIM27*) also ranked highly on the association analysis, none of the variants present in these genes segregated with the disease in the affected families. As none of these genes have previously been associated with either cutaneous sporadic melanoma or familial melanoma, the lack of variants segregating with the familial melanoma phenotype suggests that these genes are not relevant in familial melanoma genesis and development.

3.4 The identification of novel variants using a joint association-linkage analysis

The joint association-linkage analysis was performed twice using two different background datasets, one with the 1000 genomes dataset and another with the INTERVAL exomes. Each set focused entirely on the genes in the Cancer Gene Census. The complete results from both sets are attached in Supplementary Tables 4 and 5. The top 10 scoring genes for each of these tables are shown in Tables 3.4 and 3.5 respectively.

Gene	CLRT_score	P-value	LOD_score
MUC4	1203.4	0.000999	15.3
MUC16	373.2	0.000999	1.53
FAM47C	232.85	0.000999	3.64
MLL3	206.77	0.000999	0.32
NOTCH1	157.84	0.000999	0.65
HLA-A_DUP_07	154.91	0.000999	0.49
MLL2	147.9	0.00599	2.62
RNF213	146.99	0.00699	3.22
FAT1	136.62	0.043	2.6
FAT4	127.4	0.005	2.71

Table 3.5: List of top 10 scoring genes from the joint association-linkage analysis using the INTERVAL exomes dataset as the background dataset.

Each individual run for a gene returns two .vaast output files:

i) A simple file that only has the ranked list of genes with the associated p-value, LOD scores and CLRT scores.

ii) A larger file that has variant level information including the samples containing each variant and the filtered variants in each gene that were present in the background.

As each pVAAST run focussed within the region of a single gene, the output file corresponding to each run only has values for the reported gene. The background used was consistent across all runs for each attempt; this allows for the direct comparision of the output values with each other. The CLRT scores corresponding to the association and the LOD scores corresponding to the linkage were obtained for all the genes present within the Cancer Gene Census. The plot of these scores using the 1000 genomes project data as the background dataset is shown in Figure 3.2.

The CLRT scores from the association analysis are plotted on the y-axis. To correct for skewing of results, the logarithm to the base 10 of the LOD scores were estimated. This is plotted on the x-axis. In order to focus on genes with high association and/or linkage, only the genes with $\log_{10}$(LOD) greater than 0 or CLRT scores greater than 50 have been named. *MUC4* and *MUC16* are genes belonging to the Mucin family of proteins which had very high estimated CLRT and LOD scores. This is due to the presence of multiple variants in the cases that were not present in the background and were therefore not filtered out. As the scores of each variant contribute to the overall score of the gene, several low scoring mutations in *MUC4* and *MUC16* cumulatively resulted in the higher ranking of these genes. Mucins are high molecular weight proteins and have been observed to be consistently mutated without playing a role in the development of cancers. To better show the distribution of scores for the

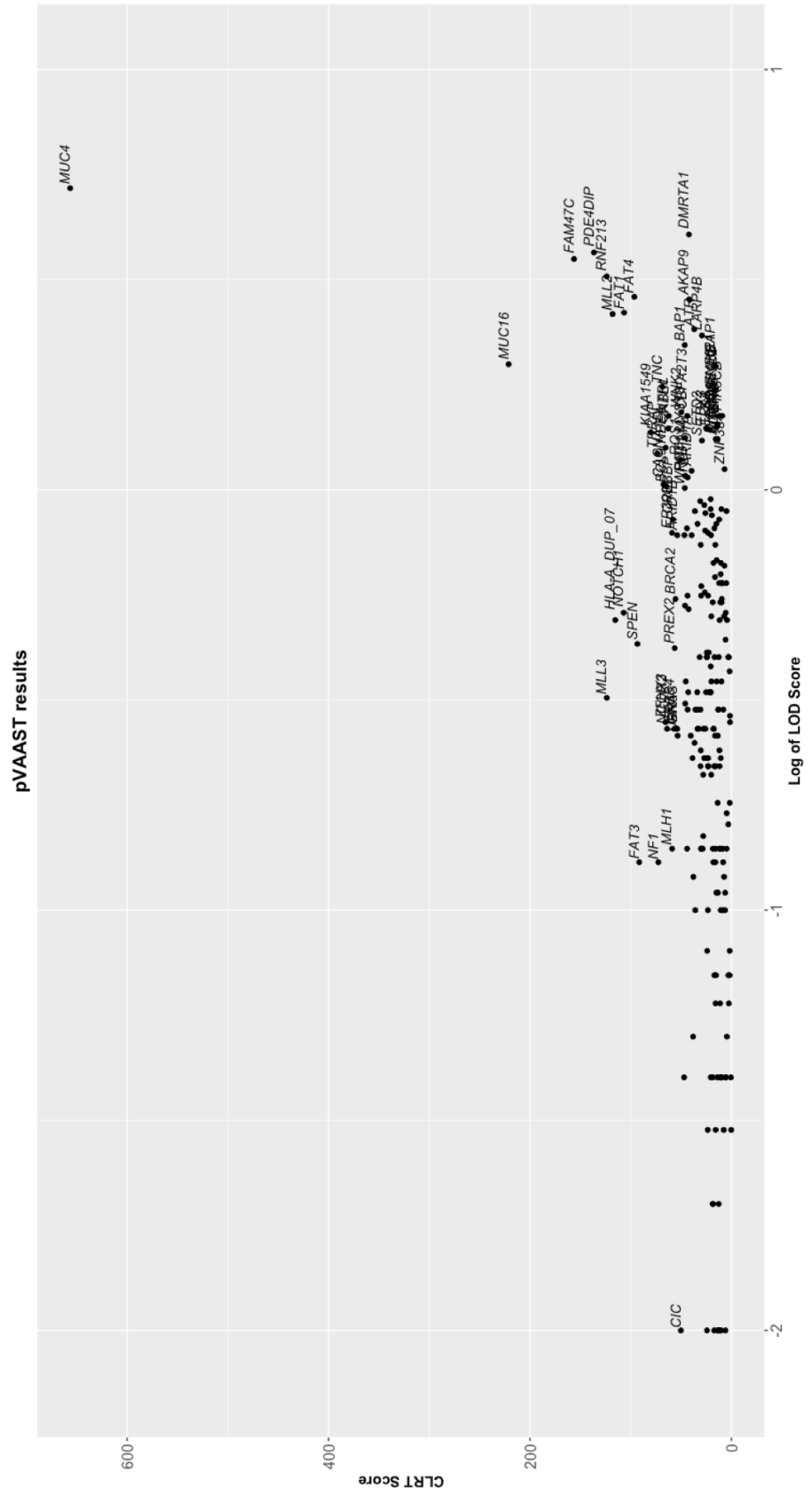


Figure 3.2: Original results from pVAAST for all genes in the Cancer Gene Census. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

other genes, these two genes were removed from the plot and a second plot was generated, shown in Figure 3.3.

In addition to the genes from the Cancer Gene Census, a single additional gene called *Doublesex And Mab-3-Related Transcription Factor A1 (DMRTA1)* was scored using pVAAST. While *DMRTA1* is in itself not a cancer gene, it lies adjacent to *CDKN2A*. A variant segregating in 10 out of 11 sequenced members of a pedigree was observed in *DMRTA1*. This was added to the list of genes as a positive control to ensure that pVAAST estimated LOD scores effectively for all genes as this variant would be expected to have a high LOD score. pVAAST estimated the LOD score *DMRTA1* to be 3.58 which was the highest LOD score estimated for a single variant, indicating that the LOD scores were being determined accurately. The potential implication of this variant in melanoma development is further discussed in Chapters 4 and 5.

Comparing the results from Figure 3.3 to Table 3.2, it can be observed that there are some genes that score well in both lists while others are significantly different. *MUC4* was the highest ranking gene in the association analysis and as discussed above, scores markedly high on pVAAST. *PDE4DIP*, the second highest ranking gene in the association analysis also ranks highly on pVAAST, with a CLRT score of 136.42 and a cumulative LOD score of 3.67 across 4 variants. Other genes that score similarly in both lists include *UBR5* (CLRT=53.4, LOD=0.26), *BAP1* (CLRT=46.34, LOD=2.21) and *GNAS* (CLRT=53.79, LOD=0.79. Interestingly, *UBR5* and GNAS score higher than BAP1 on the association score as they have more rare variants using this background dataset. However, these variants do not segregate through the pedigrees that they are present in unlike the variants in *BAP1* which completely segregate with the disease. As a result, although *BAP1* scores slightly lower on the association score, it would be considered to be more interesting and relevant to disease development compared to *GNAS* and *UBR5*, an important facet which would have been lost in a strict association analysis.

There are also several genes present in this analysis that are not in the top of the association analysis. *FAM47C,* the highest ranking gene that is not a Mucin protein, has a CLRT score of 156.23 and a LOD score of 3.542 obtained from 8 variants. This implies that there are several rare variants in this genes with most of these variants being present in smaller families. Such variants would have low individual LOD scores but cumulatively push the LOD score to greater than 3.5. *FAM47C* ranked 28th on the original association analysis discussed in Section 3.3with an uncorrected p-value of 0.003746353 and a corrected p-value of 0.088908376.

Another gene in a similar scenario is *Ring finger protein 213* (*RNF213*) which is ranked just below *PDE4DIP* with a CLRT score of 123.8 and a combined LOD score of 3.22 across 9

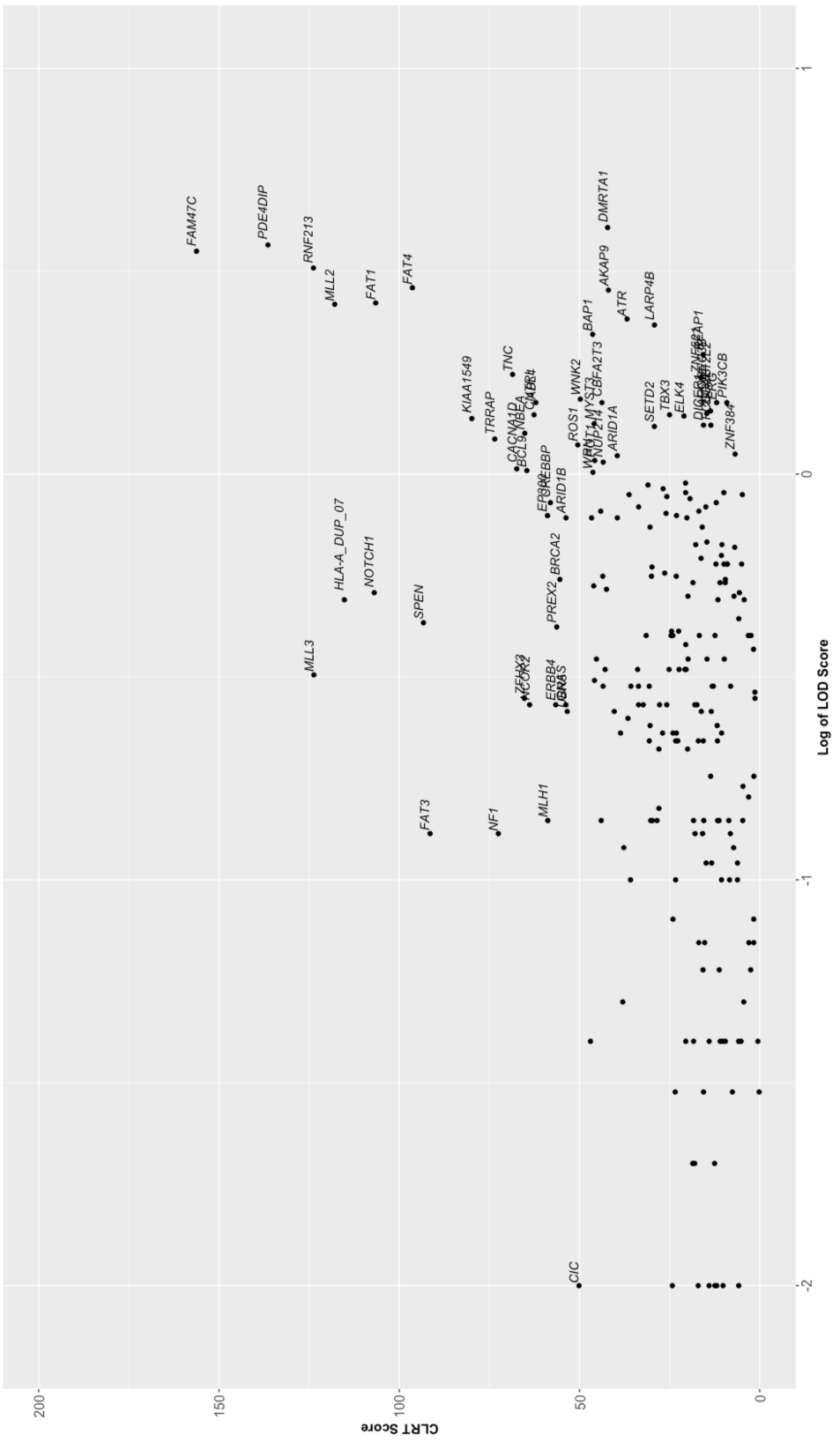


Figure 3.3: Results from pVAAST for all genes in the Cancer Gene Census excluding *MUC4* and *MUC16*. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

variants. The distribution of variants across multiple pedigrees with low number of sequenced members is similar to *FAM47C*. *RNF213*, however, ranks higher than *FAM47C* on the original association analysis discussed in Section 3.3with an uncorrected p-value of 0.001603777 and a corrected p-value of 0.075735829. Thus, while the corrected p-values of these genes might not have been originally significant, the presence of multiple variants segregating with the disease across multiple pedigrees indicate potential roles for *FAM47C* and *RNF213* in familial melanoma development.

An interesting observation from the joint-association linkage results are the presence of multiple members of the Fat atypical cadherin family of proteins. *FAT1* and *FAT4* both score highly on the joint association and linkage analysis, with high CLRT and LOD scores. *FAT3* also scores highly on the CLRT score but not on the LOD score. None of these genes scored highly on the original association analysis with *FAT1*, *FAT3* and *FAT4* ranking 94, 327 and 184 respectively. The presence of such high scores would indicate that these genes and the FAT family of protein would play a vital role in the development of familial melanoma in these pedigrees. However, on closer investigation, it was determined that these genes did not score significantly on the original association analysis as they are several thousand base pairs long and comprise of multiple functional domains. As a result, there were several low frequency variants in gnomAD and the cases leading to a high p-value. The background datasets used for the analysis on pVAAST do not have as many samples as gnomAD, resulting in higher scores for these genes. The functional relevance of the *FAT* family of genes in melanoma development is yet to be determined.

In addition to these high scoring genes, there are several genes with high LOD scores but not high CLRT scores and vice-versa such as *LARP4B* and *CIC* respectively. These genes, seen in Figure3.3, are examples of cases where an association or a linkage analysis on its own might lead to false positives as they might score highly due to the presence of several rare, non-segregating variants or one rare variant completely segregating with the disease but have no part to play in cancer development.

Similar plots and scores were obtained using the INTERVAL exomes as the background dataset. This altered the CLRT scores due to the difference in background variants and occasionally the LOD scores, if the variants were filtered for being in the background. However, the general ranking of genes and their corresponding relative CLRT scores and LOD scores remained the same. The complete scores from this set of pVAAST runs are given in Supplementary Table 5 and the plots from this set of pVAAST run are shown in Supplementary Figures 1 and 2.

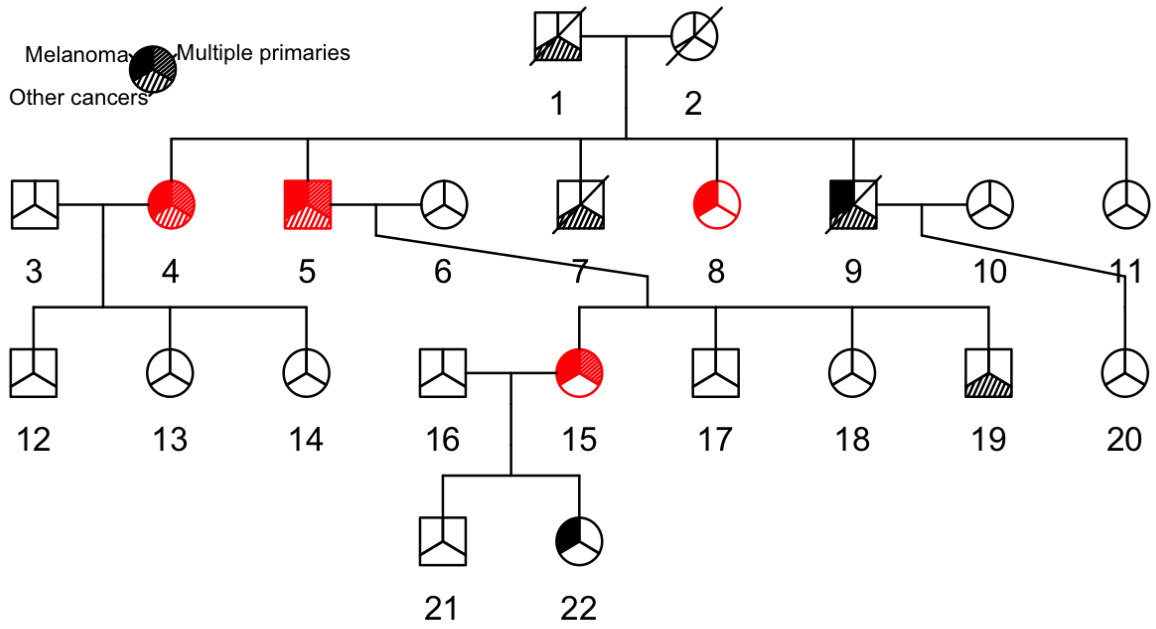


Figure 3.4: Leiden pedigree with the segregating p.Y646Ffs *BAP1* frameshift mutation. The members marked in red were sequenced from the pedigree, all of whom carried the variant.

3.5 The search for variants in known driver genes

1. *BAP1*: A four case pedigree from Leiden, shown in Figure 3.4, carried a frameshift variant encoding a p.Y646Ffs change (GRCh38 reference build, Chromosome 3, genomic position 52402825, c.2408_2409insAA). All four sequenced members carried the variant. Nonsense mutations leading to cancer development have previously been observed at this location: these mutations, however, led to the creation of premature stop-codons as opposed to the frameshift mutations observed in this study[228][229]. This variant was not observed in the gnomAD database.

2. *BRCA2*: A frameshift variant was observed in a single patient from KCL, London encoding a p.Q397Lfs variant (GRCh38 reference build, Chromosome 13, genomic position 32332667, c.1422_1423insTTAG). The frequency of this variant for non-Finnish Europeans in the gnomAD database was 0.000008822; it has not been annotated with phenotypes in ClinVar. A variant that introduced a stop codon resulting in a p.E1415*

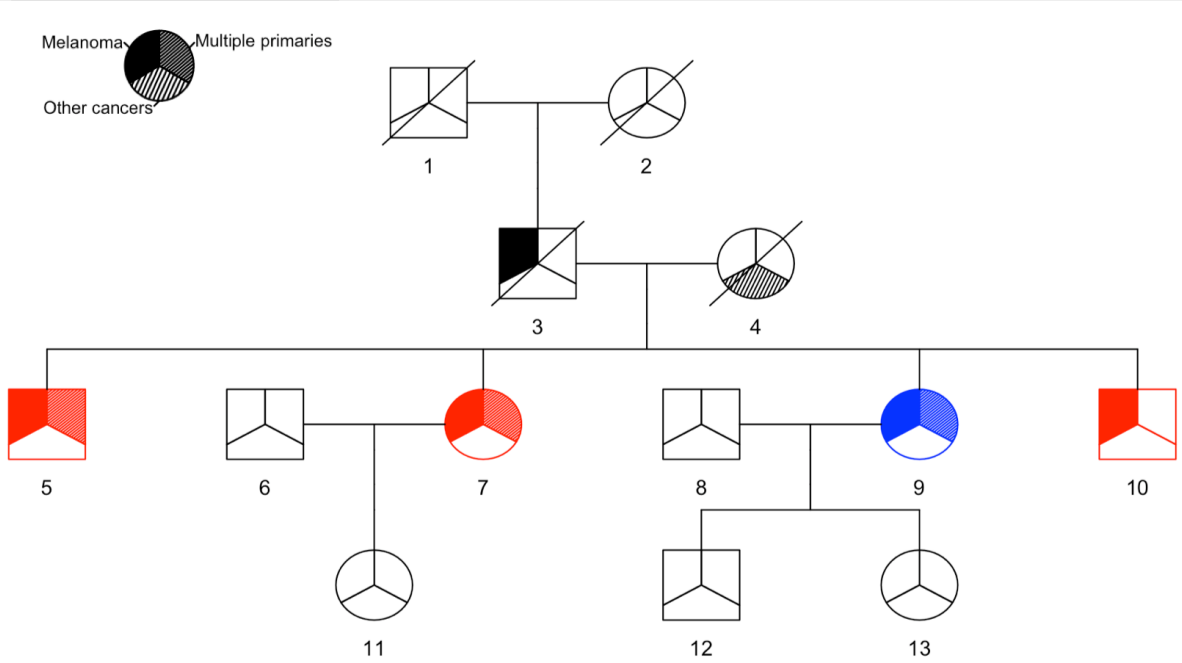


Figure 3.5: Sydney pedigree with the p.I49S missense variant in *CDKN2A* segregating in three out of four sequenced members. Whole genome sequencing for performed for the members shown in red and blue. The members marked in red carry the variant while the member shown in blue did not carry the variant and is predicted to be a phenocopy.

mutation (GRCh38 reference build, Chromosome 13, genomic position 32338598, c.4476G>T) was also observed in a patient from a family in Barcelona. This variant was annotated as being pathogenic on ClinVar for hereditary cancer-predisposing syndrome[230] (dbSNP id rs397507327). This variant was not observed in gnomAD.

3. *CDKN2A*: A p.I49S missense mutation (GRCh38 reference build, Chromosome 9, genomic position 21974682, c.417A>C) was identified in a pedigree from Sydney (shown in Figure 3.5) and was present in three out of four sequenced members. This variant has previously been observed in the context of familial melanoma[231][232] and is predicted to be deleterious based on *CDK4*/*CDK6* binding[233]. The variant is also annotated as being potentially pathogenic for hereditary cutaneous melanoma in ClinVar and was not observed in the gnomAD database.

4. *POT1*: As previously mentioned in Section 2.2.4.1, three novel *POT1* variants predisposing the families to familial melanoma development were previously identified using a dataset comprised of exome sequences of familial melanoma patients from Leeds and Leiden[143]. This dataset was a part of the cases and the samples carrying the *POT1*

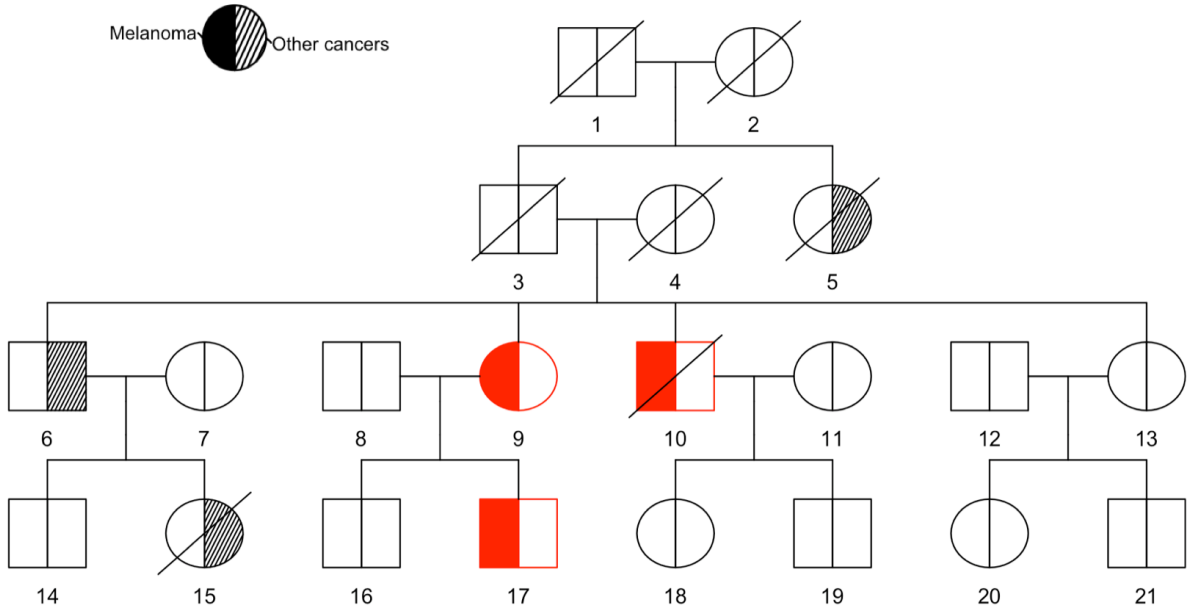


Figure 3.6: Leeds pedigree carrying the p.E312K missense variant in *MITF*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

variants were included in the analysis as positive controls to verify the ability of the applied methods in detecting disruptive variants. Additionally, a novel *POT1* missense variant was observed encoding a p.D185G change (GRCh38 reference build, Chromosome 7, genomic position 124859105, c.1153T>C) in two families: a two-case family and a single case family, both from Leeds. This variant was not observed on ClinVar (dbSNP id rs749741053) and was filtered from the gnomAD dataset.

5. *MITF*: A missense variant in *MITF* responsible for a p.E312K mutation (GRCh38 reference build, Chromosome 3, genomic position 69964940, c.934G>A) was observed to be segregating in a three member pedigree from Leeds, shown in Figure 3.6. This variant was annotated as being a risk factor for cutaneous malignant melanoma and for hereditary cancer predisposing syndrome on ClinVar (dbSNP id rs149617956). The same variant was also present in a one member of a two-member family from Leeds. This variant was also previously observed recurrently in both familial and sporadic melanoma[234] and has an allele frequency of 0.002456 for the non-Finnish European population subgroup in the gnomAD database.

No disruptive mutations were found in *CDK4* and *TERT*.

3.6 Variants with high segregation within the cases

A total of 12,923 variants were obtained after the steps described in Section 3.6. Variants in melanoma driver genes with segregation, as identified in Section 2.8, were observed again in this approach. While variants in *BAP1* and *POT1* were observed, the *MITF* variant was filtered out due to a having a gnomAD allele frequency greater than 10^{-3}. The complete set of variants is attached as Supplementary Table 6. Some of the resulting novel, segregating variants with a link to cancer development are reported here. The variants have been classified into nonsense and missense variants, depending on their effect on the protein product.

3.6.1 Nonsense mutations

1. *Ataxia telangiectasia and Rad3-related protein (ATR)* is a protein kinase that functions as a sensor and transducer for double-strand breaks caused due to to UV radiation[235]. Loss-of-function mutations in *ATR* have previously been shown to play a role in the growth of melanoma tumours[236]. A pedigree from QIMR, Brisbane with 4 sequenced members carried a p.L890*(GRCh38 reference build, Chromosome 3, genomic position 142553363, c.2669A>C) variant that encoded a stop-gain mutation in *ATR*. This variant segregated completely within the pedigree as shown in Figure 3.7. The reported L890* variant was not observed in the gnomAD database or in ClinVar.

2. *tumour protein 53-regulated apoptosis-inducing protein (TP53AIP1)* is a gene that encodes a protein which interacts with *TP53* and plays a role inTP53-mediated apoptosis[237]. Two single-case pedigrees from Leeds and one multi-case pedigree from Brisbane carry a frameshift variant encoding a p.Q22Afs change (GRCh38 reference build, Chromosome 11, genomic position 128937755, c.63_64insG).Three out of four sequenced members in the Brisbane pedigree, shown in Figure 3.8, carry the variant which was previously observed to predispose individuals from two different pedigrees to develop melanoma and is predicted to be an intermediate penetrant risk factor for melanoma onset[238]. This variant was not observed on ClinVar(dbSNP id rs141395772).

3. *Exonuclease 5 (EXO5)* is a single-stranded DNA-specific exonuclease that plays a role in DNA damage repair, with loss-of-function mutations in *EXO5* leading to increased genomic instability[239]. In a study involving the deficiency of DNA damage repair pathways across The Cancer Genome Atlas, *EXO5* was observed to be epigenetically silenced with high frequency, particularly in glioblastoma multiforme and in head and neck squamous cell carcinoma (HNSCC)[240]. A heterozygous p.R344Afs variant

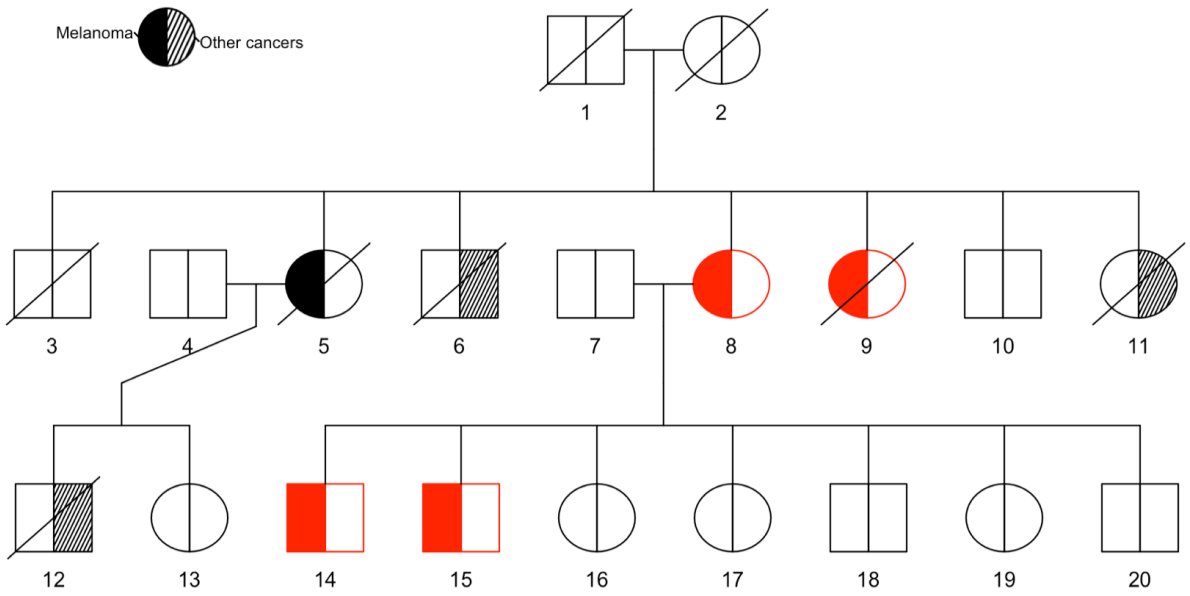


Figure 3.7: Brisbane pedigree carrying the p.L890* stop-gain mutation in *ATR*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

(GRCh38 reference build, Chromosome 1, genomic position 40515573, c.2103_2104insG) was observed in six pedigrees in the cases. Three of these pedigrees were from Leeds with single members sequenced, one was from London, and the final two were pedigrees with multiple sequenced members from Barcelona and Leeds. The Barcelona pedigree had two members sequenced, with one carrying the variant. In contrast, the Leeds pedigree, shown in Figure 3.9, had three members sequenced, all of whom carried the variant. This variant was previously observed in a study of early onset melanoma patients in Poland, where the association of the variant with the increase in melanoma risk was inconclusive[241]. This variant was also identified as a candidate that plays a role in increased susceptibility to testicular cancer[242]. The p.R344Afs was observed in gnomAD with an allele frequency of 0.0190 for non-Finnish Europeans and was not observed in ClinVar(dbSNP id rs150018949).

3.6.2 Missense mutations

1. *DMRTA1:* The variant with the largest number of affected samples in the dataset was a missense variant encoding a p.E383Q (GRCh38 reference build, Chromosome 9, ge-

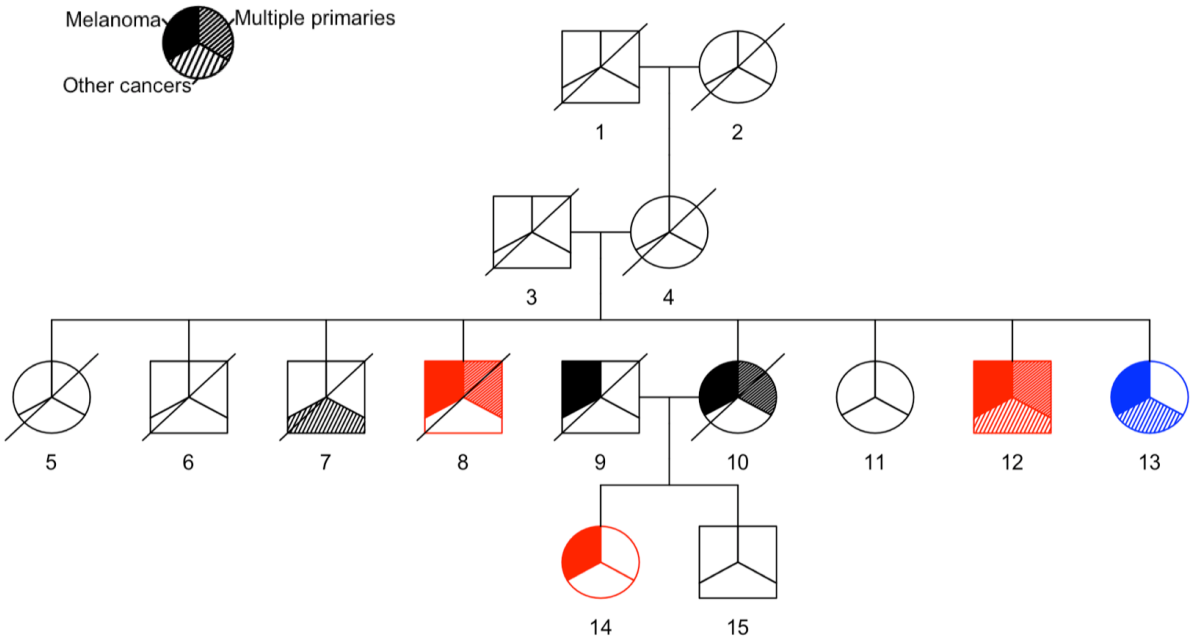


Figure 3.8: Brisbane pedigree with the p.Q22Afs frameshift variant in *TP53AIP1*. Exome sequencing for performed for the members shown in red and blue. The members marked in red carry the variant while the member shown in blue did not carry the variant and is predicted to be a phenocopy.

nomic position 22451543, c.2669G>C) change in *Doublesex- And Mab-3-Related Transcription Factor A1 (DMRTA1)*. Thirteen patients from three different families carried the variant with the following segregation: Two out of three sequenced members in a pedigree from Leiden (exome sequenced), a single sequenced member from a Leeds pedigree (exome sequenced) and ten out of eleven sequenced members in a pedigree from Sydney (whole genome sequenced). Although this variant has been annotated as not being pathogenic on both SIFT and PolyPhen, it lies adjacent to *CDKN2A* in the 9p21.3 chromosomal band. A variant with such a segregation pattern could be indicative of another, more significant variant lying within the same region, potentially be in the non-coding or intergenic region, that regulates the function of *CDKN2A*. A 233,780 base-pair deletion was eventually observed in a eleven member Sydney family 5' of *CDKN2A*. Particularly, this deletion was observed in the same ten members who also carried the *DMRTA1* variant. This is further explored in Section 3.9 in the analysis of structural variants in the whole genome sequences.

2. *AMER1*: A variant encoding a p.D233Y mutation (GRCh38 reference build, Chromosome X, genomic position 64192590, c.970C>A) was detected in a pedigree with a

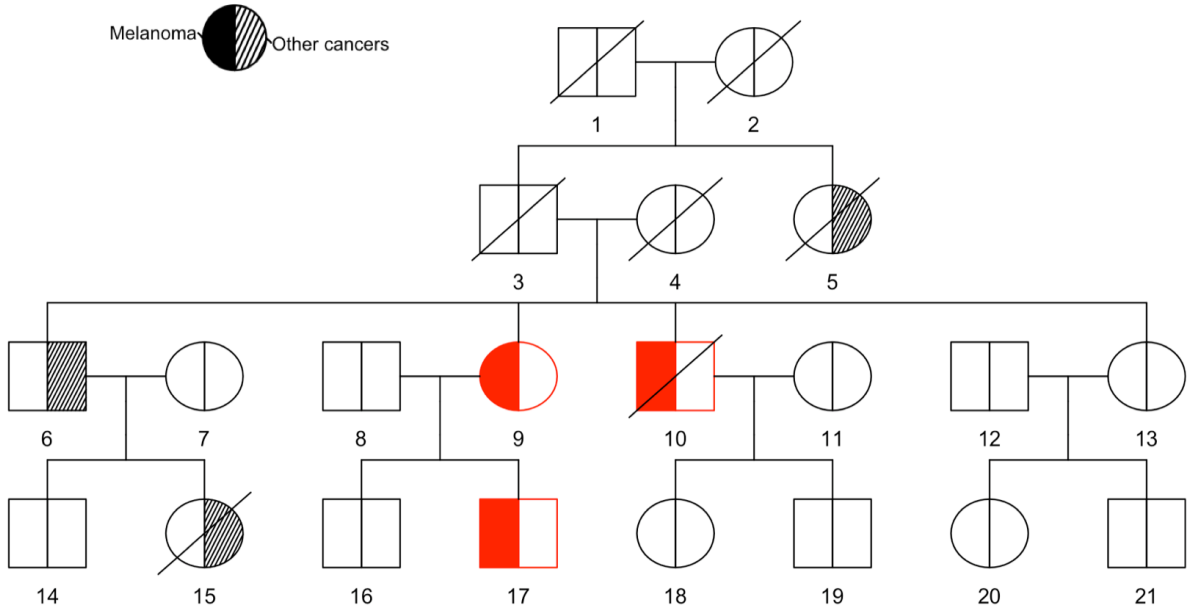


Figure 3.9: Leeds pedigree with the p.R344Afs frameshift variant in *EXO5*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant. This is the same pedigree that carries a moderate risk factor variant in *MITF* as described in 3.5.

single sequenced member from Leeds and a four-case pedigree from Sydney (shown in Figure 3.10) in *APC membrane recruitment protein 1 (AMER1)*. *AMER1*, also known as *WTX*, encodes a tumour suppressor. When mutated, this protein plays a role in the formation of pediatric kidney cancer known as Wilms tumours[243]. Germline variants in *AMER1* are also known to be causative of cranial sclerosis, a developmental disorder[244]. *AMER1* plays a role in the regulation of the WNT signaling pathway[245], a cascade that is involved in the carcinogenesis of several types of cancers including colorectal cancer[246], leukemia[247], melanoma[248] and breast cancer[249]. This variant was observed in the database with an allele frequency of 0.0004596 and was not observed in ClinVar (dbSNP id rs146685042).

3.7 Pathogenic variants in ClinVar

There were a total of 408,919 annotated variants in the ClinVar VCF file. 43,120 of these variants were also found in the cases. 18,205 variants remained after filtering for artefacts. 338 of these variants matched the restricted clinical significances for a wide range of diseases.

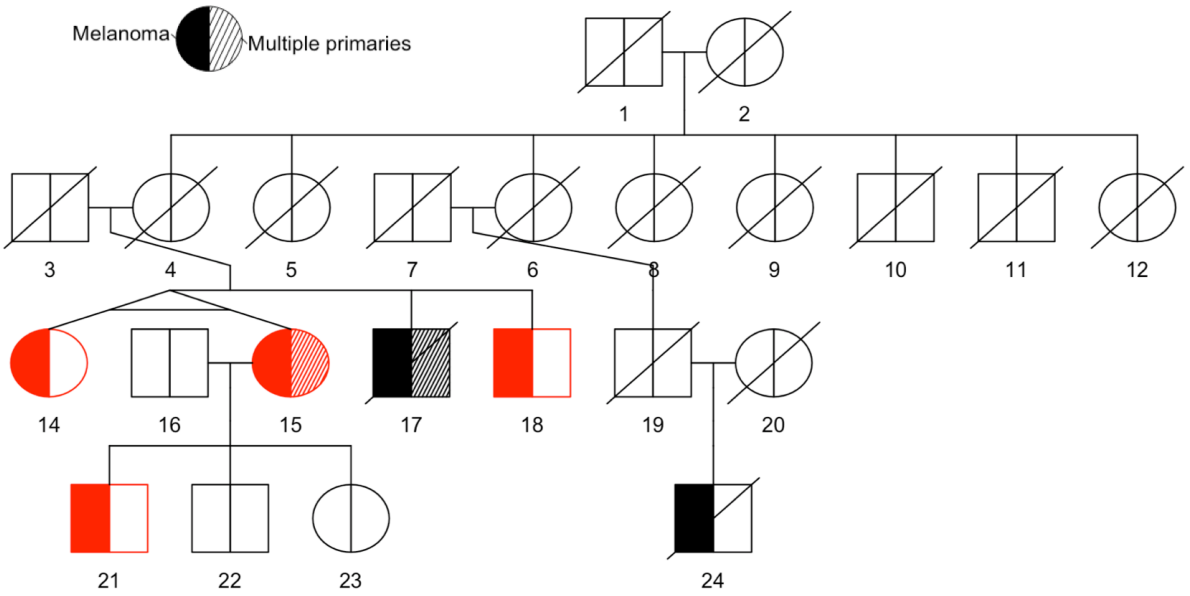


Figure 3.10: Sydney pedigree carrying the p.D233Y missense variant in *AMER1*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.

The variants most relevant to cancer and melanoma onset are described below:

1. The two *POT1* missense variants identified by Robles Espinoza et al. and the two disruptive *BRCA2* variants reported in 2.8, marked as pathogenic in ClinVar, were all observed through this approach. The *CDKN2A* p.I49S missense variant was also annotated on ClinVar as being associated with hereditary cutaneous melanoma. However, there were conflicting reports of pathogenicity regarding the clinical significance. As a result, this variant was not observed in our cases as it was filtered out.

2. Several pathogenic variants linked to different types of oculocutaneous albinism were observed. A variant linked to Tyrosinase-negative oculocutaneous albinism encoding a p.Y149C mutation (GRCh38 reference build, Chromosome 11, genomic position 89178399, c.948A>G) affecting the *Tyrosinase (TYR)* gene, was observed in an individual from KCL, London. Three variants were also identified in *oculocutaneous albinism II (OCA2),* a gene previously established as a susceptible locus in a meta-analysis of cutaneous melanoma genome wide association studies (GWAS)[51]. These variants were linked to the development of Tyrosinase-positive oculocutaneous albinism. Previous reports have identified *OCA2* missense mutations in melanoma and indicate a potential role in melanoma predisposition[250]. The location of these variants in the protein is

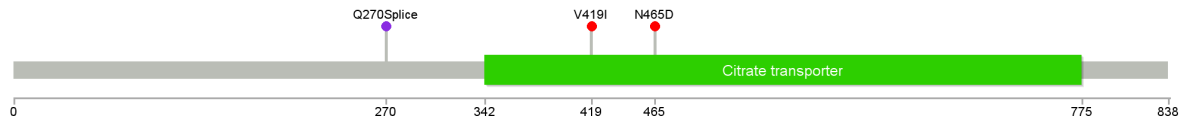


Figure 3.11: Location of pathogenic *OCA2* variants as identified in ClinVar. This plot was generated using Lollipops v1.3.2[251].

shown in Figure 3.11. These variants were as follows:

(a) A homozygous p.N465D missense variant (GRCh38 reference build, Chromosome 15, genomic position 27983383, c.1503T>C) was observed in a two member pedigree from Leiden, shown in Figure 3.12. In addition to the sequenced members, there was another member with melanoma and one more with pancreatic carcinoma. All affected members of the pedigree also had albinism. In total, there were six members in the pedigree who had albinism.

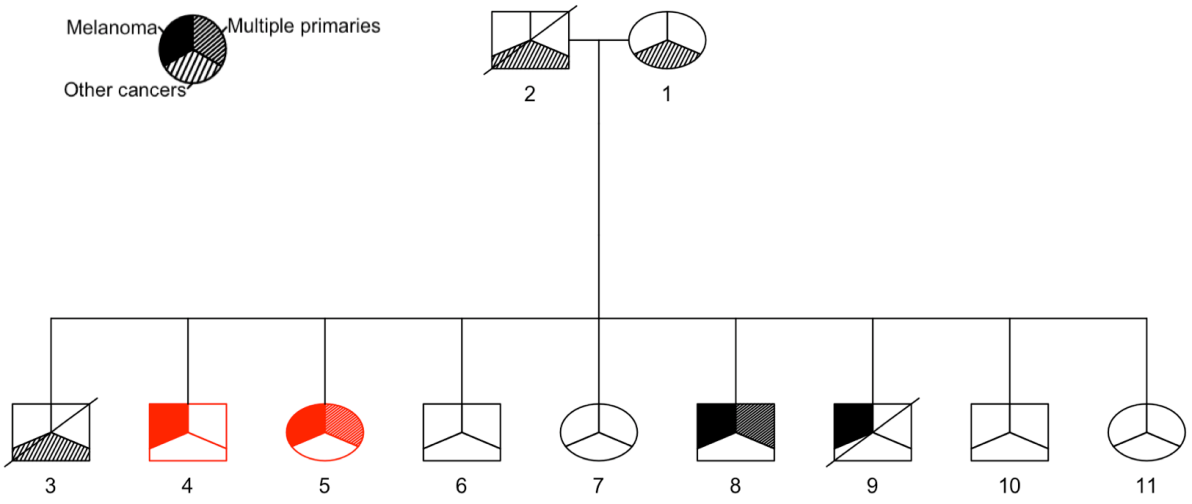


Figure 3.12: Leiden pedigree carrying the p.N465D missense variant in *OCA2*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant. All affected members also had albinism.

(b) Eleven individuals from five different pedigrees carried a p.V419I missense variant (GRCh38 reference build, Chromosome15, genomic position 27985101, c.1365C>T). Three out of four sequenced members of a Sydney pedigree, four out of four sequenced members of a second Sydney pedigree and two out of three sequenced members of a Stockholm pedigree carried the variant amongst the families with multiple sequenced members. These pedigrees are shown in Figures 3.13a, 3.13b

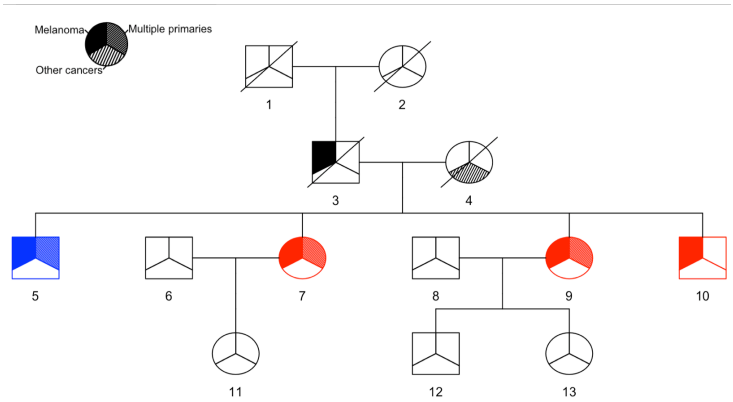
and 3.13c respectively. Two pedigrees from Leeds with single individuals sequenced also carried the variant.

(c) A canonical splice donor variant was identified at the beginning of the intron between exons 7 and 8 (GRCh38 reference build, Chromosome15, genomic position 28018396) in a single member from a Barcelona pedigree where 2 members were sequenced.

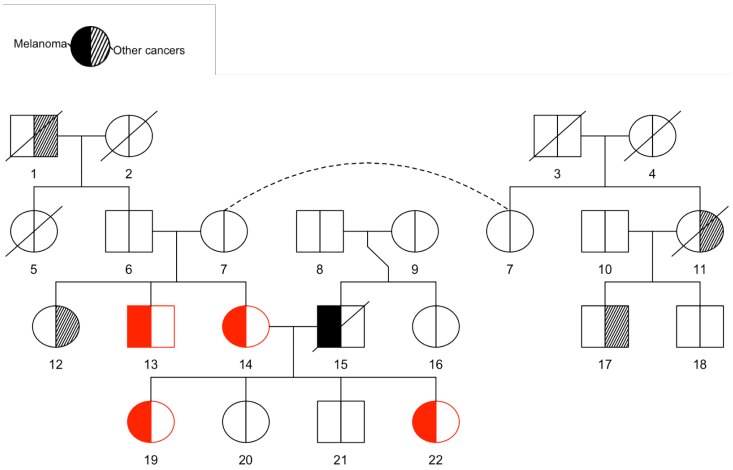
3. Multiple variants associated with different types of cancer on ClinVar were identified. Information related to these variants are given in Supplementary Table 7. While these variants are marked as pathogenic and could potentially play a role in the development of cancer in the individuals that they are present in, none of them segregate in a multi-case pedigree, i.e., none of them are present in a pedigree with several sequenced members all of whom carry the variant. This implies that the development of melanoma in the pedigrees carrying these variants cannot be attributed to these variants alone.

3.8 Analysis of non-coding variants affecting transcription factor binding motifs

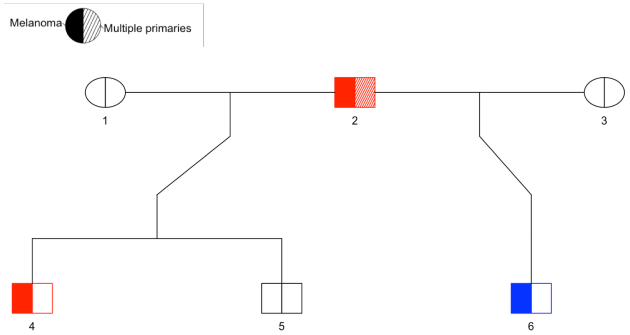
P-values were estimated and corrected for a total of 537 genes from the Cancer Gene Census. Results for the top 20 genes from the analysis of variants within the transcription factor binding motifs are shown in Table 3.6. The results for the complete set of genes are attached in Supplementary Table 8. Whilst most genes carry multiple variants distributed across several pedigrees, not every sequenced member of each pedigree carry the variant. In order to identify variants with high segregation of variant with the disease, the percentage of segregation of the variant was estimated for every pedigree at every variant. Following this, the pedigree with the maximum percentage of segregation was determined for each gene by comparing the percentages of segregation across all variants in a given gene. These results are also reported in Table 3.6.



(a) Stockholm pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree that were sequenced and carry the variant are shown in red while the member of the pedigree who was sequenced and did not carry the variant is shown in blue. This is the same pedigree that carries a segregating missense variant in *CDKN2A*, described in Section 3.5.



(b) Sydney pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree shown in red were sequenced from the pedigree. All sequenced members carry the variant.



(c) Stockholm pedigree with the p.V419I missense variant in *OCA2*. The members of the pedigree that were sequenced and carry the variant are shown in red while the member of the pedigree who was sequenced and did not carry the variant is shown in blue.

Figure 3.13: Pedigrees with the p.V419I missense variant in *OCA2*.

Ensembl ID	Gene name	Fisher's Test p-value	Corrected P-value	Maximum percentage of segregation in families (%)
ENSG00000184702	*SEPT05*	1.62E-33	8.70E-31	50
ENSG00000164362	*TERT*	2.88E-24	7.73E-22	75
ENSG00000136997	*MYC*	8.50E-21	1.52E-18	67
ENSG00000184640	*SEPT09*	2.58E-19	3.46E-17	83.3
ENSG00000072062	*PRKACA*	1.47E-18	1.58E-16	75
ENSG00000088038	*CNOT3*	1.79E-18	1.60E-16	50
ENSG00000078403	*MLLT10*	3.67E-18	2.81E-16	40
ENSG00000118046	*STK11*	4.57E-17	3.07E-15	83.3
ENSG00000175197	*DDIT3*	1.63E-15	8.78E-14	40
ENSG00000137309	*HMGA1*	1.63E-15	8.78E-14	66
ENSG00000136754	*ABI1*	1.08E-14	4.45E-13	75
ENSG00000141968	*VAV1*	1.08E-14	4.45E-13	100
ENSG00000160957	*RECQL4*	1.08E-14	4.45E-13	66
ENSG00000157933	*SKI*	1.40E-14	5.36E-13	100
ENSG00000071564	*TCF3*	7.52E-14	2.69E-12	40
ENSG00000143970	*ASXL2*	3.00E-13	9.46E-12	75
ENSG00000197122	*SRC*	3.00E-13	9.46E-12	100
ENSG00000157764	*BRAF*	5.07E-13	1.51E-11	60
ENSG00000261652	*C15orf65*	2.46E-12	6.61E-11	33
ENSG00000162367	*TAL1*	2.46E-12	6.61E-11	50

Table 3.6: List of the top twenty genes associated with variants in transcription factor binding motifs within the Cancer Gene Census. The values in the "Maximum percentage of segregation in families" represents the value of the highest percentage of segregation in any pedigree with a variant in the gene where there are at least two sequenced members and two individuals carrying the variant.

Interestingly, previously known driver genes including *TERT* and *MYC* are second and third respectively on the list of genes . *TERT* in particular is prominent due to the history of variants in TFBM within the promoter which lead to the progression of both familial and sporadic melanoma. However, the family with the maximum of percentage of segregation for the variants in these genes are 75% and 67% respectively, indicating that none of the families

have a variant that completely segregates with the disease within these genes. Additionally, the variants in *TERT* were investigated to determine if they were present at previously determined promoter variant sites. None of the observed variants were at these positions. Following this, variants with a segregation percentage of 75% or higher were then scrutinized to determine if the TFBMs for any of these genes were disrupted by these variants. The genes with the disrupted motifs and high percentage of segregation of the variant with the disease are described below.

i)Proto-oncogene vav - *VAV1*

Rho GTPases are a set of proteins responsible for several cellular functions including proliferation, adhesion, cellular migration, contraction and secretion[252]. These proteins were originally thought to play a role in the carcinogenesis of several cancer types indirectly through increased expression levels as mutations within them were rare[253]. However, recent studies have identified several direct mutations in RhoGTPases in different cancers[254] such as a recurrent mutation in *RAC1* in melanoma patients[255]. Rho GTPases are mediated and regulated by the Rho Guanine Nucleotide Exchange Factor (Rho GEF) family of proteins. These proteins include Vav1, which is responsible for the activation of Rac1[256]. Expression of wild-type VAV1 has been identified in several types of cancers including neuroblastoma, melanoma, pancreatic, lung and breast cancers[257]. It has also been reported to act as both an oncogene[258] and a tumour suppressor[259] under physiological contexts, similar to p53[260].

An upstream variant in *VAV1* was observed in 4 pedigrees with complete segregation in a Pennsylvania pedigree with 3 sequenced members (GRCh38 reference build, Chromosome 19 genomic position 6772561, G>A). This variant disrupted a binding site for the transcription factor EGR1, indicated in Figure 3.14.

There are 14 positions in the binding motif for EGR1 with varying levels of conservation between them. Positions 6, 8, 9 and 10 are the most conserved with these positions always carrying a guanine nucleotide at these positions. The variant described here changes the guanine at position 9 to an adenine, which is never present at this position. This would ablate the TFBM and disrupt the binding of EGR1 to this location.

ii)The Sloan Kettering Institute protein - *SKI*

The SKI protein, named after the Sloan Kettering institute where it was first described, is a proto-oncogene. *SKI* pathways have previously been reported as being responsible for the activation of β-catenin signalling[261] and for the progression of human malignant melanoma[262]. Following this, the knockdown and deficiency of *SKI* was also reported to reduce the proliferation of human melanoma tumours[263].

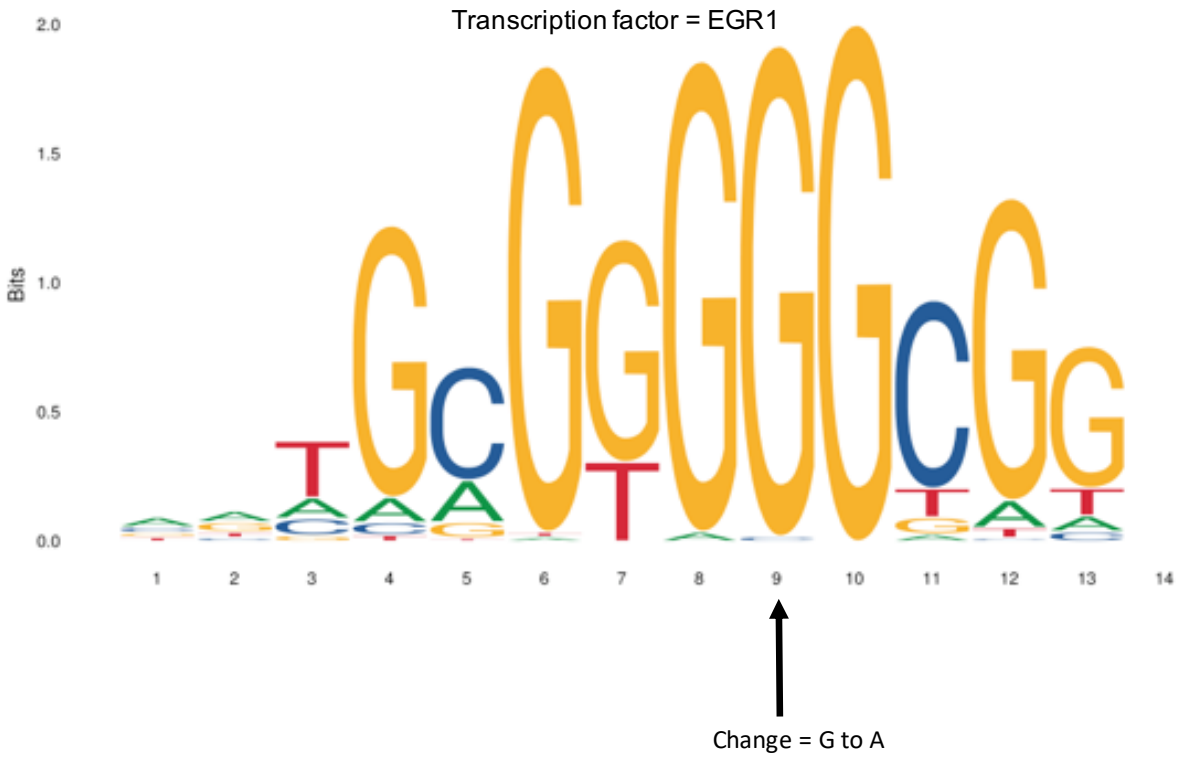


Figure 3.14: Disruption of EGR1 binding motif in *VAV1* as observed in a pedigree from Pennsylvania. Sequence logo obtained from the JASPAR[199] database.

A variant upstream to *SKI* (GRCh38 reference build, Chromosome 1 genomic position 2227654, G>A) was observed in 3 pedigrees. This included complete segregation with the disease in a Sydney pedigree with 4 sequenced members and partial segregation in a Leiden pedigree with 3 out of 6 sequenced members carrying the variant. Figure 3.15 indicates the disruption of the binding motif for GABP by this variant.

The TFBM for GABP consists of 10 bases. Positions 4 to 8 are highly conserved, represented by the base pattern GGAAG. The variant of interest is at position 8, which alters the guanine present at this position to an adenine and ablates the transcription factor binding site. Both GABP ablation and recruitment have been previously been associated with melanoma development. Disruption of GABP binding motif has previously been reported in the context of melanoma in association with recurrent mutations of the subunit D of the succinate dehydrogenase complex *(SDHD)* [200], as discussed in Section 2.11.1.3. Recruitment of GABP has been established in the reactivation of mutant *TERT* with promoter variants, leading to deviant expression of *TERT* in multiple cancer types including melanoma[264].

iii) Proto-oncogene tyrosine-protein kinase - *SRC*

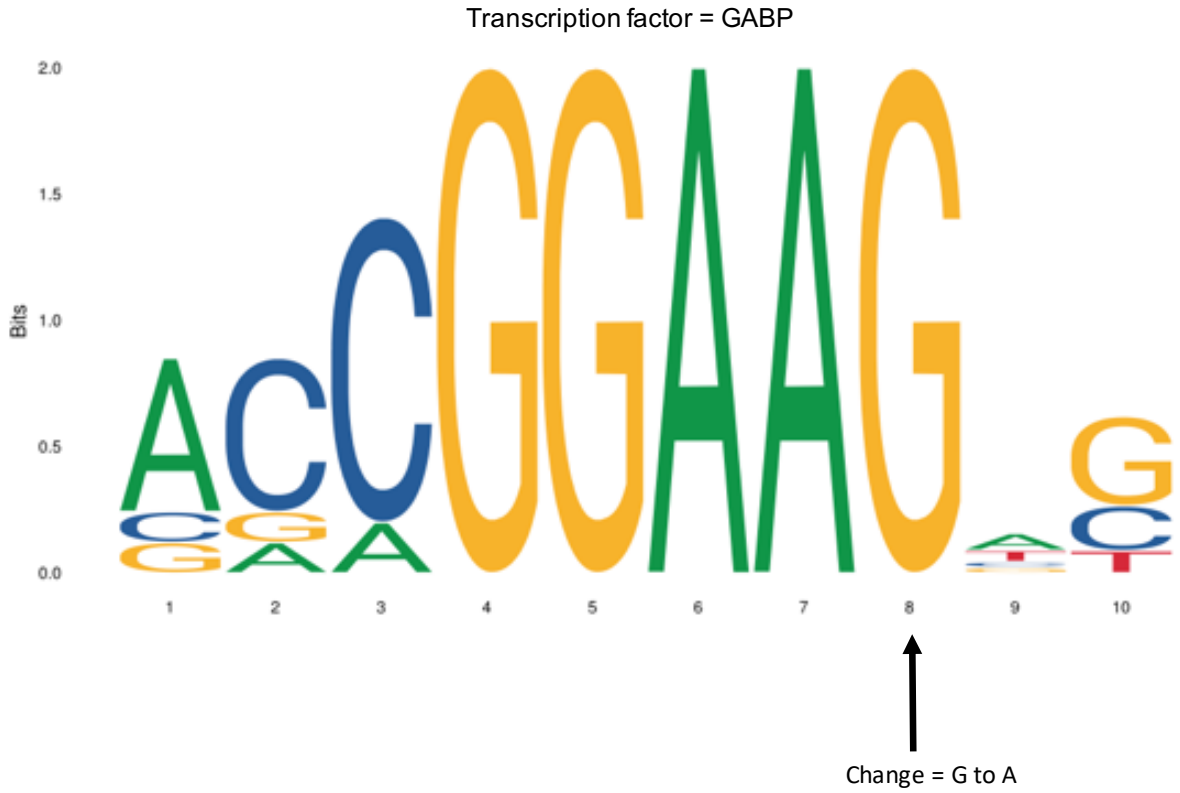


Figure 3.15: Disruption of GABP binding motif in *SKI* as observed in 3 pedigrees. Sequence logo obtained from the JASPAR[199] database.

Proto-oncogene tyrosine-protein kinase is a non-receptor tyrosine kinase protein that in humans is encoded by the *SRC* gene. It is involved in the progression and metastasis of several cancer types including breast, pancreatic, colon and brain cancer[265]. The *SRC* pathway is also active in melanoma[266] and *SRC* inhibitors are seen as potential therapeutic agents in the treatment of melanoma[267] with several studies and clinical trials focussing on *SRC* inhibitors in solid tumours[268]. In particular, a combination of *SRC* and *MEK* inhibition was reported to suppress the growth and invasion of melanoma cells[269].

A pedigree from Leeds with 2 sequenced members carried an intronic variant (GRCh38 reference build, Chromosome 20 genomic position 37396591, G>C) . This variant is shown in Figure 3.16 and affects a binding site for the transcription factor reported to be MSN2 on JASPAR.

The binding motif for MSN2 comprises of 5 positions, with positions 2 to 4 being highly conserved; these positions always have a guanine nucleotide. Adenine is present with the highest frequency in position 1 with guanine and cytosine being present at lower frequencies. The observed mutation changes a guanine at the highly conserved position 2 to a cytosine,

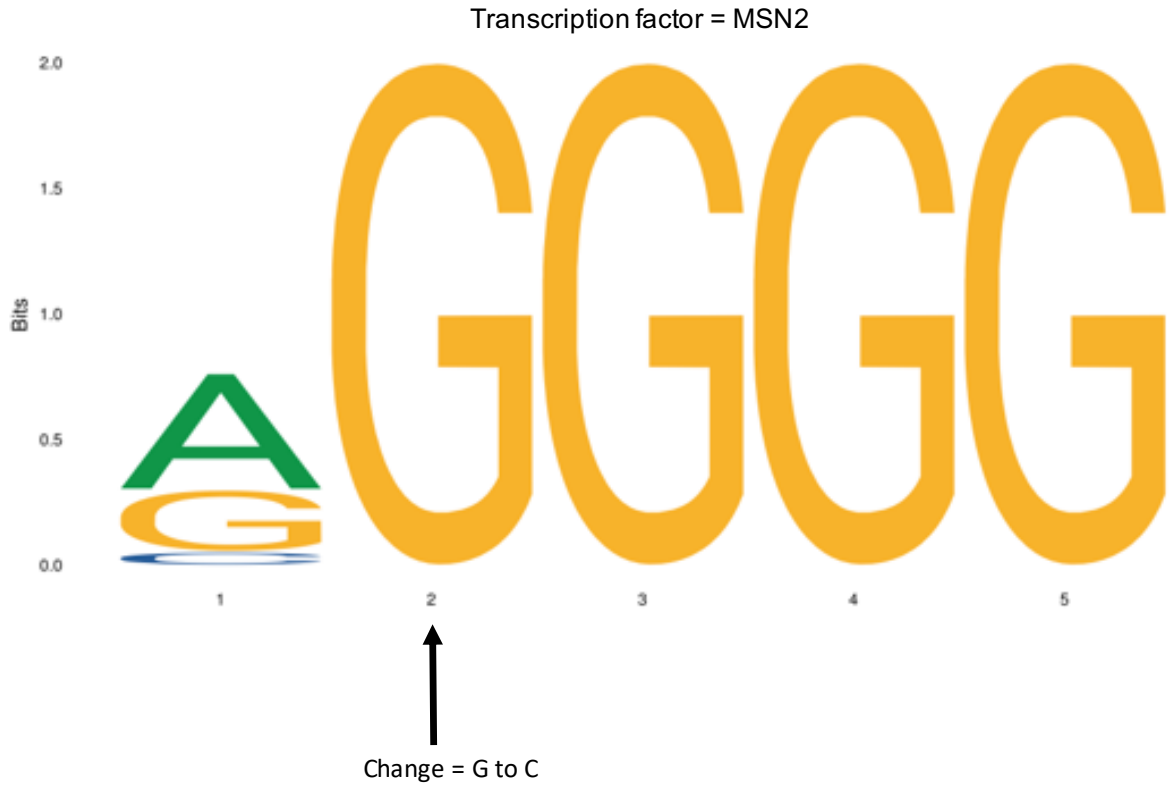


Figure 3.16: Disruption of MSN2 binding motif in *SRC* as observed in a pedigree from Leeds. Sequence logo obtained from the JASPAR[199] database.

potentially disrupting the binding of MSN2. Whilst MSN2 is a transcription factor that plays a role in stress response in yeast, it isn't reported to be active in *homo sapiens*. However, the 5 base binding motif of MSN2 is the same as the central binding motif of EGR1, comprising of several repeating guanine units. This is shown through bases 6 to 10 in Figure 3.14. This mutation in *SRC* would also disrupt and impact this binding motif, indicating that the affected transcription factor could be EGR1 as opposed to MSN2. Additionally, while the binding matrix for this variant refers to MSN2 on JASPAR, Ensembl has annotated this region as having an EGR1 binding motif.

3.9 Structural variant analysis

A 233,780 bp deletion was identified in chromosome 9 (GRCh38 reference build, 22209075-22442855) and is shown in Figure 3.17.

This deletion was detected to be 213,774 bases upstream of *CDKN2A,* a prominent melanoma driver gene (discussed in Section 1.5). The deletion was observed in a large pedigree from

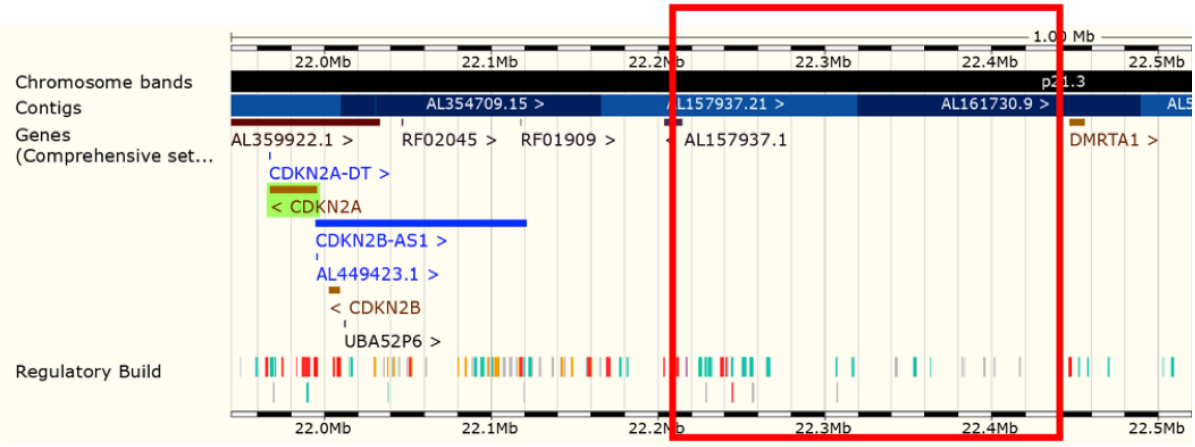


Figure 3.17: A 233,780 bp deletion upstream of *CDKN2A* observed in a Sydney pedigree with 11 sequenced members. The coloured bars indicate the relative locations of genes in chromosome 9 including *CDKN2A*, which is highlighted in green. The red box highlights the deleted region within the chromosome. This lies between *CDKN2A* and *DMRTA1*. Ten out of eleven sequenced members carried the deletion. The remaining sequenced member was confirmed to be a phenocopy by our collaborators in Sydney. Adapted from Ensembl.

Australia, within individuals residing in both Sydney and Brisbane. Twenty individuals with melanoma were identified within the pedigree with four individuals having multiple primary melanomas during their diagnosis. Lung cancer, pancreatic cancer, breast cancer and colon cancer were also observed in individuals from the pedigree. Eleven members with melanoma were chosen to be sequenced within the family, including three out of the four individuals with multiple primary melanomas. Ten out of eleven sequenced members from the pedigree carried the deletion. Other members with melanoma were sequenced by our collaborators in Australia, who confirmed the deletion in these members as well. The high segregation of the disease with the deletion, combined with the knowledge that it potentially disrupts enhancers and transcription factors that regulate expression of *CDKN2A* makes it a highly compelling candidate for a driver variant. Additionally, while roughly 40% of familial melanoma pedigrees carry germline mutations in *CDKN2A,* such variants have only been identified in the context of SNPs affecting the coding region of the gene. The role of structural variants in *CDKN2A* are still unexplored due to a lack of large scale whole genome germline sequencing of familial melanoma patients.

The identification of this deletion led to the investigation of other variants which segregated with the disease in a similar manner within the pedigree. A missense variant was identified in *DMRTA1* which encoded a p.E383Q (GRCh38 reference build, Chromosome 9, genomic position 22451543, c.2669G>C) . This variant was previously discussed in Section 3.6.2 as it

was a variant that segregated with the disease in the most number of affected members. The p.E383Q variant was observed in the same ten members who carried the deletion upstream of *CDKN2A*. *DMRTA1* has not been associated with melanoma development previously but is located adjacent to *CDKN2A* on Chromosome 9, as shown in Figure 3.17, bordering the deletion. This indicates that the missense variant is in linkage disequilibrium with the deleted region. In addition to the Sydney pedigree, individuals from two other pedigrees also carried the missense variant. However, these individuals were exome sequenced and not whole genome sequenced. Although the deletion could not be identified within the other members carrying the *DMRTA1* missense variant, the presence of the deletion in the Sydney pedigree implies that the other members could also carry the deletion.

Potentially interesting structural deletions were also observed in *AT-Rich Interaction Domain 1B* (*ARID1B)* and *Cut Like Homeobox 1 (CUX1).* However, mutations in *ARID1B* and *CUX1* have not been previously associated with familial melanoma. Additionally, these variants were present in a large number of individuals within the families that carried the deletions, but they did not completely segregate with the disease. Thus, these deletions may not as impactful or significant as they are for *CDKN2A*. Additionally, the absence of a suitable control dataset for structural variants complicates the process of discovering of novel, rare structural variants. However, the presence of such variants within this dataset warrants further investigation into the importance and role of structural variants in carcinogenesis.