Chapter 2

Dataset description and methods used for the generation and analysis of the familial melanoma datasets

2.1 Introduction

This chapter introduces the process of selection for the sequencing dataset used in the project. The pedigrees were initially chosen and sequenced as part of four distinct datasets, two exome and two whole genome. The first exome dataset, referred to as the primary exome dataset in this chapter, was sequenced and partly analysed before I started working on the project. The remaining three datasets were chosen and sequenced after I started the project, for which I collaborated with members of our melanoma consortium called GenoMEL, described in Section 2.2.1. The description of these pedigrees as four distinct datasets is purely for the distinction of the varied choices of pedigree selection, sequencing methodology, technology used and the institutions they were sequenced in. Eventually, these cohorts were merged into a single dataset and analysed as one large dataset for the rest of the project, with the exception of only whole genome sequences being considered for the noncoding and structural variant analysis. This chapter includes a description of assembly and sequencing of all four datasets and the eventual process used to merge them into a single dataset and perform variant calling on them. Additional methods on the filtering of the variants are also described in detail. Following this, all the methods for analysing the dataset including an association analysis on the coding region variants, a joint association and linkage analysis, secondary exonic analyses, and studies on variants disrupting transcription factor binding motifs and large structural alterations are

also described in this chapter. The results from all of these approaches are presented in the following chapter.

2.2 Dataset description and assembly

2.2.1 An introduction to GenoMEL

GenoMEL is a melanoma genetics consortium comprising researchers and investigators from 24 institutions across the world focussing on the identification of genes that increase the risk of both familial and sporadic melanoma. It is the largest colletion of familial melanoma data in the world and was started in the early 1990s by Professor Julia Newton Bishop from the University of Leeds, with the number of collaborators constantly increasing every year since then. The consortium also investigates the interaction of these genetic factors with environmental factors and the relevance of the inheritance of these genes to familial melanoma risk. The consortium emphasises open knowledge sharing and transfer of key knowledge related to melanoma genetics research. This project was funded as part of the MELGEN Early Training Network under GenoMEL with data being provided by 8 of the 24 GenoMEL institutions studying familial melanoma.

2.2.2 Cohort description

A total of 308 patients diagnosed with familial melanoma from 133 different pedigrees from across the world were sequenced as a part of this project, making it the largest dataset of its kind to date. These patients were selected to be *CDKN2A* and *CDK4* negative to increase the chances of finding a novel familial melanoma predisposition driver gene. An example of the type of pedigree chosen to be sequenced is provided in Figure 2.1. The pedigree in this figure comprises of 16 members, 6 of whom were diagnosed with melanoma.

The pedigrees that were sequenced as part of this dataset were identified by collaborators in 9 different institutions across the world, shown in Table 2.1.

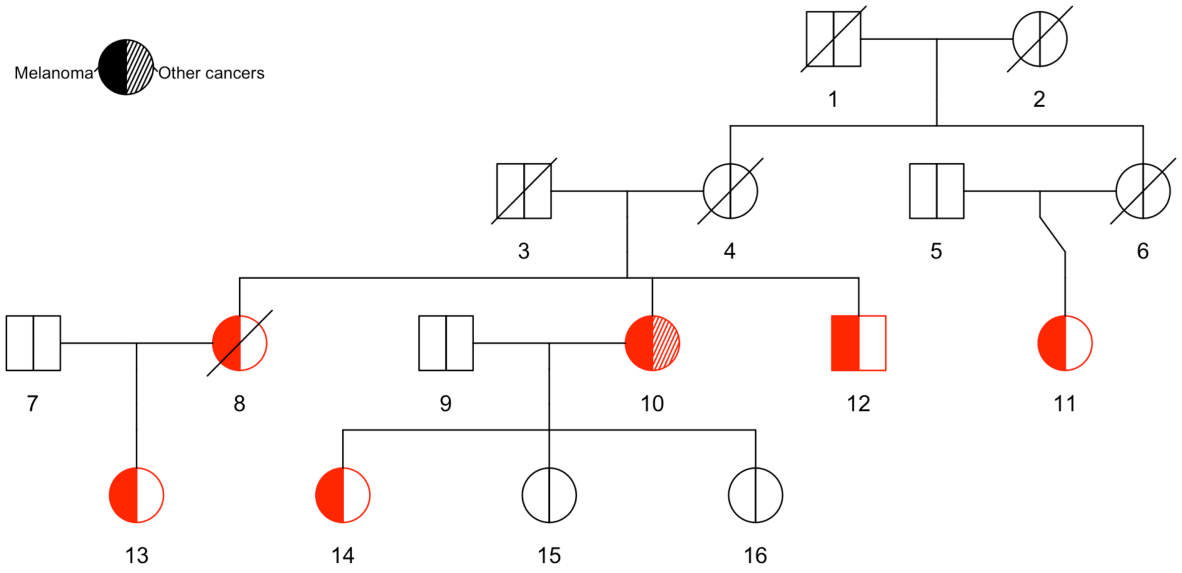


Figure 2.1: An example of a pedigree sequenced as part of the study. Circles indicate female individuals while squares indicate male individuals. A diagonal line across the symbol indicates that the individual is deceased. The members of the pedigree marked in red were the patients sequenced from this pedigree. In this case, there were 6 affected members in the pedigree, all of whom were sequenced.

Institution	Location	Lead Principal Investigator
University of Pennsylvania	Pennsylvania, United States of America	Dr. Peter A. Kanetsky
University of Sydney	Sydney, Australia	Professor Graham Mann
The QIMR Berghofer Medical Research Institute	Brisbane, Australia	Professor Nicholas Hayward
Leiden University Medical Center	Leiden, The Netherlands	Dr. Remco van Doorn and Dr. Nelleke Gruis
University of Leeds	Leeds, United Kingdom	Professor Tim Bishop and Professor Julia Newton-Bishop
Karolinska Institutet	Stockholm, Sweden	Dr. Veronica Höiom
Rigshospitalet	Copenhagen, Denmark	Dr. Karin Wadt
Institut d'Investigacions Biomediques August Pi I Sunyer	Barcelona, Spain	Dr. Susana Puig
Kings College	London, United Kingdom	Dr. Veronique Bataille

Table 2.1: The different collaborative institutions and their corresponding lead investigators who helped provide samples for this project and their respective locations.

Additional criteria based on the number of primary melanomas, age of onset and number of affected members in each pedigree were also used to finalize the list of patients chosen for sequencing. DNA from these patients was collected by our collaborators at institutions mentioned in Table 2.1 and sent to the Wellcome Sanger Institute for sequencing. The information regarding the sequencing of each dataset, including sequencing platforms, baits and read lengths for the whole genome and exome sequences are provided in Sections 2.2.3 and 2.2.4 respectively. The 308 patients were sequenced using a mixture of exome and whole genome sequencing and were initially sequenced as four individual datasets:

1. Pilot whole genome dataset - Consisting of 123 whole genome sequences from 32 pedigrees.

2. Secondary Leiden whole genome dataset - Consisting of 28 whole genome sequences from 6 pedigrees.

3. Primary exome dataset - Consisting of 80 exome sequences from 67 pedigrees.

4. Secondary exome dataset - Consisting of 77 exome sequences from 28 pedigrees.

The distribution of samples and families in each dataset based on the origin of the samples are provided in Table 2.2.

The origin of cases with multiple primary melanomas and early age of onset (<40 years of age) are provided in Table 2.3. Across all datasets, 29.06% of patients were detected to have multiple primary melanomas while 27.92% of patients had an early age of onset.

The average number of people affected and sequenced from each pedigree across all the datasets is given here:

- Pilot whole genome dataset = 6 affected and 4 sequenced.

- Secondary Leiden whole genome dataset = 5 affected and 5 sequenced.

- Primary exome dataset = 4 affected and 2 sequenced.

- Secondary exome dataset = 4 affected and 3 sequenced.

This shows that the whole genome datasets have a higher average number of people affected and sequenced compared to the whole exome datasets. This is also due to the stricter criteria of selection for patients and families imposed on these datasets. In the compiled overall dataset, 5 people were affected on average in every family and 2 were sequenced across all the datasets.

Location	Number of families (number of samples)	Sequence type	Dataset
Pennsylvania	4(15)	Whole genome	Pilot whole genome dataset
Sydney	9(56)	Whole genome	Pilot whole genome dataset
Brisbane	6(27)	Whole genome	Pilot whole genome dataset
Leiden	2(6)	Whole genome	Pilot whole genome dataset
Leeds	1(2)	Whole genome	Pilot whole genome dataset
Stockholm	1(2)	Whole genome	Pilot whole genome dataset
Denmark	4(10)	Whole genome	Pilot whole genome dataset
London (KCL)	5 (5)	Whole genome	Pilot whole genome dataset
Leiden	6(28)	Whole genome	Secondary Leiden whole genome dataset
Leeds	65(75)	Exome	Primary exome dataset
Leiden	2(5)	Exome	Primary exome dataset
Barcelona	10(20)	Exome	Secondary exome dataset
Pennsylvania	1(3)	Exome	Secondary exome dataset
Brisbane	4(15)	Exome	Secondary exome dataset
Stockholm	2(6)	Exome	Secondary exome dataset
Sydney	7(28)	Exome	Secondary exome dataset
London (KCL)	3(3)	Exome	Secondary exome dataset
Leiden	1(2)	Exome	Secondary exome dataset

Table 2.2: Distribution of samples by location, type of sequence and dataset.

Dataset	Proportion of cases with multiple primary melanomas (MPM)	Proportion of early onset cases	Number of cases for which the information was unavailable
Pilot whole genome dataset	36/116	30/116	7 for MPM, 7 for early onset cases
Secondary Leiden whole genome dataset	5/24	11/24	4 for MPM, 4 for early onset cases
Primary exome dataset	3/9	2/9	77 for MPM, 77 for early onset cases
Secondary exome dataset	15/54	12/48	23 for MPM, 29 for early onset

Table 2.3: Distribution of patients with multiple primary melanomas and early age of onset in each dataset.

2.2.3 Whole genome sequences - Sample selection and sequencing

2.2.3.1 Pilot whole genome dataset

In order to determine mutations causative of familial melanoma, the genomes of 123 individuals from 32 families were sequenced. These samples were obtained from our collaborators in Sydney, Pennsylvania, Stockholm, Leiden, Denmark, Leeds and Brisbane who are part of the GenoMEL consortium (https://genomel.org). The dataset also included 8 samples obtained from King's College, London who are not a part of the GenoMEL consortium. These 8 samples were selected due to the presence of MPMs and/or early age of onset as opposed to multiple members of the family affected. The distribution of the 123 individuals across these locations is given in Table 2.2. Informed consent was obtained by each institution (Sydney: HREC/13/CIPHS/71, Pennsylvania: 14.03.0033 Protocol MCC 17751, Sweden: 03-471, 03-713, Leeds: 99/3/045, Leiden: Protocol No. P00.117-gk2, Copenhagen: Protokol af 7. juni 2012, version 3, Brisbane: P452 (H0204-013), London: 07/HO802/84). The distribution of the families and samples are given below in Table 2.2. The criteria for selecting these particular samples varied depending on their origin of the families. These criteria are elucidated in Table 2.4.

Genomic DNA (500 ng) was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. Adapter-ligated libraries were am-

Families	Criteria
All families	Presence of multiple primary melanomas and an early age of onset (<40 years) with 2 or more DNAs available to sequence per family.
European and American families	5 or more cases with 2 or more DNAs available to sequence per family.
Australian families	6 or more cases with 2 or more DNAs available to sequence per family.

Table 2.4: Criteria for selection of whole genome samples in the pilot dataset.

plified by 6 cycles of PCR and subjected to DNA sequencing using the HiSeqX platform (Illumina) according to manufacturer's instructions. Read lengths of 150bp were obtained for this dataset

2.2.3.2 Secondary Leiden whole genome dataset

An additional 29 samples from 6 pedigrees were obtained from our collaborators at the Leiden University Medical Center. The selection criteria for the samples were the same as the ones previously mentioned in Section 2.2.3.1, i.e., 5 or more cases in each pedigree with at least 2 or more DNA samples available to sequence. Ethical approval for sequencing was obtained (Leiden: Protocol No. P00.117-gk2). Genomic DNA was sheared and amplified in the same manner as Section 2.2.3.1. However, this was done using the HiSeq2500 instead of the HiSeqX platform. Read lengths of 100 bp were obtained for this dataset. One sample was removed for low average coverage (<9X across the genome), resulting in a total of 28 samples from 6 pedigrees.

2.2.4 Exome sequences - Sample selection and sequencing

2.2.4.1 Primary exome dataset

The primary criteria for selection of the families for sequencing within this dataset were less stringent compared to the whole genome datasets. These conditions are given in Table 2.5.

The pedigrees for this study were recruited by collaborators from the University of Leeds and the Leiden University Medical Centre. Cases from these pedigrees were confirmed to be negative for *CDKN2A* and *CDK4* mutations. 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 blood using standard methods

Families	Criteria
All families	Presence of multiple primary melanomas in multiple members and/or an early age of onset (<40 years)
European families	3 or more cases with 2 or more DNAs available to sequence per family.

Table 2.5: Criteria for selection of exome samples in the primary exome dataset.

from our respective collaborators at these institutions, shown in Table 2.1.

5 µg of genomic DNA were sent for sequencing at the Wellcome Sanger Institute and exonic regions were captured with the Agilent SureSelect Target Enrichment System. Paired-end reads of 75 base pairs (bp) were generated on the HiSeq 2000 platform. A subset of samples from the Leeds cohort were sequenced at the Beijing Genomics Institute (BGI), using the Illumina HiSeq2000 platform, which generated 90 bp paired-end reads.

This dataset was previously a part of a larger dataset used by Dr Carla Daniela Robles Espinoza at the Wellcome Sanger Institute and it led to the discovery of *POT1* as a familial melanoma driver gene[143]. Two of the families from this dataset were also sequenced in the whole genome dataset. However, both the exome and whole genome sequences of these families were included in the analysis to confirm that any variant seen within these families was present in both versions.

2.2.4.2 Secondary exome dataset

The final subset of 77 samples from 28 pedigrees were obtained from collaborators in Barcelona, Sydney, Pennsylvania, Stockholm, Leiden, Leeds and Brisbane who are part of the GenoMEL consortium. Similar to the pilot whole genome dataset, samples with MPMs and/or early age of onset (<40) were again obtained from King's College, London who are not a part of the GenoMEL consortium. The distributions of the 77 samples across these locations, number of samples and number of families are given in Table 2.2. Informed consent was obtained by each institution in the same way as mentioned in Section 2.2.3.1. The criteria for the selection of these pedigrees was again less stringent compared to the whole genome samples. These criteria are shown in Table 2.6. Genomic DNA (500 ng) was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. These samples were then sequenced at the Wellcome Sanger Institute using the Illumina HiSeq 2000 platform with read lengths of 75 bp.

Families	Criteria
All families	Presence of multiple primary melanomas and an early age of onset (<40 years) with 2 or more DNAs available to sequence per family.
European and American families	4 or more cases with 2 or more DNAs available to sequence per family.
Australian families	5 or more cases with 2 or more DNAs available to sequence per family.

Table 2.6: Criteria for selection of exome samples in the secondary exome dataset.

2.3 Alignment of DNA sequence data and variant calling

Samples from the different datasets were sequenced at different times. They were however eventually aligned and processed into a single, larger dataset. The sequences were aligned to the latest version of the reference build of the human genome which was the GRCh38 build of the human genome. A Burrows-Wheeler aligner (BWA-MEM)[162] was used by the core sequencing facility to align the sequences. This was followed by the estimation of sequencing depth in every sample across all positions. To ensure good coverage across all samples, a coverage threshold of 15X was established. A single sample which was whole genome sequenced and belonged to the secondary Leiden whole genome dataset was removed from the dataset due to low mean sequencing depth, i.e., less than 15 reads. Variant calling was then performed using GATK Haplotype Caller[163] which employs a 'joint calling' approach. An intermediate file called a Genomic VCF (gVCF) file is create for each sample which contains the genotype information for that sample across all positions. The caller then jointly calls genotypes across all samples from each gvcf file to create a single multisample file. This ensures that the genotype calls are available for all samples at all positions, regardless of whether a specific sample contains a variant at a given loci or not. This output is termed as a Variant Calling Format file (VCF). The multisample VCF file containing all the variants across all the samples was then annotated with the predicted consequences for each variant using a tool called Variant Effect Predictor[164] which was established by Ensembl. This resulted in the first complete set of variants which comprised of all coding and non-coding mutations, prior to further processing or filtering. The final step in the initial variant calling process was to then filter for variants that were predicted to disrupt or alter the protein produced by different genes. The different consequences that were retained for this step and their predicted impact on the gene are shown in Table 2.7. The second complete set of variants was obtained as the output of this step.

Variant Effect Predictor Consequence	Description
Protein altering variant	A variant that affects the protein through a change in the codons.
Missense variant	A non-synonymous variant that changes an amino acid without affecting protein length.
Inframe deletion	An inframe non-synonymous variant which removes bases from the coding sequence.
Inframe insertion	An inframe non-synonymous variant which incorporates bases into the coding sequence.
Transcript amplification	Amplification of a region containing a transcript.
Start lost	A non-synonymous variant that alters a base in the canonical start codon.
Stop lost	A non-synonymous variant that alters a base in a stop codon resulting in longer transcripts.
Frameshift variant	A non-synonymous variant that disrupts the reading frame of the protein through the addition or removal of multiple adjacent bases.
Stop gained	A non-synonymous variant which disrupts a codon in such a way as to introduce a stop codon which results in shorter transcripts.
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
Transcript ablation	Deletion of a region containing a transcript.

Table 2.7: The list of predicted protein altering consequences and their impact on the protein, as obtained from the Ensembl variation website[165].

Additional filtering, annotation and processing of variants were then performed on these two sets of variants which are explored in detail in the following chapters. A summary of the workflow up to this point is given in Figure 2.2.

2.4 Exploration of population stratification bias within the dataset

Inherent differences in the ancestry of different population subgroups, especially with respect to the frequency of specific variants and alleles may lead to a bias in population based studies. As this project eventually involved association studies and the comparison of familial melanoma cases to unaffected controls, there was a possibility that an association could be identified with a particular loci due to such a bias within the population and not due to the influence of the loci on the disease status. Such a scenario is defined as population stratification. In order to determine the possible presence of population stratification within the dataset, it was deemed necessary to establish and confirm the reported population subgroups of the cases as accurate.

Genotype and variant information for different population subgroups were obtained from the 1000 genomes project and used as the control dataset for this assessment[166]. While there were several different nationalities reported within the 1000 genomes project, the subgroups were largely classified into: European, Indian, South American, Chinese and African. These variants were filtered to have an allele frequency between 0.2 and 0.8 as variants with extreme frequencies could also potentially bias the accurate estimation of the population subgroup for each sample. These variants were also filtered to ensure that they were in linkage disequilibrium with each other. Once this subgroup of variants, determined to be variable across the population, were identified from the controls, the same variants were also extracted from the complete dataset of familial melanoma cases.

A principal component analysis (PCA) was performed for all samples on this subset of variants using PLINK v1.9. To reliably differentiate between the population subgroups, the first three principal components/eigen vectors were needed (named PC1, PC2 and PC3 respectively). Figure 2.3 highlights the distribution of population clusters as a snapshot of a 3D plot.

The three principal components are represented with the three dimensions of the figure. Each nationality was assigned its own colour for the plot with larger population subgroups sharing similar shades. The familial melanoma cases were all marked in black and are denoted

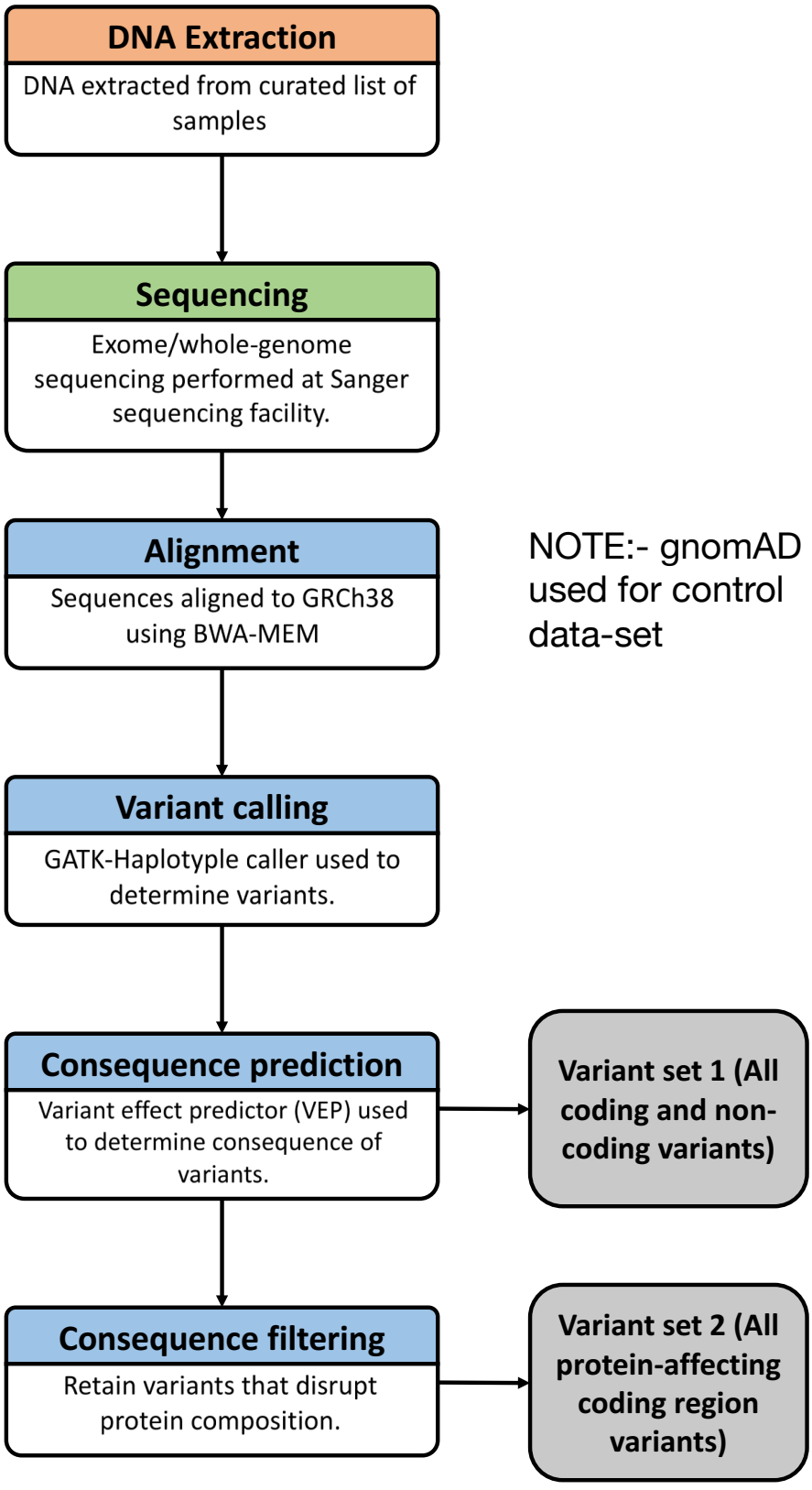


Figure 2.2: Workflow describing the steps involved in the generation of candidate variant sets in the search for novel driver genes and variants involved in melanoma susceptibility.

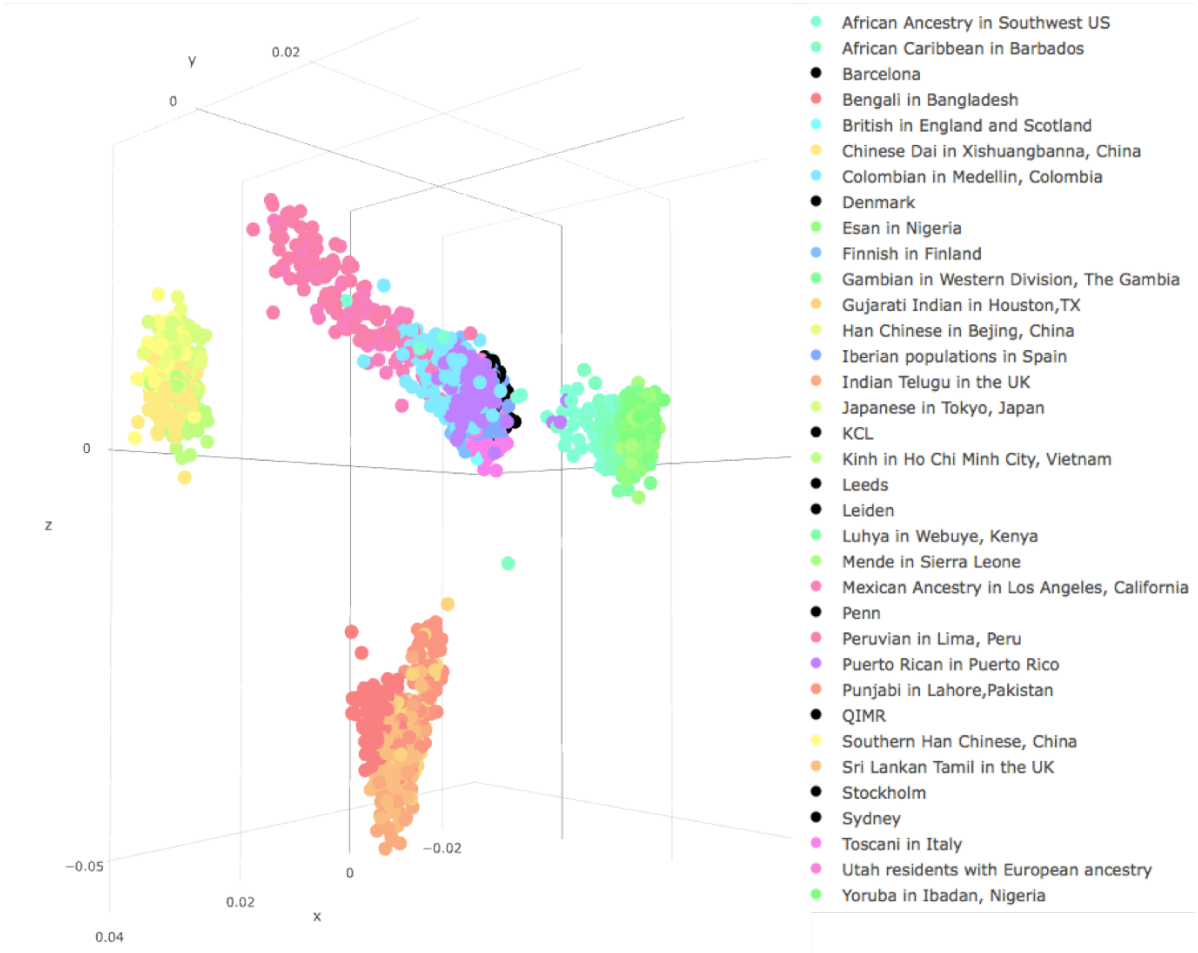


Figure 2.3: Principal component analysis to verify ethnicity.

using a small circle. It is evident that all the individuals from the Indian subcontinent, coloured red to orange in the figure, cluster together. Similarly, there are two distinct green clusters of the Chinese and the African subpopulations as well. The larger cluster at the top is a mixture of the North American, the central American and the European population subgroups. The central and south American populations are however distinctly clustered on the x dimension of the 3D plot which leaves all the Europeans clustered together. The familial melanoma cases are also clustered together with the European population. Therefore, the different population subgroup clusters are observed as being independent of each other when all three principal components are considered together.

Importantly, the cases (marked in black) cluster along with the European subgroup in all three subplots. The familial melanoma cases were all reported to be of European origin when they were sampled and sequenced; this is confirmed through the above analysis. This also implies that there is no population stratification within the dataset and that association analysis can be performed between the familial melanoma cases and a set of unaffected, European origin controls.

2.5 Estimation of polygenic risk scores

2.5.1 Introduction

The origin of a disease with genetic roots may be traced either to the presence of a few single, highly penetrant rare alleles or due to a high burden of common, low risk alleles. The estimation of polygenic risk scores is a measure used to determine if the presence of the disease in a particular individual is more likely due to the former scenario or the latter. Polygenic risk scores are metrics which are a numerical measure of the impact of a combination of genetic variants and their associated weights on a given trait. GWAS studies have helped in determining significant variants with a high association to the phenotype of interest. However, polygenic risk scores are considered to be a better approach when the trait is predicted to be affected or determined by a combination of a large number of variants with lesser impact on the trait than a singular high impact variant. These variants may or may not be statistically significant, and as such may not be identified in a standard GWA study. Such traits which are determined by not a single variant but by a combination of multiple variants in different genes are called as polygenic traits.

Each variant that affects a polygenic trait is assigned a weighted score, which is proportional to the association of the variant with the trait. Higher weighted scores correspond to

a larger association with the trait of interest with negative scores corresponding to negative association. Each allele at these variant positions are also given a coefficient corresponding to the impact of the allele on the association at the variant position. If an allele has a higher coefficient, it implies that the presence of the allele increases the impact of the variant on the trait. The risk score of each variant is determined as the product of the coefficient corresponding to the alleles and the weighted score of each variant. This would vary based on the genotypes present in any given sample at the loci. The polygenic risk score of a sample is finally estimated the sum of the risk scores of all the variants affecting the trait[167].

Common examples of polygenic traits include skin colour, eye colour and hair colour. In addition to these normal phenotypes, several disorders are also thought to follow the polygenic model including type 2 diabetes and coronary heart disease[168, 169]. In recent years, polygenic risk scores have also been used in determining patient risk for schizophrenia, bipolar disorders and for certain types of cancer including breast cancer and prostate cancer[170–172].

In order to reliably determine the burden of common, low risk alleles within the familial melanoma cases, a secondary set of sporadic cases and unaffected controls was required. The aim of this approach was to compare the risk scores of the familial melanoma cases and compare these to the risk scores of sporadic cases and unaffected controls to observe if the familial melanoma cases have a higher risk score in general. If they did, it would imply that the presence of melanoma within these pedigrees could possibly be explained by the higher burden of common mutations. If there was no discernible distinction between the different groups, this would implicitly point to the presence of rare, highly penetrant variants within the familial melanoma pedigrees.

2.5.2 Methods

A previously conducted genome wide association study (GWAS) on melanoma and a follow-up meta-analysis study[51] established 20 loci, shown in Table 2.8 as possible SNPs which increase predisposition to melanoma. This GWAS analysis included a estimation of risk bestowed by each of the listed variants.

The individual risk scores of these variants, as estimated by Law et al[51] is used in the estimated of the polygenic risk score of each sample. Each polygenic risk score is determined by combining the risk scores of all 20 SNPs based on the genotypes of the alleles present at those positions within the sample. The polygenic risk scores for all 123 familial melanoma cases within the pilot dataset were estimated using this approach. These scores were calculated using PLINK v1.941 and also verified manually. The secondary dataset required for the

rs_ID	Gene	Chromosome	Position	Reference allele	Alternate allele	Beta value
rs12410869	Intergenic	1	150883677	G	T	-0.130
rs1858550	Intergenic	1	226420403	C	A	-0.143
rs6750047	*RMDN2*	2	38049406	A	G	0.088
rs7582362	*FLACC1*	2	201311571	A	G	0.113
rs380286	*CLPTM1L*	5	1320132	G	A	0.152
rs250417	*SLC45A2*	5	33952273	G	C	-0.891
rs6914598	*CDKAL1*	6	21163688	T	C	0.108
rs1636744	Intergenic	7	16944656	C	T	0.105
rs7852450	*MTAP*	9	21825076	T	C	-0.212
rs10739221	Intergenic	9	106298549	T	C	0.120
rs2995264	*STN1*	10	103909085	G	A	0.144
rs498136	Intergenic	11	69552350	A	C	0.116
rs1393350	*TYR*	11	89277878	G	A	0.198
rs73008229	*ATM*	11	108316962	G	A	-0.188
rs4778138	*OCA2*	15	28090674	A	G	-0.178
rs12596638	*FTO*	16	54081917	G	A	0.143
rs75570604	*FANCA*	16	89780269	G	C	0.600
rs6088372	*RALY*	20	33998942	C	T	0.267
rs408825	*MX2*	21	41371569	C	T	-0.141
rs2092180	*PLA2G6*	22	38175556	A	G	-0.116

Table 2.8: Single Nucleotide Polymorphisms chosen for the polygenic risk score analysis.

comparison of these scores were obtained through the help of collaborators in the University of Leeds, namely Prof. Timothy Bishop and Dr Mark Iles. The genotypes of 1800 sporadic melanoma cases, 148 familial melanoma cases and 489 controls at these 20 loci were provided for this purpose. The polygenic risk scores for these samples were also estimated using the same method as used for the familial melanoma cases in the pilot dataset. The risk scores and their implications for the dataset are discussed in Section 3.2.

2.6 The determination of novel variants through association analysis

2.6.1 Selection of a control dataset

The cases chosen to be analysed for this project comprised familial melanoma patients. During the design of the project, it was decided that there would be no sequencing of unaffected

family members from the same pedigrees. The rationale behind this is that familial melanoma is a result of multiple factors (both genetic and environmental) which determines the onset of disease. An individual with a rare allele with high penetrance would have an increased risk of developing melanoma but might not actually develop melanoma. For example, *CDKN2A* is the single most important familial melanoma locus, with 45% of familial melanoma cases being attributed to germline mutations in *CDKN2A*[173]. However, the penetrance of *CDKN2A* mutations varies between 0.3 to 0.67, depending on the age of the carriers, indicating that even the most common familial melanoma gene is not completely penetrant[174]. In such a situation, sequencing unaffected members of the family as matched controls and filtering the common variants could result in the potential loss of the high-risk allele that is causative of the disease within the family.

To compensate for the absence of family-matched controls, a neutral, population matched control set of exome/whole genome sequence data was required. The Exome Aggregation Consortium (ExAC)[175] led by the Broad Institute was originally chosen for this purpose. ExAC was established as a consortium of exome sequencing projects involving unrelated individuals from several population genetics studies across the world. It contains aggregated variant level statistics on 60,706 individuals with population frequencies on each variant and allele provided. The distribution of samples across different population subgroups in ExAC is shown in Table 2.9. As the cases consisted entirely of patients of non-Finnish European ethnicity, the same ethnic group was chosen from ExAC to be used as the control data set. This comprised data from 33,370 individuals.

Population	Description	Genomes	Exomes	Total
AFR	African/African American	1,888	3,315	5,203
AMR	Latino	2,254	3,535	5,789
EAS	East Asian	2,016	2,311	4,327
FIN	Finnish	2,084	1,223	3,307
NFE	Non-Finnish European	18,740	14,630	33,370
SAS	South Asian	6,387	1,869	8,256
OTH	Other (population not assigned)	275	179	454
	Total	33,644	27,062	60,706

Table 2.9: Distribution of samples across different population groups in ExAC.

During the course of the project, a larger data set of controls called the Genome Aggregation Database (gnomAD)[175], also curated by the Broad Institute, became publicly avail-

able. gnomAD is a global collaboration of collated summary data from several large-scale sequencing projects. The total number of samples in the gnomAD data set 138,632 individuals compared to 60,706 in ExAC. As the gnomAD data set includes both exome and whole genome sequencing data, it was identified as the more suitable control and was chosen instead of ExAC. The distribution of samples across different population subgroups in gnomAD is given in Table 2.10.

Population	Description	Genomes	Exomes	Total
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632

Table 2.10: Distribution of samples across different population groups in gnomAD.

The non-Finnish European population subgroup was again chosen from gnomAD comprising 7,509 whole genomes. The sequences were all originally aligned to Genome Reference Consortium Human Build 37 (GRCh37). However, as the cases were sequenced and aligned to Genome Reference Consortium Human Build 38 (GRCh38), the gnomAD VCF file consisting of the aggregated variant information across all samples was lifted over from GRCh37 to GRCh38 using CrossMap v0.2.5[176]. The parameters used for this procedure are given in Table 2.11.

The aggregated coverage data for the samples across the genome was also downloaded from the gnomAD repository. The gnomAD dataset included information on 240,779,968 variants across all chromosomes and samples. Data from gnomAD v2.0.2 were used for this purpose.

Note: gnomAD does not include any variants on the Y chromosome . As a result, variants

Parameters	Description
python CrossMap.py	Execution of python script for CrossMap,
vcf	Indicating that the format of the input and the required format for the output is a VCF file.
hg38ToHg19.over.chain.gz	Location of chain files; chain file describe genome-wide pairwise alignments of positions between assemblies. In this case, the chain file is a mapping of alignments between GRCh38 and GRCh37.
input.vcf.gz	Location of input vcf file to be lifted over.
hs37d5.fa	Reference fasta file for the target output genome build, in this case, GRCh37.
output_hg19.vcf	Name of output VCF file to be generated.

Table 2.11: Parameters used for running CrossMap to lift the aligned gnomAD sequences from GRCh37 to GRCh38.

from the Y chromosome were excluded from this analysis.

2.6.2 Initial filtering of variants

2.6.2.1 Control variants from gnomAD

The lifted over variants from the gnomAD VCF file were annotated using the Variant Effect Predictor (VEP) tool from Ensembl. This was performed in order to annotate information including the affected gene and the consequences of the variant on protein function. These consequences include intergenic variants, intronic variants, nonsense mutations and loss of function mutations. Following this, the variants were filtered to retain non-synonymous, nonsense and loss of function mutations. The parameters used for this performing this function are given in Table 2.12.

The list of the consequences chosen to be retained for this purpose was the same as the ones previously mentioned in Table2.7. Genomic locations that were annotated with multiple alternate alleles (multiallelic variants) were split into multiple entries, each entry carrying information on one alternate allele. As gnomAD includes variant data on several population subgroups, as shown in Table 2.10, only the variants present in the non-Finnish European population subgroup were retained.

Parameter	Definition
-i	Input VCF file.
-o	Output VCF file.
-filter "Consequence in missense_variant,inframe_deletion, inframe_insertion, transcript_amplification,stop_lost, frameshift_variant,stop_gained, splice_acceptor_variant, start_lost,protein_altering_variant, splice_donor_variant,transcript_ablation"	List of consequences to be matched. If the annotated consequence is not present in this list, the variant is filtered.
--force_overwrite	Overwrites the output VCF if it already exists to create a new VCF file.
--only_matched	Only writes variants where the annotated consequence exactly matches the consequences provided in the filter step.

Table 2.12: List of parameters used for filtering variants based on their predicted consequence on protein function using VEP.

2.6.2.2 Case variants

Variants from the cases were previously annotated and processed, as described in Section 2.3. These variants were chosen as the initial dataset. Multiallelic variants were split as described in Section 2.6.2.1. There were a total of 131,840 variants in the cases.

During the generation of a VCF file for the sequences, each sample is annotated with a Genotype Quality (GQ) for every position. The GQ is a measure of confidence in the genotype assigned to the sample represented through a Phred-scaled score, which is derived from the probability of error. For instance, a Phred score of 30 indicates that there is a 1 in 1000 or 0.1% rate of error. A higher GQ indicates a higher confidence in the genotype of a sample for a given variant. The median GQ scores were calculated across all cases at every locus. The distribution of median GQ scores across the variants is shown in Figure 2.4. The first quartile(Q1) of the GQ distribution was at 90 while the third quartile (Q3) was at 99. Individual variants with median GQ < 30 were removed from the cases.

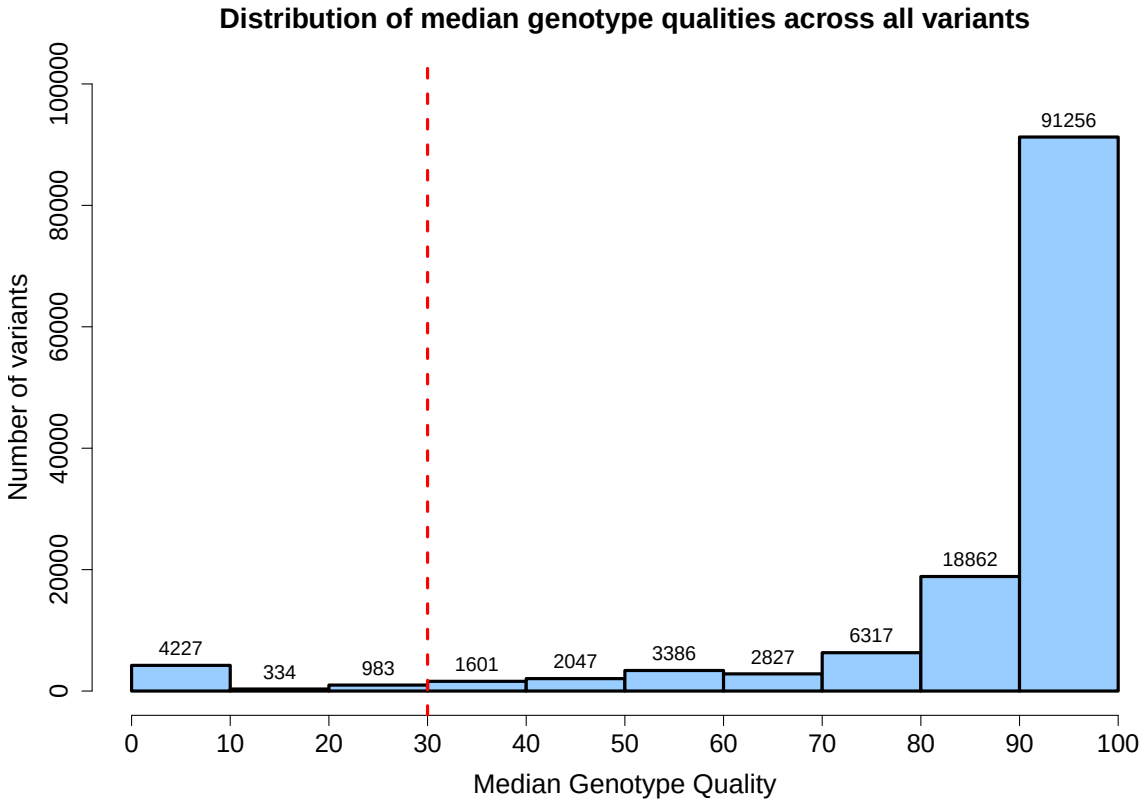


Figure 2.4: Distribution of median genotype qualities across all samples and positions. The median GQ score is denoted on the x-axis while the number of variants with that GQ score are given on the y-axis. The threshold for filtering variants based on the GQ scores was set at 30, this is denoted with the red line. 5,544 variants had median GQ score less than 30 and were removed from the set of variants, resulting in 126,296 variants in the cases.

2.6.3 Joint processing of case and control variants

Following the initial processing, a joint set of variants was established by merging the case and control variant sets together. This set was progressively filtered through several procedures for quality control which are shown in Figure 2.5 and are described here.

2.6.3.1 Annotation and filtering based on frequency of variants in gnomAD

In order to improve the power of the study in detecting important, rare mutations and to reduce the background variant burden, variants were filtered based on a population allele frequency. Prior to this, to focus on low-frequency variants with high penetrance, an sequence artefact filter was used for the cases: If a variant was contained in more than 8 families after the allele frequency filter, it was likely that it is a sequencing artifact as opposed to being involved in the development of the disease . Such variants were therefore removed from the cases. 88,994 variants remained in the cases after this step.

Variants in the cases were annotated with the allele frequencies of the variants from gnomAD. If the case variant was present in gnomAD, the frequency of mutation from the gnomAD dataset was annotated to the variant. If the case variant was not present in gnomAD, it was annotated as being absent. 32,987 of the case variants were not present in the gnomAD dataset while 56,007 variants were present in the gnomAD dataset. The distribution of population allele frequencies for these 56,007 case variants is shown in Figure 2.6. Figure 2.6 A shows the distribution for all allele frequencies while Figure 2.6 B shows the distribution for a subset of variants with allele frequencies less than 0.01. A similar distribution is shown for the control variants from gnomAD in Figure 2.7. A cut-off of one in a thousand (10^{-3}) was set as the threshold for allele frequency to select for rare mutations. Variants with a gnomAD allele frequency > .001 were removed from both the cases and the controls. Variants from the cases which were absent in gnomAD were retained.

Variants from the cases were also filtered based on the number of affected families. The distribution across all variants of the number of families is shown in Figure 2.8. The first quartile(Q1) for the family counts was 1 family while the third quartile(Q3) was 16. The median number of affected families for a variant was 2.

2.6.3.2 Annotation and filtering based on coverage of samples in cases and controls

Despite the removal of samples with low median coverage across the exome/genome in Section 2.3, variant loci with low coverage across most cases and controls were still encountered. Such variants were removed as they would yield unreliable variant calls. Variants from the

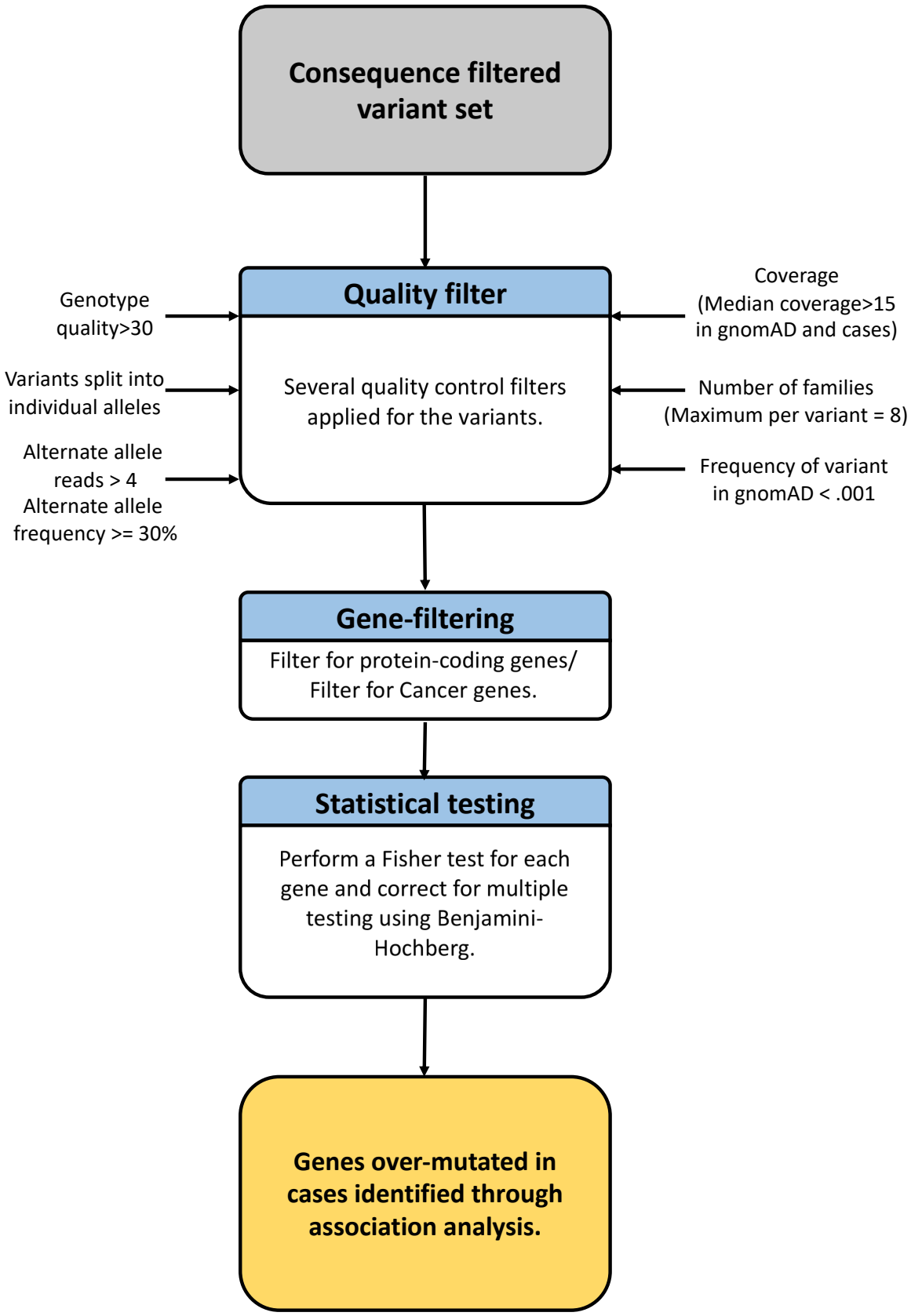


Figure 2.5: Overview of the steps involved in determination of genes with an increase burden of mutations in cases through an association analysis.

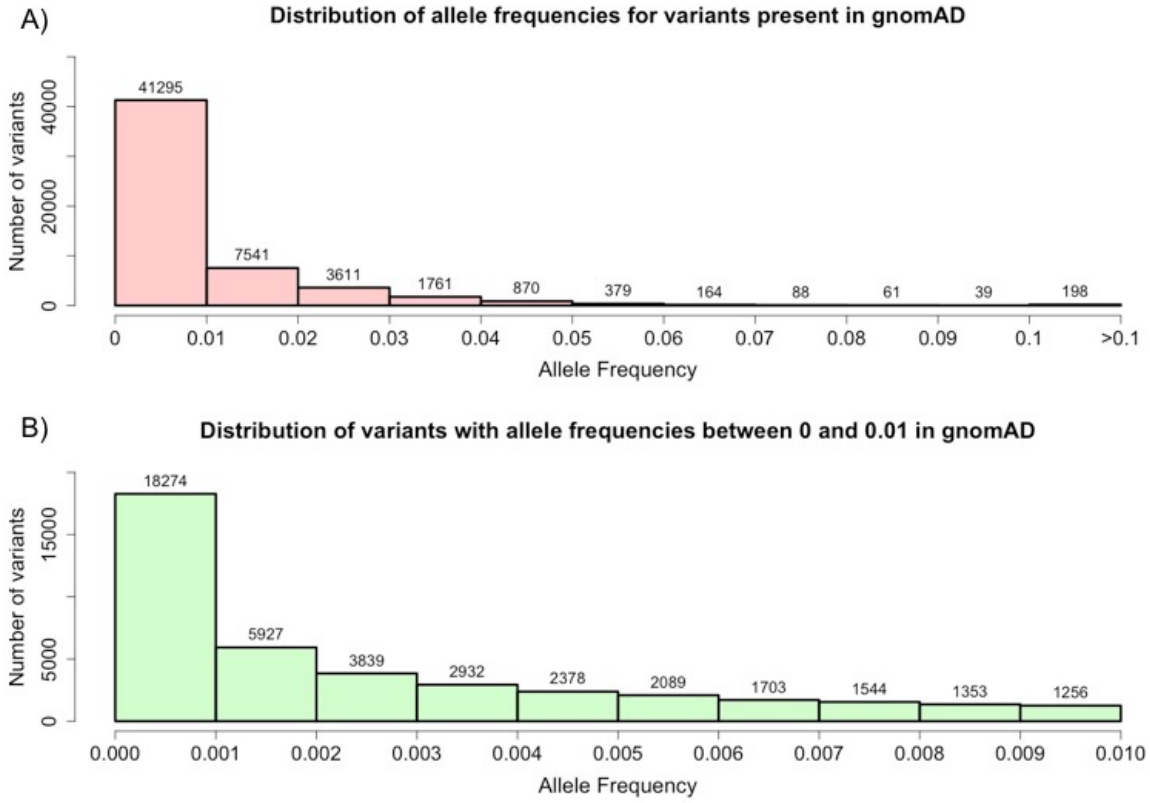


Figure 2.6: Distribution of allele frequencies for variants in cases A) This shows the distribution of allele frequencies for the 56,007 variants from the cases which were also mutated in gnomAD. 74% of the variants from the cases were present in less than 1% of the population in gnomAD. B) As the vast majority of variants had a low allele frequencies in gnomAD (<.01), the distribution of allele frequencies between 0 and 0.01 were plotted in this Figure. The distribution is again skewed as roughly 44% of the 41,295 variants had an allele frequency less than .001 (less than 1 in a 1000 people carried the variant). This was the chosen cut-off for the allele frequency.

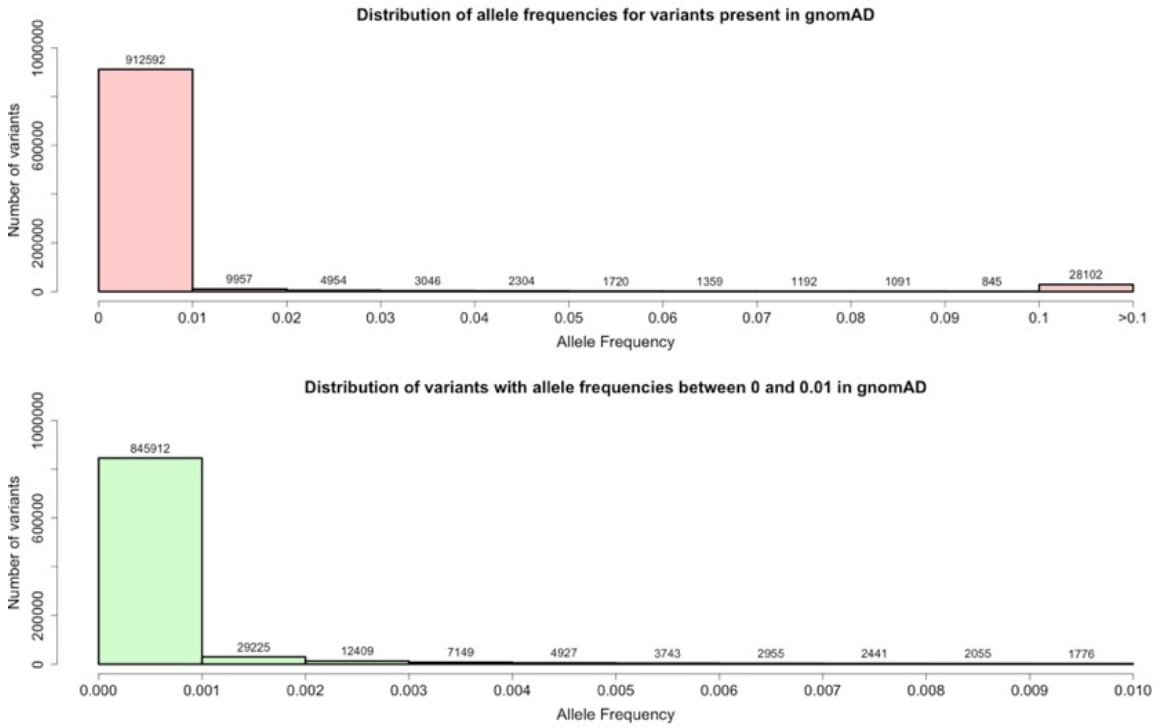


Figure 2.7: Distribution of allele frequencies for variants in controls.A) This shows the distribution of allele frequencies for the 967,162 variants from the gnomAD dataset. 94% of the variants occur in less than 1% of the population in gnomAD. B) This figure shows the distribution of allele frequencies between 0 and 0.01 for variants from gnomAD. 93% of variants occur at a very low frequency of less than 1 in a thousand.

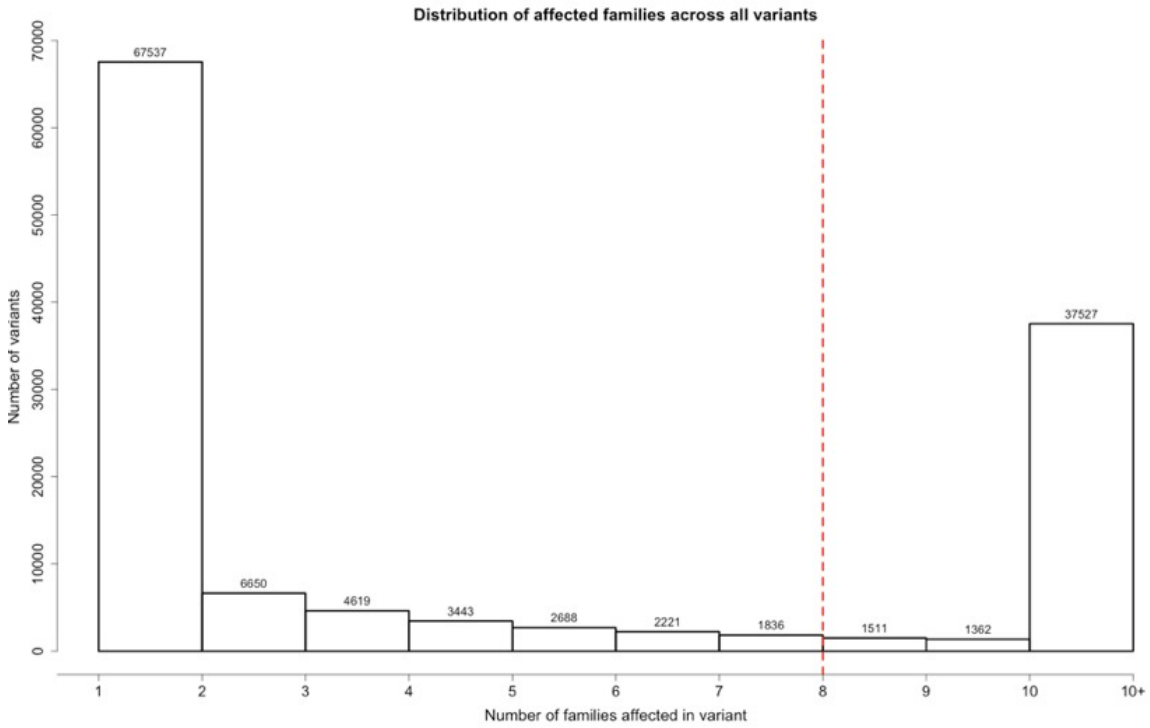


Figure 2.8: Distribution of affected family counts for all variants in cases. More than 50% of the variants in the cases were present in less than 2 families. An artefact threshold of 8 families was used, represented as the red line in the figure.

gnomAD dataset were first annotated with the coverages, which were obtained as a downloadable file. Individual coverages for the samples in gnomAD were not available; summary information across all the samples was provided for each variant. The following summary statistics were available (The number before the x indicates the number of reads covering each variant position): i) Mean coverage across all samples ii) Median coverage across all samples iii) Fraction of samples with a coverage $\geq$ 1x, 5x, 10x, 20x, 30x, 50x and 100x. A similar representation of coverages was established for the variants in the cases to enable a valid comparison with the controls. The read depth of case samples at all variant positions were determined by using the samtools depth command from Samtools v1.9[177]. In this manner, metrics identical to the gnomAD coverages were established for the cases. The coverages from the cases and the controls were annotated to the joint set of variants. Variants with a median coverage of at least 15 in both the cases and the gnomAD dataset were retained for further analysis, i.e., at least half of our cases had a coverage of at least 15x, and at least half our our controls had a coverage of at least 15x for the variant locus to our filters. The complete workflow of this process is shown in Figure 2.9.

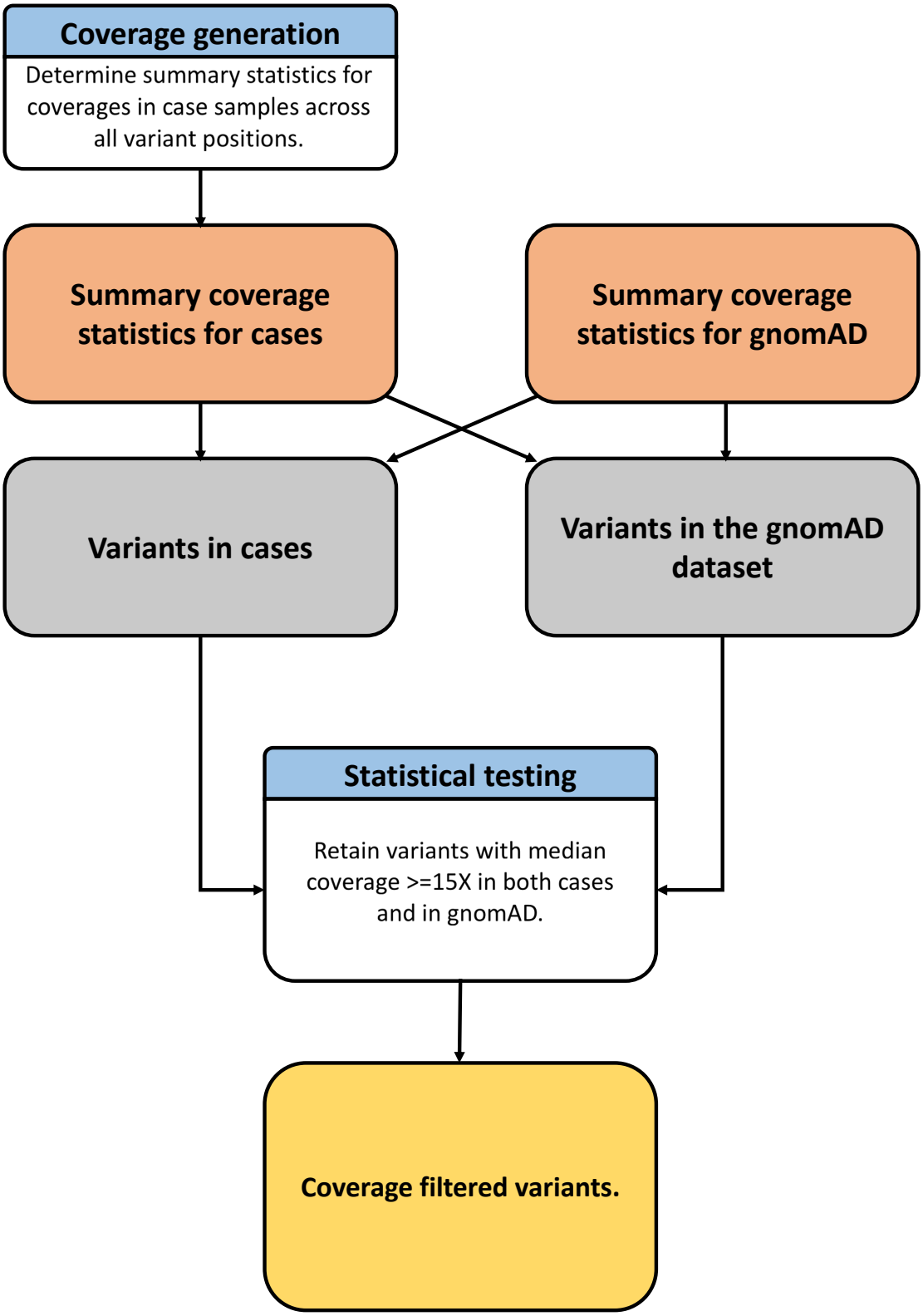


Figure 2.9: Description of coverage generation, annotation and filtration of variants in cases and controls.

2.6.3.3 Annotation and filtering case variants based on alternate allele read depth and frequency

Having removed poorly covered samples and poorly covered individual variants, the next filter applied was to maintain variant quality uniformly for every variant locus and sample. This filter was based on the allelic depths(AD) and the alternate allele read frequency. Specifically, the allelic depth refers to the number of reads supporting each allele at a variant position. The total number of reads is therefore the sum of the allelic depths for every allele at the position. The allelic frequency for every allele is estimated as the allelic depth for the allele/total number of reads at the variant position.

For a heterozygous variant, the allelic read frequency of both the reference and the alternate allele would be expected to be between 0.3-0.6. However, there were some cases where the total number of reads covering a position was greater than the applied threshold (15 reads) but the alternate allele frequency was much lower than 0.3. Such variants are of low quality and would increase the chance of miscalled variant. To account for this, the total number of reads, the allelic depths and the alternate allele read frequencies allele reads were determined at every variant position in the cases. Variants were retained if:

1. The total depth at the locus was at least 15x

2. The allelic depth for the alternate allele was at least 4x.

3. The alternate allele frequency (reads supporting alternate allele / total reads) was at least 0.3.

2.6.3.4 Annotation of cancer gene status

Previously identified familial melanoma genes (including *CDKN2A*, *BAP1* and *POT1*) were also mutated in other cancers, either through somatic or germline mutations. It is therefore expected that any mutation that is involved in the the emergence of melanoma would exist in a gene that is similarly affected in other cancers. A list of genes known to be affected in cancers, the Cancer Gene Census (CGC)[178],was utilized to affix significance to such variants.

The CGC is a regularly updated set of genes with additional data on the types of cancers affected, the type of mutations carried in these cancers. As of the version dated October 22nd 2018, the CGC comprised 719 genes in total. These genes were split into two tiers depending on their impact on cancer development and the evidence available to support the relevance of these genes:

1. Tier 1 - Genes in Tier 1 are considered to be the gold standard for cancer genes and they have compelling, documented evidence to support the relevance of the gene to cancers, including activity that drives cancer and activity that promotes oncogenic transformation. There are 574 genes annotated as Tier 1 genes. Examples include *AMER1*, *ATR*, *BAP1*, *CDKN2A* and *POT1*.

2. Tier 2 - Genes in Tier 2 have limited evidence to strengthen their claim as an important cancer gene but are still considered to play an salient role in cancer development. There are multiple reasons why these genes are not annotated in Tier 1: Lack of sufficient evidence, extremely rare cases, low burden of mutations or genes involved in cancer only through fusion. There are 145 genes annotated to be Tier 2 genes. Examples include *A1CF, CDKN1A*, *FAT3, SKI* and *ZEB1*.

The variants in cases and controls were previously annotated with the affected gene through VEP. These variants were annotated with CGC tier, if present. The tier list of the gene within the CGC was also noted. This was later used to filter the list of genes to focus on cancer genes for a component of the association analysis.

2.6.3.5 Calculation of total number of affected samples in genes

The power afforded by the low number of cases was too small to allow variant by variant association testing. Therefore, variant counts were collapsed for every gene to get gene counts in cases and controls instead. The gnomAD dataset comprised of individual unrelated samples, by contrast the cases consisted of related family members. This meant that the likelihood of two related family members carrying a mutation would be much higher than two unrelated control samples carrying the same mutation. In order to account for this, it was decided that the total number of affected families would be used to count cases, instead of the total number of affected individuals. This was chosen to prevent an overestimation of case counts due to the relatedness of samples. For each variant, every affected family in the cases was counted exactly once. The total number of affected families were estimated for every gene through this process.

For the variants in the gnomAD data set, the provided Genotype Count (GC), defined as the count of individuals for each genotype, was used. The GC was provided for every population subgroup including non-Finnish Europeans (GC NFE) which was used to determine the total number of controls with a mutation. GC NFE counts were summed across all the variants present in a gene to determine the total number of control samples carrying a variant for every gene. This is an approximation which assumes that the same individual does not carry two

different variants within the same gene. This is a reasonable assumption since the variants have been filtered for a frequency of 1/1000 which would make it extremely unlikely for the same person to carry two such variants within the same gene.

2.6.4 Statistical testing to determine ranked list of genes

The processed variants from the cases and controls were set up with the following data structure which was used for further analysis:

1. The name of affected gene, its associated HUGO Gene Nomenclature Committee (HGNC) symbol and its corresponding Ensembl stable id.

2. Total number of families with and without a member carrying a filtered variant in the gene (as described in Sections 2.6.2 and 2.6.3).

3. Total number of non-Finnish European samples in gnomAD with and without a filtered variant in the gene (as described in Sections 2.6.2 and 2.6.3).

4. Presence of the gene within the CGC and tier list status, if present.

A 2x2 contingency table was created for every gene using this data. An example of a contingency table is given in Table 2.13.

Contingency table	Cases	Controls
With variants	4	51
Without variants	131	7458

Table 2.13: A 2x2 contingency table as identified for *POT1*.

These tables were used as the input for a Fisher's Exact test. A Fisher's Exact test is a statistical test that determines if there is an association between two categorical variables. The null hypothesis is that the two variables are independent, which in this case would be the variant status (number of people with a variant in the gene vs number of people without a variant in the gene) compared to the disease status (cases vs controls). Deviations from the null hypothesis would indicate that the presence of variants in the gene are associated with the disease status. A p-value is produced as the output of the Fisher's Exact test. As we expect to find an increase in the proportion of members with variants in the cases compared to the controls, a one-sided Fisher's Exact test is more appropriate. Along with the Fisher's Exact test, an odds ratio(OR) is also be estimated. An OR is a measure of quantifying the level of

association between two properties, i.e., the ratio of the probability of occurrence of the first property in the presence of the second property compared to the probability of occurrence of the first property in the absence of the second property. Assuming that the disease status and the variant burden are our two properties of interest, the OR could be one of three outcomes:

1. OR<1 : The variant burden in the gene is associated with a lower probability of disease occurrence.

2. OR=1 : The variant burden in the gene is independent of the disease occurrence.

3. OR>1 : The variant burden in the gene is associated with a higher probability of disease occurrence.

For genes with a higher burden of mutations in the cases compared to controls, the OR would be >1 which would associate the presence of variants in the gene with a higher probability of disease occurrence. Thus, a one-sided Fisher's Exact test would show if a variant in a gene was associated with the disease, and the odds ratio would indicate the extent of association. The OR and the p-value are computed for the contingency table for all genes.

Thousands of genes present in alternative scaffolds, pseudogenes and non-coding genes were included in the analysis. These would be very unlikely to play a role in cancer development but would still affect the identification of overburdened genes, particularly when correcting for multiple testing. Two different filters were therefore applied on the list of genes based on protein product : one was restricted to all the genes within the CGC while the other was restricted to all known protein-coding genes as defined on Ensembl. A one-sided Fisher's exact test (coded in RStudio) was then applied on the contingency tables for all of the genes on these lists, thus, producing p-values for every gene. These p-values were corrected for false discovery rate using the Benjamini-Hochberg method. The two lists were then sorted based on the corrected p-value, yielding two ranked lists of genes. These results are discussed in Section 3.3.

2.6.5 Limitations of an association analysis

The use of the gnomAD as a control dataset has helped identify protein-coding genes associated with familial melanoma occurrence. However, this approach has some limitations, which are listed here.

1. Information on individual sample genotypes are not available for gnomAD variants; aggregated variant level information is instead presented across all samples. While this is

sufficient for an association analysis, this comes with the caveat that it is not possible to identify a sample having multiple variants within the same gene. This risk is minimized due to a conservative variant frequency threshold of 1/1000, it is however still present.

2. The total number of cases with a variant in a gene is measured by the total number of families with individuals who have at least one variant in the gene. Families with a higher number of sequenced members have a higher probability of having a mutation. By contrast, the controls consist of unrelated samples. Additionally, a variant that is present in all sequenced members of a pedigree with 11 members would normally be considered to be a much stronger candidate than a variant that is present in a pedigree with 2 members. However, such variants cannot be distinguished in this scenario.

3. Another drawback to determining the number of cases as the number of families carrying a variant in the gene is the loss of information regarding the segregation of variants within the pedigree. For example, in a family with 4 members affected (all of whom have been sequenced), a variant that is present in all four affected members is much more likely to be involved in melanoma development compared to a variant that is only present in only one of the members. An addition of linkage analysis is required to account for this, this is discussed in Section 2.7.

4. While this approach determines an increased burden of mutations in genes, it is restricted to non-synonymous coding region mutations. It does not account for the potential role of non-coding region mutations or structural mutations in the development of familial melanoma. The investigation of non-coding and structural mutations in melanoma development and its resulting outcomes are discussed in Sections 2.11,2.12,3.8 and 3.9.

2.7 Linkage analysis

2.7.1 An introduction to linkage analysis

Association studies, particularly GWAS, are normally aimed at identifying a set of common risk alleles with low impact on the disease. As a result, they do not explain the cause of a disease in a large percentage of cases, particularly for the disorders caused by rare, highly penetrant mutations. This issue can be addressed through linkage analysis, which is primarily used to detect and identify variants with large effect size or impact on the disease. Linkage, in the context of genetics, is defined as the propensity for a group of genetic regions present on the same chromosome to be transmitted together from a parent to an offspring during meiosis. This is represented with a logarithm of odds (LOD) score. A LOD score is a statistical estimate of the likelihood of two regions being inherited through linkage as compared to two regions being inherited through chance; the higher the LOD score, the stronger the linkage. Linkage analysis refers to the set of methods that use linkage to help determine the segments of the chromosome which segregate along with the disease phenotype through affected and unaffected individuals belonging to the same family.

Linkage analysis was originally used in the identification of genomic regions with strong linkage signals. Unlike GWAS, it did not require a comprehensive set of markers and could determine co-segregation of the trait with the marker on a larger scale. Genetic markers for disorders including cystic fibrosis [179] and Huntington's disease [180] were originally determined through linkage analysis. The two most prominent familial melanoma driver genes identified to date, *CDKN2A* and *CDK4* were also initially identified through linkage analysis[181]. Other major cancer genes that were discovered through linkage analysis include *BRCA1* and *BRCA2*[182, 183], which are collectively responsible for 90% of hereditary breast cancers globally[184].

However, linkage analysis did originally have several drawbacks, particularly in the context of identifying a genomic region with strong linkages to a disease where no causal gene could be identified. It could not help in the determination of the exact variant responsible for the presence of linkage signals within a given region. If the study focussed on exonic regions, linkage peaks in regulatory regions present in the non-coding part of the genome could not be identified. Innovations and advances in next-generation sequencing, particularly with cheap sequencing costs of whole genome sequencing have helped resolve these issues and enabled a joint association-linkage approach.

2.7.2 The joint association and linkage approach - pVAAST

Variant Annotation, Analysis and Search Tool (VAAST) is a probabilistic tool developed in 2011 which was aimed at identifying disease causing genes from genome sequences[185]. Originally developed for personal genomes, an updated version of the software called pVAAST (pedigree VAAST) was released in 2014 for the analysis of genetic data from high throughput sequencing of related individuals [186]. pVAAST takes the germline sequences of affected individuals, affected/unaffected relatives of these individuals and unaffected controls as its input. The variants in the cases are compared to variants in the controls using a composite likelihood ratio test (CLRT) [185]. In a CLRT, variants in genes are grouped together along with the information on the frequency of the variants in the cases and controls. A composite likelihood score is then estimated for these variants based on the observed frequencies of the variant in the cases and controls.

Alongside the association, pVAAST also computes linkage scores in the form of LOD scores for pedigrees with more than one affected individual. The estimated LOD scores are unlike typical LOD scores as they compute linkage of entire genes with an associated trait as opposed to individual variants. The LOD scores of each gene are cumulative across all variants present in the given gene across all pedigrees. This helps provide a single linkage peak for each gene across all cases compared to individual linkages for each family for each gene. These LOD scores are combined with the variant frequency information, annotated consequences for each variant using a built-in dataset and the association analysis to provide a prioritized list of genes and variants where the genes are ranked based on their association with the phenotype, shown in Figure 2.10.

In order to run pVAAST on the data, the following input files were required:

i) Variant files in VCF format for cases.

ii) Variant files in VCF format for the controls/background.

iii) A list of genes on which scoring is to be performed in a general feature format (.gff3).

iv) The human reference genome in the Fasta format.

v) The framework file for the different parameters used in running pVAASt in a control format (.ctl).

vi) A pedigree file containing the following information for each affected individual - the family id, the sample id for the individual in each family, the id of father of the individual, the id of the mother of the individual, gender (represented as 1 for male and 2

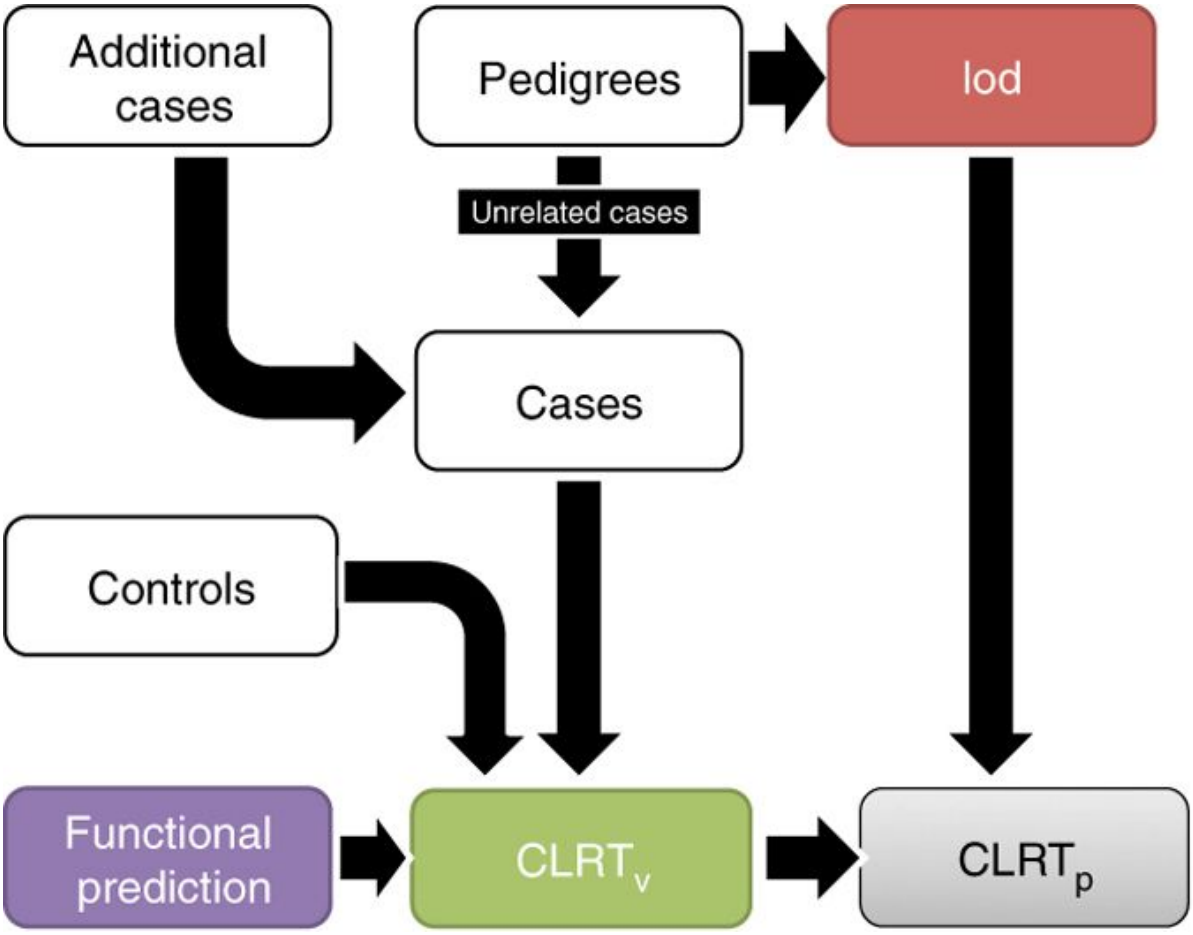


Figure 2.10: A graphical representation of the scoring pVAAST. Reproduced with permission from [186].

for female) and affected status for melanoma (represented as 1 for unaffected and 2 for affected). This file is to be in the pedigree format (.ped).

The following output files are produced by pVAAST during the process of analysing the data:

i) Variant information for each individual in the cases in the Genome Variation Format (.gvf) defined by the Sequence Ontology Group.

ii) Annotated variant files generated for each individual (.vat). This is performed using the inbuilt variant annotator tool and includes information on variant id, the position, the affected gene and the effect of the variant. The format is similar to the .gvf file and the output of this step is usually a .vat.gvf file.

iii) A condenser file containing condensed representation of the variants across (.cdr) with one file for each pedigree, one file for all singleton families combined and one for the background samples.

iv) A Vaast file containing the output of the pVAAST runs including CLRT and LOD scores (.vaast).

In order to maximise computational efficiency, each pVAAST run was performed using variant information from all pedigrees for a single gene. This was repeated for all genes in the cancer gene census list as it was not computationally feasible to run this across the entire genome. This allowed for parallelization of the pVAAST runs as the CLRT scores and the LOD scores for each gene could be directly compared with each other. The generation of the input files, the intermediate files and the output files along with the results are described in Section 2.7.3. The parameters used for each step are given in Supplementary Table 3.

2.7.3 Methods

The filtered list of variants used in the association analysis from Section 2.6.3.3 were used for the joint association and linkage analysis through pVAAST. A bed file containing a list of unique positions from these variants was generated. As pVAAST requires a VCF file as an input; a VCF file with the variants at the filtered positions was generated for the cases. All the files used by pVAAST for annotation and filtration of variants were built on and aligned to GRCh37 reference build. As a result, the VCF file for the variants from the cases was lifted over from GRCh38 to GRCh37 using CrossMap[176]. The new VCF file with the variant positions corresponding to the GRCh37 reference build was sorted and indexed using the

Tabix software. The next step in the procedure involved the generation of .gvf files from the multi-sample VCF file. This was necessary as pVAAST used .gvf files as the primary input for all downstream steps, primarily for variant annotation and condenser file generation. GVF files were generated for every sample in the multi-sample VCF file using the build in vaast converter tool available as part of the pVAAST package. Each GVF file was then sorted in place, meaning that no duplicate files were created in the sorting process.

A PED file for containing the pedigree information for all the families with multiple sequenced members was manually created as a text file. pVAAST has certain inbuilt requirements for the structure of the pedigree : there could be no consanguineous marriages, pedigrees could only have the extended family on one side (either the father or the mother but not both) and only two-generation nuclear families could be analysed in the recessive model. No pedigrees or samples were removed in this process but unaffected extended members of some pedigrees had to be pruned to account for these conditions. Pedigrees with a single sequenced member, or singleton families, were grouped together as unrelated affected individuals as linkage analysis would not have been possible for such individuals. The cases were therefore separated into two groups in this manner.

The sorted GVF files of each individual were then annotated with their impact on the genes that they were present in using another pVAAST program called Variant Annotation Tool. This process required .gvf files, the human reference genome and the list of genes and their positions as the input files. The .gvf files were generated in the previous step of the process while the reference genome file (FASTA format) and the list of genes (.gff3 format) were provided by pVAAST. This produced .vat.gvf files, described in Section 2.7.2, for every affected individual.

In order to run pVAAST across each pedigree, we need to group variants from samples belonging to the same pedigrees together. This is performed through a process called variant selection utilizing the Variant Selection Tool (VST) which is a part of the pVAAST package. It performs set operations on the .vat.gvf files such as intersection, union, complement and difference of variants within a given set of gvfs. The .vat.gvf files for all sequenced members from each pedigree are used as the input for VST. The union of variants across all the samples in the pedigree and the output produced is a condenser file or a .cdr file, described in Section 2.7.2. As the singleton families are considered to be unrelated affected individuals, all singleton families were grouped together and a joint .cdr file was produced for them.

pVAAST provided a set of background genomes as part of its package based on the 1000 genomes dataset comprising of 1,303 sample. This was used as the background population for the first set of pVAAST runs. However, to validate these results, a secondary background

dataset with more samples was required. As pVAAST required genotype information for individual samples, data from ExAC or gnomAD could not be used as the background. A set of 4,070 individuals were exome sequenced as part of the INTERVAL study of which the Wellcome Sanger Institute was a collaborative member[187]. These individuals were not enriched for any familial cancers or any other genetic disorders. Due to the presence of high quality sequence data and individual level genotypes, these samples were chosen as the second background dataset. A VCF file with variants from these samples were obtained from our collaborators within the Sanger Institute. These variants were filtered for artefacts and variant frequency similar to the cases using gnomAD. They were then processed similar to the cases to generate a CDR file.

The penultimate step for running pVAAST involved the generation of a parameter file or a .ctl file. This file contains the location of all the .ped files and .cdr files for each pedigree to be considered. The location of the .cdr for the additional cases or the unrelated affected individuals was also provided in this file. The inheritance model for the phenotype of familial melanoma was defined as a dominant model. Additional parameters that were provided within this file include the genotyping error rate, filtering of gene scores based on CLRT and LOD scores, the mode of scoring the gene each gene based on CLRT and LOD scores and the *de novo* mutation rate. This file remained unchanged for every pVAAST run, thus allowing the scores from each run to be compared.

Once the .ctl file had been produced, pVAAST was run across all the pedigrees. The input for each run included the .ctl file for the cases, the background .cdr file for comparing variant frequencies, the region of the genome within which pVAAST calculated the scores and the feature file containing information on all genes present in the genome. Each run of pVAAST resulted in a .vaast output file containing the CLRT and LOD scores for all variants and genes within the specified region. In order to parallelize the process, each pVAAST run was restricted to a single gene. pVAAST was run for all the genes in the Cancer Gene Census as it was not computationally feasible to run it across all the protein-coding genes in the genome. This was done using both the original 1000 genomes project background file and the INTERVAL exomes background file. The results from these runs are given in Section 3.4.

2.7.4 Limitations

A joint association-linkage approach has helped determine novel genes and variants involved in familial melanoma onset which would not have been possible to discern through a straightforward association analysis. However, such an approach still has its limitations:

i) A single run of pVAAST for 135 families in the region of a single gene requires 95-105 hours of computational time to process. As a result, genome wide runs of pVAAST are currently not feasible if there are multiple pedigrees to be analysed.

ii) Extended pedigrees cannot be analysed using this approach as it currently exists. Large pedigrees have to be pruned significantly to be analysed. Additionally, families with multiple affected individuals but low number of sequenced individuals would yield low LOD scores which would impact the scoring of genes.

iii) The annotation provided by the Variant Annotation Tool for the consequences of the variant is not as accurate as the Variant Effect Predictor. This is because VEP is updated more regularly. There is also currently no suitable method for comparing the results of VAT with VEP directly.

iv) Whilst the INTERVAL exomes provide a matching background and were suitable for this approach, a larger dataset with additional samples would provide a much more stringent comparison of variants and provide more accurate results. Additionally, the method has estimated high scores for large genes with multiple functional domains but little to no significance in cancer development such as the *MUC* and *FAT* family of genes. This indicates that the background dataset is not powerful enough to discern and filter out variants in such genes even after stringent quality filters for the variants.

2.8 The search for variants in known driver genes

Previous studies have helped determine several driver genes involved in the development of familial melanoma, as described in Chapter 1. These genes include *BAP1*, *BRCA2*, *CDK4*, *CDKN2A*, *MITF*, *POT1* and *TERT*. The existence of any variants in these genes would explain the presence of familial melanoma in the pedigrees carrying such variants and thereby make it unlikely that these pedigrees also carried other novel, high-penetrant causative variants. Variants in these driver genes were therefore analysed concurrently with the association analysis described in Chapter 3, to determine if they were causative of disease onset in any of the families. Such variants were then annotated with clinical relevance to disease onset, particularly with respect to their connection with hereditary cancer. This information was obtained from ClinVar[188]. The results from this investigation is presented in Section 3.5. Analysis of potential pathogenic variants in all other genes as defined on ClinVar is discussed later in Section 2.10.

2.9 Variants with high segregation within the cases

Variants which are present in all sequenced members of our pedigree are more likely to be causative of the disease for the pedigree. Due to low power, such variants/genes might not be found carrying a higher variant burden in an association study but might still be responsible for the emergence of melanoma within the pedigree. To discern the presence of such variants, variants from Section 2.6.3 were then filtered based on the proportion of samples carrying the variant in the families with the variant. This was represented through a value called the segregation proportion which was defined as: Segregation proportion (SP) = Total number of individuals carrying the variant/Total number of sequenced individuals in pedigrees where the variant is present

A variant is defined as completely segregating within a family if its SP =1, i.e., every sequenced member in the family carries the variant. While complete segregation is ideal for the determination of novel variants, phenocopies are also known to occur in the context of cancers including melanoma [189]. In order to account for phenocopies, variants with high segregation were determined as follows:

i) Variants were removed if there were no affected pedigrees with multiple sequenced members.

ii) Variants were retained if SP $\geq$ 0.85 , i.e., at least 85% of the samples sequenced in every pedigree containing the variant carried the variant.

iii) Variants were also retained if they were present in a family such that the number of members carrying the variant was at worst one lesser than the total number of sequenced members in the pedigrees. This was to account for cases where the SP would be less than 0.85 but the segregation is still high enough within a single pedigree to warrant further investigation. An example of such a case is the *CDKN2A* missense mutation described in Section 2.8: Three out of four sequenced members carry the mutation; the SP for this variant would therefore be 0.75. However, the variant is still considered interesting as the member without the variant is believed to be a phenocopy.

2.10 Pathogenic variants in ClinVar

A rare-variant association analysis would help in the identification of genes with an increased burden of rare variants in familial melanoma patients. However, there are still a few cases

Clinical significance	Interpretation
Benign/Likely benign	Variants that are not considered to affect disease onset and progression.
Pathogenic/Likely pathogenic	Variants that are considered to affect disease onset and progression.
Uncertain significance	Variants whose impact on the disease are unknown.
Drug response	Variants that disrupt the efficacy of a drug without affecting the disease.
Association	Variants identified in genome-wide association studies.
Risk factor	Variants which contribute to the pathogenicity of a disease without being causative.
Protective	Variants that reduce the pathogenicity of a disease.
Affects	Variants that are not related to disease but are linked to specific disruptive phenotypes.
Conflicting data from Submitters	Variants submitted by a single consortium but with conflicting interpretations of the significance.
Other	Variants that do not fit under any of the above categories such as variants with functional significance but no clinical significance, literature reports with no supporting evidence etc. belong here.

Table 2.14: Classification of clinical significance of variants in ClinVar. The interpretations marked in bold are obtained from the guidelines recommended by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology[190]. Adapted from the online documentation of ClinVar.

where the the development of melanoma within the family is caused due to a single, highly pathogenic variant with a prior role to cancer development. Such mutations might not be identified through a rare variant association analysis and thus, the reason for melanoma development in such pedigrees may be undetected. If there were known phenotypes for the variants observed in our cases, then we may be able to better link the variant with the disease. Data from ClinVar, a repository of variant phenotype relationships[188], was utilized for this purpose. In addition to classifying the clinical significance of variants as being benign or pathogenic, ClinVar also contains meta information on the variant (including information on the protein product and transcript) and the supporting evidence for the classification of the variant. The classification of clinical significance for a variant is shown in Table 2.14 while the different review statuses provided by ClinVar for the variant based on the supporting evidence provided is shown in Table 2.15

Variants from ClinVar were downloaded as a VCF file (version dated September 30th,

Description	Review status
Practice guideline	practice guideline
Reviewed by an expert panel	reviewed by expert panel
Two or more submitters with assertion criteria and evidence provided the same interpretation.	criteria provided, multiple submitters, no conflicts
Multiple submitters provided assertion criteria and evidence (or a public contact) but there are conflicting interpretations. The independent values are enumerated for clinical significance.	criteria provided, conflicting interpretations
One submitter provided an interpretation with assertion criteria and evidence.	criteria provided, single submitter
The allele was not interpreted directly in any submission; it was submitted to ClinVar only as a component of a haplotype or a genotype.	no assertion for the individual variant
The allele was included in a submission with an interpretation but without assertion criteria and evidence.	no assertion criteria provided
The allele was included in a submission that did not provide an interpretation.	no assertion provided

Table 2.15: Review status classification of supporting evidence for variants in ClinVar in descending order of quality. Adapted from the online documentation of ClinVar.

2018). Variants in our cases at locations common with the ClinVar VCF were identified. Sample and family information were annotated to our case variants; the information on the variants from the cases was then merged with the information from ClinVar into a single file, one line per variant. Variants with more than 8 affected families were removed as artefacts, as described in Section 2.6.3. To restrict the analysis to interesting variants, the reference and alternate alleles from ClinVar were compared to the reference and alternate alleles from the cases. Only variants with matching alleles were retained. Finally, the variants were restricted to having one of the following clinical phenotypes from Table 2.14: pathogenic/likely to be pathogenic, risk factor or protective. These were chosen as they were most likely to be disruptive to the protein product and to lead to disease onset. The results from this analysis are discussed in Section 3.7.

2.11 Non-coding variants

2.11.1 Background

2.11.1.1 Introduction

Previous analyses of cancer genomes have been restricted to the coding region due to an interest in the determination of SNPs that disrupt protein function and the restrictive cost of whole genome sequencing[191]. Large scale exome sequencing studies have helped identify several key mechanisms and genes involved in the development of familial melanoma[143][145]. The advent of next-generation sequencing technologies has drastically reduced sequencing costs, from around $14 million in 2006 to $1,500 in 2016, which has enabled cheap sequencing of whole genomes[192]. This has provided new avenues for the investigation of the importance of sequence variation data in disease onset. A facet of analysis that remains unexplored in this context are variations in the non-coding region of the genome. It is estimated that 1.5% of the human genome encodes a gene, which leaves ~98% of the genome as non-coding DNA[193]. While the role of these regions in genetic regulation was unknown for a long time, it is becoming increasingly evident that the non-coding genome encompasses key regulatory elements which play an important role in the transcription and translation of proteins.

Non-coding elements can broadly be classified into cis-regulatory elements and trans-regulatory elements. Cis-regulatory elements (CRE) are generally found in the vicinity of the gene that they modulate and control gene regulation through intramolecular interactions, i.e, the components are active in the same gene[194]. Such elements include promoters, silencers and nearby regulatory elements. Modification of sites in cis-regulatory elements therefore directly impact the activation of a gene. CREs can also be distal, i.e., hundreds of kilobases away from the gene of interest. Enhancers and insulators are examples of distal cis-regulatory elements which activate and repress transcription of the gene respectively[195]. Such elements interact with the promoter/gene in the three-dimensional structure of the genome through chromatin looping regulated by proteins including CTCF and cohesion, as shown in Figure 2.11[196]. Trans-regulatory elements do not directly interact with the gene or the promoter, rely on intermolecular interactions with cis-regulatory elements, and usually encode for transcription factors[194].

2.11.1.2 Transcription factors and sequence logos

Transcription factors (TF) are a family of proteins that bind to the DNA, usually the promoter of a gene, in a sequence specific manner and either activate or repress the transcription of a

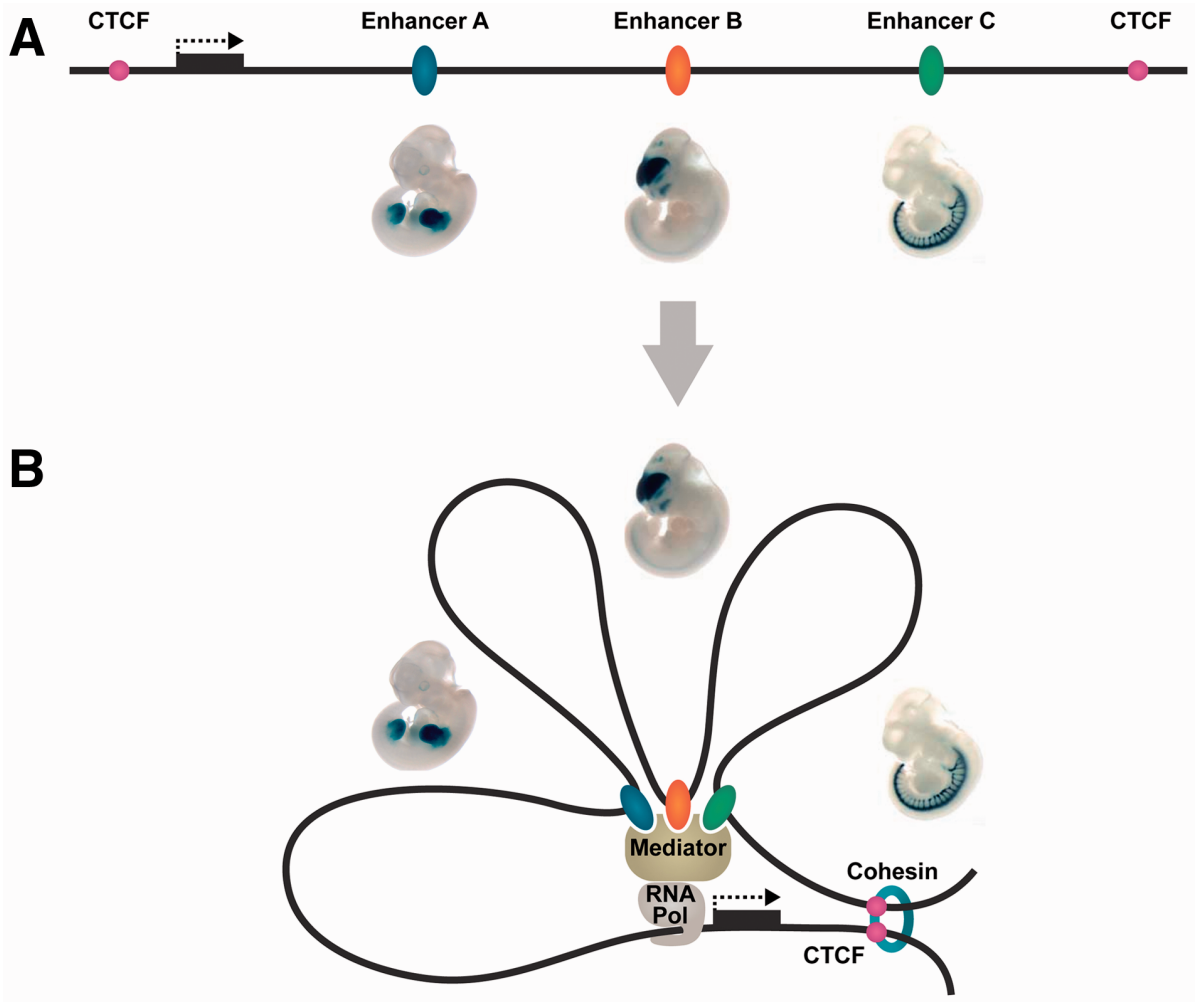


Figure 2.11: Depiction of chromatin looping to show enhancer-promotion interaction moderated through mediators, CTCFs and cohesion. Figure A shows the linear arrangement of proteins and enhancers on the chromatin while Figure B shows the interaction of the different enhancers with the RNA polymerase at the promoter through a mediator protein by chromatin looping. Figure reproduced with permission from reference [196].

gene[197]. The region of the DNA that each transcription factor binds to is defined as a transcription factor binding site (TFBS). A single transcription factor can regulate the transcription of several genes and binds to multiple locations across the genome. Non-coding variants can disrupt normal regulation of transcription by either creating or distorting the interaction between transcription factors and the DNA, usually in the promoter. The conserved sequence that represents the bases across all transcription factor binding sites for a given transcription factor are known as transcription factor binding motifs (TFBM). These motifs are visually represented through sequence logos[198]. An example of a sequence logo is shown in Figure 2.12.

Figure 2.12: An example of a sequence logo for a transcription factor binding motif. Obtained from the JASPAR[199] database.

The x-axis of the sequence logo represents the different bases across the TFBM while the y-axis represents the combined frequency of all nucleotides through bits. A bit measures the total amount of information present at every position of the sequence and is associated with the answer to a binary question[198]. In the example shown in Figure 2.12, the base at the first position of the motif is always a cytosine. This means that two binary questions need to be answered: Is it a purine? If not, is it a cytosine or a thymine? The answers to these questions are represented as bits of information. If there are multiple possible bases at a given position such as position 7 in Figure 2.12, the amount of bits available at the position changes accordingly. The height of every nucleotide at each position constitutes the relative frequency of that nucleotide at that position.

As shown in the figure, not all the positions across a transcription factor binding site are equally conserved, some positions are more conserved than others. If a base is highly conserved across all transcription factor binding sites for a transcription factor, it suggests that the base is essential for the transcription factor to interact with and bind to the DNA. Variants in positions that are highly conserved, such as position 1 in Figure 2.12, would impair the function of a transcription factor more than a variant in a position that is less conserved such as position 7.

2.11.1.3 The role of non-coding variants that modify the function of transcription factors in cancer

Variants in transcription factor binding sites that play a role in carcinogenesis have previously been observed both in the context of familial melanoma and in other cancers. Recurrent variants have been observed in the promoter of *Telomerase Reverse Transcriptase (TERT)* originally in sporadic and familial melanoma[131] and eventually in other cancers[140]. A germline variant observed in familial melanoma was responsible for the creation of a binding motif for the ETS family of transcription factors. This led to the recruitment of TFs including T-cell factors (TCFs) to the promoter region of *TERT* which led to increased expression. The mutation observed in the familial melanoma pedigree was 57 bases upstream of the transcription start site while the three mutations in the sporadic melanoma cases were observed at 124, 138 and 146 base pairs upstream of the transcription start site respectively. A follow-up study determined that such germline TERT promoter mutations were quite uncommon in familial melanoma[132]. Disruption and ablation of transcription factors have also been observed in melanoma. Recurrent promoter mutations in *SDHD* were found to disrupt the TFBS for two ETS transcription factors, GABPA and GABPB1, considered to be key regulators of melanoma driver genes including *TERT*[200].

Murine double minute 2 (MDM2) is a protein encoded by the gene *MDM2*. Promoter variants of *MDM2* result in increased binding affinity to the Sp1 transcription factor which results in the increased expression of *MDM2*[201]. Increased expression of *MDM2* represses the activity of the p53 pathway, accelerating cancer development. Variants in the promoter of *MDM2* have been associated with increased tumour formation in several types of p53-related cancers including Li-Fraumeni syndrome[202] and breast cancer[203].

While the *TERT* promoter mutations are in cis-regulatory elements that are in the vicinity of the gene, variants in distal regulatory elements have also been identified as playing a role in cancer onset. A binding motif for the myeloblastosis family of transcription factors (*MYB*) was created through somatic variants in enhancers upstream of an oncogene called T cell acute

lymphocytic leukaemia 1 (*TAL1*). This leads to the overexpression of *TAL1* which results in T-cell acute lymphoblastic leukaemia[204].

2.11.2 Methods

In order to focus on the prospective importance of TFBM disruption in familial melanoma development, the methods and results in the following sections are restricted to the variants within the whole genome sequenced individuals. Only variants that were present within known TFBMs in *Homo sapiens* were utilised for this analysis. For this purpose, the start and end sites of TFBMs along with the JASPAR binding matrices for all known transcription factors in *Homo sapiens* was obtained from the Ensembl Regulation Database v91. Ensembl includes information on transcription factors present in alternative chromosome haplotypes in addition to the normal human chromosomes. To restrict our analysis to relevant variants, only motifs in chromosomes 1-22, X and Y were considered. A bed file comprising the chromosome, the start position of the motif and the end position of the motif was generated from the list of motifs ; the bed file was then sorted on the chromosome and the position. Variants in the whole genome sequenced cases that were located within these TFBM regions were identified. Information on the chromosome, position, reference allele, alternate allele and the consequence of these variants were extracted from the VCF file for the cases and stored independently. After the removal of duplicate variants, this file was sorted based on the chromosome and nucleotide position of the variant. Multi-allelic variants were split into one variant per alternate allele. Information regarding the number of individuals, families and segregation of variant within all sequenced members of the families at each position was annotated to this file. Variants that were present in more than 5 families were removed as they were considered to be sequencing artefacts.

The whole genome non-Finnish European samples from gnomAD were chosen as the control set for this analysis. This comprised summary genotype information for every variant from 7,509 whole genome sequences. Variants from the controls which were present in the binding motifs as determined from Ensembl were identified. Information on the chromosome, position, reference allele, alternate allele and the consequence of these variants were extracted from the control VCF file. This file was sorted based on the chromosome and position of the variants after duplicate variants were removed. Multi-allelic variants were split into one variant per alternate allele. As gnomaD only provided information on summary statistics for samples and not individual genotypes, this resulted in additional duplicate variants as this reported one variant per affected transcript of gene. Every variant at a given position had

the same number of affected samples regardless of affected transcript, such duplicate variants were therefore removed. Variants from the cases and controls were then jointly processed.

The mean and median coverage of every position within the cases and controls were determined. Positions with median coverage below 10 reads were removed from both sets of variants. In order to focus on rare variants with a potential impact on the function of TFs, variants in the cases were annotated with the allele frequencies of the variants from gnomAD. If the case variant was present in gnomAD, the frequency of mutation from the gnomAD dataset was annotated to the variant. If the case variant was not present in gnomAD, it was annotated as being absent. Variants with gnomAD frequency > 0.05 were removed from the cases and controls. Variants that were not present in gnomAD were retained in the cases. Several steps that were performed for the association analysis of variants in the coding regions as discussed in Chapter **??**, including the annotation of cancer gene status (Section 2.6.3.4), calculation of total number of affected samples for cases and controls (Section 2.6.3.5) and generation of 2x2 contingency tables based on sample counts for every gene (Section 2.6.4) were replicated for this analysis. P-values were generated for every gene from these tables using the Fisher's exact test, also as described in Section 2.6.4. These values were then corrected for false discovery rate using the Benjamini-Hochberg method. A workflow for these steps is shown in Figure 2.13.

The complete results from this methodology is discussed in Section 3.8.

2.12 Structural variants

2.12.1 Background

2.12.1.1 Introduction

There are two major types of modifications in the human genome that are known to play a role in cancer development. They can classified based on the size of the modification into small variants and large variants.

1. Small variants consist of single base alterations (single nucleotide polymorphisms) and small insertions or deletions of base pairs (indels). Previously, indels were considered to be any variation that were between 1000 bp [205] to 10,000 bp[206] in length but recent studies have identified indels as variants that are less than 50 bp long[207].

2. Large variants comprise of structural variants in the chromosome which change the structure of the affected segment of the genome. They are between fifty to millions of

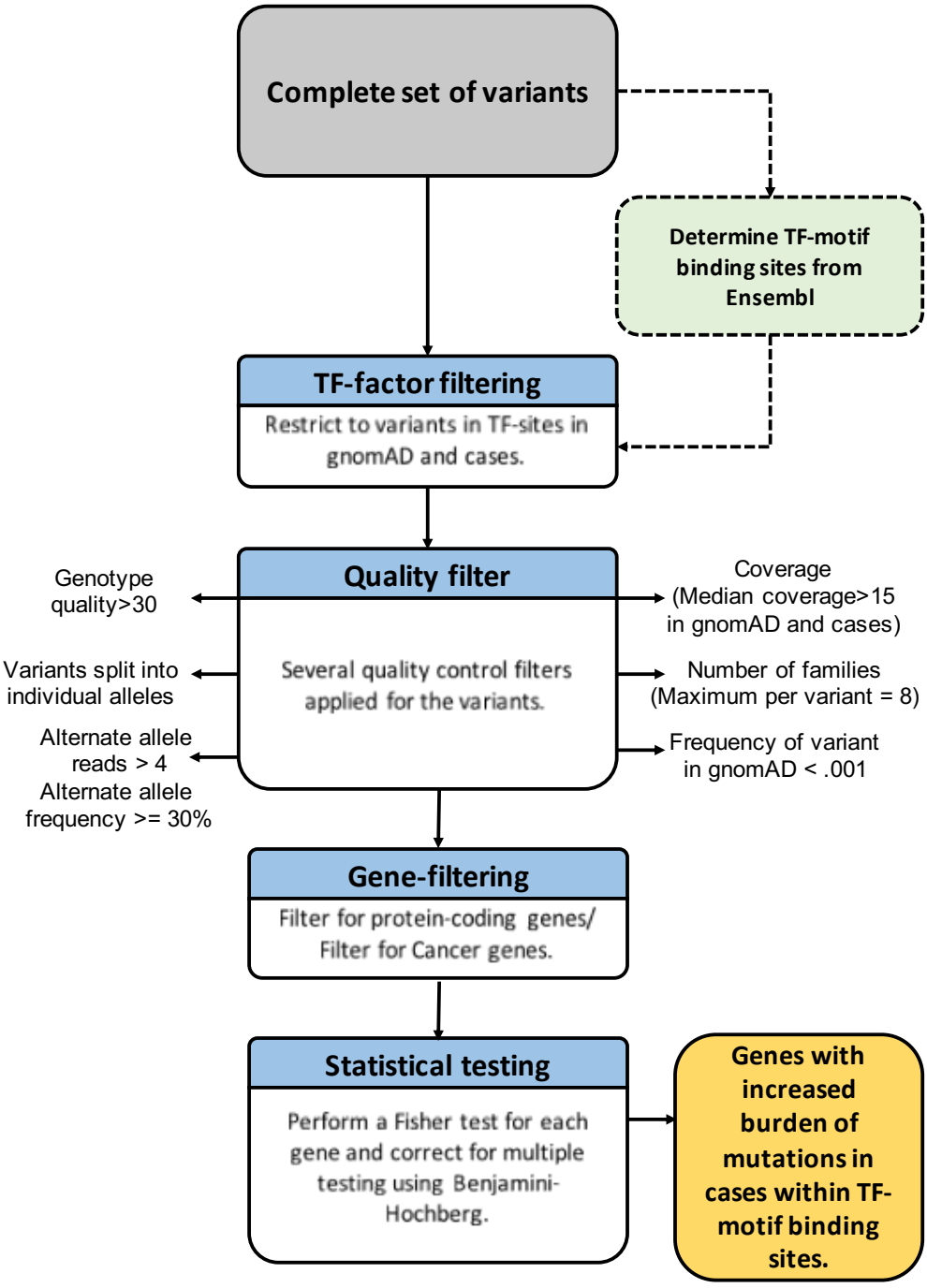


Figure 2.13: Overview of steps involved in the identification of genes with an increase burden of variants within transcription factor binding motifs in cases through an association analysis.

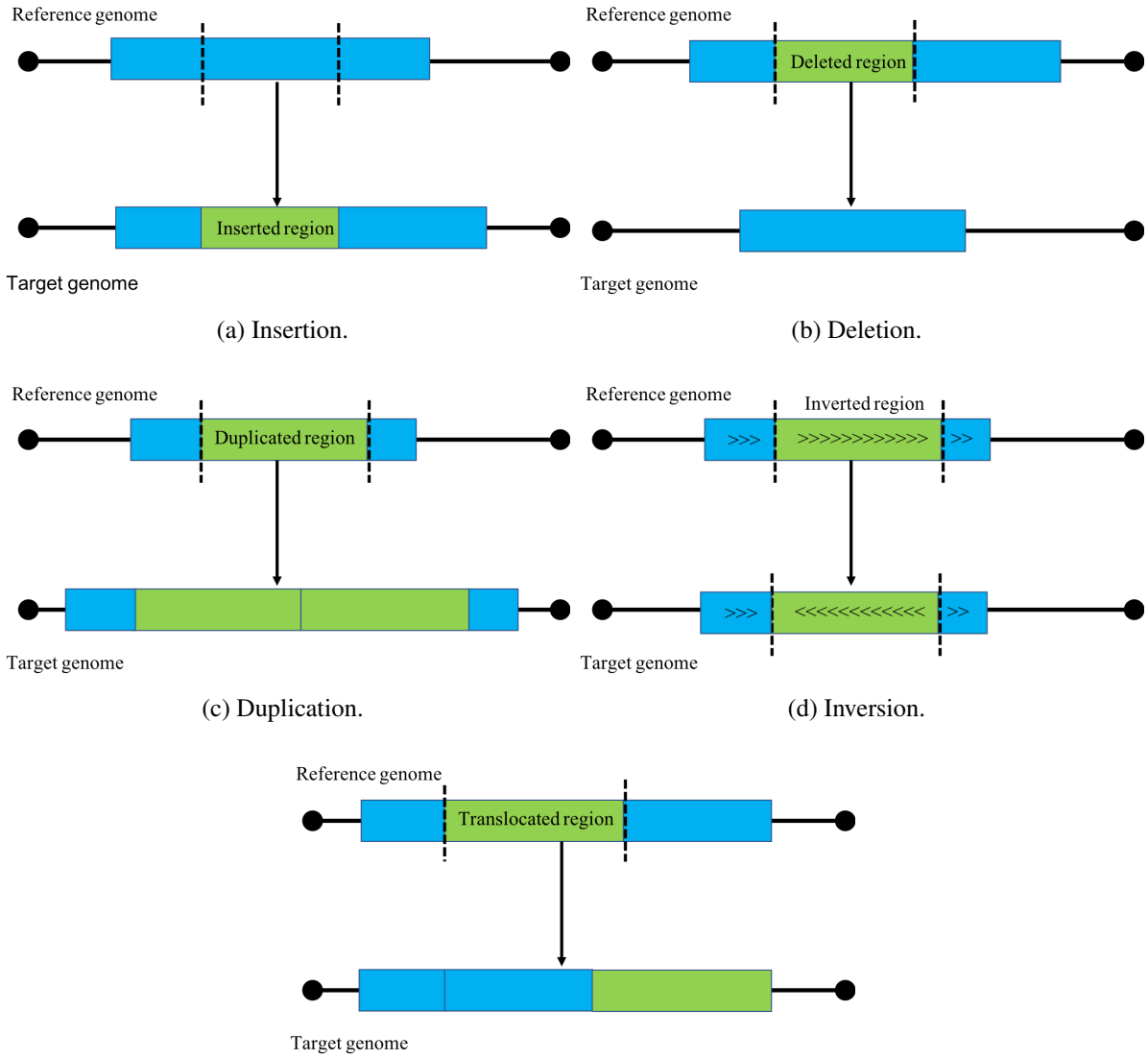
base pairs long. Some structural variant events such as chromothripsis can cluster and disrupt entire chromosomes[208].

There are several types of structural variants that occur within the human genome. The most prevalent types of structural variants are shown in Figure 2.14 and described below:

- Insertion : An insertion is a structural variant caused due to the addition of a segment of DNA between two neighbouring bases in the genome. This is shown in Figure 2.14a.
- Deletion : A deletion is a structural variant caused due to the removal of a segment of DNA between two neighbouring bases in the genome. This is shown in Figure 2.14b.
- Duplication : A duplication is a structural variant where a segment of DNA is replicated and is then inserted alongside the original segment. This is shown in Figure 2.14c.
- Inversion : An inversion is a structural variant that is caused due to the reversal of a segment of the DNA within the genome. This is shown in Figure 2.14d.
- Translocation : A translocation is a structural event where a segment of the DNA is moved to another region of the genome. This is shown in Figure 2.14e. Although Figure 2.14e shows the translocated region to be close to the original position, translocations may occur within or between chromosomes.

2.12.1.2 Structural variants in genetic disorders

Structural variants have been known to play a role in the development of several diseases. The most prominent example of a genetic disorder caused by a structural change is Huntington's disease, encoded by the Huntintgtin gene (*HTT*). A section of the gene comprises trinucleotide repeats of CAG. A normal copy of the gene contains up to 26 copies of the CAG repeats. However, an increase in the number of copies beyond 26 gradually increases the risk and penetrance of the disease, with increased risk of transferring the disease to offsprings as well[209]. A copy of *HTT* with 36-39 CAG repeats have reduced penetrance of disease while copies with 40 or more CAG repeats are considered to be almost completely penetrant[210]. Another notable example of structural disorders is the development of Down syndrome which is a genetic disease caused due to the presence of an additional copy of chromosome 21[211]. Disorders like Down syndrome which are related to extra chromosome copies are defined as trisomy disorders. Other trisomy disorders include Edwards syndrome[212] and Patau syndrome[213].



(a) Insertion.

(b) Deletion.

(c) Duplication.

(d) Inversion.

(e) Translocation.

Figure 2.14: The different types of common structural variants within the human genome. Each sub-figure has the reference genome without the structural variant shown on top. The target genome with the structural variant is shown at the bottom. The black dotted lines on the reference genome indicate the region where breakpoints would be present and should be predicted.

The earliest discovery of structural variants in the context of cancer development was during in the study of cancer cells by Theodor Boveri who associated the growth of cancer cells with observations of segmented chromosomes[214]. Multiple experiments involving fluorescence in situ hybridization (FISH) experiments led to the identification of several gene fusions and amplifications in cancer such as the *BCR-ABL* fusion in chronic myeloid leukemia and *HER2* over-expression in breast carcinomas[215]. The growth and development of microarray technologies furthered the understanding of the role of structural variants in cancer. Comparative genomic hybridization, originally used to identify copy number alterations through FISH, helped in the analysis of amplifications and deletions of genetic regions in solid tumours[216]. SNP genotyping arrays have also been used to determine copy number variations in cancers in several studies such as the Cancer Genome Atlas (https://www.cancer.gov/tcga). In spite of all such improvements, a precise estimation of structural breakpoints was not feasible till the advent of next-generation sequencing technologies. Such technologies also enabled the detection of copy neutral variations such as inversions and translocations.

2.12.1.3 Determination of structural variants in next-generation sequencing data

Structural changes are primarily identified from next-generation sequencing data through errors in the alignment of the target genome reads to the reference genome. Depending upon the type of error, these reads are classified into two types, shown in Figure 2.15:

- Discordant read-pairs. Since the paired-end NGS technique sequences both ends of each DNA fragment with library insert sizes specific to a given library preparation method and size selection procedure, the two paired reads will be generated at an approximately known distance in the sample genome. A signature of a discordant read-pair is formed when the mapping span and/or orientation of the read-pairs crossing the breakpoint are inconsistent with the reference genome. Specifically, both reads of the pair can be mapped to the reference genome, but they may map to different chromosomes or different orientations, or their coordinates may not agree with the insert size.

- Splitting reads. A sequence read that spans a breakpoint in a structural variant is called a splitting read. If both splitting parts of a read can be mapped and its mate is uniquely mapped to the reference genome, the splitting read is further masked as a soft-clipped read by some mapping algorithms such as Burrow-Wheeler Alignment(BWA) tool. Otherwise, it is categorized as an un-mapped read. The splitting reads used by current SV detection tools are all soft-clipped reads, and the term “splitting reads” is generally referred as soft-clipped reads.

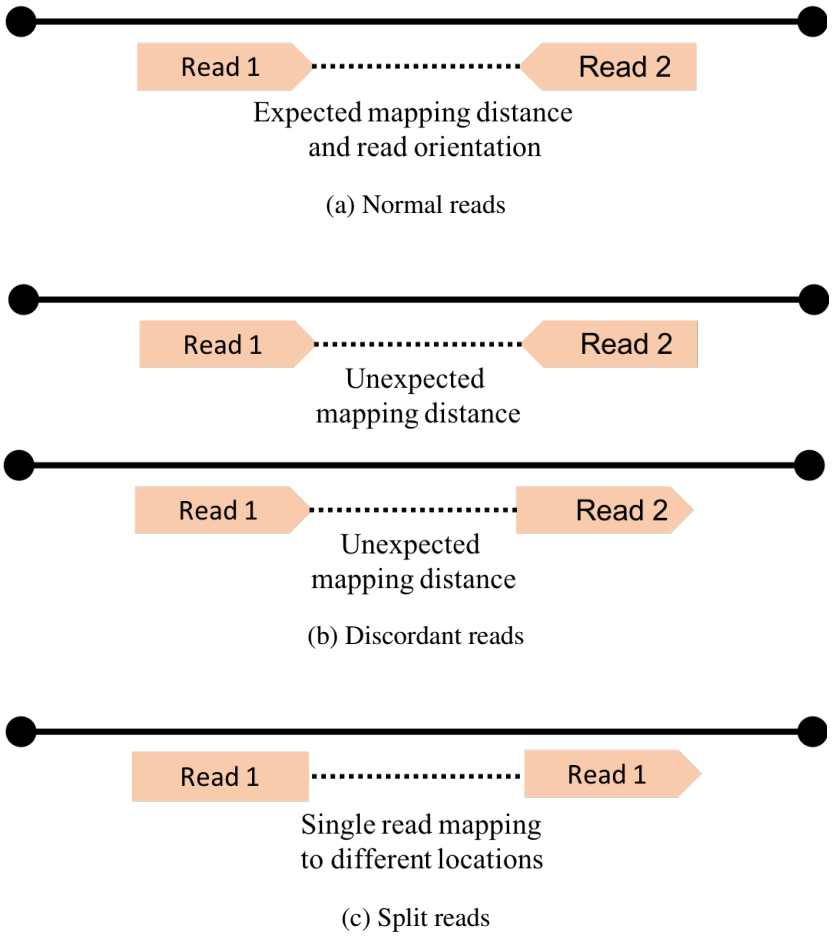


Figure 2.15: Different types of read errors used in the identification of structural variants.

The different softwares that are available for the identification of structural variants from next generation sequencing data are distinguished by two factors. The first factor is the type of read error used in the identification of structural variants, i.e., discordant read pairs or split reads. The second factor is the type of sequence that can be analyzed i.e., exome or whole genome sequences. Softwares such as BreakDancer[217], HYDRA[218] and SVDetect[219] use discordant reads for the identification of structural variants while other softwares like CREST[220] use split reads. Recent approaches have also tried to combine the approaches and to use information from both the type of errors in order to determine breakpoints. Examples of such softwares include DELLY[221] and LUMPY[222]. This section of the thesis involves the identification and analysis of structural variants that potentially play a role in cancer development within the sequenced cases.

Parameter	Description
-mw 4	Minimum weight across all samples for a call (number of reads)
-tt 0	Trim threshold
-pe id:sample_id, bam_file:lsample_id.bam	Sample id and its corresponding bam file for a paired-end reads file
histo_file:sample_id.hist	Statistics of insert size across the bam file
mean:461.115724915	Mean insert size
stdev:83.3662594786	Standard deviation of insert size
read_length:151	Read length
min_non_overlap:151	Minimum number of bases that do not overlap, usually equals read size.
min_mapping_threshold:20	Minimum mapping threshold for reads
-sr id:sample_id, bam_file:sample_id_2.bam	Sample id and its corresponding bam file for a split reads file

Table 2.16: Parameters used for the identification of structural variants using LUMPY.

2.12.2 Methods

Cases from the pilot whole genome dataset were considered for the analysis of structural variants. As structural variants consist of large scale changes that affect both the coding and the non-coding region, exomes were considered to be not as informative as whole genome samples. During the time of the analysis, the secondary Leiden whole genome dataset had not yet been sequenced. A similar structural variant analysis approach was eventually performed on these samples by our collaborators in Leiden.

The standard version of Lumpy (v0.2.13)[222] was used to generate structural variants across the 123 whole genome samples from the pilot whole genome dataset. The parameters chosen for this command are given in Table 2.16.

Individual VCF files for each sample were produced as the output from Lumpy, containing information on the structural variants detected in these samples. These VCF files were sorted, zipped and indexed based on the location of the variants. Information regarding the location, length and type of every structural variant were extracted from each sample and saved as individual text files. The number of supporting reads for each variant were also determined and annotated to these files. Variants from each sample were merged into a single joint VCF file. Ensembl contains information on individual haplotypes of chromosomes in addition to entire chromosomes. Variants that were identified as being present in such haplotypes which were

not in chromosomes 1-22, X, Y were removed. Generic breakpoints which cannot be classified as inversions, insertions or deletions are marked as "BND" variants by Lumpy. Such variants were also removed as the exact structural variant could not be established. This filtered set of variants was then sorted on the chromosome and the variant position. Some variants started and ended within 250 bp of each other in different samples but largely overlapped with each other. The longest such variants were identified within the overlapping regions and chosen as the representative variant for these regions. The median number of supporting reads were also identified across all samples carrying the variant. A total of 86,697 structural variants were identified in this manner. Variants that were greater than 1 million base pairs were removed as they were considered to be structural aberrations created through sequencing artefacts. This step resulted in the removal of 59,627 variants, resulting in 27,070 variants. The number of samples and families carrying each structural variant were established, with the fraction of samples in each family carrying the variant also being established. Variants that were present in more than five families were removed as they were considered to be artefacts. Only variants that were present in over 75% of the samples in the families with the variant were retained which comprised 773 variants. Depending on the relative location of the structural variant to the closest gene, each variant was annotated with one of six possible orientations, shown in Figure 2.16. They are:

1. Gene contained within the variant (Figure 2.16a).
2. Variant contained within the gene (Figure 2.16b).
3. Variant overlaps with the beginning of the gene (Figure 2.16c)
4. Variant overlaps with the end of the gene (Figure 2.16d)
5. 5' variant (Figure 2.16e)
6. 3' variant (Figure 2.16f)

The strand and orientation of the structural variant and each gene is taken into account for this purpose. In addition to the location of the variant, the distance to the gene was also determined. To focus on structural variants that potentially disrupt cancer pathways, every affected gene was finally annotated with information from the CGC list if they were present in it, including their tier list. Variants that were present downstream of the gene, annotated as being a 3' variant, were removed as the probability of such variants affecting the expression of the gene was lower. This resulted in a final set of 307 variants. Variants in genes associated

with melanoma that were identified through this approach are discussed in Section 3.9. The complete list of structural variants is shown in Supplementary Table 10.

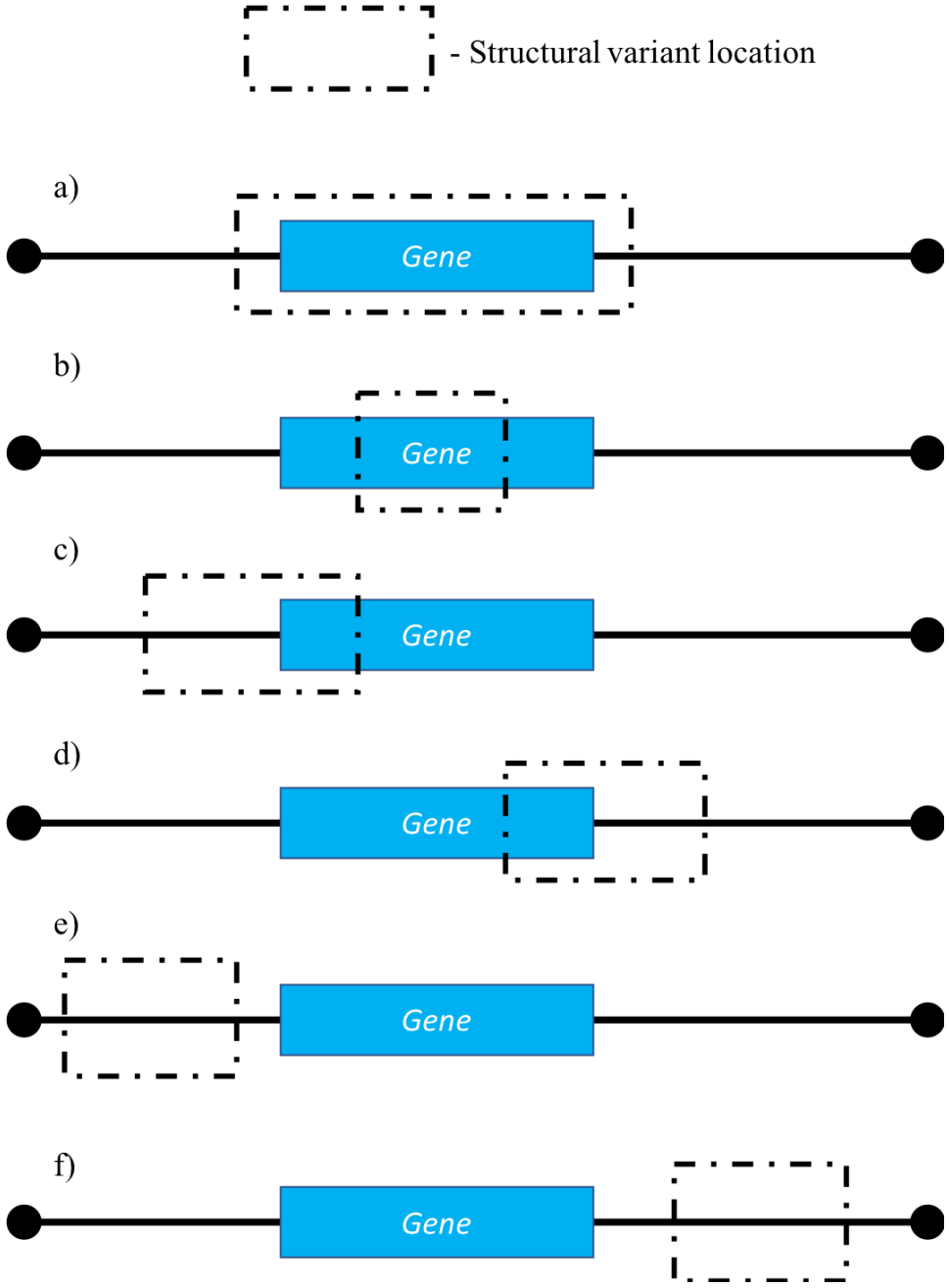


Figure 2.16: The relative locations of structural variant to gene of interest are shown here. a) The structural variant completely encompasses the gene of interest. b) The structural variant is completely contained within the gene of interest. c) The structural variant overlaps with the 5' end of the gene of interest. d) The structural variant overlaps with the 3' end of the gene of interest. e) The structural variant does not overlap with the gene and is present upstream of the gene of interest. f) The structural variant does not overlap with the gene and is present downstream of the gene of interest.