The effect of patient and tumour genetics on survival from melanoma

Ernest Mangantig

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Medicine

January, 2017

Acknowledgements

I would like to acknowledge the Malaysian Ministry of Higher Education and Universiti Sains Malaysia for funding my PhD study which has made this thesis possible.

I would like to express my gratitude to everyone who have helped me directly or indirectly to complete this thesis. I would especially like to thank:

Professor Jenny Barrett, my main supervisor, for giving me the opportunity to join the Melanoma research group in Leeds as a PhD student, for introducing me to the fascinating field of statistical genetics and providing me training, for always being available and taking such good care of me. I am deeply grateful for your excellence guidance, continuous support and encouragement.

Professor Tim Bishop, my co-supervisor and the Head of Section of Epidemiology and Biostatistics, for creating a pleasant learning atmosphere in our section, and for your helpful comments and suggestions to improve my thesis.

Dr Mark Iles, my co-supervisor, for sharing with me your imputed SNP data and brilliant R codes, and also for your valuable comments and suggestions to improve my thesis.

Dr Jeremie Nsengimana for performing the pre-processing of the gene expression data and explaining the details to me.

Dr John Davies for performing the pre-processing of the genotype data, for sharing your inspiring PhD journey stories and helpful tricks in R programming.

Dr Jon Laye for taking your time to teach me the histopathology of melanoma.

May Chan for your continuous help with the data and showing me beautiful places around Yorkshire.

Katie Cairns and Sally O'Shea for your kind help in printing my thesis and made it possible for me to submit it.

Sally O'Shea, Faye Elliott, Jenna Strathdee, and Hara for all your motivations and the merry lunch time we have spent together.

My Malaysian friends in Leeds, Fairuz, Fatimah, Syikin, Jun, Mae, and Stephanie for taking care of me like a family in Leeds and cooking all those delicious foods when I was homesick. I wish you all the best in life.

My family, especially my sisters Elvina, Nonny, Nora, Diana, and Levia for their love and never-ending support.

Finally, my husband Tan Yew Chin, for your love, incredible understanding and support for these past few years of my PhD journey, and for taking care of me when I was unwell.

Abstract

The American Joint Committee on Cancer (AJCC) staging system based on tumour histopathological characteristics (Breslow thickness, ulceration and mitotic rate) is used in clinical practice to assess melanoma patients' prognosis. Although a reasonable predictor of deaths from melanoma (area under the curve 0.68), AJCC staging does not always provide an accurate assessment of individual risk. Recent studies have shown that gene expression levels within the tumour are associated with survival from melanoma, and there is also preliminary evidence that a patient's genotype may influence survival. Combining clinical data with gene expression data may improve prediction, but so far no study has analysed the combined effects of the three different types of factor on melanoma-specific survival (MSS).

This study used patient and tumour characteristics, whole-genome gene expression levels, and genome-wide single nucleotide polymorphism (SNP) data to identify predictors of MSS in a training set of ~2000 patients from the Leeds Melanoma Cohort (LMC). The selected clinical and genomic predictors were combined to build melanoma prognostic models in a test set of 190 patients from LMC, using several approaches. In addition, heritability of survival from melanoma and of Breslow thickness (the most important predictor of MSS) was estimated using genome-wide SNP data.

In the training set, five established clinical predictors (age, sex, tumour site, Breslow thickness and presence of ulceration) were associated with MSS; in addition penalized Cox regression identified 16 gene expression levels and 13 SNPs predictive of MSS. In the test set, the selected genomic predictors did not substantially improve on the predictive performance of the clinical factors. The 16 gene expression levels were also predictive, but were highly correlated with the clinical predictors, especially Breslow thickness, suggesting gene expression influences MSS through clinical predictors. The heritability analyses in this study provided some evidence that germline SNPs influence Breslow thickness.

Table of Contents

List of Tables

List of Figures

Abbreviations

AJCC	American Joint Committee on Cancer
ALM	Acral lentiginous melanoma
AUC	Area under the receiver operating characteristic curve
cDNAs	Complementary DNAs
Chr	Chromosome
Cvpl	Cross-validated partial log-likelihood
DNA	Deoxyribonucleic acid
eQTL	expression Quantitative Trait Loci
FDR	False discovery date
GCTA	Genome-wide Complex Trait Analysis
GRM	Genetic relationship matrix
GWAS	Genome-wide association studies
h^2	Heritability
HR	Hazard ratio
HWE	Hardy-Weinberg equilibrium
LCLs	Lymphoblastoid cell lines
LD	Linkage disequilibrium
LDH	Lactate dehydrogenase level
LMC	Leeds Melanoma Cohort
LMM	Lentigo maligna melanoma
LRT	Likelihood ratio test
MAF	Minor allele frequency
mm	Millimetre
mRNA	Messenger RNA
MSS	Melanoma-specific survival
n	Number of samples
NCBI	National Centre for Biotechnological Information
NM	nodular melanoma
ONS	Office for National Statistics
PCA	Principal components analysis
PCR	Principal components regression

PLS	Partial least squares
QC	Quality control
RNA	Ribonucleic acid
SD	Standard deviation
SE	Standard error
SNP	Single nucleotide polymorphism
SSM	Superficial spreading melanoma
TILs	Tumour-infiltrating lymphocytes
TNM	Tumour-node-metastasis
V_e	Residual variance
V_G	Genetic variance
V_P	Phenotype variance

Chapter 1 Introduction

1.1 Background of the research

1.1.1 Melanoma

Melanoma is a type of skin cancer arising from melanocytes, the pigment-producing cells. A diagnosis suggestive of melanoma can be motivated by naked eye inspection based on ABCDE rules but clinical diagnosis requires histopathological confirmation. The ABCDE mnemonic is A for asymmetrical nevi, B for irregular border, C for multiple colours, D for large (>5mm) diameter and E for evolving size, shape and colour. When lesions are removed, histopathological investigation confirms both the diagnosis of melanoma and the histological subtype of this cancer. There are four major subtypes of melanoma with different appearance and histological characteristics: superficial spreading melanoma (SMM), nodular melanoma (NM), lentigo maligna melanoma (LMM) and acral lentiginous melanoma (ALM) (Burns *et al*., 2004).

SSM is the commonest type of melanoma which spreads horizontally and grows in diameter. The commonest sites of SSM are female leg and male back. NM is the second most common subtype and has a nodular appearance that grows vertically and grows in depth. It is a fast growing tumour compared to other subtypes and the common site is trunk. LMM appear particularly on chronically sun-exposed sites such as the face, mostly on the upper cheek, temple or forehead. It has a flat, brown or black and irregular pigmentation appearance and has a long period of horizontal growth phase over months or years. In time, a raised central nodule will develop which indicates transition to the vertical growth phase. ALM is the rarest subtype and found mainly on the sole of the foot and on the palm of the hand and has a large, thinly pigmented appearance. Other less common types of melanoma are subungual melanoma (found on the nail), mucosal melanoma (found in the oral cavity, on the genital mucosa and in the perianal area) and ocular melanoma (found on the eye) (Burns *et al*., 2004).

1.1.2 Epidemiology of melanoma

Melanoma predominantly occurs among fair-skinned individuals and it is the fifth most common cancer in the United Kingdom. In Europe, the most recent estimated annual age-standardized incidence and mortality rates of melanoma on the skin were 11.1 per 100,000 and 2.3 per 100,000, respectively (Ferlay *et al*., 2013). The incidence of melanoma varies across countries, with highest incidence rates reported in Australia and New Zealand followed by countries in North America (United States and Canada) and then in Northern Europe, particularly Scandinavia. Incidence has increased in the last 50 years, but recent reports indicate stabilization in the countries with highest rates but a continued increase in rates elsewhere, especially Southern and Western Europe. The increase in incidence rates is likely due in part to improved screening and early detection of cancer, but is primarily due to changes in sun-related behaviour such as the increasing trend of holidaying in sunny places (Erdmann *et al*., 2013).

The established risk factors for melanoma are nevi phenotypes (characterized by the presence of large numbers of nevi and atypical nevi), pigmentation phenotypes (characterized by fair skin, inability to tan, red hair and freckling) and presence of family history of melanoma (Gandini *et al*., 2005a; Gandini *et al*., 2005b; Chang *et al*., 2009a; Olsen *et al*., 2010). The main environmental risk factor associated with melanoma risk is sun exposure (ultraviolet radiation), especially recreational sun exposure such as sunbathing (Chang *et al*., 2009b). Also, the use of indoor tanning beds has been reported as a risk factor for melanoma (Boniol *et al*., 2012).

Linkage studies using families with melanoma have identified major high-penetrance genes that increase risk for melanoma, such as germline mutations in *CDKN2A* and a less common mutation in *CDK4* genes (Kamb *et al*., 1994; Hussussian *et al*., 1994; Zuo *et al*., 1996). Large families study by Goldstein *et al*. (2006) reported about 40% of high risk families have germline *CDKN2A* mutations and 2% of the families carried germline *CDK4* mutations. However, these mutations are rare and only explain a small proportion of melanoma cases overall.

Over the past few years, genome-wide association studies (GWAS) have identified more common low-penetrance genetic variants associated with melanoma phenotypes (physical characteristics that associated with melanoma such as nevi and pigmentation) and melanoma risk. Several genetic variants in the region of *CDKN2A/MTAP*, *PLA2G6* and *IRF4* were associated with development of nevi and melanoma risk (Bishop *et al*., 2009; Falchi *et al*., 2009; Duffy *et al*., 2010a). For pigmentation, genetic variants identified in the regions of *MC1R*, *ASIP*, *OCA2*, *SLC45A2*, *TYR* and *TYRP1* were associated with this trait and also with melanoma risk (Duffy *et al*., 2010b). Various inherited variants identified in the region of other loci such as *TERT* (Rafnar *et al*., 2009), *CASP8*, *ATM*, *CCND1, MX2* (Barrett *et al*., 2011), *ARNT*, *PARP1* (MacGregor *et al*., 2011) and *FTO* (Iles *et al*., 2013) were associated with melanoma risk. However, the mechanisms of action of these variants on melanoma risk is still unknown.

1.1.3 Survival from melanoma

Melanoma has a good prognosis in patients with a thin tumour, but it can be lethal in patients with advanced disease where cancer has spread to the lymph nodes and/or other body organs. The 10-year survival rate for patients with thin, non-ulcerated melanomas with mitosis less than 1/mm^2 (stage IA) is about 93%, but the survival rate falls to 39% for patients with thick and ulcerated melanomas (stage IIC). For patients with stage III melanomas (without nodal or intralymphatic metastases), their 5-year survival rates range from 70% to 40% depending on the number of metastatic nodes. Among patients with stage IV melanomas (distant metastatic), their 1-year survival rates range from 62% to 33% depending on the site of distant metastases (Balch *et al*., 2009).

For melanoma, the American Joint Committee on Cancer (AJCC) staging system is used to evaluate a patient’s prognosis and to determine appropriate treatment and follow-up (Garbe *et al.,* 2010). The final version of the AJCC staging system was based on analysis of 30,946 patients by Balch *et al*. (2009). The final AJCC staging uses three primary tumour histopathological characteristics (Breslow thickness, ulceration and mitotic rate), lymph node metastasis, site of distant metastasis and serum lactate dehydrogenase level (LDH) as criteria for the tumour-node-metastasis (TNM) classification (Table 1.1) and group staging (Table 1.2) for melanoma (Balch *et al*., 2009).

Table 1.1 TNM classification for cutaneous melanoma

Classification		
T	Thickness (mm)	Ulceration status/Mitoses
T1	≤ 1.00	a: without ulceration and mitosis < 1/mm^2 b: with ulceration or mitosis ≥ 1/mm^2
T2	1.01 – 2.00	a: without ulceration b: with ulceration
T3	2.01 – 4.00	a: without ulceration b: with ulceration
T4	> 4.00	a: without ulceration b: with ulceration
N	Number of metastasis nodes	Number of metastatic burden
N0	0	NA
N1	1	a: micrometastasis* b: macrometastasis†
N2	2 – 3	a: micrometastasis* b: macrometastasis† c: in transit metastasis/satellites without metastatic nodes
N3	4+ metastatic nodes, or matted nodes, or in transit metastases/satellites with metastatic nodes	
M	Site	Serum LDH
M0	No distant metastases	NA
M1a	Distant skin, subcutaneous, or nodal metastases	Normal
M1b	Lung metastases	Normal
M1c	All other visceral metastases Any distant metastases	Normal Elevated

NA: not applicable

*Micrometastases are diagnosed after sentinel node biopsy

†Macrometastases are defined as clinically detectable nodal metastases confirmed pathologically

Table from Balch *et al*. (2009).

Table 1.2 Clinical staging for cutaneous melanoma

Staging	T	N	M
Stage I			
IA	T1a	N0	M0
IB	T1b T2a	N0 N0	M0 M0
Stage II			
IIA	T2b T3a	N0 N0	M0 M0
IIB	T3b T4a	N0 N0	M0 M0
IIC	T4b	N0	M0
Stage III	Any T	N > N0	M0
Stage IV	Any T	Any N	M1

Table from Balch *et al*. (2009).

Although the AJCC stage is a powerful prognostic tool, assessment of individual risk and identification of patients who might benefit from aggressive therapy based on the AJCC staging system are still insufficient. For example, among patients with thin melanoma (Breslow thickness ≤1mm), at least 6% relapsed and 4% died from this disease even if diagnosed at a very early stage (stage IA and IB) (Gimotty *et al*., 2004). This shows that the current AJCC stage still cannot fully explain the variation in survival and there is a need to identify new prognostic biomarkers to refine the prognosis for individual patients beyond the AJCC staging.

1.2 Prognostic factors for primary cutaneous melanoma

1.2.1 Clinical and histopathological characteristics

The established prognostic factors for reduced melanoma survival are tumour thickness (with worse prognosis for thicker melanomas), the presence of ulceration, high mitotic rate, site of primary melanoma (melanoma on the trunk, head and neck have worse prognosis than melanoma on the extremities), male sex and age (with worse prognosis in older patients) (Thorn *et al*., 1994; Lindholm *et al*., 2004; Leiter *et al.,* 2004; Buettner *et al*., 2005; Balch *et al*., 2009).

Of all the prognostic factors, tumour thickness (also known as Breslow thickness) is the strongest predictor for melanoma survival (Balch *et al*., 2009). The thickness is measured in millimetre (mm) from the surface of the tumour to the deepest point where the tumour penetrates the skin layers (Breslow, 1970). Ulceration in melanoma is defined as absence of intact epidermis. The presence of ulceration strongly influences survival, but Breslow thickness remains the stronger predictor (Balch *et al*., 2009). For patients with ulcerated tumours, their survival is lower compared to those without ulceration in the same thickness group, but survival is similar compared to those without ulceration in the next thickness group according to the T classification in Table 1.1 (Balch *et al*., 2009). Mitotic rate reflects the proliferation of the tumour and is measured as the number of mitoses per mm^2. Mitotic rate was a strong independent predictor of survival after tumour thickness, and was added in the final version of the AJCC staging replacing Clark invasion level for defining the T1a and T1b subcategory in the TNM classification (Balch *et al*., 2009).

Other tumour factors that have been reported to be associated with melanoma survival but are not part of the AJCC staging are the presence of tumour-infiltrating lymphocytes (TILs), the presence of vascular or lymphatic invasion and the presence of tumour regression. TILs are lymphocytes that infiltrate tumours and disrupt the tumour cells. TILs can be categorized as absent (no lymphocytes mixed with melanoma cells), non-brisk (focal infiltration) or brisk (involves the entire base of the tumour). Several studies have reported that presence of TILs associated with better survival in primary melanoma (Azimi *et al*., 2012; Thomas *et al*., 2013), and higher TILs grade remained as a significant predictor for MSS independent of age, sex, tumour site and the AJCC staging (Thomas *et al*., 2013).

A few studies have reported the association of the presence of vascular or lymphatic invasion in primary melanoma with poor prognosis (Kashani-Sabet *et al*., 2002; Xu *et al*., 2012). However, as this factor is not routinely reported in primary melanoma histology report, it has not been assessed in large studies for primary melanoma such as the AJCC database.

Tumour regression refers to the interaction between the tumour cells with the host immune system that leads to replacement of tumour tissue with non-malignant tissue. The prognostic value of regression in primary melanoma remains unclear, with some studies reporting that presence of regression is associated with poor prognostic outcome in patients with thin melanomas (Shaw *et al*., 1992; Guitart *et al*., 2002) and others finding no evidence that regression influences survival (Brogelli *et al*., 1992; Burton *et al*., 2011).

1.2.2 Gene expression signatures

Gene expression is the level of transcription of the deoxyribonucleic acid (DNA), a process where ribonucleic acid (RNA) is copied from DNA. DNA microarray experiments measure the messenger RNA (mRNA) expression level of thousands of genes simultaneously. DNA microarrays are small platforms of glass or silica containing single-stranded DNA sequences, called probes. From cell culture or tissue samples, mRNA is extracted and used to derive complementary DNAs (cDNAs). The cDNAs are then labelled using fluorescent dyes and hybridized to DNA microarrays. After washing off unbound or weakly bound material, the platform is scanned to detect label's signal. The signal intensity gives an approximation of the relative proportion for each gene labelled sequence and therefore an estimate of gene expression in the sample (Nguyen *et al*., 2002).

In melanoma, numerous studies used microarray data to investigate gene expression patterns related to survival in metastatic melanomas such as Mandruzzato *et al.* (2006), Alonso *et al.* (2007), John *et al.* (2008), Bogunovic *et al.* (2009), and Mann *et al.* (2013) in Table 1.3. However, microarray data have not been widely used to identify gene expression patterns related to survival in primary melanoma.

The first gene expression profiling in primary melanoma was reported by Winnepenninckx *et al.* (2006), who identified a 254-gene expression signature that was associated with 4-year distant metastatic-free survival in 58 patients using a pangenomic 44,000 60-mer oligonucleotide microarray, validated in 17 independent primary melanoma patients. The authors then compared the predictive accuracy of the gene signature with standard prognostic factors used in tumour staging (tumour thickness and presence of ulceration) and found similar prognostic accuracy (29% misclassification rate using 254-gene signature and 28% misclassification rate using prognostic factors). Although the gene signature did not show better prognostic value than the standard measures, this study had shown the potential use of gene expression as prognostic factor for primary melanoma. Among the genes identified in the gene signature were those involved in DNA replication such as minichromosome maintenance genes. Twenty-three of the genes were examined at protein level using immunohistochemical analysis, and found to be associated with overall survival (calculated from time of diagnosis to time of death from any cause) in independent patients. In multivariable model that included Breslow thickness, age, ulceration and sex, *MCM4* and *MCM6* remained associated with overall survival.

Conway *et al.* (2009) reported the second gene expression profiling in primary melanoma using formalin-fixed, paraffin-embedded tissue and identified osteopontin (*SPP1*) as predictive of relapse-free survival (calculated from time of diagnosis to time of first relapse) in the training set. The finding was validated in an independent test set in an unadjusted analysis. Unlike the training set, the predictive ability of *SPP1* was not maintained in the validation set when adjusted for the most important histological predictors, Breslow thickness, presence of ulceration and mitotic rate in the multivariate analysis. In tumour cells, *SPP1* is involved in cell adhesion, chemotaxis, prevention of apoptosis, invasion and migration. Rangel *et al*. (2008) had previously reported that increased *SPP1* expression (examined using immunohistochemical staining method) was associated with reduced relapse-free survival and disease-specific survival (calculated from time of diagnosis to time of death from melanoma) in primary melanoma.

In an examination of 57 stage IV metastatic melanomas, Jonsson *et al*. (2010) identified gene signatures for four tumour subtypes, which they called *proliferative*, *high immune response*, *pigmentation* and *normal-like*, using unsupervised hierarchical clustering. When used to predict overall survival, a significant difference was observed between the four subtypes, with the proliferative subtypes having worse survival compared to the other three groups. The four group classifier was validated in an independent sample of 44 stage III and IV melanomas using in-house data and expression datasets from four independent studies and publicly available cell line data. The results from the validation, however, varied, with the prediction capability holding up in some datasets but not in others.

In 223 primary melanomas, Harbst *et al.* (2012) replicated the study by Jonsson *et al*. (2010) and found that the 4-class molecular structure in metastatic melanoma also exists in primary melanoma. The authors further refined the four subtypes into two subtypes which can distinguish low grade tumours (high immune/normal like) and high grade tumours (proliferative/pigmentation), and noted that low grade tumours have higher expression of immune genes whereas high grade tumours have higher expression of proliferation and DNA damage signalling genes. The authors also found that the high grade tumour group was significantly associated with increased tumour thickness, mitotic rate, ulceration and poorer overall survival and relapse-free survival. When validated in two expression datasets from independent studies, the 2-molecular grade remained as an independent prognostic factor alongside AJCC stage.

A more recent study by Nsengimana *et al.* (2015) replicated the 4-class gene signature in Jonsson *et al*. (2010) and the 2-grade gene signature in Harbst *et al.* (2012). Both gene signatures were predictive of MSS (calculated from time of diagnosis to time of death from melanoma) and, as previously reported, also correlated with Breslow thickness, presence of ulceration and mitotic rate. The 2-grade gene signature also remained an independent predictor when adjusted for AJCC stage. When assessed for its predictive ability, the area under the receiver operating characteristic curve (AUC) for the AJCC alone was 0.66, 0.68. and 0.75 for relapse, deaths from melanoma, and all-causes deaths, respectively. The inclusion of the 2-molecular grade in a combined model with AJCC stage improved the AUC by 2%-4% for relapse (0.68), deaths from melanoma (0.72) and all deaths (0.78) compared to AJCC stage alone. Although the improvement in the AUC was not dramatic, this study suggested an added prognostic value of molecular classification in predicting melanoma survival.

In a cross-validation of gene expression signatures by Schramm *et al.* (2012), it was reported that gene signatures identified from different studies consistently contain immune-related genes, suggesting that immune-related genes may have an important role in melanoma progression and survival outcomes. This motivated Sivendran *et al.* (2014) to explore the association of immune-genes with melanoma outcomes. They selected 446 candidate immune-genes from reported studies and identified a 53-gene panel of immunoregulatory genes in 40 stage II and stage III primary melanomas associated with non-progression (whether patient progressed to unresectable stage III or stage IV), relapse-free survival (calculated from date of diagnosis to date of recurrence) and disease-specific survival (calculated from date of diagnosis to date of death from melanoma) in multivariable analysis. The associations were validated in 48 independent patients.

In another more recent study, Gerami *et al.* (2015) selected a set of 28 genes associated with melanoma metastasis, based on data mining of publicly available datasets. The gene set was measured in 268 primary melanomas and using radial basis machine modelling, the authors classified patients into low risk and high risk groups. A significant difference was noted in disease-free survival (calculated from time of diagnosis to time of any metastasis or local regional recurrence including involvement of sentinel nodes, in transit metastasis or distant metastasis) between the two risk groups. When included in a multivariable analysis with AJCC stage, Breslow thickness, ulceration, mitotic rate and age, the defined signature was an independent predictor of metastatic risk suggesting that the gene signature provides additional information.

Collectively, these studies reported gene signatures that are associated with survival outcomes in primary melanoma, and recent studies support the prognostic potential for adding gene signature into the current staging system. However, the use of a gene signature as a prognostic tool in clinical practice is so far not implemented. Further study to validate the published models are needed before implementation of gene signature in clinical practice.

Table 1.3 Gene expression signatures associated with survival in metastatic melanoma

Study	Tissue examined	Group compared	Gene signature	Survival associations and performance of gene signature
Mandruzzato *et al*., 2006	43 stage III and IV metastatic melanomas	Survived vs died	70 differentially expressed genes	40 over-expressed genes associated with long overall survival and 30 over-expressed genes associated with short overall survival.
Alonso *et al*., 2007	34 vertical growth phase melanomas (21 with nodal metastasis vs. 13 without)	Metastatic melanoma vs without metastasis	243 differentially expressed genes	Only three genes (*SPP1*, *CDH2* and *SPARC*) were validated for prognosis and showed association with disease-free survival. However, this was only demonstrated by univariable analysis.
John *et al*., 2008	29 stage III and IV metastatic melanomas	≥24 months of time taken to tumour progression (TTP) from stage III to stage IV vs <24 months TTP stage III to stage IV	21 most significantly differentially expressed genes were selected as a gene signature from 2140 differentially expressed genes	No prognostic impact shown for the gene signature.

Bogunovic *et al*., 2009	38 stage III and IV metastatic melanomas	Prolonged survival (≥1.5 years) vs shorter survival (<1.5 years)	266 differentially expressed genes	Gene signature was an independent predictor of post-recurrence survival alongside mitotic index in multivariable analysis.
Mann *et al*., 2013	79 stage III nodal metastatic melanomas	Survival < 1 year vs survival > 4 years after surgery	46-gene expression signature	Gene signature with strong overrepresentation of immune response genes was associated with survival, and validated in two independent stage III melanoma datasets.

The studies listed in Table 1.3 are restricted to those that identified gene signatures derived from gene expression data in metastatic melanoma and relate the gene signatures with survival outcomes. Otherwise, the list would have been longer.

1.2.3 Germline genetic variants

Many studies focusing on the genetic susceptibility to melanoma have been carried out but there is limited research looking at the influences of genetic variants on survival from melanoma. Rendleman *et al.* (2013) have showed for the first time that genetic variants associated with melanoma risk or melanoma-associated phenotypes may have roles in melanoma outcomes and that there is a potential for combining histopathological characteristics with genetic variant information for prognostic prediction. In their study, Rendleman *et al.* (2013) examined the effect of 108 susceptibility variants identified from previous GWAS of susceptibility on melanoma survival. They found two SNPs (rs7538876 in *RCC2* and rs9960018 in *DLGAP1*) significantly associated with both reduced overall survival and recurrence-free survival, and also demonstrated that inclusion of the two SNPs with tumour stage and histological subtype slightly improved the prediction of 3-year recurrence compared to a model with tumour stage and histological subtype alone (AUC 82% versus 78%). This study however requires further validation in an independent sample.

Few studies based on candidate gene analysis have reported associations between melanoma survival and genetic variants. In addition to its major role in pigmentation, the *MC1R* gene has non-pigmentary biological functions such as apoptosis and DNA repair through *MITF*, the major regulator of melanocyte development. This motivated the Leeds Melanoma Group to explore the role of *MC1R* variants on melanoma survival using a large collaborative dataset (Davies *et al*., 2012). Results from Davies *et al*. (2012) suggest that there is a survival benefit for melanoma patients with inherited *MC1R* susceptibility variants, and a similar result was also noted by Taylor *et al.* (2015a).

The Leeds Melanoma Group has reported the association of higher serum vitamin D levels with increased survival within thickness group (Newton-Bishop *et al.*, 2009); this also motivated the group to examine the causal association of SNP rs228679, in *GC* gene coding for the vitamin D-binding protein, which is associated with lower serum Vitamin D levels with survival from melanoma (Davies *et al*., 2014a). In the meta-analysis of seven cohorts, Davies *et al.* (2014a) found that inheritance of the minor allele of the *GC* SNP associated with increased risk of death from melanoma. Recently, Orlow *et al*.

(2016) reported significant association between vitamin D receptor SNP rs2239182 with MSS, suggesting that vitamin D receptor gene may influence survival from melanoma.

In another study, Davies *et al*. (2014b) explored the association of genetic variants in the *PARP1* gene, involved in DNA repair, with melanoma survival. The meta-analysis results from this study shows that inheritance of the minor allele in the *PARP1* variant rs2249844 was associated with increased risk of death from melanoma. The association between *PARP1* variant and MSS were also reported by Law *et al.* (2015b).

To date, there is no published study exploring the association of genetic variants at genome-wide level with melanoma survival, mainly due to the limited availability of large melanoma cohorts with reliable follow-up and genetic data available, hence the lack of power to detect association at a genome-wide significance level. Results based on candidate gene analyses suggest that genetic variants may have a role in melanoma survival. Therefore, this study will aim to identify genetic variants that may predict melanoma survival by using genome-wide data.

1.2.4 Other factors

1.2.4.1 Somatic mutation status

Mutations in the *BRAF* and *NRAS* genes lead to activation of the mitogen-activated protein kinase pathway that contributes to tumour growth in melanoma (Swick and Maize, 2012). Most studies focus on the common V600E *BRAF* mutation but other mutations with less defined properties are also found in *BRAF*. Several studies (Maldonado *et al*., 2003; Devitt *et al*., 2011; Thomas *et al.*, 2015) have reported the association of *BRAF* and *NRAS* mutations in tumours with melanoma survival, but evidence in the literature has not been established.

For *BRAF* mutations, a meta-analysis showed that patients with BRAF-mutant tumours have increased risk of death compared to those with *BRAF* wild-type tumours (Maldonado *et al*., 2003). In contrast, smaller studies of primary melanomas reported no independent association of *BRAF* mutation status with survival (Shinozaki *et al*., 2004; Akslen *et al*., 2005; Ellerhorst *et al*., 2011).

Similarly, for *NRAS* mutations, one study reported that patients with NRAS-mutant tumours had increased risk of death from melanoma and increased risk for recurrence (Devitt *et al*., 2011). In other studies, no association was found between *NRAS* mutation status and overall survival, relapse-free survival, and MSS (Akslen *et al*., 2005; Ellerhorst *et al*., 2011). In a more recent report by Thomas *et al.* (2015) based on 912 primary melanoma patients, no association was found between *NRAS* or *BRAF* mutation status and MSS when adjusted for age, sex, site, AJCC stage, TIL grade and study centre. However, when subgroup analysis was conducted, the authors found independent association of *NRAS* or *BRAF* mutations with MSS in higher risk tumours (T2b or higher stage), but significant associations were not seen in lower risk tumours (T2a or lower), which may indicate that mutational status has a prognostic role in higher AJCC stage primary melanomas.

1.2.4.2 Serum vitamin D level

Vitamin D is a fat soluble hormone and is synthesized in the skin in response to sun exposure. Apart from its role in calcium and phosphate intestinal absorption and bone homeostasis, Vitamin D also has an important role in cell growth, differentiation, and apoptosis, and in the regulation of tumour/immune system interaction (Deeb *et al*., 2007; Fleet *et al*., 2012). The Leeds Melanoma Group has reported the association of higher serum vitamin D levels with lower risk of relapse and lower Breslow thickness at diagnosis (Newton-Bishop *et al*., 2009). Another study by Nurnberg *et al.* (2009) supported the possible role of vitamin D in melanoma progression, as they found that serum vitamin D levels were significantly reduced in stage IV melanoma patients as compared to stage I melanoma patients. However, the concern with this association is that higher vitamin D levels might be acting as a marker of healthier lifestyles rather than contributing directly to melanoma survival, as thinner, healthier and more active people were shown to have higher vitamin D levels, leaving open the possibility that the association with high vitamin D levels might be due to other aspects of healthier lifestyles.

1.3 Methods for combining clinical and genomic data

A motivation for integrating different types of data as predictor variables is to allow for a more thorough exploration and modelling of complex traits to identify key factors that can explain or predict disease risk or outcomes (Ritchie *et al*., 2015). Several studies in melanoma have suggested the potential benefit of combining gene expression or genetic variant data with the standard clinical and histopathological factors to improve prognostic prediction for melanoma patients (Harbst *et al*., 2012; Nsengimana *et al*., 2015; Rendleman *et al*., 2013). However, no studies so far have explored the combined effect of patient and histopathological characteristics, gene expression and large-scale genetic variation on survival from melanoma. Hence, this study will combine the three different types of data to build survival models and subsequently prognostic prediction models for melanoma patients. A survival model is a model that can be used to identify factors that influence the time to an event. A prognostic model is the use of multiple prognostic factors in combination to predict the risk of developing future clinical outcomes in individual patients (Steyerberg *et al*., 2013).

A review by Ritchie *et al.* (2015) summarized two main approaches for combining multiple -omics data, which are multi-staged analysis and meta-dimensional analysis. The multi-staged approach involves integrating information in a linear or a hierarchical manner. This approach is suitable when the complex-trait aetiology is assumed to be hierarchical, such that variation at the DNA level will lead to changes in gene expression, leading to changes in protein and finally affecting the phenotype. The basic approach of multi-stage analysis is to find associations first between different types of data then subsequently between the data types and the phenotype of interest. For example, the first step is to find SNPs that are associated with the phenotype. Second, the significant SNPs are then tested for association with another level of -omic data such as gene expression levels. Third, the SNPs that associated with gene expression levels which are called eQTL SNPs are then tested for association with the phenotype of interest (Ritchie *et al.*, 2015).

An example of a recent study that applied a multi-staged integrative approach is by Huang *et al.* (2015), where they developed susceptibility models for childhood asthma based on mediation effect (genetic effect on disease risk mediated through gene expression) or alternative effect (genetic effect through other biological pathways or environmental risk factors). Under the mediation effect model, Huang *et al.* (2015) first found a set of expression quantitative trait loci (eQTL) SNPs and then jointly modelled the eQTL SNPs with the corresponding gene expression on the occurrence of asthma by using logistic regression analysis, adjusting for covariates. Using the integrative approach, they successfully identified novel susceptibility genes for childhood asthma and also confirmed several previously reported susceptibility genes. An advantage of this integrative approach over the standard GWAS analysis is that it incorporates biological information into the model which may give more insights into how the genetic variants and gene expression interact to influence the phenotype.

In a recent study, Gusev *et al.* (2016) introduced a new approach to identify gene expression that associated with complex traits in individuals without directly measured expression levels by imputing the cis-genetic component gene expression, and then relating the imputed gene expression to the trait of interest. This new approach uses a smaller set of individuals with both gene expression and genotype data available as a training panel to impute the cis-genetic component gene expression (using only SNPs within 1Mb of gene present in the GWAS data). The imputed gene expression is a linear model of genotypes with weight based on correlation between SNPs and gene expression in the training data while accounting for linkage disequilibrium (LD) among cis-SNPs (Gusev *et al*., 2016). The advantage of this new approach is that it could also be implemented when only GWAS summary statistics are available. From GWAS summary statistics, z score or standardized effect size of SNP for the trait is computed. The imputed gene expression from summary statistics is the linear combination of the z score with weight precompiled from the reference panel (Gusev *et al*., 2016). Using the new approach, Gusev *et al.* (2016) imputed gene expression using summary statistics from three recent GWAS for various traits (high-density lipoprotein, low-density lipoprotein, total cholesterol, triglycerides, height and BMI) to identify new expression-trait associations, and found 665 significant gene-trait associations, of which 69 did not overlap with genome-wide significant SNPs in the corresponding GWAS. The authors further evaluated the significance of the 69 new expression-trait associations conditional on the SNP-trait effects at the locus using permutation test, and found the permutation test was significant for 54 genes, indicating the potential use of this approach to identify new expression-trait associations. The potential advantages of this approach over standard GWAS are that the “gene” is a more interpretable biological unit than the SNP, lower number of cis-genetic component gene expressions reduced multiple testing burden, and combining cis-SNPs into a single predictor may capture heterogeneous signal better than individual SNPs.

The meta-dimensional approach involves integrating multiple different data types simultaneously to build a multivariable model associated with the outcome, as it is assumed that multiple levels of molecular variation contribute to disease aetiology in a complex way. This approach can combine different types of data by combining multiple data matrices for each sample into one large matrix before constructing a model using suitable statistical methods, or by generating multiple models first using different types of data, and then generating a final model from the multiple models (Ritchie *et al.*, 2015).

An example of study that applied a meta-dimensional approach is by Mankoo *et al.* (2011), who integrated copy number variation, methylation, microRNA and gene expression data to predict time to recurrence and survival in ovarian cancer. Their approach involves data reduction within each type of -omics data using Cox lasso regression analysis, which also does variable selection, and then combining the selected variables from each type of data in a standard Cox regression analysis. In another study, Gentles *et al.* (2015) integrated clinical prognostic index (a score created using selected clinical predictors from a larger study; age, sex and stage) with molecular prognostic index (a score created using selected nine gene expression levels from a training dataset) to create a composite risk model score (integrating clinical prognostic index and molecular prognostic index) to stratify patient overall survival into low- and high-risk for early-stage non-small cell lung cancer. Results in Gentles *et al.* (2015) shows that the composite risk model score (Hazard ratio (HR)=3.43, 95% confidence interval=2.18 to 5.39, P-value<0.001) has greater prognostic power for survival compared to using the molecular prognostic index alone (HR=2.28, 95% confidence interval=1.48 to 3.53, P-value<0.001) in stage I patients in the validation cohorts, indicating that integration of gene expression with clinical data improves the prognostic model compared with using either data alone.

The major challenges when using the meta-dimensional approach to combine different types of -omics data are the high-dimensionality of the genomic data, where the number of samples are much smaller than the number of predictor variables, and high correlation between the variables. Another challenge is to identify the best method to combine multiple types of data in a meaningful way as different types of data have different scales. The advantage of the meta-dimensional approach is that it allows for exploration of interactions between different types of -omics data that may be missed in a single type of data analysis (Ritchie *et al.*, 2015).

In data with high-dimensionality and multicollinearity, standard regression methods are subject to instability of coefficients and model over-fitting, which occurs when model predicts the outcome for the samples within the data extremely well but performs poorly in other data (Hastie *et al*., 2001). Model regularization techniques are increasingly used in high-dimensional settings. The techniques control the model complexity to prevent over-fitting the model to the training data, hence improving the model's generalizability to new samples. Model regularization is based on penalization of the model complexity through penalty terms (Hastie *et al*., 2001; Okser *et al*., 2014). Finding the optimal value of the penalty is crucial to prevent over-fitting. Cross-validation, in which some proportion of the dataset is used to build the model and a subset is used for testing the model, can be used to find the optimal value of the penalty (Hastie *et al*., 2001).

When combining clinical and genomic data to build survival prediction models, Bovelstad *et al.* (2009) has shown that penalized regression methods using ridge regression (L_2 penalty) and lasso (L_1 penalty) outperform univariable and other multivariable regression methods such as principal components regression (PCR), supervised PCR, partial least squares (PLS) regression, and supervised PLS regression in multiple genomic datasets. To deal with the high-dimensional problem, Bovelstad *et al.* (2009) explored the use of dimension reduction techniques, based on shrinkage methods (penalized regression) and linear combinations (PCR and PLS). In the PCR method, Bovelstad *et al.* (2009) used principal components analysis (PCA) to find a linear combination of the genomic variables (the combination that captures most of the variance present in the data), and then including the principal components as covariates with clinical variables in a multivariable Cox regression. The PCR has no guarantee that the selected principal components are associated with survival because the survival times are not used when forming the principal components. Hence, supervised PCR was conducted using pre-selected genes that correlated to survival first using univariate selection, and then apply the PCA to this subset. PLS regression is similar to PCR, but unlike PCR, PLS construct the linear combinations using the outcome variable for guidance, so the linear combinations is the product of the variance of the components and the correlation between the components and the survival. For the supervised PLS regression, it also involves initial gene selection step before applying the PLS method.

Various strategies to integrate different types of data were discussed in Ritchie *et al.* (2015), but there is still no gold standard method for data integration. Therefore, methods to combine different types of data will be explored in this study.

1.4 Aims and Objectives

1.4.1 Aims

This study aims to explore the inter-relationships between different types of factors associated with survival from melanoma and to build survival models and subsequently prognostic prediction models for melanoma patients.

1.4.2 Research Objectives

i. To determine the relationship of patient and tumour characteristics, gene expression levels, and genetic variants with survival from melanoma.

ii. To estimate the proportion of variance in survival from melanoma and other important factors such as Breslow thickness that can be explained by genetics. Heritability analysis using the genome-wide complex trait analysis (GCTA) tool by Yang *et al.* (2011) will be applied to use unrelated individuals and genome-wide SNP data in the analysis.

iii. To determine the inter-relationships between different types of factors associated with melanoma survival (such as gene expression levels with clinical predictors, genetic variants with clinical predictors, gene expression levels with genome-wide SNPs, and susceptibility SNPs with expression levels of nearby genes).

iv. To combine different types of factors (patient and tumour characteristics, gene expression levels and genetic variants) to build melanoma survival models and subsequently prognostic prediction models. Different statistical methods to combined different types of data will be explored.

v. To compare different statistical methods to combine different types of data to improve the methods for integrating clinical and genomic data in survival model.

Chapter 2 Materials and methods

2.1 Study population

Patients in this study are from the Leeds Melanoma Cohort (LMC) consisting of 2,184 melanoma cases recruited between 2001 and 2012 from various centres within the Yorkshire and Northern region of the United Kingdom and followed up for survival. Cases were identified and ascertained by clinicians, pathologists and the cancer registry.

The cohort study collected detailed clinical and genetic data for each patient. After consenting to participate in the study, cases were asked to complete postal and telephone questionnaires, to undergo physical examination of their skin and to give their blood samples for the extraction of DNA and other measurements such as vitamin and protein levels. For those who gave consent to the use of their stored tissue, blocks and slides were obtained from the local hospital for investigation. Slides were prepared from blocks for histological examination and tissue was taken from the tumour for extraction of DNA and RNA for allele screening and expression arrays.

Cases were followed up annually for up to 5 years (*active* follow-up). Data from the GP records and linkage to Office for National Statistics (ONS) death data were also used (*passive* follow-up). At the annual follow-up, the consenting cases were asked to complete a further questionnaire by telephone and to consider giving another blood sample to be used to determine which proteins or other constituents of blood are important for prognosis for melanoma patients.

2.2 Study variables

The study variables used in this study are the survival outcome, and the three different types of factors; patient and tumour characteristics, gene expressions, and genotypes.

2.2.1 Survival outcome

MSS is the outcome of interest in this study. Patients who are still alive or died from other causes are censored observations. Survival time for event of interest was calculated from the time of diagnosis to time of death from melanoma. For censored observations, survival time was calculated from death of diagnosis to last follow-up time for those who are still alive or died from other causes. Survival status was determined by looking at the definitive cause of death from the ONS and death certificate.

2.2.2 Patient and tumour characteristics

The following patient and tumour characteristics were included as potential prognostic factors in the survival models:

- Age at diagnosis
- Sex
- Breslow thickness
- Mitotic rate
- Type of melanoma (superficial spreading/ nodular/ lentigo maligna, acral lentiginous/ unclassified and other)
- Tumour site (limbs/ head or neck/ trunk/ other)
- Presence of ulceration (no/yes)
- Presence of lymphocytic infiltration (no/yes)
- Presence of lymphatic or vascular infiltration (no/yes)
- Presence of histological regression (no/yes)
- Hair colour at age 18 (black or brown/ red/ blonde) – self-reported
- Eye colour (brown/ green or hazel/ blue/ grey or pale blue) – self-reported

2.2.3 Genotype data

Genome-wide genotype data are available for 1,994 cases in the LMC. Patients were genotyped at ~800K variants using the Illumina Human Omni Express Exome chip.

2.2.3.1 Sample handling and DNA preparation (performed by lab technician)

DNA was extracted from blood samples and the DNA concentration and quantity was examined for quality control (QC) before processing on the Illumina arrays.

2.2.3.2 Sample QC (performed by Dr John Davies)

Genotypes were called using Illumina's BeadStudio software. Samples were excluded for any of the following reasons:

i. Sex as inferred from genotyping not matching reported sex
ii. Call rate of less than 97% (of the total number of SNPs on the array)
iii. Evidence on non-European ancestry from PCA
iv. Evidence of first-degree relationship with another individual
v. Samples from individuals who had not given appropriate consent

After excluding samples using the exclusion criteria above, only 1907 samples were eligible for analysis.

2.2.3.3 SNP QC (performed in PLINK)

The following criteria were used for SNP exclusion:

i. Missing rate of more than 3%
ii. Minor allele frequency (MAF) of less than 5%
iii. Hardy-Weinberg equilibrium (HWE) test P-value of less than 10^{-4}

Additional genome-wide genotype data were obtained for two melanoma cohorts from Cambridge, UK (n=494) and Houston, US (n=1522). The additional genotype data were used for heritability analysis described in Chapter 4. Patients from the Cambridge cohort were genotyped using the Illumina Human Omni Express Exome chip for ~800 SNPs. In the Houston cohort, patients were genotyped using Illumina Human Omni Quad chip for about ~100K SNPs. Similar criteria were applied for SNP QC in the two cohorts.

2.2.4 Gene expression data

The gene expression data used in this study are from 699 samples on the Illumina Whole Genome DASL chip that includes 29,354 probes and has between one to eight probes representing each gene from 20,819 genes.

Samples were not randomly chosen from the cohort for the microarray analysis. Individuals with tumour thickness more than 0.75mm and with the longest follow-up period were selected in order to increase the power of the study to identify biomarkers associated with melanoma outcomes. Individuals with thinner tumours were not selected as they are usually cured by excision of the tumour.

2.2.4.1 Sample preparation (performed by lab technician)

Samples for expression arrays were obtained from the primary tumour tissue. Formalin-fixed paraffin-embedded tissue blocks were used to extract the RNA. Sampling of the tumour from tissue blocks was performed using tissue microarray needle inserted horizontally through the most invasive part of the tumour and containing the fewest immune and stromal cells to reduce contamination during extraction of the RNA. A cross-section of the tumour was taken for hematoxylin and eosin staining to accurately identify a suitable area to sample from the tissue block. The RNA was extracted from the tissue core using an RNA kit.

2.2.4.2 Microarray data pre-processing (performed by Dr Jeremie Nsengimana)

The pre-processing of the microarray data was conducted within the Genomestudio software supplied by Illumina. The steps included examination of the number of probes detected in each sample, followed by background correction to remove non-specific signal (i.e. the expression that is identified by the negative control probes that are present on the array) from total signal, and then normalization to remove non-biological variation between samples.

The first step in the microarray data pre-processing is the removal of unsuccessful samples based on the number of genes detected in each sample. However, no samples were excluded from the expression dataset as explained in the next paragraph.

The normalization method for the expression data was done within R software using the R package SWAMP. This package allows adjustment to be made to correct for the differences between chips, and hence no samples were excluded in the dataset. Quantile normalization was applied on the expression data. The method ranked the genes in each sample and calculated the average gene expression across samples, then used this value as a reference sample. Each gene value was then replaced by the average expression on each sample and the genes' order was reverted back to the original order to obtain the normalized signals. Several normalization methods (average method, rank invariant, cubic spline and quantile) were available. The quantile normalization method was chosen as this method produced the most similar distributions of gene intensities between the samples compared to other normalization methods.

2.2.4.3 Probe QC

Probe QC was performed to filter probes detected in a small proportion of samples (which may indicate low expression) and probes with low variance (which are likely to be affected by noise). Filtering probes detected in a low proportion of samples (chosen threshold less than 5% of samples) was based on the proportion of samples detecting each probe at P-value<0.05. The detection P-value is calculated within the Illumina software and provides the significance level that the signal is detected (probability that the signal from a given probe is greater than the signal from the background noise). There were 5.5% probes on the expression data with low proportion of samples detected. After removal of probes detected in a low proportion of samples, there were 27,730 probes retained in the data.

Filtering of probes with low variance was based on the distribution of expression levels of all probes in the data. The variance of each probe on the Whole Genome DASL was calculated and a histogram of the variance distribution was plotted to determine the cut-off value for low variance in the data. After filtering probes based on low proportion of samples detected, 134 out of 27,730 probes with low variance (variance < 0.045) were excluded, further reducing the probes to 27,596.

For further analysis, the expression level of each probe on the Whole Genome DASL arrays was transformed to a $\log_2$ scale so that an increase in one unit corresponds to a doubling of expression levels. Then, the $\log_2$ expression level for each gene was standardised so that all genes have a mean of 0 and variance of 1 to convert all genes to a comparable scale. Each probe was analysed individually as suggested by Illumina.

Chapter 3 Survival analyses of single types of variable

The aim in this chapter is to:

i. Determine the relationship of patient and tumour characteristics, gene expression levels, and germline genetic variants with MSS

3.1 Introduction

The established clinical prognostic factors in primary melanoma as described in Chapter 1 are age, sex, tumour site, tumour thickness, presence of ulceration, and mitotic rate. A prognostic factor is defined as a measure at a given starting point such as diagnosis of disease, that is associated with a subsequent endpoint such as death (Riley *et al.*, 2013). The advances in technology allow for generation of -omics data such as genomic, transcriptomic, and proteomic, that could be used to identify new prognostic biomarkers that may have better predictive accuracy than clinical data alone.

This chapter explores the relationships between clinical predictors, such as patient and tumour characteristics, and -omics data, such as gene expression levels and genetic variants, with MSS. For the purpose of building prognostic prediction models, samples used in this study were split into a training set that will be used to develop models and a test set that will be used to assess the predictive performance of the models. Analyses in this chapter were conducted mainly in a training set (to identify the important predictors for MSS) that excludes samples in the test set which will be used in Chapter 6 (to build the prognostic models and to assess predictive performance).

3.2 Methods

3.2.1 Samples

Samples used for analysis in this chapter are as described in Chapter 2. Analyses in this chapter excluded samples from patients with survival analysis exclusion criteria (patients with multiple melanomas, who were recruited into the study more than two years after diagnosis, or were missing cause of death) and those in the test set that have been kept aside for analysis in Chapter 6.

3.2.2 Study variables

3.2.2.1 Clinical predictors

Twelve patient and tumour characteristics were included as potential prognostic factors for MSS which include age (in years), sex (categorized as female and male), Breslow thickness (in mm), type of melanoma (categorized as superficial spreading, nodular, lentigo maligna, acral lentiginous, unclassified, and other), tumour site (categorized as limbs, head/neck, truck, and other), presence of ulceration (categorized as no and yes), mitotic rate (per mm^2), presence of lymphocytic infiltration (categorized as no and yes), presence of lymphatic or vascular infiltration (categorized as no and yes), presence of histological regression (categorized as no and yes), hair colour at age 18 (categorized as black/brown, red, and blonde), and eye colour (categorized as brown, green/hazel, blue, and grey/pale blue). For the categorical variables, the category with the best expected outcome (based on the literature) were coded as the reference group. Samples with missing data were excluded from the final analysis.

3.2.2.2 Whole-genome gene expression

Whole-genome gene expression data used in this chapter are from 699 samples using Illumina Whole Genome DASL chip which consists of 29,354 probes. As explained in Chapter 2 (§ 2.2.4), these samples were not randomly selected from the cohort but were enriched for patients with thicker tumours.

Patients with gene expression data were randomly split into a training set (2/3) and a test set (1/3) by Dr. Jeremie Nsengimana. A total of 464 patients were used as a training set and 235 as a test set. The data split based on Dr. Nsengimana’s analysis was used, to be consistent with other analyses by researchers in the group. The QC for gene expression data was as described in Chapter 2. After QC, 424 patients from the training set (excluding patients with survival analysis exclusion criteria) and 27,596 probes were retained for analysis.

3.2.2.3 Genome-wide single nucleotide polymorphism (SNP)

Genome-wide SNP data and QC used in this chapter are as described in Chapter 2. After SNP QC, the genome-wide SNP data still contains a very large number of SNPs (> 500,000 SNPs). Therefore, results from survival genome-wide association analysis conducted by Dr. John Davies were used to screen SNPs across the genome as potential predictors for MSS and to reduce dimensionality.

A total of 7414 SNPs from across the genome with P-values < 0.01 in univariable Cox models (excluding patients in the test set) were selected as potential predictors. The threshold P-value < 0.01 was chosen to include as many SNPs as possible, but also reduce the set of potential predictors for the penalized Cox regression analysis. After QC, 1543 patients (excluding patients with survival analysis exclusion criteria and those in the test set) and 5651 SNPs were retained for further analysis.

3.2.3 Statistical analysis

The survival outcome in these analyses was MSS. The survival time was calculated from the time of diagnosis until the time of death from melanoma or until last follow-up time for censored observations who did not experience the event of interest (death from melanoma) in this study. Three main survival analyses were conducted in this chapter as follows:

1. Survival analysis to determine the relationship between clinical predictors and MSS: The associations of 12 clinical predictors with MSS were first explored in the whole cohort (n=1985) using Cox proportional hazards regression analysis for three models; univariable model, adjusted for age and sex, and further adjusted for Breslow thickness, as age, sex, and Breslow thickness are the most important predictors for survival in primary melanoma. The next analysis explored the association of five established clinical predictors with MSS in 1795 patients (excluding test set samples). A multivariable Cox regression was performed to determine the effect of the predictors on MSS in multivariable analysis.

2. Survival analysis to determine the relationship between gene expression levels and MSS: After QC, lasso penalized Cox regression analysis was performed in 424 patients to identify the important gene expression levels that predict MSS. After selecting a set of gene expression levels important for MSS from the penalized Cox model, univariable Cox regression was performed to identify the effect of each expression level on MSS. A univariable analysis of each probe was also performed to compare the top probes in univariable Cox models and the selected probes in the penalized model.

3. Survival analysis to determine the relationship between genetic variants and MSS: After QC, lasso penalized Cox regression analysis was performed in 1543 patients to identify the important SNPs that predict MSS. After selecting a set of SNPs important for MSS from the penalized Cox model, univariable Cox regression was performed to identify the effect of each SNP on MSS. A univariable analysis of each SNP was also performed for comparison with the penalized model.

3.2.3.1 Descriptive analysis

Descriptive statistics were calculated on clinical and tumour characteristics and survival variables for all patients in the cohort. Numerical variables were assessed for normality; the median and range were reported for variables with a skewed distribution, otherwise mean and standard deviation (SD) for normally distributed variables. For categorical variables, the frequency and percentage of each category were reported.

3.2.3.2 Univariable and multivariable Cox regression

A common approach in modelling time to event data is to use a Cox proportional hazards model which has the form:

$$h(t|X) = h_0(t)\exp(X\beta)$$

where $h_0(t)$ is baseline hazard function at time t, X is the vector of explanatory variables and β is the vector of coefficients for each variable. The parameter $\boldsymbol{\beta}$ is estimated using the maximum likelihood estimate from the partial likelihood:

$$L(\beta) = \prod_{r\epsilon\, D} \frac{\exp(X_i\beta)}{\sum_{j\in R_r} \exp(X_j\beta)}$$

where D is the set of indices for the failures and R_r is the set of indices for patients at risk at time t_r.

The hazard ratio exp(β) is the ratio of the rate at which patients in the two (or more) groups defined by the corresponding covariate are experiencing an event (Hosmer and Lemeshow, 1999). The Cox model involves the assumption that the regression effect β is constant over time; this is the proportional hazards assumption.

Univariable Cox regression analysis was performed to provide a preliminary idea of which variables had possible prognostic importance. Multivariable Cox regression was performed for the selected established clinical predictors to identify the effect of each predictor on MSS when adjusting for other predictors. The regression coefficient (β), hazard ratio (HR), standard error (SE) of the β and P-value were reported.

3.2.3.3 Penalized Cox regression

In a high-dimensional data setting, reliable estimation of parameters β using a Cox model is no longer possible when the number of predictors p is much larger than the number of samples N (p >> N). One of the strategies for dealing with high-dimensional data is penalized regression, which is a regularization or shrinkage method that introduces some constraint on the parameters to be estimated (Hastie *et al.*, 2001).

For generalized linear regression, Tibshirani (1996) described the least absolute shrinkage and selector (lasso) technique as a variable selection method using L_1 penalty; this adds a penalty to the log-likelihood, comprised of the sum of the absolute values of the coefficients ($\sum|\beta_j|$). This method is attractive in the high-dimensional setting because it can shrink some of the coefficients to zero and thus can be used for variable selection. For survival analysis, Tibshirani (1997) proposed a similar strategy for L_1-penalized Cox regression to estimate β which minimizes the penalized log-likelihood function:

$$\hat{\beta}(\lambda) = \arg\min_{\beta}[-\log\{L(\beta)\} + \lambda\|\beta\|_1]$$

where $L(\beta)$ is the partial likelihood and λ>0 is the regularization parameter which controls the amount of regularization applied to the estimate (Tibshirani, 1997). To increase the computational efficiency of the estimation method used in Tibshirani (1997), a more efficient algorithm was proposed by Park and Hastie (2007) for both generalized linear models and Cox models. Another commonly used penalized regression method is ridge regression which uses L_2 penalty that comprises the sum of square of the coefficients ($\sum \beta_j^2$). Although ridge regression shrinks the coefficients towards zero, it does not select variables as lasso does because it does not set any coefficients to zero.

For a case where there are only two predictors, Figure 3.1 shows some insight of why lasso often produces coefficients that are exactly zero and ridge regression does not. The constraint region when using the lasso constraint ($\sum|\beta_j|$≤t) is the rotated square (left diagram). The lasso solution finds the first point that touches the constraint region which sometimes occurs at a corner, and this corresponds to a zero coefficient. With ridge regression ($\sum \beta_j^2$≤t), the constraint region is the circle (right diagram), which has no corners for the contours to hit, hence does not result in zero coefficients.

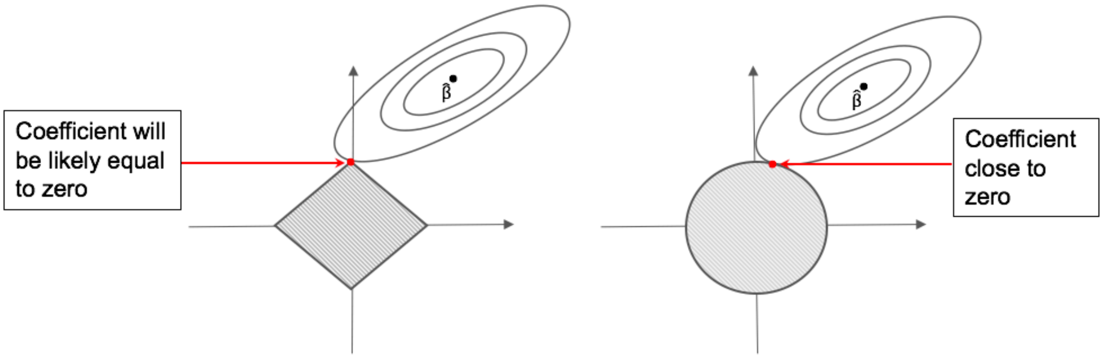

Coefficient will be likely equal to zero
$\hat{\beta}$
$\hat{\beta}$
Coefficient close to zero

Figure 3.1 Estimation diagram for the lasso and ridge regression

(Based on Tibshirani, 1996).

Coxpath algorithm

In this chapter, the method proposed by Park and Hastie (2007) for fitting L_1-penalized Cox regression was applied to select important gene expression levels and SNPs for MSS. Park and Hastie (2007) introduced the Coxpath algorithm that implements the “predictor-corrector” method to determine the entire path of the coefficient estimates as the regularization parameter λ varies $\{\hat{\beta}(\lambda): 0 < \lambda < \lambda_{max}\}$. The algorithm starts from λ_{max}, the largest lambda that makes any coefficients non-zero and computes a series of solution sets, each time estimating the coefficients with a smaller λ based on the previous estimate. With each step, the penalty is lowered and this results in more coefficients becoming non-zero. The iteration stops when the set of non-zero coefficients is not augmented anymore and this is usually when the λ = 0, reducing to the standard Cox proportional hazards model estimates.

The estimate of the coefficient and the step size of λ between iterations are determined by the predictor-corrector method which consists of three steps: determining the step size in λ, predicting the corresponding change in the coefficients, and correcting the error in the previous prediction. This algorithm provides the exact order of the active set changes which can be used to identify the order in which the variables enter or leave the model. The *coxpath* function by Park and Hastie (2007) was implemented in the R package *glmpath*.

K-fold cross-validation

Once a path of solutions is calculated, an optimal λ value needs to be selected so that the corresponding model will be the best model while avoiding over-fitting the data. The most widely used method to select the best model is cross-validation. There are other methods for doing this such as using Akaike or Bayesian information criteria, but these methods have been shown to over-fit in survival predictions using high-dimensional data (Schumacher *et al.*, 2007).

In K-fold cross-validation, data is split into *K* equal parts, then one part is set aside as a test set (to evaluate model performance, for example by calculating the prediction error) and other remaining parts are used as a training set (to build a model). This procedure will be done for *k* = 1, 2,…,K by keeping one fold outside as a test set in each cycle. After cycling through the procedure, the errors from each *k* will be combined to obtain the prediction error of the overall model. Repeating this process for a grid of λ values, the value that minimizes the prediction error will be selected to find the best model. For a continuous outcome, squared error is used to evaluate a model's test performance, and for a categorical outcome, misclassification error could be used (Hastie *et al.*, 2001).

For survival data, the use of squared error is inappropriate due to censoring. In Cox regression setting, the cross-validated partial log-likelihood (cvpl) introduced by Verweij and van Houwelingen (1993) is often used to evaluate the predictive ability of a survival model. Verweij and van Houwelingen (1993) used a leave-one-out cross-validation method to compute the cvpl which measures how well every observation i can be predicted using the other observations.

This calculates the contribution of observation i to the partial log-likelihood:

$$l_i(\beta) = l(\beta) - l_{-i}(\beta)$$

where $l(\beta)$ is the full log-likelihood model and $l_{-i}(\beta)$ is the log-likelihood when the observation i is left out. Cvpl is given by:

$$cvpl = \sum_{i=1}^{n} l_i(\hat{\beta}_{-i})$$

where $\hat{\beta}_{-i}$ is the value of β that maximizes the log-likelihood when observation i is left out. A large value of the cvpl indicates a model that fits new observation well (Verweij and van Houwelingen, 1993). Leave-one-out cross-validation (K = *n*), where *n* is the sample size, takes a lot of computation time, because it requires fitting n models, each containing $n - 1$ observations. Some studies such as Bovelstad *et al*. (2007) have modified the leave-one-out cross-validation to K-fold cross-validation to select the optimal tuning parameter for the L_1- and L_2 penalized Cox regression in their analyses. Their method splits the data into a training set and a test set, and instead of calculating the cvpl when observation i is left out, it calculates the cvpl when the *k*th fold, *k* = 1, 2, …, K, is left out.

Model selection for lasso penalized Cox models in this chapter was performed using an internal cross-validation *cv.coxpath* function in the *glmpath* package which computes the cross-validated (minus) log-partial likelihood for *coxpath*. A 10-fold cross-validation was performed for a grid of values of λ; the optimal value of λ is chosen to correspond to the value which has the smallest cross-validation error.

3.3 Results

3.3.1 Clinical predictors and survival

Table 3.1 shows the descriptive analysis for clinical and tumour characteristics of patients in the whole cohort. A total of 199 patients with multiple melanomas, recruited into the study more than two years after diagnosis or missing cause of death were excluded from the survival analysis. A total of 1985 individuals in LMC with mean age of 54.3 years (SD=13.7) were included for survival analysis. The number of deaths from melanoma in this analysis is 349 (17.6%). The median survival time for those who have died is 3.2 years, and the median follow-up time for survivors is 7.5 years. Patients in this analysis are cases with primary melanoma, and the majority (56%) were in AJCC stage I. The majority of patients were female (56.6%), most had superficial spreading melanoma (57.8%), the most common site of tumour was on the limbs (44.1%), and most had no ulceration (79.6%).

Table 3.2 shows the exploratory analysis of the association of clinical predictors with MSS in the whole cohort of 1985 patients. In univariable analysis, age, sex, tumour type, tumour site, Breslow thickness, mitotic rate, presence of ulceration, presence of TILs, and presence of vascular infiltration showed significant association with MSS. Factors that remained significant after controlling for three established predictors (age, sex, and Breslow thickness) are melanoma type (acral lentiginous, with worse prognosis than the most common type), tumour site (trunk and other, with worse prognosis than limbs), mitotic rate, presence of ulceration, presence of TILs, and presence of vascular infiltration.

Table 3.3 shows the association of five established clinical predictors (age, sex, tumour site, Breslow thickness, and presence of ulceration) with MSS in the cohort excluding 190 patients that will be used for a test set in Chapter 6 (n=1795). All predictors were significantly associated with MSS in univariable analysis and remained significant in multivariable analysis except for some differences between tumour sites.

Table 3.1 Clinical and tumour characteristics of patients in the whole cohort (n=1985)

Variables	n missing	n (%)
Age (years)	0	54.3 (13.7)
Sex	0	
Female		1123 (56.6)
Male		862 (43.4)
Tumour type	0	
Superficial spreading		1148 (57.8)
Nodular		402 (20.3)
Lentigo maligna melanoma		34 (1.7)
Acral lentiginous		58 (2.9)
Unclassified		165 (8.3)
Other		178 (9.0)
Tumour site	0	
Limbs		875 (44.1)
Head/neck		196 (9.9)
Trunk		698 (35.1)
Other		216 (10.9)
Breslow thickness (mm)	49	1.5 (0.2 – 20)*
Mitotic rate (per mm^2)	293	2 (0 – 83)*
Presence of ulceration	15	
No		1569 (79.6)
Yes		401 (20.4)
Presence of TILs	440	
No		232 (15.0)
Yes		1313 (85.0)
Presence of histological regression	499	
No		1205 (81.1)
Yes		281 (18.9)
Presence of vascular infiltration	263	
No		1600 (92.9)
Yes		122 (7.1)
Hair colour at age 18	81	
Black/brown		1308 (68.7)
Red		227 (11.9)
Blonde		369 (19.4)
Eye colour	85	
Brown		309 (16.3)
Green/hazel		567 (29.8)
Blue		803 (42.3)
Grey		221 (11.6)
AJCC stage	4	
Stage I		1094 (56.0)
Stage II		588 (30.1)
Stage III		266 (13.6)
Stage IV		6 (0.3)
Follow-up time for patients who are still alive[†]	0	7.50 (0.45 – 14.69)*
Survival time for patients who have died[†]		3.20 (0.38 – 14.48)*
Survival status	0	
Alive/censored		1636 (82.4)
Died from melanoma		349 (17.6)

*Median (range) [†]MSS time calculated in years

Table 3.2 **Association of clinical predictors with MSS in the whole cohort (n=1985*)**

Variable	n	Univariable Cox model				Multivariable Cox model (Adjusted for age and sex)				Multivariable Cox model (Adjusted for age, sex and Breslow thickness)			
		β	HR	SE	P-value	β	HR	SE	P-value	β	HR	SE	P-value
Age (years)	1985	0.04	1.04	0.01	2.0 x 10^{-16}	-	-	-	-	-	-	-	-
Sex	1985												
Female	1123	-	-	-	-	-	-	-	-	-	-	-	-
Male	862	0.59	1.81	0.11	3.7 x 10^{-8}	-	-	-	-	-	-	-	-
Tumour type	1985												
Superficial spreading	1148	-	-	-	-	-	-	-	-	-	-	-	-
Nodular	402	0.83	2.30	0.12	1.4 x 10^{-11}	0.70	2.02	0.12	1.6 x 10^{-8}	0.20	1.22	0.13	0.14
Lentigo maligna melanoma	34	-0.16	0.85	0.51	0.75	-0.71	0.49	0.51	0.17	-0.78	0.46	0.51	0.13
Acral lentiginous	58	1.15	3.17	0.24	1.4 x 10^{-6}	0.98	2.67	0.24	4.5 x 10^{-5}	0.77	2.16	0.24	1.4 x 10^{-3}
Unclassified	165	-0.02	0.98	0.23	0.92	0.06	1.06	0.23	0.81	-0.30	0.74	0.27	0.27
Other	178	0.42	1.52	0.19	0.03	0.38	1.47	0.19	0.04	-0.49	0.63	0.24	0.05
Tumour site	1985												
Limbs	875	-	-	-	-	-	-	-	-	-	-	-	-
Head/neck	196	0.66	1.94	0.19	4.9 x 10^{-4}	0.35	1.42	0.19	0.07	0.12	1.13	0.20	0.55
Trunk	698	0.70	2.01	0.13	2.1 x 10^{-7}	0.55	1.73	0.14	1.2 x 10^{-4}	0.49	1.62	0.14	6.4 x 10^{-4}
Other	216	1.50	4.47	0.15	2.0 x 10^{-16}	1.22	3.40	0.16	1.1 x 10^{-14}	0.63	1.88	0.18	6.3 x 10^{-4}
Breslow thickness (mm)	1936	0.22	1.24	0.01	2.0 x 10^{-16}	0.21	1.24	0.01	2.0 x 10^{-16}	-	-	-	-
Mitotic rate (per mm^2)	1692	0.05	1.05	0.004	2.0 x 10^{-16}	0.05	1.05	0.004	2.0 x 10^{-16}	0.03	1.03	0.01	2.1 x 10^{-8}
Presence of ulceration	1970												
No	1569	-	-	-	-	-	-	-	-	-	-	-	-
Yes	401	1.38	3.97	0.11	2.0 x 10^{-16}	1.21	3.36	0.11	2.0 x 10^{-16}	0.80	2.23	0.12	1.1 x 10^{-10}
Presence of TILs	1545												
No	232	-	-	-	-	-	-	-	-	-	-	-	-
Yes	1313	-0.69	0.5	0.14	1.9 x 10^{-6}	-0.80	0.45	0.14	3.5 x 10^{-8}	-0.71	0.49	0.14	1.1 x 10^{-6}

Presence of vascular infiltration	1722												
No	1600	-	-	-	-	-	-	-	-	-	-	-	-
Yes	122	1.34	3.82	0.15	2.0 x 10^{-16}	1.34	3.83	0.15	2.0 x 10^{-16}	0.90	2.46	0.16	1.7 x 10^{-8}
Presence of histological regression	1486												
No	1205	-	-	-	-	-	-	-	-	-	-	-	-
Yes	281	-0.08	0.93	0.16	0.63	-0.17	0.84	0.16	0.29	-0.13	0.88	0.16	0.44
Hair colour at age 18	1904												
Black/brown	1308	-	-	-	-	-	-	-	-	-	-	-	-
Red	227	-0.31	0.73	0.19	0.10	-0.21	0.81	0.19	0.25	-0.23	0.79	0.19	0.23
Blonde	369	-0.23	0.79	0.15	0.12	-0.16	0.85	0.15	0.27	-0.10	0.90	0.15	0.49
Eye colour	1900												
Brown	309	-	-	-	-	-	-	-	-	-	-	-	-
Green/hazel	567	-0.22	0.81	0.17	0.19	-0.13	0.88	0.17	0.45	-0.18	0.84	0.17	0.30
Blue	803	-0.12	0.89	0.16	0.44	-0.14	0.87	0.16	0.38	-0.18	0.84	0.16	0.26
Grey	221	-0.07	0.93	0.20	0.73	-0.05	0.95	0.20	0.79	-0.04	0.96	0.21	0.86

*Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis and missing cause of death

TILs: Tumour-infiltrating lymphocytes

MSS: Melanoma-specific survival

Table 3.3 Association of selected established clinical predictors with MSS in the training set (n=1795[a])

Predictors	n	Univariable Cox model				Multivariable Cox model[b,c]			
		β	HR	SE	P-value	β	HR	SE	P-value
Age (years)	1795	0.04	1.04	0.01	1.3×10^{-15}	0.03	1.03	0.01	8.4×10^{-8}
Sex	1795								
Female	1022	-	-	-	-	-	-	-	-
Male	773	0.67	1.95	0.12	8.8×10^{-9}	0.33	1.38	0.13	0.01
Tumour site	1795								
Limbs	794	-	-	-	-	-	-	-	-
Head/neck	171	0.73	2.08	0.20	2.7×10^{-4}	0.11	1.12	0.22	0.61
Trunk	642	0.71	2.02	0.14	1.0×10^{-6}	0.44	1.55	0.15	4.2×10^{-3}
Other	188	1.48	4.41	0.17	2.0×10^{-16}	0.37	1.45	0.21	0.07
Breslow thickness (mm)	1752	0.21	1.24	0.01	2.0×10^{-16}	0.16	1.18	0.02	2.0×10^{-16}
Presence of ulceration	1780								
No	1447	-	-	-	-	-	-	-	-
Yes	333	1.36	3.89	0.12	2.0×10^{-16}	0.75	2.12	0.14	4.7×10^{-8}

[a] Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis, missing cause of death and those in the test set
MSS: Melanoma-specific survival
[b] n= 1747, 48 patients were excluded due to missing values
[c] Proportional hazards assumption was checked and not violated (Table 1 in Appendix II)

3.3.2 Gene expression levels and survival

Table 3.4 shows 16 gene expression levels with non-zero coefficients selected by penalized Cox regression at the chosen penalty, in the order in which probes entered the model. However, two of the 16 selected probes (in *HLA-DQB2* and *CIAPIN1*) have very small coefficients in the penalized model. In univariable analysis, all selected probes were highly associated with MSS. For eight selected probes, high expression was protective for survival and for the remaining eight probes high expression was associated with increased risk of death. Eleven of the selected probes were among the top 20 probes associated with MSS in univariable Cox analysis as shown in Table 3.5.

The selected probes show low to moderate pairwise correlation (Pearson's correlation) as shown in Table 3.6. The highest correlation was observed between probe ILMN_1778401 (*HLA-B*) and ILMN_1764109 (*C1R*) with r=0.61. For the top 20 probes in univariable analysis, four pairs show strong correlation and the rest mostly show moderate correlation (Table 3.7).

Table 3.4 16 probes selected by penalized Cox model of MSS at cross-validated penalty in the training set (n=424[a])

Probe	Gene	Chr	Penalized Cox model	Univariable Cox model			
			β	β	HR	SE	P-value
ILMN_1701441	*LPAR1*	9	-0.05	-0.47	0.62	0.07	9.9 x 10^{-12}
ILMN_3249501	*ZNF697*	1	0.17	0.62	1.85	0.09	2.8 x 10-^{12}
ILMN_1749829	*DLG1*	14	0.10	0.53	1.70	0.09	1.3 x 10^{-9}
ILMN_1731206	*NKD2*	5	-0.05	-0.44	0.65	0.07	2.3 x 10^{-10}
ILMN_1764109	*C1R*	12	-0.03	-0.57	0.56	0.08	1.1 x 10^{-11}
ILMN_2056167	*OSTC*	4	0.03	0.46	1.58	0.08	5.2 x 10^{-8}
ILMN_3238435	*SNORA12*	10	-0.05	-0.49	0.61	0.07	2.0 x 10^{-11}
ILMN_1695959	*C21orf63*	21	-0.06	-0.44	0.65	0.07	1.7 x 10^{-9}
ILMN_1741648	*HLA-DQB2*	6	-0.002	-0.45	0.64	0.08	3.5 x 10^{-9}
ILMN_1784238	*SEC22B*	1	0.04	0.43	1.54	0.07	1.8 x 10^{-10}
ILMN_1778401	*HLA-B*	6	-0.02	-0.42	0.66	0.07	8.4 x 10^{-9}
ILMN_1759729	*NDUFA8*	9	0.03	0.64	1.89	0.11	4.5 x 10^{-9}
ILMN_2344221	*IGSF5*	21	0.04	0.39	1.48	0.08	1.0 x 10^{-6}
ILMN_2095633	*FGF22*	19	-0.03	-0.45	0.64	0.10	4.8 x 10^{-6}
ILMN_1700547	*CHST9*	18	0.01	0.36	1.43	0.08	1.7 x 10^{-5}
ILMN_1735199	*CIAPIN1*	16	0.001	0.61	1.85	0.12	4.1 x 10^{-7}

[a] Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis, missing cause of death and those in the test set

Chr: Chromosome

Table 3.5 Top 20 probes associated with MSS in univariable Cox analysis (n=424[a])

Probe	Gene	Chr	Univariable Cox model			
			β	HR	SE	P-value
ILMN_3249501	*ZNF697*	1	0.62	1.85	0.09	2.76×10^{-12}
ILMN _1701441	*LPAR1*	9	-0.47	0.62	0.07	9.87×10^{-12}
ILMN _1764109	*C1R*	12	-0.57	0.56	0.08	1.10×10^{-11}
ILMN _3238435	*SNORA12*	10	-0.49	0.61	0.07	2.04×10^{-11}
ILMN _1768227	*DCN*	12	-0.55	0.58	0.08	5.53×10^{-11}
ILMN _2334210	*ITGB4*	17	-0.55	0.57	0.09	1.57×10^{-10}
ILMN _1784238	*SEC22B*	1	0.43	1.54	0.07	1.78×10^{-10}
ILMN _1731206	*NKD2*	5	-0.44	0.65	0.07	2.28×10^{-10}
ILMN _2313079	*NLRP1*	17	-0.50	0.61	0.08	1.07×10^{-9}
ILMN _1749829	*DLG1*	14	0.53	1.70	0.09	1.31×10^{-9}
ILMN _1763837	*ANPEP*	15	-0.47	0.62	0.08	1.62×10^{-9}
ILMN _1695959	*C21orf63*	21	-0.44	0.65	0.07	1.74×10^{-9}
ILMN _1670305	*SERPING1*	11	-0.52	0.60	0.09	2.31×10^{-9}
ILMN _1757415	*C1orf163*	1	0.71	2.04	0.12	3.19×10^{-9}
ILMN _1741648	*HLA-DQB2*	6	-0.45	0.64	0.08	3.49×10^{-9}
ILMN _1794612	*UBA7*	3	-0.50	0.61	0.09	3.76×10^{-9}
ILMN _1759729	*NDUFA8*	9	0.64	1.89	0.11	4.48×10^{-9}
ILMN _1737650	*DIO2*	14	-0.45	0.64	0.08	5.91×10^{-9}
ILMN _1702787	*SEMA4A*	1	-0.45	0.64	0.08	7.33×10^{-9}
ILMN _1778401	*HLA-B*	6	-0.42	0.66	0.07	8.37×10^{-9}

[a] Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis, missing cause of death and those in the test set

Chr: Chromosome

11 probes highlighted were selected by the penalized Cox regression in Table 3.4

Table 3.6 Pairwise correlations between 16 probes in the training set using Pearson's correlation (n=424)

	ILMN_1701441	ILMN_3249501	ILMN_1749829	ILMN_1731206	ILMN_1764109	ILMN_2056167	ILMN_3238435	ILMN_1695959	ILMN_1741648	ILMN_1784238	ILMN_1778401	ILMN_1759729	ILMN_2344221	ILMN_2095633	ILMN_1700547	ILMN_1735199
ILMN_1701441	1	-0.36	-0.43	0.45	0.61	-0.23	0.33	0.34	0.35	-0.1	0.32	-0.12	-0.15	-0.02	-0.11	-0.16
ILMN _3249501		1	0.2	-0.28	-0.31	0.07	-0.16	-0.1	-0.33	0.4	-0.19	0.09	0.04	-0.09	0.08	0.2
ILMN _1749829			1	-0.3	-0.38	0.33	-0.37	-0.25	-0.25	0.14	-0.25	0.22	0.08	-0.17	0.1	0.27
ILMN _1731206				1	0.35	-0.14	0.33	0.28	0.53	-0.29	0.18	0.05	-0.16	0.23	-0.02	-0.19
ILMN _1764109					1	-0.31	0.37	0.36	0.44	-0.1	0.61	-0.21	-0.23	-0.07	-0.17	-0.22
ILMN _2056167						1	-0.31	-0.36	-0.18	0.24	-0.3	0.33	0.19	-0.11	0.23	0.11
ILMN _3238435							1	0.25	0.34	-0.07	0.28	-0.25	-0.07	0.14	-0.1	-0.18
ILMN _1695959								1	0.2	-0.11	0.3	-0.04	-0.22	0.05	-0.2	-0.11
ILMN _1741648									1	-0.2	0.46	-0.13	-0.32	0.26	-0.14	-0.27
ILMN _1784238										1	-0.04	0.06	0.07	-0.22	0.02	0.12
ILMN _1778401											1	-0.17	-0.33	0.04	-0.21	-0.15
ILMN _1759729												1	0.09	-0.06	0.08	0.25
ILMN _2344221													1	-0.09	0.17	0.17
ILMN _2095633														1	0.01	-0.17
ILMN _1700547															1	0.05
ILMN _1735199																1

No correlation (|r|<0.1) in black text; Mild correlation (0.1≤|r|<0.3) in blue text; Moderate correlation (0.3 ≤|r|< 0.7) in green text

Table 3.7 Pairwise correlations between the top 20 probes in the training set using Pearson's correlation (n=424)

	ILMN _3249501	ILMN _1701441	ILMN _1764109	ILMN _3238435	ILMN _1768227	ILMN _2334210	ILMN _1784238	ILMN _1731206	ILMN _2313079	ILMN _1749829	ILMN _1763837	ILMN _1695959	ILMN _1670305	ILMN _1757415	ILMN _1741648	ILMN _1794612	ILMN _1759729	ILMN _1737650	ILMN _1702787	ILMN _1778401
ILMN_3249501	1	-0.36	-0.31	-0.16	-0.29	-0.24	0.4	-0.28	-0.38	0.2	-0.31	-0.1	-0.34	0.22	-0.33	-0.32	0.09	-0.22	-0.33	-0.19
ILMN _1701441		1	0.61	0.33	0.71	0.37	-0.1	0.45	0.53	-0.43	0.63	0.34	0.43	-0.35	0.35	0.46	-0.12	0.51	0.24	0.32
ILMN _1764109			1	0.37	0.79	0.42	-0.1	0.35	0.66	-0.38	0.66	0.36	0.71	-0.42	0.44	0.52	-0.21	0.54	0.32	0.61
ILMN _3238435				1	0.4	0.33	-0.07	0.33	0.34	-0.37	0.38	0.25	0.35	-0.27	0.34	0.32	-0.25	0.37	0.26	0.28
ILMN _1768227					1	0.45	-0.14	0.46	0.63	-0.48	0.73	0.35	0.57	-0.37	0.46	0.49	-0.2	0.62	0.32	0.51
ILMN _2334210						1	-0.28	0.48	0.42	-0.36	0.43	0.41	0.25	-0.27	0.42	0.3	-0.1	0.43	0.37	0.32
ILMN _1784238							1	-0.29	-0.14	0.14	-0.11	-0.11	-0.09	0.21	-0.2	-0.14	0.06	-0.21	-0.22	-0.04
ILMN _1731206								1	0.47	-0.3	0.44	0.28	0.27	-0.21	0.53	0.33	0.05	0.55	0.34	0.18
ILMN _2313079									1	-0.3	0.64	0.31	0.56	-0.35	0.58	0.51	-0.24	0.54	0.53	0.59
ILMN _1749829										1	-0.4	-0.25	-0.37	0.36	-0.25	-0.36	0.22	-0.39	-0.23	-0.25
ILMN _1763837											1	0.28	0.51	-0.32	0.47	0.44	-0.2	0.6	0.37	0.5
ILMN _1695959												1	0.24	-0.17	0.2	0.28	-0.04	0.37	0.12	0.3
ILMN _1670305													1	-0.36	0.35	0.57	-0.14	0.38	0.37	0.53
ILMN _1757415														1	-0.28	-0.37	0.14	-0.34	-0.21	-0.28
ILMN _1741648															1	0.42	-0.13	0.49	0.59	0.46
ILMN _1794612																1	-0.24	0.32	0.34	0.48
ILMN _1759729																	1	-0.24	-0.12	-0.17
ILMN _1737650																		1	0.24	0.39
ILMN _1702787																			1	0.37
ILMN _1778401																				1

No correlation (|r|<0.1) in black text; Mild correlation (0.1≤|r|<0.3) in blue text; Moderate correlation (0.3 ≤|r|< 0.7) in green text; Strong correlation (|r|>0.7) in red text

3.3.3 Genetic variants and survival

Table 3.8 shows 13 SNPs with non-zero coefficients selected by penalized Cox model at the chosen penalty in order of probes that entered the model. In univariable analysis, all SNPs were highly significantly associated with MSS. The minor alleles of four SNPs were associated with better survival, while the minor alleles of the remaining 9 SNPs were associated with increased risk of death. When compared to the top 20 SNPs in the univariable analysis, 9 SNPs from the penalized model were among the top 20 SNPs (Table 3.9).

Of the 13 SNPs selected by the penalized model, three pairs of SNPs showed high correlation as shown in Table 3.10. Two pairs were in almost complete LD, with $r \geq 0.96$ and were located very closely in the genome (RS2902554 and RS9957831; RS2392477 and RS10233832), but none of the other pairs were correlated ($|r|<0.1$). In contrast, for the top 20 SNPs from univariable analysis, there were 24 pairs that were strongly correlated as shown in Table 3.11 This shows that using univariable Cox analysis to select potential predictors for MSS will results in selecting many highly correlated SNPs, whereas penalize regression is quite successful at selecting independent SNPs.

Table 3.8 13 SNPs selected by penalized Cox model of MSS at cross-validated penalty in the training set (n=1543[a])

SNPs	Chr	Position	Gene*	MAF	Penalized Cox model	Univariable Cox model			
					β	β	HR	SE	P-value
RS17837209	13	51851122	*FAM124A*	0.09	0.16	0.68	1.98	0.13	1.2×10^{-7}
RS9957831	18	27451322	unknown	0.17	0.02	0.51	1.67	0.11	1.1×10^{-6}
RS4768090	12	45901586	*LOC105369743*	0.30	0.04	0.44	1.55	0.09	1.8×10^{-6}
RS2902554	18	27466177	unknown	0.16	0.05	0.52	1.68	0.11	1.2×10^{-6}
RS5770310	22	49767410	unknown	0.47	0.02	0.39	1.48	0.09	8.4×10^{-6}
RS10233832	7	36991362	unknown	0.34	-0.01	-0.46	0.63	0.10	9.6×10^{-6}
RS17379771	5	37813591	*GDNF*	0.38	0.02	0.41	1.50	0.09	8.4×10^{-6}
RS16956192	17	6613374	*SLC13A5*	0.11	0.04	0.56	1.75	0.12	5.1×10^{-6}
RS2392477	7	36985541	*ELMO1*	0.34	-0.02	-0.46	0.63	0.10	9.6×10^{-6}
RS6689263	1	20882523	unknown	0.13	-0.02	-0.77	0.46	0.18	2.6×10^{-5}
RS11639902	16	71755852	*PHLPP2*	0.38	-0.01	-0.43	0.65	0.10	1.9×10^{-5}
RS12519276	5	37873327	unknown	0.45	0.01	0.40	1.49	0.09	1.1×10^{-5}
RS10941528	5	40962863	*C7*	0.23	0.002	0.42	1.53	0.10	1.5×10^{-5}

[a] Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis, missing cause of death and those in the test

Chr: Chromosome

* Information obtained from http://www.ncbi.nlm.nih.gov/SNP/

Table 3.9 Top 20 SNPs associated with MSS in univariable Cox analysis ((n=1543[a])

SNPs	Chr	Position	Gene*	MAF	Univariable Cox model			
					β	HR	SE	P-value
RS17837209	13	51851122	*FAM124A*	0.09	0.68	1.98	0.13	$1.21\ \mathrm{x}10^{-7}$
RS3825409	13	51825548	*FAM124A*	0.12	0.59	1.81	0.12	$6.24\ \mathrm{x}10^{-7}$
RS9957831	18	27451322	unknown	0.18	0.51	1.67	0.11	$1.14\ \mathrm{x}10^{-6}$
RS2902554	18	27466177	unknown	0.17	0.52	1.68	0.11	$1.18\ \mathrm{x}10^{-6}$
RS11617293	13	51823185	*FAM124A*	0.12	0.58	1.78	0.12	$1.25\ \mathrm{x}10^{-6}$
RS10502535	18	27484742	unknown	0.17	0.51	1.66	0.11	$1.66\ \mathrm{x}10^{-6}$
RS4768090	12	45901586	*LOC105369743*	0.30	0.44	1.55	0.09	$1.83\ \mathrm{x}10^{-6}$
RS2863737	18	27491494	unknown	0.17	0.51	1.66	0.11	$1.96\ \mathrm{x}10^{-6}$
RS10880801	12	45899396	*LOC105369743*	0.39	0.41	1.51	0.09	$4.58\ \mathrm{x}10^{-6}$
RS4576884	12	45892993	*LOC105369743*	0.39	0.41	1.51	0.09	$4.58\ \mathrm{x}10^{-6}$
RS2408323	12	45881521	*LOC105369743*	0.30	0.42	1.52	0.09	$4.67\ \mathrm{x}10^{-6}$
RS16956192	17	6613374	*SLC13A5*	0.11	0.56	1.75	0.12	$5.08\ \mathrm{x}10^{-6}$
RS9807803	18	27454863	unknown	0.20	0.45	1.57	0.10	$7.41\ \mathrm{x}10^{-6}$
RS5770310	22	49767410	unknown	0.47	0.41	1.50	0.09	$8.40\ \mathrm{x}10^{-6}$
RS17379771	5	37813591	*GDNF*	0.38	0.39	1.48	0.09	$8.40\ \mathrm{x}10^{-6}$
RS17252076	13	51807093	*FAM124A*	0.13	0.50	1.65	0.11	$8.85\ \mathrm{x}10^{-6}$
RS2392477	7	36985541	*ELMO1*	0.35	-0.46	0.63	0.10	$9.57\ \mathrm{x}10^{-6}$
RS10233832	7	36991362	*ELMO1*	0.34	-0.46	0.63	0.10	$9.60\ \mathrm{x}10^{-6}$
RS12519276	5	37873327	*GDNF-AS1*	0.45	0.40	1.49	0.09	$1.11\ \mathrm{x}10^{-5}$
RS3977734	18	27460959	unknown	0.20	0.44	1.56	0.10	$1.32\ \mathrm{x}10^{-5}$

[a] Excluding patients with multiple melanomas, recruited into the study more than 2 years after diagnosis, missing cause of death and those in the test

Chr: Chromosome 9 SNPs highlighted were selected by the penalized Cox regression in Table 3.8

* Information obtained from http://www.ncbi.nlm.nih.gov/SNP/

Table 3.10 Pairwise correlations between 13 SNPs in the training set using Pearson’s correlation (n=1543)

	RS17837209	RS9957831	RS4768090	RS2902554	RS5770310	RS10233832	RS17379771	RS16956192	RS2392477	RS6689263	RS11639902	RS12519276	RS10941528
RS17837209	1	-0.02	0.03	-0.03	-0.01	-0.01	0.02	0.05	0	-0.03	0.03	0.01	0.02
RS9957831		1	0.02	0.96	0.03	-0.04	0.03	0.01	-0.05	-0.03	-0.04	0.08	0.02
RS4768090			1	0.01	0.01	-0.01	0	0.01	-0.01	-0.03	0	-0.04	0.08
RS2902554				1	0.04	-0.03	0.03	0.01	-0.04	-0.03	-0.03	0.08	0.02
RS5770310					1	-0.02	0.03	0.01	-0.02	0	-0.04	0.04	0
RS10233832						1	-0.02	0.01	0.97	0.01	0	0	-0.02
RS17379771							1	-0.03	-0.01	-0.03	-0.03	0.61	0.05
RS16956192								1	0.02	-0.03	0	-0.02	0.01
RS2392477									1	0	0	0.01	-0.03
RS6689263										1	0.03	0.01	-0.02
RS11639902											1	0	-0.02
RS12519276												1	0.07
RS10941528													1

No correlation (|r|<0.1) in black text; Moderate correlation (0.3 ≤|r|< 0.7) in green text; Strong correlation (|r|>0.7) in red text

Table 3.11 Pairwise correlations between the top 20 SNPs in the training set using Pearson's correlation (n=1543)

	RS17837209	RS3825409	RS9957831	RS2902554	RS11617293	RS10502535	RS4768090	RS2863737	RS10880801	RS4576884	RS2408323	RS16956192	RS9807803	RS17379771	RS5770310	RS17252076	RS2392477	RS10233832	RS12519276	RS3977734
RS17837209	1	0.79	-0.02	-0.03	0.8	-0.04	0.03	-0.03	0.06	0.06	0.03	0.05	-0.02	0.02	-0.01	0.67	0	-0.01	0.01	-0.02
RS3825409		1	-0.01	-0.02	0.98	-0.03	0.02	-0.02	0.06	0.06	0.01	0.03	-0.01	0.01	-0.01	0.81	0	-0.01	0	-0.02
RS9957831			1	0.96	-0.01	0.93	0.02	0.95	0.02	0.02	0.01	0.01	0.92	0.03	0.03	0	-0.05	-0.04	0.08	0.91
RS2902554				1	-0.02	0.95	0.01	0.93	0.01	0.01	0	0.01	0.88	0.03	0.04	-0.01	-0.04	-0.03	0.08	0.89
RS11617293					1	-0.03	0.02	-0.02	0.06	0.06	0.01	0.03	-0.01	0.01	0	0.8	0	0	0	-0.02
RS10502535						1	0.01	0.97	0.01	0.01	0	0.01	0.85	0.02	0.03	-0.01	-0.03	-0.02	0.07	0.86
RS4768090							1	0.02	0.81	0.81	0.99	0.01	0.03	0	0.01	0.01	-0.01	-0.01	-0.04	0.03
RS2863737								1	0.01	0.01	0.01	0.01	0.87	0.02	0.03	-0.02	-0.04	-0.04	0.07	0.87
RS10880801									1	1	0.81	0	0.03	0.04	0	0.06	-0.01	-0.01	0	0.03
RS4576884										1	0.81	0	0.03	0.04	0	0.06	-0.01	-0.01	0	0.03
RS2408323											1	0	0.03	0	0.01	0.01	-0.01	-0.01	-0.04	0.02
RS16956192												1	0.02	-0.03	0.01	-0.01	0.02	0.01	-0.02	0.02
RS9807803													1	0.04	0.03	0	-0.05	-0.04	0.08	0.99
RS17379771														1	0.03	0.03	-0.01	-0.02	0.61	0.04
RS5770310															1	-0.01	-0.02	-0.02	0.04	0.03
RS17252076																1	0.01	0	0.04	0
RS2392477																	1	0.97	0.01	-0.05
RS10233832																		1	0	-0.04
RS12519276																			1	0.08
RS3977734																				1

No correlation (|r|<0.1) in black text; Moderate correlation (0.3 ≤|r|< 0.7) in green text; Strong correlation (|r|>0.7) in red text

3.4 Discussion

3.4.1 Clinical predictors and survival

Results in the multivariable analyses are consistent with the literature on clinical predictors of melanoma survival, which include age, sex, tumour site, tumour type, Breslow thickness, presence of ulceration, and mitotic rate (Thorn *et al*., 1994; Lindholm *et al*., 2004; Buettner *et al*., 2005; Balch *et al*., 2009). The evidence for the association between presence of TILs and survival varies in the literature for primary melanoma. The main reason for this inconsistency is thought to be due to inter-observer variability between the pathologists that assessed the tumour slide (Thomas *et al*., 2013). For the presence of vascular infiltration, most evidence of association with survival has been observed in metastatic melanoma.

Only five clinical predictors (age, sex, tumour site, Breslow thickness, and presence of ulceration) were eventually considered for inclusion in the multivariable analysis of all predictors which will be used to build prognostic prediction models for melanoma survival in Chapter 6. Although mitotic rate was highly significant in univariable analysis and after adjusting for the three established predictors, it was not included for further analysis as data is unavailable for many individuals in the test set. In the AJCC staging system, mitotic rate information is used mainly for patients with thin tumours (less than 1 per mm^2) to determine whether the patient is in stage IA or IB (Balch *et al*., 2009). As the LMC is enriched for cases with thicker tumours, mitotic rate is less relevant than Breslow thickness and presence of ulceration for patients in this cohort. Presence of TILs was also not included for further analysis because predictors subject to inter-observer variability are not suitable for building a prognostic prediction model, as they may give rise to different predictive ability when tested in new data or future individuals (Moons *et al*., 2012).

3.4.2 Gene expression levels and survival

Until recently, due to limitations in obtaining fresh tissue for microarray profiling analysis, there are not many studies reporting gene expression signatures for survival outcomes in primary melanoma. As reviewed in Chapter

1 (§ 1.2.2), only a few studies have conducted gene expression profiling in primary melanoma to identify new prognostic biomarkers for melanoma outcomes (Winnepennincxk *et al.*, 2006; Conway *et al*., 2009; Jewell *et al*., 2010; Sivendran *et al*., 2014; Gerami *et al*., 2015) or to identify molecular classifications for melanoma which may be related to prognosis (Harbst *et al*., 2012; Nsengimana *et al*., 2015).

In metastatic melanoma, gene expression profiling was used to predict survival outcomes (Mandruzzato *et al*., 2006; John *et al.*, 2008; Bogunovic *et al*. 2009; Mann *et al*., 2013; Cirenajwis *et al*., 2015), tumour progression and metastasis (Haqq *et al*., 2005; Jaegar *et al*., 2007; Riker *et al*., 2008), and immunotherapy response (Johnson *et al*., 2015). A cross-validation of gene expression signatures from different studies (Winnepennincxk *et al.*, 2006; John *et al*, 2008; Bogunovic *et al*. 2009; Conway *et al*., 2009; Jonsson *et al*., 2010) by Schramm *et al.* (2012) found that most of the studies contain immune-related genes in their gene signatures, which suggests that immune-related genes may have an important role in melanoma progression and outcome.

The penalized Cox regression analysis using whole-genome gene expression data in this chapter selected 16 gene expression levels that are associated with MSS (*LPAR1*, *ZNF697*, *DLG1*, *NKD2*, *C1R*, *OSTC*, *SNORA12*, *C21ORF63*, *HLA-DQB2*, *SEC22B*, *HLA-B*, *NDUFA8*, *IGSF5*, *FGF22*, *CHST9,* and *CIAPIN1*), of which three were immune-related genes (*C1R*, *HLA-DQB2, and HLA-B)*. Descriptions of the selected genes are shown in Table 3.12. Little is known about the effect of these genes in cancer progression, but expression levels of three genes (*HLA-B*, *CIAPIN1*, and *NKD2*) were reported to be associated with prognosis in other cancers.

HLA-B is involved in immunity and its association with uveal melanoma has been reported by Blom *et al.* (1997)**.** Their study reported that low expression of *HLA-A* and *-B* was associated with improved survival in uveal melanoma, however, this has not been validated in a larger sample.

Cytokine-induced anti-apoptosis molecule 1 (*CIAPIN1*) is a newly identified apoptosis-related molecule that has been shown to be a mediator of the RAS signalling pathway. *CIAPIN1* expression has been reported to be associated with several cancers such as gastric cancer, liver cancer, esophageal cancer, lung cancer, lymphoma and kidney cancer and it was suggested that *CIAPIN1* might involved in tumorigenicity and carcinogenesis. Shi *et al.* (2010) reported that high expression level of *CIAPIN1* in tumour tissue was associated with longer survival in 273 patients with colorectal cancer (P = 0.0002 from Kaplan-Meier survival analysis). In multivariable analysis, expression of *CIAPIN1* remained as prognostic factors for colorectal cancer survival alongside cancer stage, distant organ metastasis, regional lymph node metastasis and local recurrence. Chen *et al.* (2012) examined the association of the expression of *CIAPIN1* in tumor tissue with survival in patients with pancreatic cancer and found that loss of *CIAPIN1* expression directly correlated with decreased survival. While increased expression of CIAPIN1 in colorectal and pancreatic cancer seems to improved survival, the opposite effect was observed for melanoma survival in this study (Table 3.4).

Naked cuticle homolog 2 (*NKD2*) is involved in the Wnt signalling pathway. In a study by Zhao *et al.* (2015), they identified *NKD2* as a novel suppressor of osteosarcoma tumor growth and metastasis in both mouse and human osteosarcoma. They identified downregulation of *NKD2* in metastatic osteosarcoma cells and re-expression of *NKD2* correlated with downregulation of signalling pathways that drive cell motility, angiogenesis and growth signalling.

Table 3.12 Description of the genes selected by penalized Cox model

Gene	Description	Gene type	Biological process/molecular functions	Tissue relevance
LPAR1	Lysophosphatidic acid receptor 1	Known protein coding	Activation of MAPK activity; G-protein coupled receptor signalling pathway	Unknown
ZNF697	Zinc finger protein 697	Known protein coding	Regulation of transcription; DNA template	Unknown
DLG1	Disc, large homolog 1 (Drosophila)	Known protein coding	Regulation of transcription from RNA polymerase II promoter	Increased expression in melanoma
NKD2	Naked cuticle homolog 2	Known protein coding	Wnt signalling pathway; calcium ion bonding; protein binding; growth factor binding	
C1R	Complement component 1	Known protein coding	Immune response; proteolysis; complement activation	Unknown
OSTC	Oligosaccharyltransferase complex subunit	Known protein coding	Protein binding	Unknown
SNORA12	Small nucleolar RNA	Non-coding RNAs	Unknown	Unknown
C21ORF63	Unknown	Unknown	Unknown	Unknown
HLA-DQB2	Major histocompatibility complex, class II, DQ beta 2	Known protein coding	Regulation of immune response; antigen binding; T cell receptor signalling pathway; interferon-gamma-mediated signalling pathway; MHC class II receptor activity; antigen binding	Unknown
SEC22B	SEC22 homolog B, vesicle trafficking protein	Known protein coding	Protein transport; Protein binding	Unknown
HLA-B	Major histocompatibility complex, class I, B	Known protein coding	Regulation of immune response; interferon-gamma-mediated signalling pathway; antigen binding	Unknown
NDUFA8	NADH: ubiquinone oxidoreductase subunit A8	Known protein coding	NADH dehydrogenase (ubiquinone) activity; protein complex binding	Unknown
IGSF5/ JAM4	Immunoglobulin superfamily member 5	Known protein coding	Protein binding	Unknown

FGF22	Fibroblast growth factors 22: large family of 22 signalling molecules	Known protein coding	Responsible for regulating a range of cellular processes including proliferation, survival, migration, differentiation and response to injury	Increased expression in stroma cells
CHST9	Carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9	Known protein coding	Carbohydrate metabolic process	Unknown
CIAPIN1	Cytokine induced apoptosis inhibitor 1	Known protein coding	Apoptotic process; methylation; protein binding; methyltransferase activity; electron carrier activity; ion binding	Unknown

Source: www.ensembl.org

3.4.3 Genetic variants and survival

The penalized Cox model using SNP data selected 13 SNPs that are important for MSS. The association of these SNPs with melanoma or other cancers has never been reported before and there is little known about the selected SNPs in the literature. The selected SNPs however, might just be tagging another causal variant in LD. Therefore, the gene that is located nearest to the SNP is more relevant than the selected SNPs. Table 3.13 shows the gene that is located nearest to the selected SNPs and its functions. Of the 13 SNPs, only nine were located near to a gene. However, the associations of the nine genes with melanoma has never been reported before.

To date, no genome-wide study of melanoma survival has been published. Several candidate gene studies, as reviewed in Chapter 1 (§ 1.2.3), have reported significant associations between melanoma survival and genetic variants such as *MC1R* variants (Davies *et al*., 2012; Taylor *et al*., 2015a), vitamin D receptor variants (Davies *et al*., 2014a; Orlow *et al*., 2016), and *PARP1* variants (Davies *et al*., 2014b; Law *et al*., 2015b). There were also other significant associations reported for other variants, such as variants in angiogenesis (development of new blood vessels) and lymphangiogenesis (development of new lymphatic vessels) genes (Park *et al*., 2013), variants in nucleotide excision repair genes (Li *et al*., 2013), and variants in fanconi anemia pathway genes (Yin *et al*., 2015), but these have not been replicated.

Table 3.13 Gene located nearest to the selected SNPs and its functions

SNP	Functional class of the SNP	Chr	Gene	Biological process/molecular functions of the gene
RS17837209	Intron variant	13	*FAM124A*	Protein binding
RS9957831	unknown	18	unknown	unknown
RS4768090	Intron variant	12	*LOC105369743*	unknown
RS2902554	unknown	18	unknown	unknown
RS5770310	unknown	22	unknown	unknown
RS10233832	Intron variant	7	*ELMO1*	Protein binding
RS17379771	Utr variant 3 prime	5	*GDNF*	Protein binding; transport activity; growth factor activity
RS16956192	Intron variant	17	*SLC13A5*	Transporter activity
RS2392477	Intron variant	7	*ELMO1*	Protein binding
RS6689263	unknown	1	unknown	unknown
RS11639902	Intron variant	16	*PHLPP2*	Epidermal growth factor receptor signalling pathway; protein binding
RS12519276	Intron variant	5	*GDNF-AS1*	unknown
RS10941528	Intron variant	5	*C7*	Complement activation; protein binding

Chr: Chromosome
Source: http://www.ncbi.nlm.nih.gov/SNP/

In summary, most significant -omic (gene expression and SNP) predictors for melanoma survival identified in this chapter have not been reported in the literature before. The selected gene expression levels are highly significantly associated with MSS in the training set; however, the selected SNPs did not meet genome-wide significance. The significant predictors from different types of data will be combined to determine the combined effect of clinical and -omics data on MSS and to build prognostic prediction models for melanoma survival in Chapter 6.

Chapter 4 Heritability analysis

The aims in this chapter are to:

i. Estimate the heritability of survival from melanoma

ii. Estimate the heritability of Breslow thickness

4.1 Introduction

4.1.1 Heritability analysis

Heritability summarizes how much of the variation in a trait is due to variation in genetic factors. High heritability implies strong resemblance between genetically related people for a specific trait, whereas, low heritability implies a low level of resemblance. The total phenotypic variance (V_P) of a trait can be partitioned into two components, the genetic variance (V_A) and the residual variance (V_R) which includes environmental effects. Heritability in this chapter will refer to the narrow-sense heritability (h^2), which is the ratio of additive genetic to the total phenotypic variance: $h^2=V_A/V_P$ (Visscher *et al*., 2008).

Heritability can be estimated from related and unrelated individuals. Classical methods for estimating heritability in human traits are based on studying twins or families. In twin studies for example, heritability can be estimated by comparing the concordance rates of monozygotic and dizygotic twin pairs. If more closely related individuals are more similar for the trait under study, this demonstrates evidence of heritability (Thomas, 2004). Recently developed methods can estimate heritability from genome-wide genotype data in seemingly unrelated individuals (Yang *et al*., 2011). Relatedness can be estimated based on genetic similarities between individuals and measured by the kinship coefficient. Yang *et al*. (2011) developed a tool called genome-wide complex trait analysis (GCTA) that can be used to estimate heritability in unrelated individuals based on SNPs data (see §4.2.5). The GCTA tool estimates the variation explained by common SNPs, though is likely to underestimate the overall heritability as it is very unlikely that all the causal variants are all captured by the SNPs used in GWAS particularly those with low MAF.

4.1.2 Survival from melanoma

The strongest known influences on survival from melanoma are tumour characteristics such as tumour thickness, presence of ulceration and mitotic rate. In addition, demographic variables such as age at diagnosis and sex also influence a patient's survival (Balch *et al*., 2009). This chapter aims to determine whether genetic variants are also related to variation in survival from melanoma.

4.1.3 Breslow thickness

Breslow thickness is a measure of tumour thickness in melanoma. It is measured in millimetres from the surface of the melanoma to the deepest point where the tumour penetrates the skin layers (Breslow, 1970). Breslow thickness is one of the most important prognostic factors in melanoma; individuals having a thinner tumour at diagnosis have better prognosis compared to those with a thicker tumour (Balch *et al*., 2009).

Several factors influence Breslow thickness at diagnosis such as awareness of risk in the population, patients' behavioural characteristics, and how fast the tumour is growing. In a population with high incidence of melanoma where awareness of risk is high such as in Australia, people tend to be diagnosed early and present with thinner tumours at diagnosis (Baade *et al*., 2012). At diagnosis, men tend to have thicker tumours compared to women (Balch *et al.,* 2009). Further factors that may influence Breslow thickness are gender differences in patients' behavioural characteristics in using health screening services and seeking medical care, which may influence why there is a difference in Breslow thickness between males and females. Another factor is how fast the tumour is growing, which could be under genetic control. In a study by Liu *et al*. (2006) which assessed melanoma rate of growth, they identified patients with thicker tumours to have faster growing melanomas. However, no study so far has shown evidence that genetics influence tumour thickness. Hence, this chapter aims to determine how much genetic factors contribute to the variation in Breslow thickness at presentation between individuals.

4.2 Methods

4.2.1 Samples

Samples used for analysis in this chapter are from the LMC as described in Chapter 2. For heritability analysis of Breslow thickness, two additional cohorts from Cambridge and Houston were used (see Chapter 2).

4.2.2 Phenotype data

4.2.2.1 Survival from melanoma

Survival from melanoma was analysed as a dichotomous trait. Samples were treated as a case-control study to carry out the heritability analysis where cases are those who have died from melanoma within a specific time period and controls are those who survived for at least the same length of time. Samples in the cohort were followed-up passively after recruitment into the study as described in Chapter 2. The research team receive information about patient deaths when the ONS send updated information every four months; otherwise patients are still alive until last follow-up.

Survival data were available for 2184 samples in the LMC. Of these samples, 1907 individuals had genotype data for analysis, and 5 individuals had unknown cause of death. Individuals with missing melanoma survival status were excluded from the analysis leaving 1902 samples. The number of individuals who were alive and the number who had died from melanoma were calculated for different lengths of follow-up period to decide on appropriate lengths of follow-up time to observe reasonable numbers in each category.

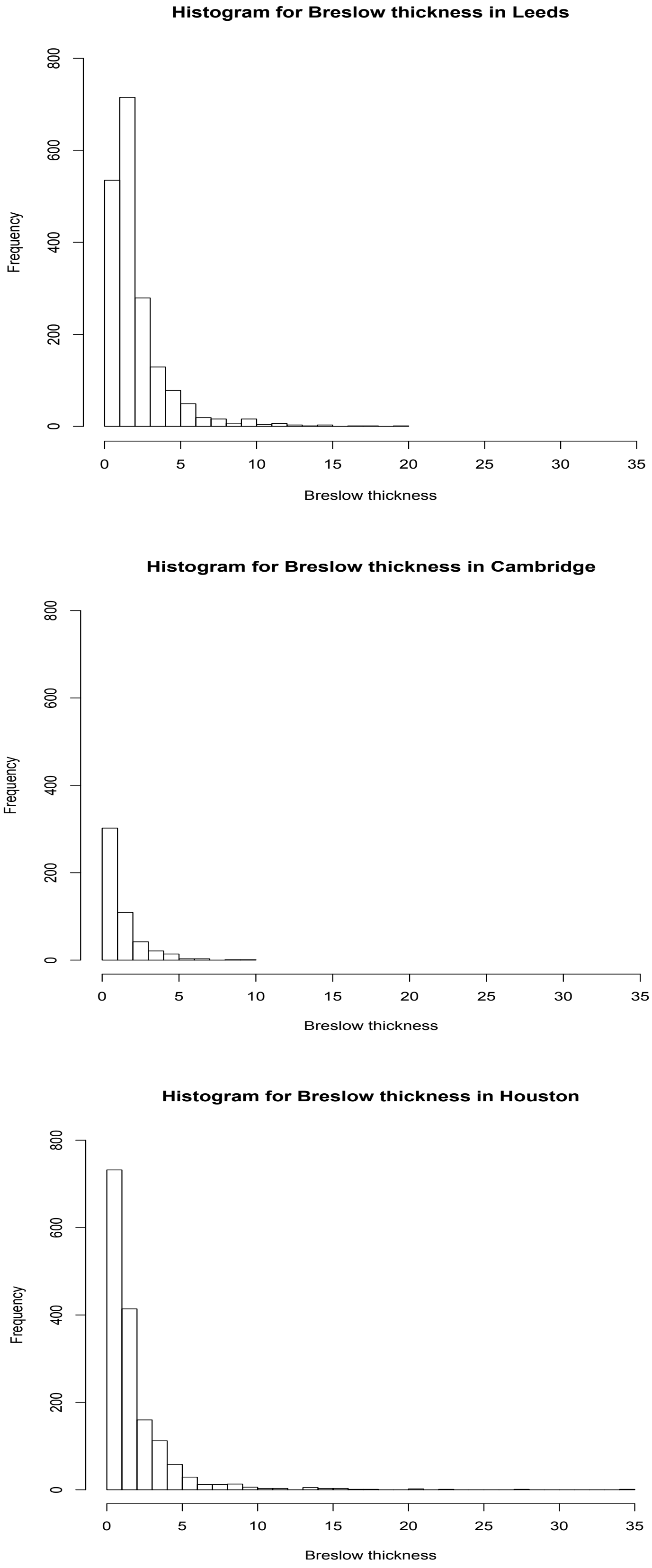
From Table 4.1, 5-year and 10-year follow-up periods were chosen as the cut-off times to determine the numbers of individuals who survived and died from melanoma. The 5-year follow-up time was chosen as 79% of the cohort had been followed up for this time and 237 deaths had occurred within this period. As there were not many individuals who had been followed up for more than 10 years, this time was chosen as the second cut-off time, to maximize the number of individuals who had died while still retaining 44% of the cohort in the analysis.

Table 4.1 Number of patients who are known to have survived or died from melanoma by time in the LMC (n=1902)

Time	Number of patients known to be alive at that time	Number of patients known to have died by that time	Total followed up for that time
5 years	**1257**	**237**	**1494**
6 years	1081	277	1358
7 years	899	291	1190
8 years	764	301	1065
9 years	633	305	938
10 years	**517**	**313**	**830**
11 years	397	316	713
12 years	214	318	532
13 years	74	322	396
14 years	9	324	333

4.2.2.2 Breslow thickness

Breslow thickness was analysed as a quantitative trait. A total of 2129 samples from Leeds, 496 samples from Cambridge and 1572 samples from Houston have Breslow thickness information. However, only 1858, 494 and 1552 samples from Leeds, Cambridge and Houston, respectively had both Breslow thickness and genotype data, so could be included in the analysis. Figure 4.1 shows that the Breslow thickness distributions in the Leeds and Houston cohorts look quite similar, whereas patients in the Cambridge cohort have fewer thicker tumours. As the distributions in all cohorts were skewed, Breslow thickness was log-transformed to make the distribution more normally distributed, and the log-transformed values were then used in the heritability estimation.



Histogram for Breslow thickness in Leeds

Histogram for Breslow thickness in Cambridge

Histogram for Breslow thickness in Houston

Figure 4.1 Histogram for Breslow thickness distribution in Leeds, Houston, and Cambridge

4.2.3 Genotype data

Genotype data used in this chapter were as described in Chapter 2. Genotype data for samples from the Cambridge and Houston cohorts were extracted for the subset of samples with Breslow thickness data only. After dropping samples according to the exclusion criteria described in Chapter 2, the number of remaining samples eligible for analysis in the Leeds, Cambridge and Houston cohorts were 1907, 494 and 1522, respectively.

4.2.4 QC measures

QC was performed in PLINK as described in Chapter 2. After QC, SNP thinning was also performed in PLINK to exclude SNPs in high LD using a cut-off of $r^2 < 0.2$. LD thresholds that have been used in other studies of heritability studies range from 0.1 to 0.3 (Ehret *et al.*, 2012; Taylor *et al.*, 2015b).

Speed *et al.* (2012) pointed out that LD patterns contribute to SNP-based h^2 estimate; causal variants in regions of strong LD tend contribute to overestimation of h^2, whereas those in low LD contribute to underestimation. Using a very high r^2 threshold retained more SNPs in LD, and h^2 could be overestimated, while using a very low r^2 threshold could remove SNPs with non-redundant information, and h^2 could be underestimated. Therefore, an r^2 threshold of 0.2 was chosen in this analysis as it could provide a balance of these problems. After QC and SNP thinning, there were 92,968 SNPs retained in the Leeds cohort, 92,801 SNPs retained in the Cambridge cohort, and 100,119 SNPs retained in the Houston cohort. Only autosomal SNPs were used in the heritability analysis (Table 4.2).

Table 4.2 QC for samples with genotype data in the Leeds, Cambridge and Houston cohort

	Leeds cohort	Cambridge cohort	Houston cohort
No. of samples	1907	494	1522
No. of SNPs in genotype data	880,209	914,986	1,012,457
No. of SNPs after QC (Missing rate <3%, HWE test P-value $>10^{-4}$ and MAF >5%)	556,632	578,587	718,174
No. of SNPs after SNP thinning ($r^2<0.2$)	92,968	92,801	100,119
No. of autosomal SNPs	92,781	92,615	99,801

4.2.5 Statistical analysis

After QC and SNP thinning, the GCTA tool by Yang *et al*. (2011) was used to estimate the proportion of variation explained by all SNPs for survival from melanoma and for Breslow thickness. The method estimates the variance in the trait explained by genotyped SNPs in LD with unknown causal variants. The first step in the GCTA tool is to calculate the pairwise genetic relationships between individuals using autosomal SNPs. This creates a genetic relationship matrix (GRM). The next step is to exclude one of any pair of closely related individuals using a recommended cut-off kinship coefficient of 0.025 for genetic relatedness, which corresponds to cousins two to three times removed. The new GRM is a GRM of unrelated individuals (Yang *et al*., 2011). The last step was to fit a mixed linear model to estimate the phenotypic variance captured by all SNPs in the form of:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{g} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of phenotype values, $\boldsymbol{\beta}$ is a vector of fixed effects of variables $\mathbf{X}$ such as age, and sex, $\mathbf{g}$ is a vector of random genetic effects, and $\boldsymbol{\varepsilon}$ is a vector of residuals. The variance-covariance matrix of $\mathbf{g}$ is $\mathbf{A}\boldsymbol{\sigma}_{\mathbf{g}}^{2}$**,** where **A** is the GRM. The phenotypic variance ($\mathbf{V}$) can be partitioned into the variance explained by the genetic factors $(\boldsymbol{\sigma}_{\mathbf{g}}^{2})$ and the residual variance $(\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{2})$:

$$\mathbf{V} = \mathbf{A}\boldsymbol{\sigma}_{\mathbf{g}}^{2} + \mathbf{I}\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{2}$$

The GCTA outputs give estimates for the genetic variance, $\boldsymbol{\sigma_g^2}$ (shown as V_G in the table of results), residual variance, $\boldsymbol{\sigma_\varepsilon^2}$ (shown as V_e in the table of results), and phenotype variance, $\mathbf{V}$ (shown as V_p in the table of results). The estimate of heritability is calculated as the ratio of genetic variance and phenotype variance ($h^2=V_G/V_p$). A P-value from the likelihood ratio test (LRT) is also produced which tests whether the heritability is greater than zero.

The GCTA tool can also be used to estimate the heritability of a dichotomous trait, such as disease phenotype. For this, the heritability is interpreted on a liability scale. This liability model assumes that case/control status is determined by whether or not an individual's liability, an unobservable normally distributed random variable, lies above or below a threshold. The estimated heritability on the observed binary scale can be transformed to the underlying liability scale by using linear transformations to estimate the proportion of variance on the liability scale. For the linear transformations, the disease prevalence of case/control study needs to be specified in the GCTA (Lee *et al*., 2011).

For the analysis to estimate heritability in survival from melanoma, the survival status was analysed as case-control study where cases are those who have died from melanoma and controls are those who have survived within the chosen time period. Two analyses were conducted to estimate the heritability of survival from melanoma within 5-year and 10-year follow-up periods. The proportion of cases who died from melanoma (17%) in the Leeds cohort was specified to transform the heritability on the observed scale to the liability scale.

For the estimation of heritability in Breslow thickness, two main analyses were performed; separate analyses within the Leeds, Cambridge and Houston cohorts, and a combined cohort analysis using the combined samples from the three cohorts to increase the power to detect heritability. In the combined cohort analysis, only those SNPs common to the three datasets were used. There were 518,421 SNPs common to the three cohorts but only 91,217 SNPs passed the second QC and SNP thinning process, and 91,072 autosomal SNPs were used in the analysis. Heritability estimation in the combined cohort analysis was adjusted for centre as differences between the cohorts could be a potential confounding factor. As age and sex are associated with Breslow thickness (Nagore *et al*., 2006), another analysis further adjusting for age and sex was performed to potentially improve the power to detect heritability by removing non-genetic components of variance. The log-transformed Breslow thickness was adjusted for the effect of age and sex by fitting a linear regression model, then using the residual from the regression analysis as the new adjusted Breslow thickness value in GCTA. The adjusted Breslow thickness value was also used in the individual cohort analysis.

We used the online calculator to calculate the power to detect whether heritability is greater than zero for our data when using the GCTA tool (Visscher *et al*., 2014).

Several subgroup analyses were conducted for the estimation of heritability of Breslow thickness using the combined cohort as follows:

1. Heritability analysis using SNPs from even chromosomes only and SNPs from odd chromosomes only. This analysis was conducted as the initial results showed some evidence of heritability and therefore, it is of interest to partition the heritability into group of chromosomes to determine whether there is a polygenic effect or whether the heritability is clearly largely due to one region
2. Heritability analysis in individuals with thicker tumours only (Breslow thickness > 1.0mm) to determine whether variation in Breslow thickness is more heritable within those with thicker tumours at presentation. Very thin tumours were excluded from this analysis as they may not be informative.
3. Heritability analysis by sex to compare the heritability estimates between males and females.
4. Heritability analysis using different thresholds (r^2=0.3 and r^2=0.4) for SNP thinning to determine their effect on heritability estimates within individual cohorts.

4.3 Results

4.3.1 Heritability of survival from melanoma

Table 4.3 shows the estimate of heritability of 5-year and 10-year survival in melanoma. Of the 1494 individuals that had been followed-up for 5 years, 1257 survived and 237 died from melanoma (Table 4.1). Only 1479 individuals were included in the analysis after excluding 15 individuals who were related to other members of the cohort (kinship coefficient >0.025). Of the 830 individuals with up to 10-years follow-up, 517 survived and 313 died from melanoma. Only 822 individuals were retained for the analysis after removing 8 individuals due to relatedness. The proportion of phenotypic liability variation in 5-year and 10-year survival from melanoma explained by common SNPs was estimated to be 71% and 40%, respectively, but neither estimate was significantly greater than zero. However, this analysis does not have enough power to detect heritability, and thus cannot show clear evidence of heritability in survival from melanoma.

As we don't actually know how much of the variation in survival can be explained by SNPs, we assumed that SNPs could only explain a small proportion of the variations in survival, and calculated the power to detect h^2 if true heritability is 15% in Table 4.3. Lower percentage was used for this calculation as heritability is likely to be underestimated by GCTA.

Table 4.3 Estimation of heritability in survival from melanoma

GCTA outputs	Within 5-years follow-up time	Within 10-years follow-up time
V_G (SE)	0.04 (0.03)	0.07 (0.09)
V_e (SE)	0.09 (0.03)	0.16 (0.09)
V_p (SE)	0.13 (0.01)	0.24 (0.01)
V_G/V_p (SE)	**0.31 (0.24)**	**0.30 (0.42)**
*The estimate of variance explained on the observed scale is transformed to that on the liability scale	**0.71(0.55)**	**0.40(0.56)**
logL	742.429	180.042
logL0	741.596	170.769
LRT	1.67	0.55
df	1	1
P-value	**0.09**	**0.23**
n	1479	822
Power to detect h^2 if true heritability is 15%	6.1%	6.0%

*using linear transformations

4.3.2 Heritability of Breslow thickness

4.3.2.1 Individual cohort analysis

The sample characteristics for the samples included in this analysis are shown in Table 4.4. A total of 1858 individuals (56.9% females) were included in the Leeds cohort with mean age of 55 years (SD=13.3). In the Cambridge cohort, only 494 individuals (47.8% females) had both Breslow thickness and genotype data. The mean age of participants from the Cambridge cohort was similar to the Leeds cohort at about 55 years (SD=10.5). In the Houston cohort, there were 1552 individuals (41.2% females) with both Breslow thickness and genotype data. Samples in the Houston cohort were slightly younger with mean age of 52 years (SD=14.5).

Table 4.4 Characteristics of samples with both Breslow thickness and genotyped data in Leeds, Cambridge and Houston cohorts

	Leeds cohort (n=1858)	Cambridge cohort (n=494)	Houston cohort (n=1552)
Age, years			
Mean (SD)	54.6 (13.3)	55.4 (10.5)	52.2 (14.5)
Min, Max	17, 90	22, 69	16.1, 94.1
Sex			
Male, n (%)	801 (43.1)	258 (52.2)	912 (58.8)
Female, n (%)	1057 (56.9)	236 (47.8)	640 (41.2)
Breslow thickness, mm Median (Range)	1.5 (0.2 – 20)	0.8 (0.03 – 9.4)	1.1 (0.1 – 35)
Breslow thickness within male Median (Range)	1.7 (0.2 – 17)	0.8 (0.03 – 6.5)	1.3 (0.1 – 28)
Breslow thickness within female Median (Range)	1.4 (0.2 – 20)	0.9 (0.2 – 9.4)	0.93 (0.13 – 35)

The numbers of individuals included in the analysis after removing related individuals were 1839, 493 and 1510 in the Leeds, Cambridge, and Houston cohorts, respectively. The estimated proportion of variance in Breslow thickness explained by common SNPs was 32% (SE=0.19) in the Leeds cohort, 0% (SE=0.71) in the Cambridge cohort, and 30% (SE=0.24) in the Houston cohort (Table 4.5). At a 5% significance level, the heritability estimate was significantly greater than zero only in the Leeds cohort.

Exploration of the relationship of Breslow thickness with age and sex showed strong association of Breslow thickness with these factors within each cohort (Table 4.6). Breslow thickness was adjusted for age and sex within each cohort and residual values were used to estimate the heritability. The estimated proportion of variance in adjusted Breslow thickness that can be explained by common SNPs was 35% (SE=0.19) in the Leeds cohort, 0% (SE=0.70) in the Cambridge cohort, and 33% (SE=0.24) in the Houston cohort (Table 4.7). The heritability estimates were higher for the adjusted Breslow thickness in the Leeds and Houston cohorts compared to the unadjusted estimates as expected, as adjusting for age and sex removed some of the variability due to non-genetic factors. To calculate the power to detect h^2 of Breslow thickness in this analysis, we also assumed a lower true heritability and used 15% true heritability in the calculation.

Table 4.5 Estimation of heritability in Breslow thickness in individual cohort analysis (Using SNP thinning r^2<0.2)

GCTA outputs	Leeds cohort	Cambridge cohort	Houston cohort
V_G (SE)	0.17 (0.1)	0.00 (0.49)	0.26 (0.21)
V_e (SE)	0.36 (0.1)	0.70 (0.49)	0.60 (0.21)
V_p (SE)	0.54 (0.02)	0.69 (0.04)	0.86 (0.03)
V_G/V_p (SE)	**0.32 (0.19)**	**0.00 (0.71)**	**0.30 (0.24)**
logL	-348.529	-161.146	-646.421
logL0	-350.120	-161.145	-647.215
LRT	3.180	0.000	1.588
df	1	1	1
P-value	**0.037**	**0.50**	**0.10**
n	1839[a]	493[b]	1510[c]
Power to detect h^2 if true heritability is 15%	14%	5%	11%

[a] 19 individuals excluded from the genotype data in Leeds, leaving only 1888 samples for analysis. 1839 samples had both Breslow thickness and genotype information.

[b] 1 individual excluded from the genotype data in Cambridge, leaving only 493 samples for analysis.

[c] 42 individuals excluded from the genotype data in Houston, leaving only 1510 samples for analysis.

Table 4.6 Association of log Breslow thickness with age and sex in each cohort

	Leeds cohort (n=1858)		Cambridge cohort (n=494)		Houston cohort (n=1552)	
	β (95%CI)	P-value*	β (95%CI)	P-value*	β (95%CI)	P-value*
Age	0.01 (0.008, 0.013)	$2x10^{-16}$	0.01 (0.003, 0.02)	$5.41x10^{-3}$	0.03 (0.02, 0.04)	$2x10^{-16}$
Sex (Female)	-0.13 (-0.19, -0.06)	$1.87x10^{-4}$	-0.27 (-0.42, -0.13)	$2.75x10^{-4}$	-0.32 (-0.41, -0.22)	$3.19x10^{-11}$

*using simple linear regression

Table 4.7 Estimation of heritability in adjusted Breslow thickness in individual cohort analysis (Using SNP thinning r^2<0.2)

GCTA outputs	Leeds cohort	Cambridge cohort	Houston cohort
V_G (SE)	0.18 (0.10)	0.00 (0.47)	0.26 (0.19)
V_e (SE)	0.34 (0.09)	0.67 (0.47)	0.54 (0.19)
V_p (SE)	0.51 (0.02)	0.67 (0.04)	0.80 (0.03)
V_G/V_p (SE)	**0.35 (0.19)**	**0.00 (0.70)**	**0.33 (0.24)**
logL	-308.400	-152.721	-588.643
logL0	-310.277	-152.719	-589.556
LRT	3.754	0.000	1.827
df	1	1	1
P-value	**0.026**	**0.50**	**0.09**
n	1839	493	1510

4.3.2.2 Combined cohort analysis

For the combined cohort analysis, the three cohorts provided 3904 individuals with both Breslow thickness and genotype data; 71 related individuals were removed, leaving 3833 individuals for the analysis. In the combined cohort, common SNPs explained about 21% (SE=0.09) of the variation in Breslow thickness when adjusted for centre only and the estimate reduced to 18% (SE=0.09) when further adjusted for age and sex effect (Table 4.8). Both estimates show some evidence of heritability in Breslow thickness (P-value=0.01 and P-value=0.03).

Table 4.8 Estimation of heritability in Breslow thickness in combined cohort analysis (Using SNP thinning r^2<0.2)

GCTA outputs	Adjusted for center	Adjusted for age, sex and center
V_G (SE)	0.15 (0.06)	0.11 (0.06)
V_e (SE)	0.54 (0.06)	0.53 (0.06)
V_p (SE)	0.68 (0.02)	0.65 (0.01)
V_G/V_p (SE)	**0.21 (0.09)**	**0.18 (0.09)**
logL	-1198.264	-1088.512
logL0	-1200.896	-1090.274
LRT	5.263	3.523
df	1	1
P-value	**0.01**	**0.03**
n	3833	3833
Power to detect h^2 if true heritability is 15%	44.4%	

4.3.2.3 Heritability analysis by groups of chromosomes

The proportion of variance in Breslow thickness than can be explained by common SNPs from odd numbered chromosomes was 8% (SE=0.06) and 11% (SE=0.06) from even numbered chromosomes (Table 4.9). Although only the heritability estimate from even chromosomes was significantly greater than zero, the estimates were similar.

Table 4.9 Estimation of heritability in Breslow thickness by groups of chromosomes

GCTA outputs	Using SNPs from odd chromosomes only	Using SNPs from even chromosomes only
V_G (SE)	0.06 (0.04)	0.08 (0.04)
V_e (SE)	0.63 (0.04)	0.61 (0.05)
V_p (SE)	0.68 (0.02)	0.68 (0.02)
V_G/V_p (SE) (Adjusted for centre)	**0.08 (0.06)**	**0.11 (0.06)**
logL	-1200.0611	-1199.321
logL0	-1200.896	-1200.896
LRT	1.67	3.15
df	1	1
P-value	**0.09**	**0.04**
n	3833	3833

4.3.2.4 Heritability analysis in individuals with thicker tumours

There were 2526 individuals with Breslow thickness > 1.0mm in the combined cohort but 45 closely related individuals were excluded. A total of 1378 individuals (423 from Leeds, 280 from Cambridge, and 675 from Houston) with Breslow thickness < 1.0mm were excluded from the analysis. Table 4.10 shows the proportion of variance in Breslow thickness of > 1.0mm that can be explained by common SNPs was 21% (SE=0.14), which was similar to the estimate from the overall combined results (Table 4.8).

Table 4.10 Estimation of heritability in Breslow thickness in individuals with thicker tumours in the combined cohort analysis

GCTA outputs	Individuals with Breslow thickness > 1.0mm
V_G (SE)	0.08 (0.05)
V_e (SE)	0.29 (0.05)
V_p (SE)	0.37 (0.01)
V_G/V_p (SE) (Adjusted for centre)	**0.21 (0.14)**
logL	-7.030
logL0	-8.083
LRT	2.107
df	1
P-value	**0.07**
n	2481
Power to detect h^2 if true heritability is 15%	21.8%

4.3.2.5 Heritability analysis by sex

There were 1951 females (excluding 4 individuals with extreme Breslow thickness value) and 1947 males (excluding 2 individuals with extreme Breslow thickness value) in the combined cohort samples. After removing closely related individuals (32 females and 39 males), only 1919 females and 1908 males were retained for analysis within each group. The proportion of variance in Breslow thickness that can be explained by common SNPs was 25% (SE=0.19) within females and 17% (SE=0.19) within males, although neither estimate is significantly greater than zero (Table 4.11).

Table 4.11 Estimation of heritability in Breslow thickness by sex in the combined cohort analysis

GCTA outputs	Within females	Within males
V_G (SE)	0.15 (0.11)	0.12 (0.13)
V_e (SE)	0.46 (0.11)	0.59 (0.13)
V_p (SE)	0.61 (0.02)	0.71 (0.02)
V_G/V_p (SE) (Adjusted for centre)	**0.25 (0.19)**	**0.17 (0.19)**
logL	-490.072	-631.707
logL0	-490.962	-632.106
LRT	1.78	0.83
df	1	1
P-value	**0.09**	**0.18**
n	1919	1908
Power to detect h^2 if true heritability is 15%	14.9%	14.8%

4.3.2.6 Heritability analysis using different r^2 threshold for SNP pruning in individual cohort

The heritability estimates when using different LD thresholds for SNP thinning are shown in Table 4.12 to 4.14 for each cohort. In the Leeds cohort, all heritability estimates using different r^2 thresholds were significant, with higher estimates when using higher r^2 thresholds. In the Cambridge cohort, using different r^2 thresholds does not make any difference to estimates as the analysis is under-powered to detect heritability in this cohort. In the Houston cohort, using higher r^2 thresholds results in lower heritability estimates but were not significant; these results were not as expected, but this could be due to the relatively small sample size and consequent instability of estimates.

Table 4.12 Estimation of heritability in Breslow thickness in Leeds cohort using different SNP thinning thresholds

GCTA outputs	r^2<0.2	r^2<0.3	r^2<0.4
V_G (SE)	0.17 (0.1)	0.18 (0.1)	0.21 (0.1)
V_e (SE)	0.36 (0.1)	0.35 (0.1)	0.33 (0.1)
V_p (SE)	0.54 (0.02)	0.54 (0.02)	0.54 (0.02)
V_G/V_p (SE)	**0.32 (0.19)**	**0.34 (0.20)**	**0.39 (0.20)**
logL	-348.529	-348.290	-348.011
logL0	-350.120	-349.812	-349.812
LRT	3.180	3.043	3.602
df	1	1	1
P-value	**0.037**	**0.040**	**0.029**
n	1839	1838	1838

Table 4.13 Estimation of heritability in Breslow thickness in Cambridge cohort using different SNP thinning thresholds

GCTA outputs	$r^2<0.2$	$r^2<0.3$	$r^2<0.4$
V_G (SE)	0.00 (0.49)	0.00 (0.53)	0.00 (0.54)
V_e (SE)	0.70 (0.49)	0.70 (0.53)	0.70 (0.53)
V_p (SE)	0.69 (0.04)	0.69 (0.04)	0.69 (0.04)
V_G/V_p (SE)	**0.00 (0.71)**	**0.00 (0.76)**	**0.00 (0.77)**
logL	-161.146	-161.145	-161.145
logL0	-161.145	-161.145	-161.145
LRT	0.000	0.000	0.000
df	1	1	1
P-value	**0.50**	**0.50**	**0.50**
n	493	493	493

Table 4.14 Estimation of heritability in Breslow thickness in Houston cohort using different SNP thinning thresholds

GCTA outputs	$r^2<0.2$	$r^2<0.3$	$r^2<0.4$
V_G (SE)	0.26 (0.21)	0.17 (0.23)	0.07 (0.23)
V_e (SE)	0.60 (0.21)	0.69 (0.23)	0.79 (0.23)
V_p (SE)	0.86 (0.03)	0.86 (0.03)	0.86 (0.03)
V_G/V_p (SE)	**0.30 (0.24)**	**0.20 (0.26)**	**0.08 (0.27)**
logL	-646.421	-647.249	-647.596
logL0	-647.215	-647.531	-647.643
LRT	1.588	0.565	0.094
df	1	1	1
P-value	**0.10**	**0.22**	**0.38**
n	1510	1510	1510

4.4 Discussion

4.4.1 Heritability of survival from melanoma

The total proportion of liability variation explained by common SNPs for 5-year and 10-year survival estimated in the Leeds cohort were 71% (SE=0.55, P-value=0.09) and 40% (SE=0.56, P-value=0.23), respectively. However, the analysis does not have adequate power to detect heritability, and the results do not provide clear evidence of heritability in survival from melanoma. To date, there is no published evidence for association of common SNPs with melanoma survival from genome-wide analysis perhaps because a large sample size is required to identify genome-wide significant SNPs. As a result, most of the published studies were based on candidate genes only such as *PARP1* (Davies *et al*., 2014b; Law *et al*., 2015a) and *MC1R* (Davies *et al*., 2012; Taylor *et al*., 2015a), and some of the studies focus primarily on the effect of melanoma risk loci on melanoma survival (Rendleman *et al*., 2013). Results from these studies suggest that inherited variants may have a role in melanoma survival, but so far none have explained much of the variation in survival from melanoma.

Since survival from melanoma is strongly influenced by tumour characteristics such as Breslow thickness, ulceration and mitotic rate (Balch *et al*., 2009), which could themselves be under genetic control, and since other non-genetic factors are known to influence survival such as age and sex (Thorn *et al*., 1994; Lindholm *et al*., 2004), there could be stronger evidence of genetic influences on these prognostic factors rather than on survival.

4.4.2 Heritability in Breslow thickness

This study provides heritability estimates for Breslow thickness from different populations. The individual cohort heritability estimates vary from 0% (SE=0.71, P-value=0.5) in Cambridge, 30% (SE=0.24, P-value=0.1) in Houston to 32% (SE=0.19, P-value=0.04) in Leeds. A statistically significant heritability estimate (at the 5% level) was only seen in the Leeds cohort, and this was consistent whether or not adjustment was made for age and sex (unadjusted P-value=0.04 and adjusted P-value=0.03). Results from the individual cohort analyses are limited by the small sample size and hence underpowered to detect heritability, especially in the Cambridge cohort. The estimated powers to detect heritability greater than zero if true heritability assumed to be 15% are 5%, 11% and 14% in Cambridge, Houston and Leeds, respectively.

In addition, differences in the Breslow thickness distribution between the cohorts are likely to contribute to the differences in the heritability estimates. In particular, the Cambridge cohort comprised mostly individuals with thin Breslow thickness. Houston and Leeds cohorts on the other hand had similar Breslow thickness distributions. Therefore, discrepancies in the heritability estimates in the cohort-level analysis could be due to differences between the Breslow thickness distributions of the three cohorts.

In the combined cohort analysis, a statistically significant heritability estimate was seen in both analysis adjusting for centre only (h^2=0.21, SE=0.09, P-value=0.01) and adjusting for centre, age and sex (h^2=0.18, SE=0.09, P-value=0.03). Although further adjustment by age and sex did not increase the heritability estimate as might be expected, the significant estimate provides further support that there is evidence of heritability in Breslow thickness.

To date, there is no published report of heritability of Breslow thickness from twin and family studies for direct comparison with the GCTA estimates. Such studies are not feasible to conduct in melanoma as it is not possible to obtain a large cohort of related individuals all diagnosed with melanoma. The newer GCTA tool based on common SNPs that are in LD with causal SNPs allows heritability to be estimated without using twin or family data. However, heritability estimates from the GCTA are likely to be underestimates of the true narrow-sense heritability, due to imperfect LD between genotyped SNPs and unknown causal variants. Lower heritability estimates from the GCTA compared to twin and family studies have been observed in other traits. For example, in height, the heritability estimate from the GCTA was 45% (Yang *et al.*, 2010), lower than the 80% variance estimated from twin studies (Silventoinen *et al.*, 2003). In pediatric obesity, measured by BMI, the GCTA tool estimated heritability of 30% for BMI, whereas an estimate from standard twin analysis was 82% (Llewellyn *et al.*, 2013).

In the subgroup analysis, no clear conclusion can be made about the heritability estimate by groups of chromosomes, by sex, in individuals with thicker tumours only, and in analyses using different LD thresholds as these analyses were underpowered to detect heritability.

In summary, results in this study are largely limited by the small sample size. No clear conclusion can be made about the evidence of heritability in survival from melanoma. This study however provides reasonable evidence of heritability of Breslow thickness; we estimated that about 21% of the phenotypic variance in Breslow thickness can be explained by common SNPs, and this is likely to be an underestimate of the overall heritability.

Chapter 5 Inter-relationships between different types of predictors

The aims in this chapter are to:

i. Determine the association of selected gene expression levels (and gene expression score) with clinical predictors
ii. Determine the association of selected SNPs (and SNP score) with clinical predictors
iii. Carry out genome-wide association studies of the selected gene expression levels (expression level of 16 genes that associated with MSS)
iv. Determine the association of top melanoma susceptibility SNPs with MSS and clinical predictors
v. Determine the association of top melanoma susceptibility SNPs with the expression levels of nearby genes
vi. Determine the association of other SNPs in the regions around melanoma susceptibility SNPs with the expression levels of nearby genes

5.1 Introduction

In this chapter the inter-relationships between different types of factors associated with melanoma, such as patient and tumour characteristics, gene expression levels, and SNP genotypes, are explored. In the survival analysis of whole-genome gene expression levels with MSS in Chapter 3, 16 gene expression levels were identified as important predictors showing significant association with MSS in the training set (see Table 3.4). However, the role of selected genes was mostly still unclear, as discussed in Chapter 3. Hence, this chapter will explore the association of the 16 selected gene expression levels with clinical predictors to further understand how the selected genes impact on survival (Aim i).

Analysis in Chapter 3 also identified 13 SNPs associated with MSS (see Table 3.8). Similar to selected gene expression levels, little information is available about the role of the selected SNPs in melanoma. Therefore, this chapter will also explore the association of the selected SNPs with clinical predictors (Aim ii).

Many studies have shown the existence of genetic variants that control gene expression in various human cells and tissues, for example, in lymphoblastoid cell lines (LCLs) (Morley *et al*., 2004; Stranger *et al*., 2005; Nica *et al*., 2011; Bryois *et al*., 2014), whole blood (Schramm *et al*., 2014), skin tissue (Nica *et al*., 2011), adipose tissue (Nica *et al*., 2011), type 2 diabetes (Morris *et al*., 2012), and colorectal cancer (Ongen *et al*., 2014). The analysis of eQTL associates genetic variants with gene expression to identify genetic variants that regulate the expression of a particular gene. The regulatory variants could be local, mapping close to the physical location of the affected gene (cis-eQTL), or distant, mapping elsewhere in the genome (trans-eQTL) (Rockman & Kruglyak, 2006). Recently, some studies have also shown that there is a measurable genetic component to the control of gene expression levels. In Wright *et al*. (2014), 777 gene expression levels in the peripheral blood (4.2% of the genes on the microarray) were shown to be heritable with mean heritability of 0.10. In Grundberg *et al*. (2012), the mean heritability estimates for gene expression levels from LCLs, skin, and adipose tissue were 0.21, 0.16, and 0.26 respectively. As the LMC only have gene expression measured on a subset of samples, it would be of interest to determine whether gene expression levels could be predicted by SNPs, as this could provide a proxy measure of the expression levels for patients where these have not been measured but with available genetic data. Therefore, this chapter will carry out GWAS of the expression levels for 16 genes associated with MSS in Chapter 3 to determine whether gene expression levels could be predicted from the genome (Aim iii).

To date, GWAS have successfully identified 20 melanoma susceptibility loci (Table 5.1), of which 15 were established earlier (Bishop *et al*., 2009; Rafnar *et al*., 2009; Barrett *et al*., 2011; MacGregor *et al*., 2011; Iles *et al*., 2013) and five were newly found in a recent large meta-analysis (Law *et al*., 2015a). The mechanism of action for most of the loci is still unclear, although several are in regions related to pigmentation (*SLC45A2, TYR, MC1R,* and *ASIP*), nevus count (*TERT, CDKN2A*, and *PLA2G6*) or DNA repair pathways (*PARP1* and *ATM*). To further understand the possible role of susceptibility loci in MSS, and whether this might act through regulation of gene expression in the tumour, the association of the top SNP reported from each of the loci in Table 5.1 with MSS and clinical predictors (Aim iv), and expression levels of the nearby genes is assessed (Aim v).

GWAS tend to highlight the most significant SNP (often referred as the top SNP) identified in each region; however, it is still unclear whether the identified SNPs are the causal SNPs. Furthermore, Barrett *et al*. (2015) found evidence for multiple independent associations in some of the melanoma susceptibility regions in a fine-mapping study. Therefore, the eQTL analysis between the top SNP and gene expression levels will be extended to include other SNPs in the susceptibility regions (Aim vi).

Six main analyses will be conducted in this chapter: (i) an association analysis for the selected gene expression levels (and gene expression score) with clinical predictors, (ii) an association analysis for the selected SNPs (and SNP score) with clinical predictors, (iii) a genome-wide eQTL analysis for the 16 selected gene expression levels, (iv) an association of top melanoma susceptibility SNPs with MSS and clinical predictors, (v) an eQTL analysis for the top SNP in each melanoma susceptibility locus, and lastly (vi) an extension of the eQTL analysis to include other SNPs in the region within each susceptibility locus.

In §5.2, the datasets used in this chapter and the analyses mentioned above are further described. In §5.3, the results of each analysis are presented, and the findings are discussed in §5.4.

Table 5.1 20 melanoma risk loci

Region	Top SNP	Gene[a]	Putative functional role of the gene in melanoma	Loci status
1q21.3	RS12410869	*ARNT*	Unclear	Established
1q42.12	RS1858550	*PARP1*	DNA maintenance pathways	Established
2p22.2	RS6750047	***RMDN2 (CYP1B1)***	Unknown	New**
2q33-q34	RS7582362	*CASP8*	Unclear	Established
5p15.33	RS380286	*TERT/ CLPTM1L*	Nevus count	Established
5p13.2	RS250417	*SLC45A2*	Pigmentation	Established
6p22.3	RS6914598	***CDKAL1***	Unknown	New**
7p21.1	RS1636744	***AGR3***	Unknown	New**
9p21	RS7852450	*CDKN2A/ MTAP*	Nevus count	Established
9q31.2	RS10739221	***TMEM38B (RAD23B, TAL2)***	Unknown	New**
10q24.33	RS2995264	***OBFC1***	Unknown	New**
11q13	RS498136	*CCND1*	Unclear	Replicated*
11q14-q21	RS1393350	*TYR*	Pigmentation	Established
11q22-q23	RS73008229	*ATM*	Role in DNA maintenance pathways	Established
15q13.1	RS4778138	*OCA2*	Unclear	Replicated*
16q12.2	RS12596638	*FTO*	Unclear	Established
16q24.3	RS75570604	*MC1R*	Pigmentation	Established
20q11.2-q12	RS6088372	*ASIP*	Pigmentation	Established
21q22.3	RS408825	*MX2*	Unclear	Established
22q13.1	RS2092180	*PLA2G6*	Nevus count	Established

[a] The top SNP is in or near the gene, hence the region was given the nearest gene's name for convenience, but it is still unclear whether this is the causal gene

* Replicated in Law *et al*. (2015a)

**New loci identified in Law *et al*. (2015a)

The top SNP given is the most significant SNP from the latest GWAS meta-analysis of melanoma susceptibility in Law *et al.* (2015a)

5.2 Methods

5.2.1 Samples

Samples used in this chapter are those from patients with both gene expression and SNP genotype data available. Of the 699 samples with whole-genome gene expression data available, only 619 have genotype data. The remaining samples have no genotype data for various reasons such as failed genotyping, no blood sample for DNA extraction, failed DNA extraction, and insufficient DNA for genotyping.

5.2.2 Study variables and analyses

The datasets used in this chapter are the Illumina whole-genome gene expression data (n=699) and the genome-wide genotype data (n=1907) as described in Chapter 2. The QC methods for each dataset were as described in Chapter 2. Imputed SNP data were also used as further described in §5.2.2.4. For the clinical predictors, only five established clinical predictors (age, sex, Breslow thickness, presence of ulceration and tumour site) were included in the analysis.

The following section provides more description of the variables and analysis methods used for each analysis in the following order:

(i) Analysis 1: Association analysis of the selected gene expression levels (and gene expression score) with clinical predictors

(ii) Analysis 2: Association analysis of the selected SNPs (and SNP score) with clinical predictors

(iii) Analysis 3: Genome-wide eQTL analysis of the 16 selected gene expression levels

(iv) Analysis 4: Association of top melanoma susceptibility SNPs with MSS and clinical predictors

(v) Analysis 5: eQTL analysis of the top SNP in each melanoma susceptibility locus

(vi) Analysis 6: eQTL analysis including other SNPs in the region within each locus

5.2.2.1 Analysis 1

To determine the association of the 16 selected gene expression levels from Chapter 3 with clinical predictors, five established clinical predictors for MSS were selected as the dependent variables: age at diagnosis in years, sex (coded as 0 for female and 1 for male), tumour site (coded as 0 for limbs and 1 for rest of the body), log-transformed Breslow thickness, and presence of ulceration (coded as 0 for no and 1 for yes). In 699 samples with gene expression data available, simple linear regression was used to explore the associations of the $\log_2$ transformed expression level of the 16 genes with age and log-transformed Breslow thickness, whereas simple logistic regression was used for the association of the $\log_2$ transformed expression levels with sex, tumour site, and presence of ulceration. In addition, a gene expression score which summarises the 16 gene expression levels, weighting them according to their effect on MSS was also used to explore the effect of the combined expression levels on the clinical predictors. The score was calculated in a way that patient with higher gene expression score have higher risk of death from melanoma. A more detailed explanation on the calculation of the gene expression score is shown in Chapter 6 § 6.2.4.1 and an example of the calculation is shown in § 6.2.3.1. Estimates from the penalized Cox regression in Chapter 3 (Table 3.4) were used to create the gene expression score for the remaining 275 samples that were not included in the training set analysis (n=424). Similar statistical methods as the individual association analysis were applied for the gene expression score, although the analysis was only conducted on samples after excluding the training set. Results significant at a level of 5% were highlighted.

5.2.2.2 Analysis 2

To determine the association of the 13 selected SNPs from Chapter 3 with clinical predictors, five established clinical predictors as in Analysis 1 were again selected as the dependent variables. In 1907 samples with genotype data available, simple linear regression was performed to explore the associations of the SNP genotypes (coded as 0, 1, and 2) with age and log-transformed Breslow thickness, and simple logistic regression analysis for the SNP genotypes with sex, tumour site, and presence of ulceration, using an additive model (trend test). A SNP score was also created to explore the association of the combined SNPs on the clinical predictors. The SNP score was also calculated in a way that patient with higher SNP score have higher risk of death from melanoma. Detailed explanation on the calculation of the SNP score is also shown in Chapter 6 § 6.2.4.1 and § 6.2.3.1. Estimates from the penalized Cox regression in Chapter 3 (Table 3.8) were used to create the SNP score on 364 samples that were not included in the training set analysis (n=1543). A similar method to the individual SNP association was used to test the association of the SNP score with the clinical predictors; the analysis was restricted to those samples not in the training set.

5.2.2.3 Analysis 3

To determine whether gene expression levels could be predicted from the genome, the association of genome-wide SNPs with 16 selected gene expression levels from melanoma tumours was explored. The genome-wide SNP data (~500K SNPs after QC) were combined with the expression data of the 16 genes in 619 samples with both expression and genotype data. A genome-wide eQTL analysis was conducted for each gene expression probe using Plink version 1.9 (Purcell *et al.*, 2007). Simple linear regression was used to test the association of SNP genotypes (coded as 0, 1, and 2) with expression level ($\log_2$ scale) for each gene using an additive genetic model (trend test). For an additive effect of a SNP, the regression coefficient represents the effect of each copy of the minor allele. For each gene, the number of SNPs that reached P-value < 5.0×10^{-8} (indicating genome-wide significance) and P-value < 1.0×10^{-5} (indicating suggestive association) were reported, as well as the most significant SNP and its P-value. For associations that reached the genome-wide significance level, further analysis adjusting for the top SNP was performed to identify any secondary association. A Manhattan plot with lines drawn at P-value 1.0×10^{-5} and 5.0×10^{-8} (if applicable) was created for each gene following the association tests.

5.2.2.4 Analysis 4

To determine the association of top melanoma susceptibility SNPs with MSS, the list of the 20 top SNPs from Table 5.1 was used as a reference. Of the 20 SNPs, only four can be extracted from the Leeds genome-wide genotype data. Therefore, imputed SNP data were used for this analysis. Genotype imputation is a method of predicting genotypes that have not been directly typed in the study sample using an external high-density reference panel of phased haplotypes (Marchini and Howie, 2010). The imputation was performed by Dr Mark Iles using IMPUTE version 2 (Howie *et al.*, 2009) and the 1000 Genomes as reference panel. Prior to imputation, SNPs were filtered for MAF < 0.03, P-value of < 10^{-4} for HWE test or missingness > 0.03, and individuals with call rates < 0.97, identified as first degree relatives and/or identified as non-European by PCA were excluded. After the imputation, only those SNPs with INFO score > 0.5 and MAF > 0.01 were selected for analysis. For the imputed SNPs, gene dosage (expected genotype count) was calculated for each SNP for use in the analysis. In 1733 patients (excluding patients with missing cause of death, having multiple melanomas, and recruited into the study 2 years after diagnosis), simple Cox regression analysis was performed to determine the association of top susceptibility SNPs with MSS assuming an additive model. Results significant at a level of 5% were highlighted.

To determine the association of top melanoma susceptibility SNPs with clinical predictors, five established clinical predictors as in Analysis 1 were again selected as the dependent variables. Imputed SNP data were also used to perform the analysis for SNPs that are not available in the genotyped dataset. Simple linear regression analysis was performed to determine the association of the top susceptibility SNPs with age and log-transformed Breslow thickness, and simple logistic regression for the association of the top susceptibility SNPs with sex, tumour site, and presence of ulceration. Results significant at a level of 5% were highlighted.

5.2.2.5 Analysis 5

To determine the association of the top melanoma susceptibility SNPs with the expression levels of nearby genes in each locus, the 20 top SNPs from Table 5.1 were again analysed. The expression levels for genes in each of the 20 melanoma susceptibility loci were identified from the Leeds whole-genome gene expression data. Four loci contain more than one mapped gene; *RMDN2/CYP1B1, TERT/CLPTM1L, CDKN2A/MTAP,* and *TMEM38B/RAD23B/TAL2*.

For these loci, the expression levels for all genes were used to perform the eQTL analysis as it is still unclear which gene is involved in these loci. For nine genes (*ARNT, CASP8, TERT, SLC45A2, AGR3, CDKN2A, TMEM38B, ATM,* and *PLA2G6*), more than one probe was available, as shown in Table 5.2, and in this case, all probes were analysed separately as probes may measure distinct transcripts and this was recommended by the manufacturer (Illumina).

Using Plink software, simple linear regression was conducted to determine the association of each melanoma susceptibility SNP with the expression level ($\log_2$ scale) of a nearby gene, assuming an additive model. Gene dosage (expected genotype count) was calculated for the imputed SNPs for use in the analysis. Results significant at a level of 5% were highlighted.

Table 5.2 Gene expression probes from the whole genome gene expression data for the 20 melanoma susceptibility loci

Gene	No. of probes	Probes
ARNT	2	ILMN_1762582, ILMN_2347314
PARP1	1	ILMN_1686871
RMDN2	1	ILMN_1812302
CYP1B1	1	ILMN_1693338
CASP8	4	ILMN_1673757, ILMN_1787749, ILMN_1809313, ILMN_2377733
TERT	2	ILMN_1796005, ILMN_2373119
CLPTM1L	1	ILMN_1752802
SLC45A2	4	ILMN_1654165, ILMN_1685259, ILMN_2246188, ILMN_2320391
CDKAL1	1	ILMN_1788022
AGR3	2	ILMN_1728787, ILMN_2050246
CDKN2A	3	ILMN_1717714, ILMN_1744295, ILMN_1757255
MTAP	1	ILMN_1753639
TMEM38B	2	ILMN_1669940, ILMN_2093980
RAD23B	1	ILMN_1722662
TAL2	1	ILMN_2135833
OBFC1	1	ILMN_1789186
CCND1	1	ILMN_1688480
TYR	1	ILMN_1788774
ATM	4	ILMN_1713630, ILMN_1716231, ILMN_1779214, ILMN_2370825
OCA2	1	ILMN_1746116
FTO	1	ILMN_2288070
MC1R	1	ILMN_1653319
ASIP	1	ILMN_1791647
MX2	1	ILMN_2231928
PLA2G6	2	ILMN_1697654, ILMN_1798955

5.2.2.6 Analysis 6

To determine the association of other SNPs in the susceptibility region with the expression levels of nearby genes, other SNPs surrounding the top SNPs were included in the analysis. SNPs within 500Kb on either side of the top SNPs were identified for this analysis as shown in Table 5.3. Imputed SNP data were also used for this analysis to increase the coverage of the SNPs in the defined region. The association of the imputed SNPs (gene dosage) within each region with the expression levels ($\log_2$ scale) of the nearby genes (using all probes available for each gene) were then explored using simple linear regression in Plink software. A Manhattan plot was created for each region following the association tests. In each plot, the top susceptibility SNP shown by red square to highlight the association of the top SNP in comparison to other SNPs. As testing the association of multiple SNPs simultaneously causes multiple testing issues, correction was made for this as further described §5.2.3.2.

Table 5.3 List of 20 melanoma risk loci and the number of imputed SNPs in each defined region[a]

Region	Top SNP	Gene	Number of genotyped and imputed SNPs surrounding the top SNP in the defined region[a]
1q21.3	RS12410869	*ARNT*	2533
1q42.12	RS1858550	*PARP1*	3361
2p22.2	RS6750047	*RMDN2 (CYP1B1)*	4590
2q33-q34	RS7582362	*CASP8*	2810
5p15.33	RS380286	*TERT/ CLPTM1L*	4582
5p13.2	RS250417	*SLC45A2*	2160
6p22.3	RS6914598	*CDKAL1*	3620
7p21.1	RS1636744	*AGR3*	4644
9p21	RS7852450	*CDKN2A/ MTAP*	3938
9q31.2	RS10739221	*TMEM38B (RAD23B, TAL2)*	3232
10q24.33	RS2995264	*OBFC1*	3353
11q13	RS498136	*CCND1*	3922
11q14-q21	RS1393350	*TYR*	2889
11q22-q23	RS73008229	*ATM*	2643
15q13.1	RS4778138	*OCA2*	1763
16q12.2	RS12596638	*FTO*	3785
16q24.3	RS75570604	*MC1R*	4344
20q11.2-q12	RS6088372	*ASIP*	2564
21q22.3	RS408825	*MX2*	3812
22q13.1	RS2092180	*PLA2G6*	3175

[a] 500 Kb on either side of the top SNP

5.2.3 Statistical methods

5.2.3.1 Simple linear regression

Simple linear regression is used to identify the relationship of one explanatory variable with a continuous outcome. The *lm* function in R software was used to determine whether clinical predictors (age and log-transformed Breslow thickness) are associated the selected gene expression levels (Analysis 1), selected SNPs (Analysis 2), and top susceptibility SNPs (Analysis 4). Plink software was used to determine whether gene expression levels can be predicted from the whole-genome SNPs (Analysis 3) and to determine whether expression levels of nearby genes are associated with top susceptibility SNPs (Analysis 5 and 6).

5.2.3.2 Simple logistic regression

Simple logistic regression is used to identify the relationship of one explanatory variable with a binary outcome. The *glm* function in R software was used to determine whether clinical predictors (sex, tumour site, and presence of ulceration) are associated with the selected gene expression levels (Analysis 1), selected SNPs (Analysis 2), and top susceptibility SNPs (Analysis 4).

5.2.3.3 Simple Cox regression

Simple Cox regression is used to identify the relationship of one explanatory variable with a survival outcome. The *coxph* function in the *survival* package in R software was used to perform the association of top melanoma susceptibility SNPs with MSS.

5.2.3.4 Multiple testing correction

Multiple testing occurs when testing the association of multiple SNPs simultaneously. Multiple test correction was performed using the method by Benjamini and Yekutieli (2001) to account for the hypotheses tested not being independent (due to LD), with a 10% false discovery rate (FDR).

5.3 Results

5.3.1 Association of selected gene expression levels (and gene expression score) with clinical predictors

Table 5.4 shows the association of each selected expression level with five clinical predictors using all samples with gene expression data in the cohort (n=699). Of the 16 selected gene expression levels, only expression level of *CHST9* did not show association with any clinical predictors, while all others show significant association at least with Breslow thickness. There were 10 gene expression levels that associated with age at diagnosis, three associated with sex, seven associated with tumour site, 15 associated with log-transformed Breslow thickness, and 11 associated with presence of ulceration. Interestingly, the directions of effect for the expression level that associated with more than one factors are always consistent. For example, expression level of *NKD2* that associated with all clinical predictors shows that doubling its expression level associated with reduced age at diagnosis, reduced log-transformed Breslow thickess, reduced the odds for male sex, reduce odds for tumour on the rest of the body, and reduced odds for presence of ulceration.

For age at diagnosis, log-transformed Breslow thickness, and presence of ulceration, several of the expression levels showed very strong association, especially with Breslow thickness. The strongest associations with P-value $<10^{-16}$ were seen between *NKD2* expression level (coefficient= -0.24, P-value=2 x 10^{-16}) and *HLA-DQB2* expression level (coefficient= -0.25, P-value=2 x 10^{-16}) with log-transformed Breslow thickness.

When using the gene expression score in 275 samples (after excluding patients from the training set used to estimate the weights), the score was significantly associated with all clinical predictors, with the strongest association also seen with Breslow thickness, followed by age at diagnosis, presence of ulceration, sex, and tumour site.

Table 5.4 Association of the selected gene expression levels (and gene expression score) with clinical predictors in all samples with gene expression data (n=699)

	Age at diagnosis (years)[a]	Sex (Female vs male)[b]	Tumour site (Limbs vs rest of the body)[b]	Log-transformed Breslow thickness[a]	Presence of ulceration (No vs yes)[b]
ILMN_1701441 (*LPAR1*)	2.3×10^{-6} (-2.30)	0.29	4.8×10^{-3} (0.79)	5.5×10^{-13} (-0.16)	2.2×10^{-5} (0.70)
ILMN_3249501 (*ZNF697*)	0.16	0.39	0.38	7.7×10^{-5} (0.09)	0.05
ILMN_1749829 (*DLG1*)	1.1×10^{-5} (2.14)	0.04 (1.16)	0.09	1.6×10^{-9} (0.14)	1.9×10^{-4} (1.36)
ILMN_1731206 (*NKD2*)	1.2×10^{-11} (-3.27)	4.2×10^{-3} (0.81)	4.4×10^{-5} (0.70)	2.0×10^{-16} (-0.24)	4.4×10^{-12} (0.54)
ILMN_1764109 (*C1R*)	3.2×10^{-4} (-1.75)	0.15	0.04 (0.85)	9.9×10^{-9} (-0.13)	2.5×10^{-3} (0.79)
ILMN_2056167 (*OSTC*)	0.01 (1.26)	0.43	0.09	7.8×10^{-6} (0.10)	0.38
ILMN_3238435 (*SNORA12*)	0.05	0.66	8.4×10^{-4} (0.75)	2.0×10^{-11} (-0.15)	2.5×10^{-5} (0.70)
ILMN_1695959 (*C21orf63*)	5.5×10^{-6} (-2.21)	0.12	0.06	2.3×10^{-4} (-0.09)	0.08
ILMN_1741648 (*HLA-DQB2*)	2.4×10^{-5} (-2.06)	0.07	6.0×10^{-4} (0.76)	2.0×10^{-16} (-0.25)	5.4×10^{-13} (0.54)
ILMN_1784238 (*SEC22B*)	0.01 (1.21)	0.17	0.94	6.6×10^{-6} (0.10)	4.2×10^{-3} (1.27)
ILMN_1778401 (*HLA-B*)	0.15	0.11	0.01 (0.82)	2.6×10^{-6} (-0.11)	3.4×10^{-4} (0.76)
ILMN_1759729 (*NDUFA8*)	0.09	0.13	1.2×10^{-3} (1.38)	3.5×10^{-5} (0.10)	4.4×10^{-3} (1.34)
ILMN_2344221 (*IGSF5*)	2.0×10^{-5} (2.08)	0.36	0.06	1.7×10^{-8} (0.13)	9.1×10^{-3} (1.23)
ILMN_2095633 (*FGF22*)	7.1×10^{-7} (-2.41)	0.02 (0.84)	0.12	4.8×10^{-8} (-0.13)	1.0×10^{-4} (0.71)
ILMN_1700547 (*CHST9*)	0.10	0.13	0.39	0.82	0.44
ILMN_1735199 (*CIAPIN1*)	0.08	0.50	0.96	9.1×10^{-6} (0.10)	0.07
Gene expression score*	3.3×10^{-5} (8.33)	0.02 (2.10)	0.04 (2.01)	2.0×10^{-7} (0.51)	3.6×10^{-4} (3.60)

[a] P-values from simple linear regression; values in bracket are coefficients for significant association with P-value<0.05

[b] P-values from simple logistic regression: values in bracket are odds ratio for significant association with P-value<0.05

*Association of gene expression score with clinical predictors was explored in 275 samples

The highlighted cells indicate significant association at P-value<0.05

5.3.2 Association of selected SNPs (and SNP score) with clinical predictors

Table 5.5 shows the association of each selected SNP with five clinical predictors in all samples with genotyped data in the cohort (n=1907). Of the 13 selected SNPs, two shows significant association with age at diagnosis, none was associated with sex, one was significantly associated with tumour site, one was significantly associated with log-transformed Breslow thickness, and two were significantly associated with presence of ulceration. However, the evidence for association for these SNPs was marginal with strongest P-value of 0.02 for the presence of ulceration. When using the SNP score in 364 samples, the score show significant association with log-transformed Breslow thickness only (P-value=0.01).

Table 5.5 Association of the selected SNPs (and SNP score) with clinical predictors in all samples with gene expression data (n=1907)

	Age at diagnosis (years)[a]	Sex (Female vs male)[b]	Tumour site (Limbs vs rest of the body)[b]	Log-transformed Breslow thickness[a]	Presence of ulceration (No vs yes)[b]
RS17837209	0.11	0.93	0.76	0.73	0.42
RS9957831	0.63	0.31	0.16	0.17	0.97
RS4768090	0.03 (1.02)	0.32	0.03 (1.17)	0.34	0.02 (1.22)
RS2902554	0.50	0.34	0.06	0.36	0.80
RS5770310	0.61	0.34	0.58	0.18	0.82
RS10233832	0.05	0.19	0.62	0.57	0.42
RS17379771	0.25	0.77	0.98	0.23	0.71
RS16956192	0.35	0.65	0.42	0.17	0.17
RS2392477	0.14	0.32	0.69	0.87	0.73
RS6689263	0.86	0.18	0.90	0.06	0.44
RS11639902	0.60	0.89	0.10	0.05	0.02 (0.82)
RS12519276	0.04 (0.88)	0.43	0.50	0.04 (0.05)	0.34
RS10941528	0.09	0.34	0.40	0.16	0.06
SNP score*	0.61	0.64	0.51	0.01 (-1.14)	0.18

[a] P-values from simple linear regression; values in bracket are coefficients for significant association with P-value<0.05

[b] P-values from simple logistic regression: values in bracket are odds ratio for significant association with P-value<0.05

*Association of SNP score with clinical predictors was explored in 364 samples

The highlighted cells indicate significant association at P-value<0.05

5.3.3 Genome-wide association study of the selected gene expression levels (expression levels of 16 genes that are associated with MSS)

Table 5.6 shows the results for genome-wide eQTL analysis for each gene, and Figure 5.1 to 5.16 shows the Manhattan plot for each analysis. The SNP with the strongest association with each expression level, together with the number of SNPs that reached P-value < 5 x 10^{-8} and P-value < 1 x10^{-5} are shown in the Table 5.6. Of the 16 gene expression levels, only those of *HLA-DQB2* and *NDUFA8* show association with SNPs reaching genome-wide significance level (P-value < 5 x 10^{-8}). For expression level of eight genes (*LPAR1, DLG1, NKD2, C1R, HLA-B, IGSF5, FGF22*, and *CIAPIN1*), the strongest association reached P-value < 10^{-7}. For expression level of six genes (*ZNF697, OSTC, SNORA12, C21orf63, SEC22B*, and *CHST9*), their strongest association reached P-value < 10^{-6}. Of these associations, four show cis-eQTL associations (*LPAR1, NKD2, HLA-DQB2*, and *HLA-B*) as the top SNPs are located on the same chromosome as the genes (results highlighted in Table 5.6). For the associations that are only marginally significant, these could be false positive associations.

For *HLA-DQB2*, the SNP RS5019296 was the most predictive of *HLA-DQB2* expression with coefficient of -0.76 (P-value=3.0 x 10^{-14}); the effect of each minor allele is estimated to reduce the expression level by 0.76 units on the $\log_2$ transformed scale. The 29 SNPs associated with *HLA-DQB2* expression at the genome-wide significance level have MAF > 5% (Table 5.7), and the Manhattan plot for eQTL analysis in Figure 5.9 shows high peak associations at chromosome 6. As the associated SNPs were located close to the region of *HLA-DQB2* in chromosome 6, these associations represent a cis-eQTL. All the associated SNPs however, were highly correlated with one another as shown in Table 5.9. When conditioned on the most significant SNP (RS5019296), the strongest association reached P-value < 10^{-6} (RS6901084), hence suggesting a secondary association (Table 5.7), despite the strong LD between these SNPs (r = 0.9). When further analysis conditioning on two SNPs (RS5019296 and RS6901084) was performed to identify other SNPs that could predict the expression level, none of the SNPs reached P-value < 10^{-5}. However, because of the strong LD in this region, it is unclear which of the

SNPs showing association in simple linear regression were regulatory in the eQTL.

For *NDUFA8*, the SNP RS17398871 shows the strongest association with the expression level, with coefficient of -0.25 (P-value=1.0 x 10^{-8}); the effect of each minor allele is estimated to reduce 0.25 unit of log2 transformed expression level (Table 5.8). As both of the SNPs identified at genome-wide significance level are located on chromosome 2 and *NDUFA8* gene located on chromosome 9, these associations represent a trans-eQTL. Both of the associated SNPs were highly correlated (r=0.99) and when adjusting for the most significant SNP, the second SNP (RS2192689) was no longer significant (P-value=0.95).

Table 5.6 Genome-wide eQTL analysis for 16 gene expression levels (n=619)

Probe	Gene (Chr)	Simple linear regression			
		No. of SNPs with P < 5 x 10^{-8}	No. of SNPs with P < 1 x 10^{-5}	Top SNP and its position on chromosome	P-value*
ILMN_1701441	*LPAR1* (9)	-	18	RS1331250 (9:32191942)	1.6 x 10^{-7}
ILMN_3249501	*ZNF697* (1)	-	7	RS2345782 (12:126373092)	2.6 x 10^{-6}
ILMN_1749829	*DLG1* (14)	-	8	RS2147356 (13: 33474475)	3.2 x 10^{-7}
ILMN_1731206	*NKD2* (5)	-	12	RS10515509 (5:139290054)	2.2 x 10^{-7}
ILMN_1764109	*C1R* (12)	-	5	RS6817112 (4:154080813)	2.0 x 10^{-7}
ILMN_2056167	*OSTC* (4)	-	2	RS185063 (16:84456705)	1.7 x 10^{-6}
ILMN_3238435	*SNORA12* (10)	-	3	RS2944776 (8:141541881)	4.5 x 10^{-6}
ILMN_1695959	*C21orf63* (21)	-	4	RS6029941 (20:35519475)	2.1 x 10^{-6}
ILMN_1741648	*HLA-DQB2* (6)	29	57	RS5019296 (6:32733446)	3.0 x 10^{-14}
ILMN_1784238	*SEC22B* (1)	-	6	RS2382215 (2:148977189)	2.5 x 10^{-6}
ILMN_1778401	*HLA-B* (6)	-	15	RS6912584 (6:28309590)	9.1 x 10^{-7}
ILMN_1759729	*NDUFA8* (9)	2	31	RS17398871 (2:58481589)	1.0 x 10^{-8}
ILMN_2344221	*IGSF5* (21)	-	13	RS12147287 (14:101554839)	1.6 x 10^{-7}
ILMN_2095633	*FGF22* (19)	-	13	RS13333251 (16:86290512)	3.2 x 10^{-7}
ILMN_1700547	*CHST9* (18)	-	7	RS1004551 (3:134519445)	4.8 x 10^{-6}
ILMN_1735199	*CIAPIN1* (16)	-	19	RS4142346 (20:39513865)	2.1 x 10^{-7}

*Additive model

Chr: Chromosome

The highlighted rows indicate associations are cis-eQTLs as the top SNPs located on the same chromosome as the genes

Table 5.7 Association of *HLA-DQB2* expression level with the 29 SNPs identified at genome-wide significance level (n=619)

SNP	Chr	Position	MAF	Simple linear regression		Adjusting for the most significant SNP (RS5019296)	
				β	P-value	β	P-value
RS5019296	6	32733446	0.46	-0.76	3.0×10^{-14}	-	-
RS1573649	6	32731258	0.46	-0.76	4.4×10^{-14}	NA	NA
RS9276598	6	32733987	0.46	-0.76	4.4×10^{-14}	NA	NA
RS7382794	6	32734030	0.46	-0.76	4.4×10^{-14}	NA	NA
RS6903130	6	32732210	0.46	-0.76	4.4×10^{-14}	NA	NA
RS6902723	6	32731960	0.46	-0.76	4.4×10^{-14}	NA	NA
RS9276595	6	32733931	0.46	-0.74	1.2×10^{-13}	NA	NA
RS9276586	6	32732937	0.46	-0.74	1.6×10^{-13}	NA	NA
RS1573648	6	32731439	0.46	-0.74	2.1×10^{-13}	NA	NA
RS2006165	6	32728787	0.47	0.65	1.3×10^{-10}	-0.68	2.4×10^{-4}
RS1023449	6	32727905	0.46	0.65	2.4×10^{-10}	0.06	0.73
RS2395256	6	32728588	0.47	0.64	2.4×10^{-10}	0.06	0.73
RS2213572	6	32719804	0.47	0.64	2.4×10^{-10}	-0.70	1.1×10^{-4}
RS2213568	6	32711576	0.47	0.64	2.4×10^{-10}	0.06	0.73
RS4248169	6	32728554	0.47	0.64	2.6×10^{-10}	-0.71	1.3×10^{-4}
RS9276558	6	32724061	0.47	0.64	2.7×10^{-10}	-0.71	1.1×10^{-4}
RS7453920	6	32730012	0.47	0.64	3.9×10^{-10}	-0.77	4.7×10^{-5}
EXM-RS9276431	6	32712247	0.47	0.63	5.4×10^{-10}	-0.01	0.98
RS2071551	6	32729459	0.47	0.63	6.5×10^{-10}	-0.77	4.8×10^{-5}
RS2301271	6	32725193	0.47	0.63	6.5×10^{-10}	-0.77	4.8×10^{-5}
EXM-RS2213567	6	32711655	0.47	0.62	7.6×10^{-10}	0.01	0.95
RS2051549	6	32730086	0.46	0.62	1.1×10^{-9}	-0.03	0.89
EXM-RS6936428	6	32739174	0.38	-0.63	1.7×10^{-9}	0.06	0.75
RS6457661	6	32737494	0.39	-0.62	4.8×10^{-9}	0.14	0.47
RS6901084	6	32736936	0.39	-0.62	4.8×10^{-9}	-0.87	3.1×10^{-6}
EXM-RS1585891	6	32736722	0.39	-0.62	5.9×10^{-9}	0.14	0.47
RS2859071	6	32703366	0.48	0.59	5.9×10^{-9}	-0.05	0.79
EXM-RS7773149	6	32706042	0.48	0.58	1.6×10^{-8}	-0.87	1.1×10^{-6}
RS9276370	6	32707295	0.48	0.58	1.6×10^{-8}	-0.14	0.44

NA: coefficient for the SNP was not estimable due to collinearity

Table 5.8 Association of *NDUFA8* expression level with the SNPs identified at genome-wide significance level (n=619)

SNP	Chr	Position	MAF	Simple linear regression		Conditioning on the most significant SNP (RS17398871)	
				β	P-value	β	P-value
RS17398871	2	58481589	0.09	-0.25	1.0×10^{-8}	-	-
RS2192689	2	58475916	0.09	-0.25	2.8×10^{-8}	0.01	0.95

Table 5.9 **Pairwise correlations between the 29 SNPs associated with *HLA-DQB2* expression level at genome-wide significance level**

	RS5019296	RS1573649	RS9276598	RS7382794	RS6903130	RS6902723	RS9276595	RS9276586	RS1573648	RS2006165	RS1023449	RS2395256	RS2213572	RS2213568	RS4248169	RS9276558	RS7453920	EXM_RS9276431	RS2071551	RS2301271	EXM_RS2213567	RS2051549	EXM_RS6936428	RS6457661	RS6901084	EXM_RS1585891	RS2859071	EXM_RS7773149	RS9276370
RS5019296	1	1	1	1	1	1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS1573649		1	1	1	1	1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS9276598			1	1	1	1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS7382794				1	1	1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS6903130					1	1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS6902723						1	1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS9276595							1	1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS9276586								1	1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS1573648									1	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	0.9	0.9	0.9	0.9	-0.8	-0.8	-0.8
RS2006165										1	1	1	1	1	1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS1023449											1	1	1	1	1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2395256												1	1	1	1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2213572													1	1	1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2213568														1	1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS4248169															1	1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS9276558																1	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS7453920																	1	1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
EXM_RS9276431																		1	1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2071551																			1	1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2301271																				1	1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
EXM_RS2213567																					1	1	-0.7	-0.7	-0.7	-0.7	1	1	1
RS2051549																						1	-0.7	-0.7	-0.7	-0.7	1	1	1
EXM_RS6936428																							1	1	1	1	-0.7	-0.7	-0.7
RS6457661																								1	1	1	-0.7	-0.7	-0.7
RS6901084																									1	1	-0.7	-0.7	-0.7
EXM_RS1585891																										1	-0.7	-0.7	-0.7
RS2859071																											1	1	1
EXM_RS7773149																												1	1
RS9276370																													1

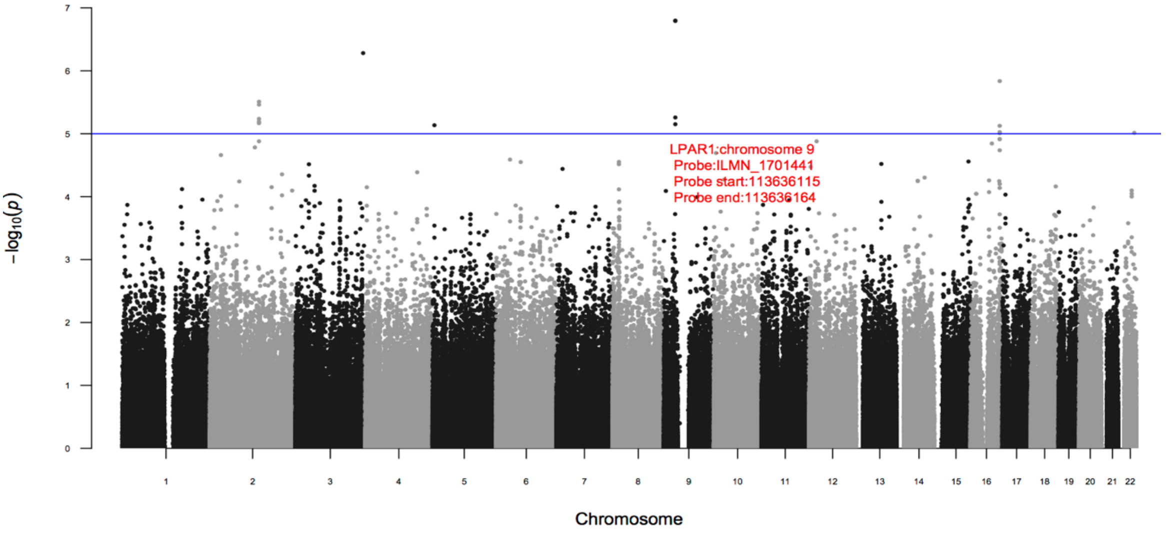


Figure 5.1 Genome-wide eQTL analysis for LPAR1 expression level (Chr 9)

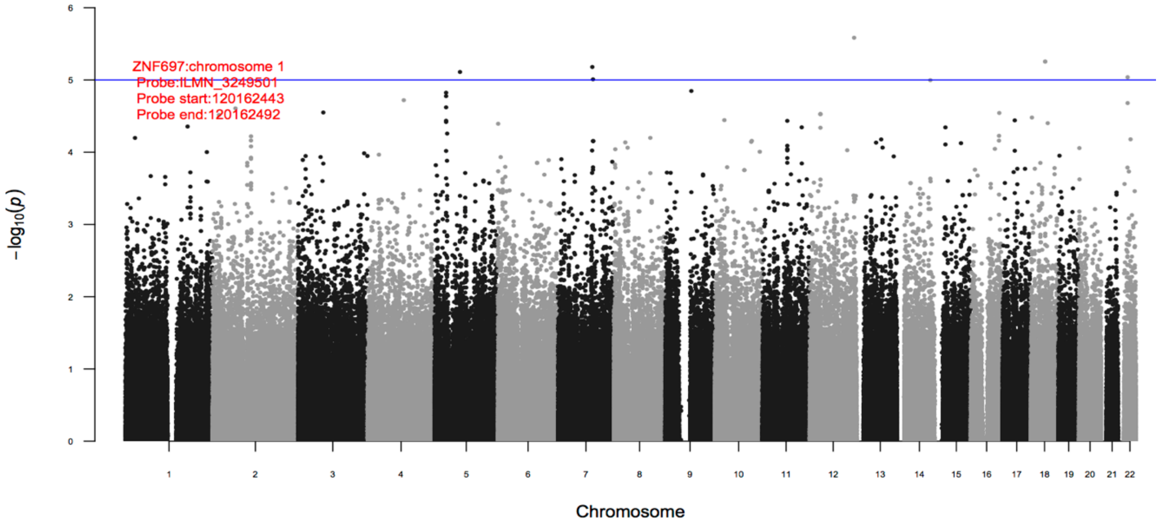


Figure 5.2 Genome-wide eQTL analysis for ZNF697 expression level (Chr 1)

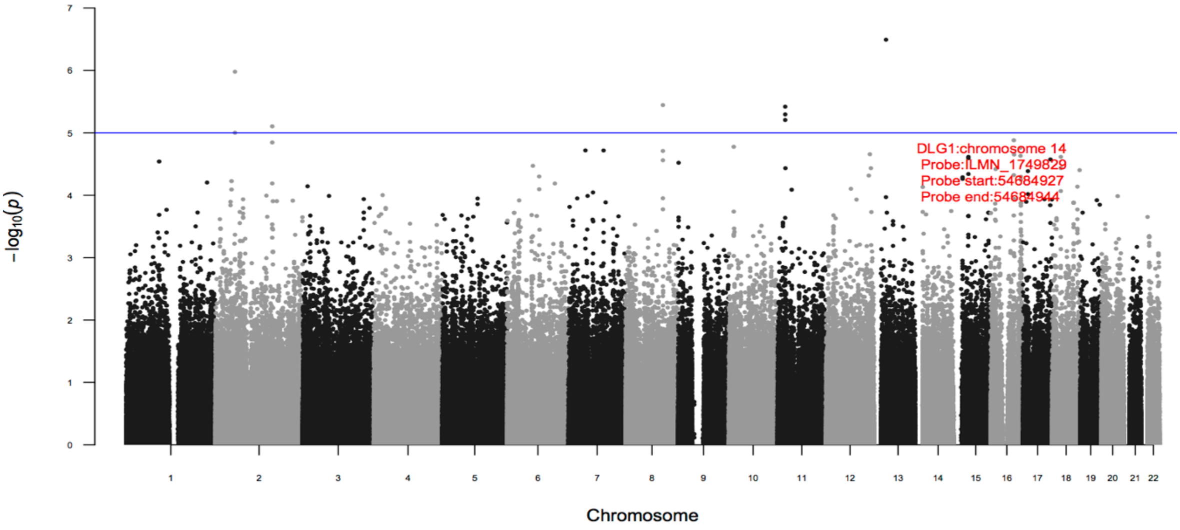


Figure 5.3 Genome-wide eQTL analysis for DLG1 expression level (Chr 14)

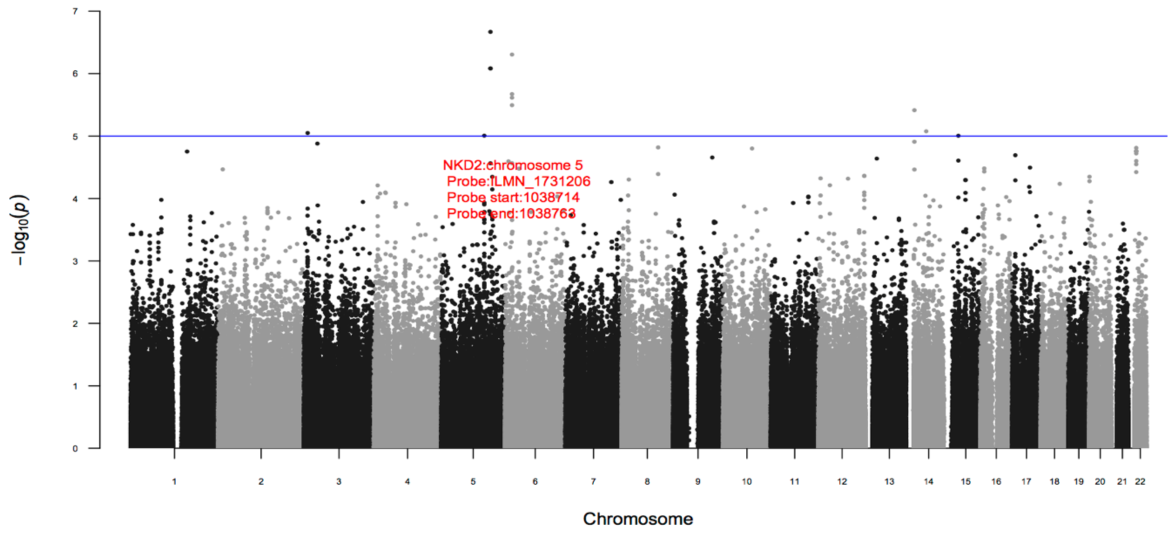


Figure 5.4 Genome-wide eQTL analysis for NKD2 expression level (Chr 5)

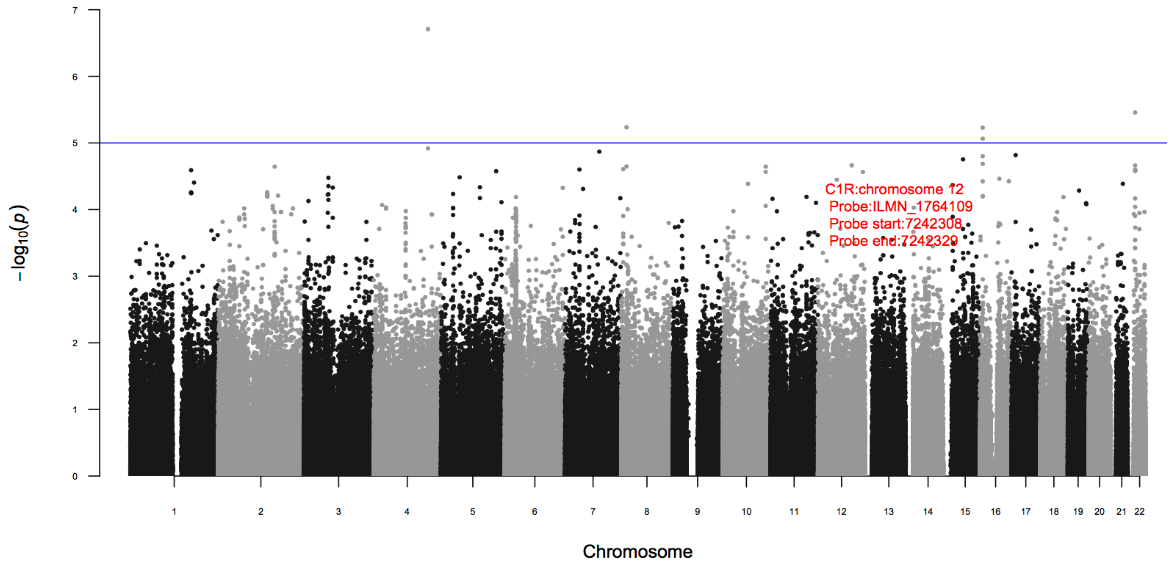


Figure 5.5 Genome-wide eQTL analysis for C1R expression level (Chr 12)

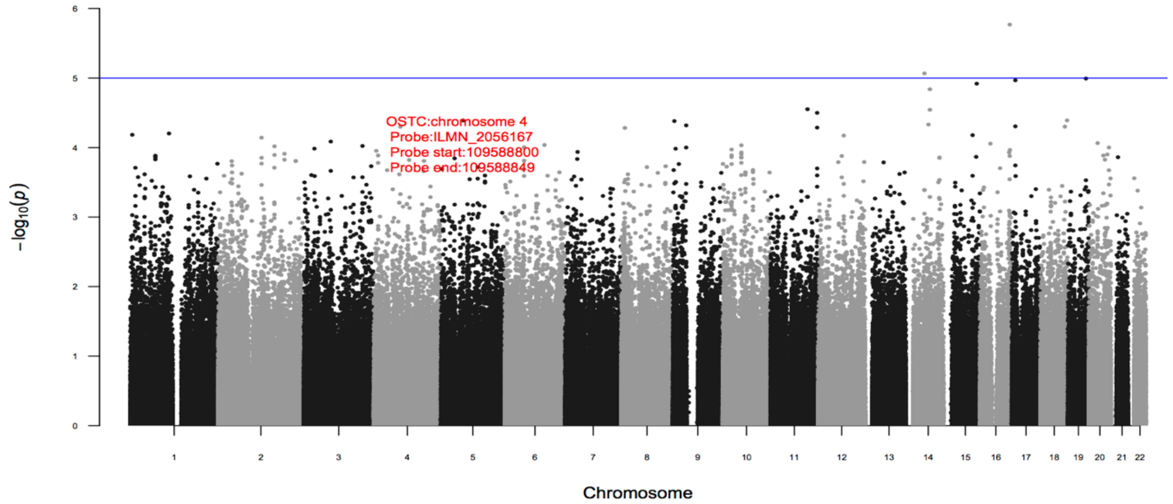


Figure 5.6 Genome-wide eQTL analysis for OSTC expression level (Chr 4)

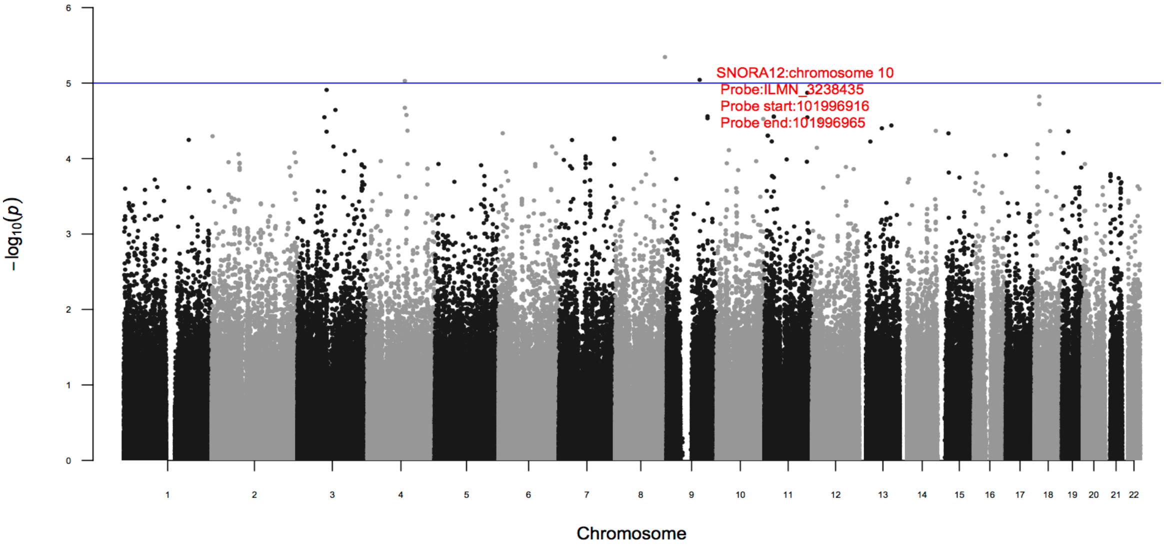


Figure 5.7 Genome-wide eQTL analysis for SNORA12 expression level(Chr 10)

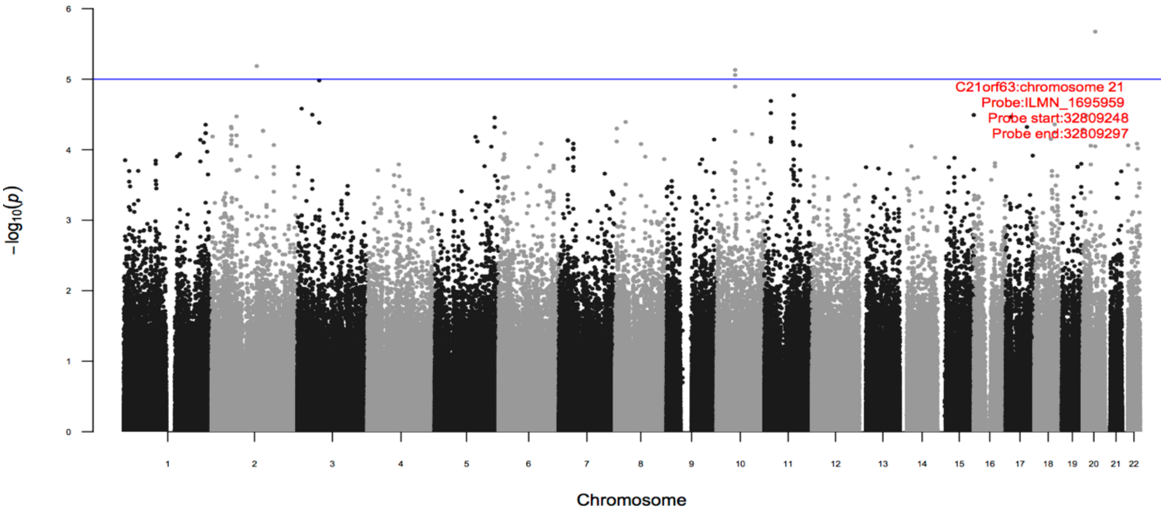


Figure 5.8 Genome-wide eQTL analysis for C21orf63 expression level (Chr 21)

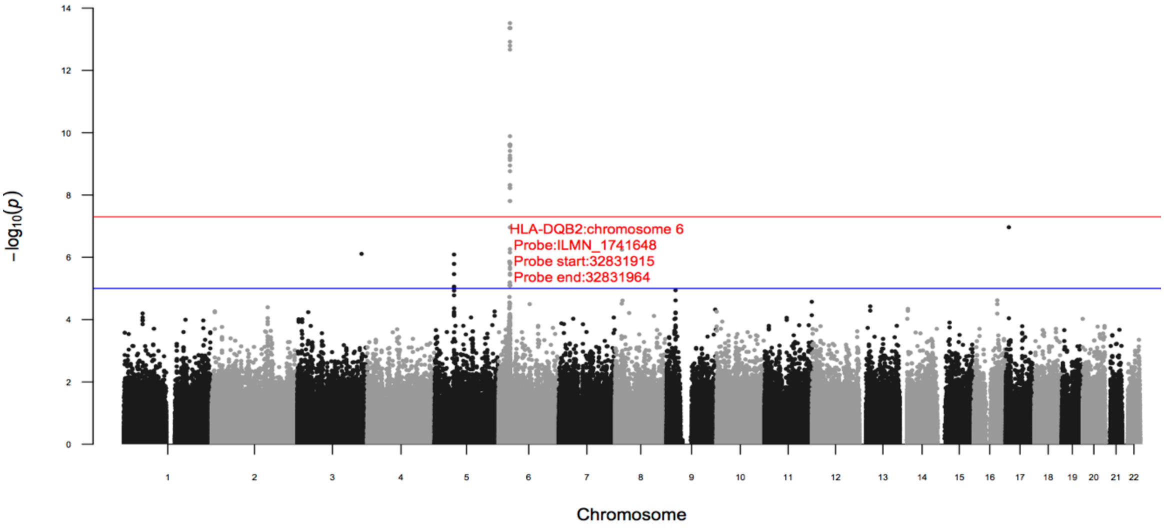


Figure 5.9 Genome-wide eQTL analysis for HLA-DQB2 expression level (Chr 6)

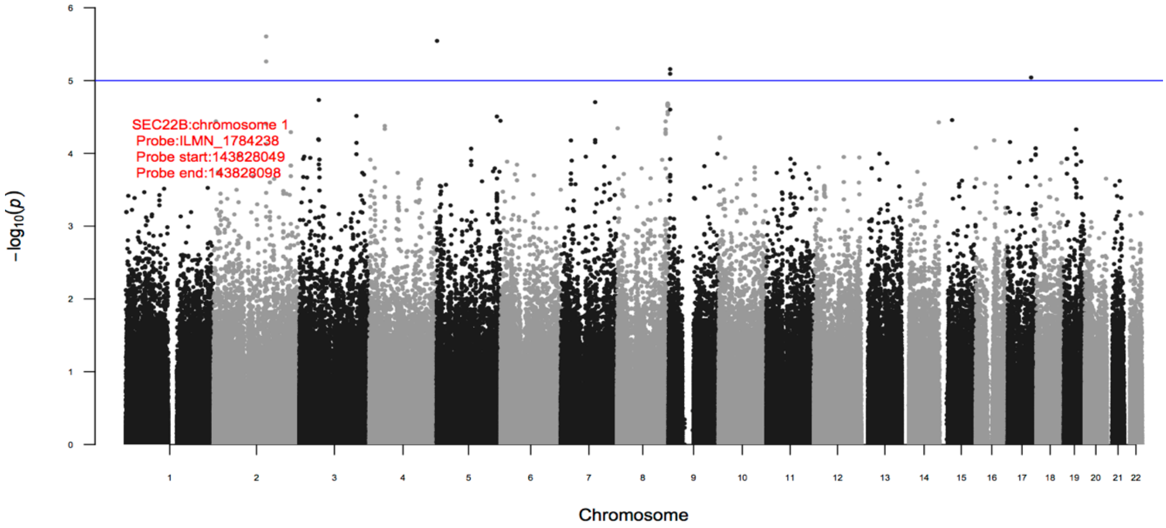


Figure 5.10 Genome-wide eQTL analysis for SEC22B expression level (Chr 1)

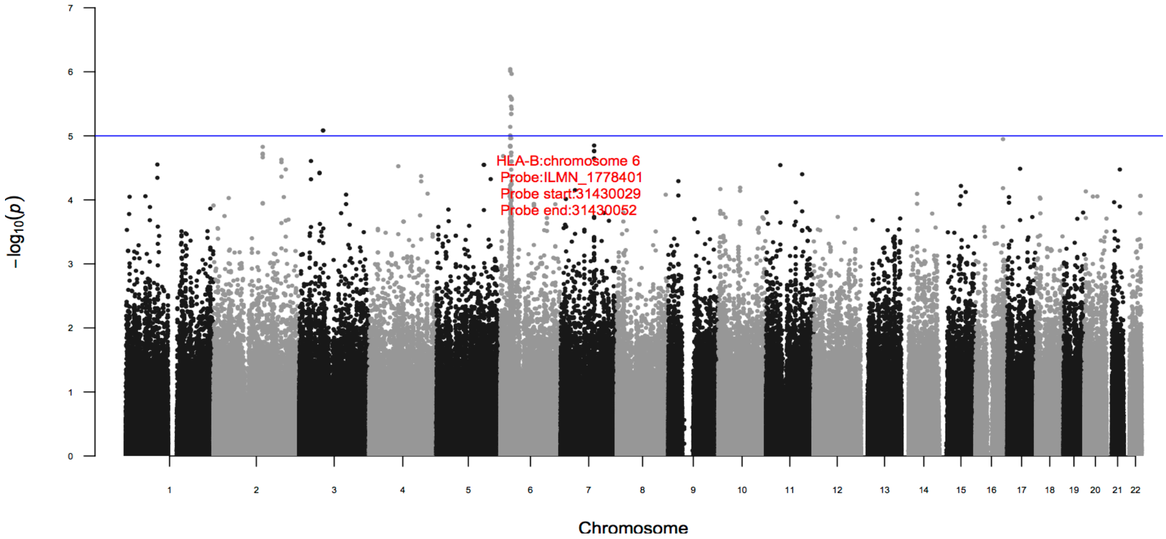


Figure 5.11 Genome-wide eQTL analysis for HLA-B expression level (Chr 6)

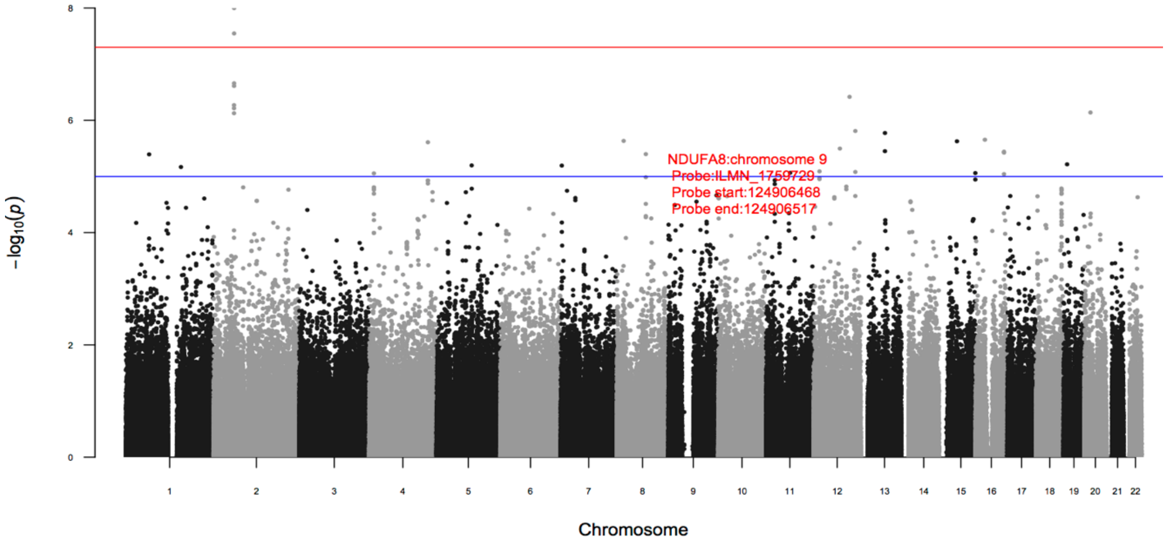


Figure 5.12 Genome-wide eQTL analysis for NDUFA8 expression level (Chr 9)

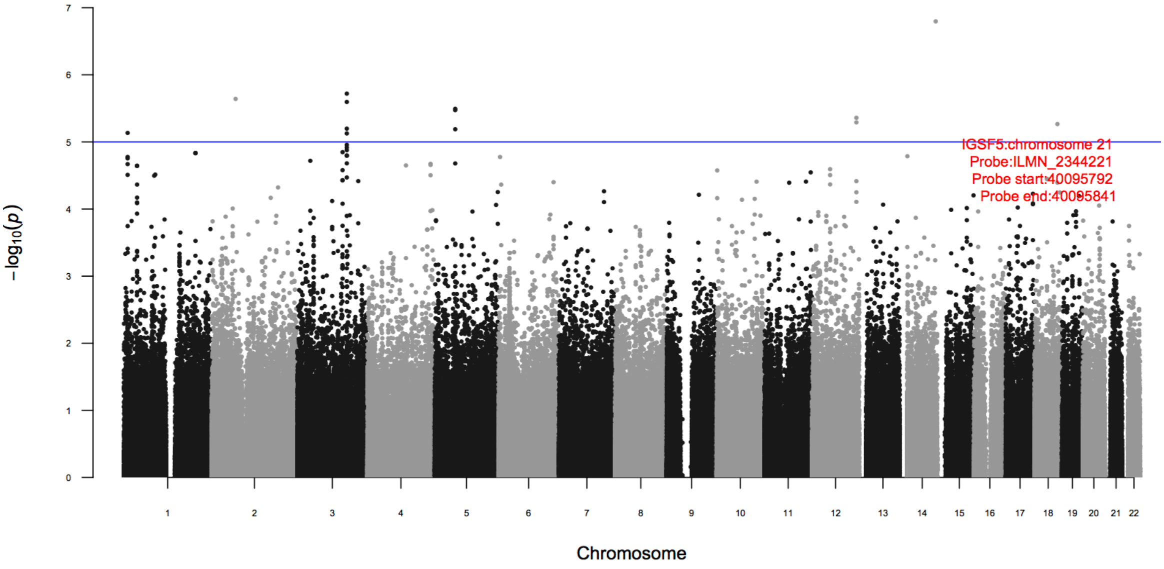


Figure 5.13 Genome-wide eQTL analysis for IGSF5 expression level (Chr 21)

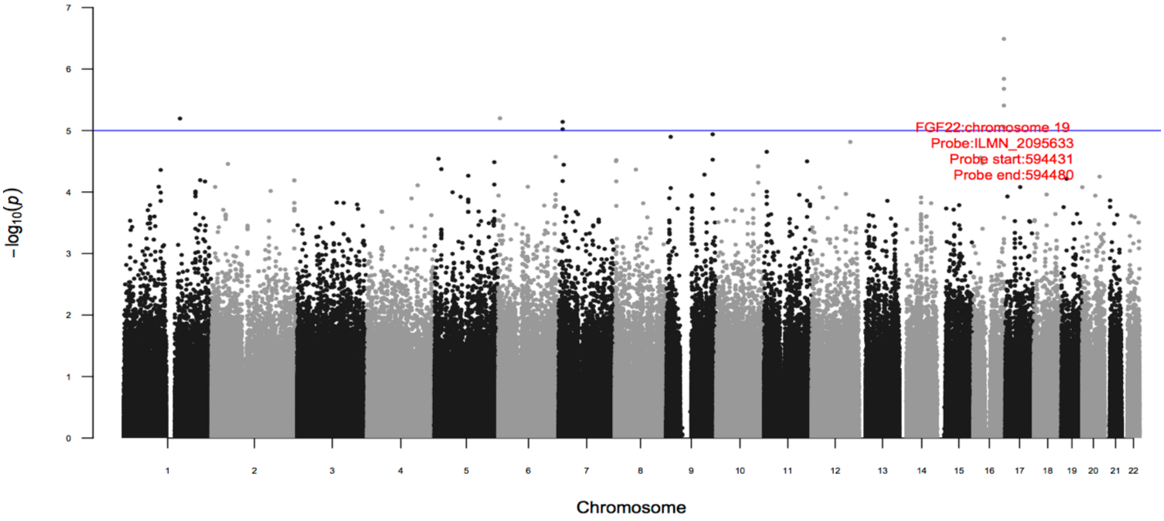


Figure 5.14 Genome-wide eQTL analysis for FGF22 expression level (Chr 19)

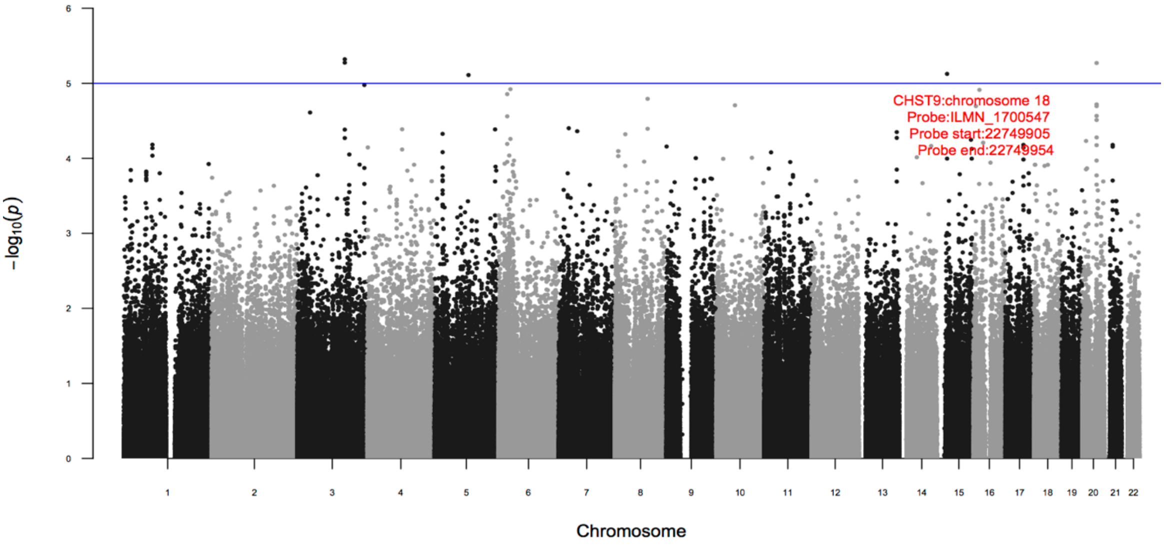


Figure 5.15 Genome-wide eQTL analysis for CHST9 expression level (Chr 18)

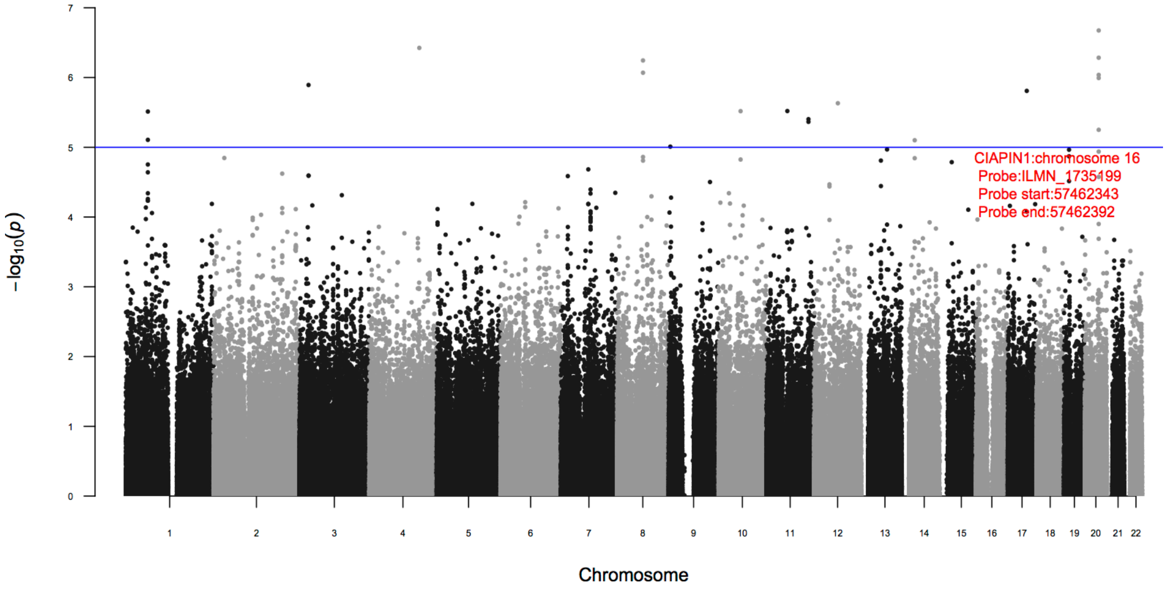


Figure 5.16 **Genome-wide eQTL analysis for CIAPIN1 expression level (Chr 16)**

5.3.4 Association of top melanoma susceptibility SNPs with MSS and clinical predictors

The associations between the top melanoma susceptibility SNPs with MSS are shown in Table 5.10. Of the 20 top SNPs, only the top SNP in *PARP1* (RS1858550) was predictive of MSS (HR=0.77, P-value=2.3 x 10^{-3}); each copy of the minor allele is associated with reduced risk of death from melanoma.

Table 5.11 shows the association of the 20 top melanoma susceptibility SNPs with five clinical predictors. As highlighted in Table 5.11, at the nominal 5% significance level two SNPs (RS498136 in *CCND1* and RS75570604 in *MC1R*) were significantly associated with age at diagnosis, three (RS1858550 in *PARP1*, RS2995264 in *OBFC1*, and RS73008229 in *ATM*) were significantly associated with sex, three (RS6750047 in *RMDN2*, RS250417 in *SLC45A2*, and RS10739221 in *TMEM38B*) were significantly associated with log-transformed Breslow thickness, and one (RS1393350 in *TYR*) was associated with presence of ulceration. None of the SNPs were associated with tumour site. For the significant associations, the highest P-value was P<$10^{-3,}$ and only seen for association with age and sex.

Table 5.10 Association of top melanoma risk SNPs with MSS (n=1733)

Loci	Top SNP (Position)	β	HR	SE	P-value
ARNT	RS12410869 (1: 150856153)	-0.06	0.94	0.09	0.48
PARP1	RS1858550 (1:226608104)	-0.26	0.77	0.09	2.3×10^{-3}
RMDN2 (CYP1B1)	RS6750047 (2:38276549)	0.01	1.01	0.08	0.89
CASP8	RS7582362 (2:202176294)	-0.07	0.93	0.09	0.42
TERT/ CLPTM1L	RS380286 (5:1320247)	-0.06	0.94	0.08	0.46
SLC45A2	RS250417 (5:33952378)	-0.57	0.56	0.51	0.26
CDKAL1	RS6914598 (6:21163919)	0.12	1.13	0.09	0.18
AGR3	RS1636744 (7:16984280)	-0.02	0.98	0.08	0.82
CDKN2A/ MTAP	RS7852450 (9:21825075)	-0.03	0.98	0.08	0.77
TMEM38B (RAD23B, TAL2)	RS10739221 (9:109060830)	-0.18	0.83	0.10	0.07
OBFC1	RS2995264 (10:105668843)	-0.08	0.93	0.14	0.60
CCND1	RS498136 (11:69367118)	-0.07	0.94	0.09	0.45
TYR	RS1393350 (11:89011046)	0.03	1.03	0.09	0.69
ATM	RS73008229 (11:108187689)	-0.08	0.92	0.12	0.50
OCA2	RS4778138 (15:28335820)	-0.10	0.90	0.13	0.44
FTO	RS12596638 (16:54115829)	-0.03	0.97	0.11	0.80
MC1R	RS75570604 (16:89846677)	0.17	1.19	0.12	0.16
ASIP	RS6088372 (20:32586748)	0.10	1.11	0.11	0.37
MX2	RS408825 (21:42743496)	-0.05	0.95	0.09	0.54
PLA2G6	RS2092180 (22:38571563)	0.11	1.12	0.08	0.17

The highlighted row indicate significant association at P-value < 0.05

Table 5.11 Association of top melanoma risk SNPs with clinical predictors (n=1907)

Loci	Top SNP (Position)	Age at diagnosis (years)[a]	Sex (Female vs male)[b]	Tumour site (Limbs vs rest of the body)[b]	Log-transformed Breslow thickness[a]	Presence of ulceration (No vs yes)[b]
ARNT	RS12410869 (1: 150856153)	0.97	0.15	0.12	0.37	0.49
PARP1	RS1858550 (1:226608104)	0.24	0.02 (0.84)	0.98	0.58	0.56
RMDN2 (CYP1B1)	RS6750047 (2:38276549)	0.89	0.97	0.53	0.04 (0.05)	0.37
CASP8	RS7582362 (2:202176294)	0.38	0.94	0.98	0.23	0.65
TERT/ CLPTM1L	RS380286 (5:1320247)	0.09	0.77	0.31	0.22	0.15
SLC45A2	RS250417 (5:33952378)	0.71	0.64	0.52	0.02 (-0.27)	0.60
CDKAL1	RS6914598 (6:21163919)	0.77	0.55	0.15	0.08	0.14
AGR3	RS1636744 (7:16984280)	0.70	0.78	0.59	0.14	0.52
CDKN2A/ MTAP	RS7852450 (9:21825075)	0.35	0.07	0.76	0.77	0.68
TMEM38B (RAD23B, TAL2)	RS10739221 (9:109060830)	0.92	0.39	0.78	0.04 (-0.06)	0.23
OBFC1	RS2995264 (10:105668843)	0.75	0.02 (1.28)	0.17	0.44	0.71
CCND1	RS498136 (11:69367118)	8.2 x 10^{-3} (-1.20)	0.73	0.89	0.27	0.42
TYR	RS1393350 (11:89011046)	0.91	0.55	0.45	0.88	0.02 (0.83)
ATM	RS73008229 (11:108187689)	0.17	9.9 x 10^{-3} (1.30)	0.08	0.80	0.80
OCA2	RS4778138 (15:28335820)	0.47	0.13	0.86	0.84	0.85
FTO	RS12596638 (16:54115829)	0.13	0.18	0.83	0.79	0.72
MC1R	RS75570604 (16:89846677)	0.03 (1.36)	0.11	0.62	0.32	0.63
ASIP	RS6088372 (20:32586748)	0.16	0.72	0.90	0.78	0.19
MX2	RS408825 (21:42743496)	0.15	0.85	0.59	0.91	0.40
PLA2G6	RS2092180 (22:38571563)	0.46	0.25	0.64	0.89	0.42

[a] P-values from simple linear regression; values in bracket are coefficients for significant association with P-value <0.05

[b] P-values from simple logistic regression: values in bracket are odds ratio for significant association with P-value <0.05

The highlighted cells indicate significant association at P-value < 0.05

5.3.5 Association of top melanoma susceptibility SNPs with the expression levels of nearby genes

Table 5.12 shows the association of top melanoma susceptibility SNPs with the expression levels of nearby genes using simple linear regression for each locus. For genes with more than one probe, all probes were used to test the association. Of the 20 SNPs analysed, five (RS12410869 in *ARNT*, RS6750047 in *RMDN2*, RS6914598 in *CDKAL1*, RS6088372 in *ASIP*, and RS408825 in *MX2* region) show significant association with gene expression at the 5% significance level.

Table 5.12 Association of top melanoma risk SNPs with expression levels of nearby genes in melanoma tumours (n=619)

Loci	Top SNP (Position)	Probe (Gene)	β (SE)	P-value
ARNT	RS12410869 (1: 150856153)	ILMN_1762582 (ARNT)	-0.23 (0.03)	8.27 x 10^{-11}
	RS12410869	ILMN_2347314 (ARNT)	-0.06 (0.03)	0.05
PARP1	RS1858550 (1:226608104)	ILMN_1686871 (PARP1)	0.05 (0.05)	0.36
RMDN2 (CYP1B1)	RS6750047 (2:38276549)	ILMN_1812302 (RMDN2)	0.01 (0.04)	0.77
	RS6750047	ILMN_1693338 (CYP1B1)	0.19 (0.07)	8.65 x 10^{-3}
CASP8	RS7582362 (2:202176294)	ILMN_1673757 (CASP8)	-0.03 (0.04)	0.47
	RS7582362	ILMN_1787749 (CASP8)	0.03 (0.09)	0.74
	RS7582362	ILMN_1809313 (CASP8)	0.01 (0.06)	0.92
	RS7582362	ILMN_2377733 (CASP8)	-0.02 (0.05)	0.74
TERT (CLPTM1L)	RS380286 (5:1320247)	ILMN_1796005 (TERT)	-0.003 (0.06)	0.96
	RS380286	ILMN_2373119 (TERT)	-0.02 (0.07)	0.77
	RS380286	ILMN_1752802 (CLPTM1L)	-0.02 (0.07)	0.79
SLC45A2	RS250417 (5:33952378)	ILMN_1654165 (SLC45A2)	0.14 (0.33)	0.68
	RS250417	ILMN_1685259 (SLC45A2)	0.04 (0.28)	0.90
	RS250417	ILMN_2246188 (SLC45A2)	0.16 (0.31)	0.61
	RS250417	ILMN_2320391 (SLC45A2)	-0.11 (0.34)	0.75
CDKAL1	RS6914598 (6:21163919)	ILMN_1788022 (CDKAL1)	-0.08 (0.04)	0.03
AGR3	RS1636744 (7:16984280)	ILMN_1728787 (AGR3)	-0.02 (0.04)	0.68
	RS1636744	ILMN_2050246 (AGR3)	-0.05 (0.04)	0.26
CDKN2A (MTAP)	RS7852450 (9:21825075)	ILMN_1717714 (CDKN2A)	-0.07 (0.08)	0.38
	RS7852450	ILMN_1744295 (CDKN2A)	0.04 (0.07)	0.57
	RS7852450	ILMN_1757255 (CDKN2A)	-0.15 (0.09)	0.14
	RS7852450	ILMN_1753639 (MTAP)	0.05 (0.05)	0.35
TMEM38B (RAD23B, TAL2)	RS10739221 (9:109060830)	ILMN_1669940 (TMEM38B)	-0.02 (0.04)	0.55

	RS10739221	ILMN_2093980 (TMEM38B)	0.07 (0.07)	0.33
	RS10739221	ILMN_1722662 (RAD23B)	0.03 (0.03)	0.43
	RS10739221	ILMN_2135833 (TAL2)	0.04 (0.04)	0.25
OBFC1	RS2995264 (10:105668843)	ILMN_1789186 (OBFC1)	0.02 (0.05)	0.69
CCND1	RS498136 (11:69367118)	ILMN_1688480 (CCND1)	0.03 (0.03)	0.41
TYR	RS1393350 (11:89011046)	ILMN_1788774 (TYR)	-0.11 (0.06)	0.06
ATM	RS73008229 (11:108187689)	ILMN_1713630 (ATM)	0.06 (0.09)	0.48
	RS73008229	ILMN_1716231 (ATM)	-0.02 (0.04)	0.72
	RS73008229	ILMN_1779214 (ATM)	0.003 (0.05)	0.95
	RS73008229	ILMN_2370825 (ATM)	0.06 (0.06)	0.28
OCA2	RS4778138 (15:28335820)	ILMN_1746116 (OCA2)	-0.23 (0.21)	0.28
FTO	RS12596638 (16:54115829)	ILMN_2288070 (FTO)	0.07 (0.07)	0.33
MC1R	RS75570604 (16:89846677)	ILMN_1653319 (MC1R)	-0.11 (0.06)	0.08
ASIP	RS6088372 (20:32586748)	ILMN_1791647 (ASIP)	-0.27 (0.12)	0.03
MX2	RS408825 (21:42743496)	ILMN_2231928 (MX2)	-0.26 (0.05)	1.32×10^{-6}
PLA2G6	RS2092180 (22:38571563)	ILMN_1697654 (PLA2G6)	0.03 (0.03)	0.22
	RS2092180	ILMN_1798955 (PLA2G6	-0.03 (0.05)	0.51

The highlighted rows indicate significant association at P-value < 0.05

5.3.6 Association of other SNPs in the susceptibility regions with the expression level of nearby genes

Table 5.13 shows the association of other SNPs in the regions (analysed as genotype dosage) with the expression levels of nearby genes in each locus. Figures 5.17 to 5.57 in Appendix I shows the Manhattan plot for association test in each locus. When including other SNPs in the region, significant associations were found in two regions (*ARNT* and *MX2*) after multiple testing corrections (10% FDR).

In the *ARNT* region, 361 SNPs were associated with the expression level from probe ILMN_1762582 after multiple testing correction (Figure 5.17). However, when conditioned on the most significant SNP (RS11204735, P-value=2.4 x 10^{-12}) in simple linear regression, no further associations were seen in the 361 SNPs at 5% significance level. All SNPs identified in the simple linear regression were highly correlated with SNP RS11204735; therefore, it is unclear which SNP is causal.

In the MX2 region, 18 SNPs showed significant association with MX2 expression level after multiple testing correction (Figure 5.55). After conditioning on the most significant SNP in simple linear regression (RS376364, P-value=6.7 x 10^{-7}), none of the SNPs remained significant at the 5% significance level. As all SNPs identified in the simple linear regression were also highly correlated with one another, it is again unclear which SNP is causal.

Table 5.13 Association of other SNPs in the susceptibility region with expression level of nearby gene (n=619)

Region (Total no. of SNPs in the region[†])	Probe	Total no. of sig SNPs[*]	Top SNP (Position)	β (SE)	P-value*
ARNT (2533)	ILMN_1762582 (ARNT)	361	RS11204735 (1:150841667)	0.23 (0.03)	2.37 x 10^{-12}
	ILMN_2347314 (ARNT)	-	RS146468719 (1:151000790)	-0.12 (0.04)	1.77 x 10^{-3}
PARP1 (3361)	ILMN_1686871 (PARP1)	-	RS114646469 (1:226601738)	-0.56 (0.20)	5.96 x 10^{-3}
RMDN2 (CYP1B1) (4590)	ILMN_1812302 (RMDN2)	-	RS10166521 (2:38116186)	-0.17 (0.06)	2.64 x 10^{-3}
	ILMN_1693338 (CYP1B1)	-	RS336031 (2:38265130)	-0.26 (0.07)	2.16 x 10^{-4}
CASP8 (2810)	ILMN_1673757 (CASP8)	-	RS6747200 (2:202418976)	0.09 (0.04)	8.82 x 10^{-3}
	ILMN_1787749 (CASP8)	-	RS11890734 (2:202384886)	-0.50 (0.17)	3.23 x 10^{-3}
	ILMN_1809313 (CASP8)	-	RS553977725 2:202303828	0.85 (0.32)	7.76 x 10^{-3}
	ILMN_2377733 (CASP8)	-	RS78982960 2:202395313	0.31 (0.09)	7.27 x 10^{-4}
TERT (CLPTM1L) (4582)	ILMN_1796005 (TERT)	-	RS145297127 (5:1343525)	-0.80 (0.21)	1.12 x 10^{-4}
	ILMN_2373119 (TERT)	-	RS148487301 (5:1318797)	-0.49 (0.16)	2.98 x 10^{-3}
	ILMN_1752802 (CLPTM1L)	-	RS183615159 (5:1187235)	1.24 (0.33)	2.27 x 10^{-4}
SLC45A2 (2160)	ILMN_1654165 (SLC45A2)	-	RS10044427 (5:33811670)	-0.32 (0.11)	5.36 x 10^{-3}
	ILMN_1685259 (SLC45A2)	-	RS10044427 (5:33811670)	-0.29 (0.10)	3.34 x 10^{-3}
	ILMN_2246188 (SLC45A2)	-	RS112493515 (5:33789210)	0.84 (0.32)	7.97 x 10^{-3}
	ILMN_2320391 (SLC45A2)	-	RS10941098 (5:33816914)	-0.26 (0.09)	4.92 x 10^{-3}
CDKAL1 (3620)	ILMN_1788022 (CDKAL1)	-	RS145289628 (6:21337898)	0.42 (0.09)	2.37 x 10^{-5}
AGR3 (4644)	ILMN_1728787 (AGR3)	-	RS41363746 (7:17140827)	-0.46 (0.11)	1.85 x 10^{-5}
	ILMN_2050246 (AGR3)	-	RS2892833 (7:16913557)	-0.16 (0.05)	4.34 x 10^{-4}
CDKN2A (MTAP) (3938)	ILMN_1717714 (CDKN2A)	-	RS112793495 (9:22018696)	-1.09 (0.34)	1.49 x 10^{-3}
	ILMN_1744295 (CDKN2A)	-	RS112793495 (9:22018696)	-0.96 (0.33)	3.43 x 10^{-3}
	ILMN_1757255 (CDKN2A)	-	RS143469042 (9:21602348)	1.62 (0.60)	7.30 x 10^{-3}
	ILMN_1753639 (MTAP)	-	RS11787926 (9:21575784)	0.49 (0.18)	5.76 x 10^{-3}
TMEM38B (RAD23B,	ILMN_1669940 (TMEM38B)	-	RS112167989 (9:109224990)	0.26 (0.09)	9.66 x 10^{-3}

TAL2) (3232)	ILMN_2093980 (TMEM38B)	-	RS7046731 (9:109270933)	-0.34 (0.12)	4.57 x 10^{-3}
	ILMN_1722662 (RAD23B)	-	RS56927240 (9:109148074)	0.53 (0.16)	1.42 x 10^{-3}
	ILMN_2135833 (TAL2)	-	RS62575303 (9:109006118)	0.40 (0.11)	4.53 x 10^{-4}
OBFC1 (3353)	ILMN_1789186 (OBFC1)	-	RS35478175 (10:105553793)	0.15 (0.06)	0.01
CCND1 (3922)	ILMN_1688480 (CCND1)	-	RS3212892 (11:69465860)	0.38 (0.11)	5.58 x 10^{-4}
TYR (2889)	ILMN_1788774 (TYR)	-	RS6483008 (11:88999996)	0.90 (0.28)	1.33 x 10^{-3}
ATM (2643)	ILMN_1713630 (ATM)	-	RS4356187 (11:108355227)	-0.21 (0.06)	5.56 x 10^{-4}
	ILMN_1716231 (ATM)	-	RS877474 (11:108382303)	-0.23 (0.10)	0.02
	ILMN_1779214 (ATM)	-	RS79942405 (11:108381644)	-0.25 (0.08)	1.25 x 10^{-3}
	ILMN_2370825 (ATM)	-	RS72992174 (11:108330256)	-0.39 (0.14)	6.22 x 10^{-3}
OCA2 (1763)	ILMN_1746116 (OCA2)	-	RS1129038 (15:28356859)	0.41 (0.15)	5.23 x 10^{-3}
FTO (3785)	ILMN_2288070 (FTO)	-	RS58687241 (16:54023335)	0.24 (0.07)	3.01 x 10-^{4}
MC1R (4344)	ILMN_1653319 (MC1R)	-	RS9939542 (16:90053048)	-0.13 (0.04)	2.37 x 10^{-3}
ASIP (2564)	ILMN_1791647 (ASIP)	-	RS6059655 (20:32665748)	0.39 (0.14)	6.02 x 10^{-3}
MX2 (3812)	ILMN_2231928 (MX2)	18	RS376364 (21:42746568)	-0.27 (0.05)	6.65 x 10^{-7}
PLA2G6 (3175)	ILMN_1697654 (PLA2G6)	-	RS11798033 (22:38653322)	0.41 (0.13)	1.05 x 10^{-3}
	ILMN_1798955 (PLA2G6)	-	RS143452361 (22:38431407)	0.77 (0.26)	3.43 x 10^{-3}

The highlighted rows indicate significant associations after multiple testing corrections using Benjamini and Yekutielli (2001) method (10% FDR)

[†] 500kb on either side of the top melanoma susceptibility SNP in each region

[*] The P-values shown were P-values before correction, and highlighted as significant if they meet the 10% FDR

5.4 Discussion

5.4.1 Understanding the gene expression and SNP associations by relating them to clinical predictors

When the associations of the 16 gene expression levels with five established clinical predictors for MSS were explored, most of the selected gene expression levels showed significant association with clinical predictors; 10 gene expression levels associated with age at diagnosis, 3 associated with sex, 7 associated with tumour site, 15 associated with log-transformed Breslow thickness, and 11 associated with presence of ulceration. When using the gene expression score, associations remained significant for all clinical predictors; higher gene expression score associated with increased age, increased log-transformed Breslow thickness, and increased odds for being male, having tumour at the rest of the body, and presence of ulceration, which are unfavourable predictors for melanoma survival (Thorn *et al*., 1994; Lindholm *et al*., 2004; Leiter *et al.,* 2004; Buettner *et al*., 2005; Balch *et al*., 2009).

As shown in Table 5.14, the significant gene expression levels that are associated with good prognosis according to one clinical predictor, also consistently have similar direction of effect on other predictors, and likewise for genes that are associated with poor prognosis. For example, doubling expression level of *NKD2* is associated with younger age at onset, decreased Breslow thickness, reduce odds for male sex, reduced odds for tumour site at the rest of the body (vs limbs), and reduced odds for presence of ulceration. This is not altogether surprising, because many of the clinical predictors are correlated with each other (see Table 6.5 in Chapter 6). Results in this analysis shows that gene expression levels are associated with important clinical predictors for melanoma survival; hence, increased expression levels for the 16 top gene expression levels identified in Chapter 3 may affect MSS through the clinical predictors.

For the 13 selected SNPs, significant associations with clinical predictors were seen for 3 SNPs only (RS4768090 with age at diagnosis, tumour site, and presence of ulceration; RS11639902 with presence of ulceration; RS12519276 with age at diagnosis, and log-transformed Breslow thickness). However, these associations were not as strong as the associations of gene expression levels with clinical predictors, suggesting that genetic variants may have no strong influence on the clinical predictors. On the other hand, these associations could only be false positives as most of the associated SNPs were only marginally significant.

Table 5.14 Summary of the significant associations of 16 gene expression levels with clinical predictors

	Significant associations with good prognostic indicators		Significant associations poor prognostic indicators	
	Genes with negative coefficient	Interpretations	Genes with positive coefficient	Interpretations
Age at diagnosis	LPAR1, NKD2, C1R, C21orf63, HLA-DQB2, and FGF22	Doubling the expression level of 6 genes associated with reduced age	DLG1, OSTC, SEC22B, and IGSF5	Doubling the expression level of 4 genes associated with increased age
Log-transfromed Breslow thickness	LPAR1, NKD2, C1R, SNORA12, c21orf63, HLA-DQB2, HLA-B, and FGF22	Doubling the expression level of 8 genes associated with decreased Breslow thickness	ZNF692, DLG1, OSTC, SEC22B, NDUFA8, IGSF5, and CIAPIN1	Doubling the expression level of 7 genes associated with increased Breslow thickness
	Significant associations with good prognostic indicators		Significant associations with poor prognostic indicators	
	Genes with OR <1	Interpretations	Genes with OR >1	Interpretations
Sex (female vs male)	NKD2 and FGF22	Doubling the expression level of 2 genes associated with reduced odds of male sex	DLG1	Doubling the expression level of 1 gene associated with increased odds of male sex
Tumour site (limbs vs rest of the body)	LPAR1, NKD2, C1R, SNORA12m HLA-DQB2, and HLA-B	Doubling the expression level of 6 genes associated with reduced odds of tumour at the rest of the body	NDUFA8	Doubling the expression level of 1 gene associated with increased odds of tumour at the rest of the body
Presence of ulceration (absence vs presence)	LPAR1, NKD2, C1R, SNORA12, HLA-DQB2, HLA-B, and FGF22	Doubling the expression level of 7 genes associated with reduced odds for presence of ulceration	DLG1, SEC22B, NDUFA8, and IGSF5	Doubling the expression level of 4 genes associated with increased odds for presence of ulceration

5.4.2 GWAS of 16 gene expression levels

When GWAS of the 16 expression levels were performed, results show that SNPs could predict the expression level of two genes (*HLA-DQB2* and *NDUFA8*) at genome-wide significance level. For the other 14 expression levels, while not reaching the genome-wide significance level, the strongest associations reached P-value<10^{-6} (for 6 genes) and P-value<10^{-7} (for 8 genes), which indicates suggestive association.

For the two expression levels with associations that attained genome-wide significance levels, there were 29 and 2 eQTL SNPs for *HLA-DQB2* and *NDUFA8*, respectively. Further analysis conditioning on the top SNP for *HLA-DQB2* shows other significant SNPs that reached P-value<10^{-6}, indicating that eQTL signals for *HLA-DQB2* could be explained by more than one SNP. For NDUFA8, no further associations were found when conditioned on the most significant SNP from the simple linear regression, suggesting that the eQTL signals for *NDUFA8* could be explained by one SNP. However, it is unclear which SNP is the causal SNP due to high LD between the SNPs.

Using the National Centre for Biotechnological Information (NCBI) eQTL browser, one of the 29 SNPs associated with *HLA-DQB2* expression level in the simple linear regression was found as an eQTL SNP in liver tissue (RS1573649, P-value=1.1 x 10^{-14} from Kruskall-Wallis test). This SNP was the second most significant SNP in the current analysis and was also highly correlated with the most significant SNP (RS5019296), suggesting that these the SNPs could have a genuine eQTL signal in melanoma tissue. For *NDUFA8* expression, no SNPs were identified as eQTLs in other tissues.

As results in this analysis show that gene expression levels in the tumour are associated with SNPs to some degree, future analysis could explore how well gene expression levels could be predicted for patients without expression data but with genotype data. A recent study by Gusev *et al.* (2016) as described in Chapter 1 has introduced a new method using a smaller set of individuals with both gene expression and genotype data available as a reference panel to impute gene expression data in patients where expression levels were not measured.

5.4.3 Investigation of melanoma susceptibility SNPs

When the associations of the 20 top melanoma susceptibility SNPs with MSS were explored, only SNP RS1858550 in *PARP1* was predictive of MSS (P-value=2.3 x 10^{-3}). When corrected for multiple testing using a Bonferroni correction for 20 tests, this association is only borderline significant at the 5% level. However, association of the risk variants in *PARP1* with melanoma survival has been previously reported by Davies *et al.* (2014b) and Law *et al.* (2015b). For susceptibility SNPs in other loci, no significant associations were seen, but it cannot be concluded from our results that these SNPs have no effect on MSS as the analysis does not have adequate power to detect weak associations.

Of the 20 top susceptibility SNPs, two SNPs (RS498136 in *CCND1* and RS75570604 in *MC1R*) were significantly associated with age at diagnosis, three (RS1858550 in *PARP1*, RS2995264 in *OBFC1*, and RS73008229 in *ATM*) were significantly associated with sex, three (RS6750047 in *RMDN2*, RS250417 in *SLC45A2*, and RS10739221 in *TMEM38B*) were significantly associated with log-transformed Breslow thickness, and one (RS1393350 in *TYR*) was associated with presence of ulceration. However, the associations for these SNPs were not strong, suggesting that risk variants may not have strong effects on the important clinical predictors for melanoma survival and some are likely to be false positives especially the SNPs with marginal significant. As SNPs on autosomal chromosomes are unlikely to be related to sex in general populations, the significant associations between the three susceptibility SNPs with sex could be false positives too.

Previous studies have explored the association of melanoma susceptibility SNPs with factors that are associated with risk of developing melanoma, such as nevus count, pigmentation and telomere length (Barrett *et al*., 2011; Iles *et al*., 2013; Law *et al*., 2015a); however, the association of susceptibility SNPs with the established clinical predictors for melanoma survival is still unclear in the literature.

When the association of the 20 top susceptibility SNPs with the expression levels of nearby genes was investigated, only five were associated with expression levels in neighbouring genes at the 5% significance level: RS12410869 was associated with *ARNT* expression (P-value=8.2 x 10^{-11}), RS6750047 with *RMDN2* expression (P-value=8.7 x 10^{-3}), RS6914598 with *CDKAL1* expression (P-value=0.03), RS6088372 with *ASIP* expression (P-value=0.03), and RS408825 with *MX2* expression (P-value=1.3 x 10^{-6}). However, using the NCBI eQTL browser, none of the 20 melanoma susceptibility SNPs were identified as eQTLs in other tissues.

When including other SNPs in the susceptibility loci, evidence of eQTLs were found in two regions (*ARNT* and *MX*2). A total of 361 SNPs in the *ARNT* region were associated with the expression level from probe ILMN_1762582 in simple linear regression, but no further associations were observed at 5% level of significance after conditioning on the most significant SNP (RS11204735, P-value=2.4 x 10^{-12}). In the *MX2* region, 18 SNPs were associated with expression levels from probe ILMN_2231928 in simple linear regression, but none were significant after conditioning on the most significant SNP (RS376364, P-value=6.7 x 10^{-7}). Using the NCBI eQTL browser, two of the 361 associated SNPs in the *ARNT* region were also eQTL SNPs in liver tissue: RS1088395 (P-value=3.9 x 10^{-11} from Kruskall-Wallis test) and RS3768015 (P-value=7.7 x 10^{-10} from Kruskall-Wallis test). Both SNPs were also highly correlated with the most significant SNP (RS11204735) in this analysis. For *MX2*, none of the associated SNPs were found as eQTL in other tissues.

In summary, results in this analysis suggest that a melanoma susceptibility SNP in *PARP1* is associated with MSS. Also, susceptibility SNPs may not have a strong effect on the clinical prognostic factors for melanoma. It is unclear which SNP has the regulatory effect on *ARNT* and *MX2* expression due to strong LD between the SNPs, but results suggest that the cis-eQTL signals for both *ARNT* and *MX2* can be explained by one SNP.

Chapter 6 Models combining different types of variable

The aims in this chapter are to:

i. Combine the selected predictors from different types of variable to build MSS models using different approaches

ii. Assess the predictive performance of the prognostic models

6.1 Introduction

A prognostic factor is defined as a measure at a given starting point such as diagnosis of disease, that is associated with a subsequent endpoint such as death (Riley *et al*., 2013). A prognostic model is the use of multiple prognostic factors in combination to predict the risk of developing future clinical outcomes in individual patients (Steyerberg *et al*., 2013). Prognostic models can be used to guide clinical decisions for a patient's treatment (Steyerberg *et al*., 2013). Current prognostic models for primary melanoma are based on the AJCC staging system, which is comprised of tumour and histology characteristics, such as tumour thickness, presence of ulceration, and mitotic rate. Several studies, as discussed in Chapter 1, have looked into incorporating gene expression data with clinical data to improve current prognostic models for primary melanoma, and reported potential use of gene expression data as predictors for melanoma outcome (Conway *et al.,* 2009; Harbst *et al.,* 2012; Nsengimana *et al.*, 2015). There are also a few studies that suggested the potential use of genetic variants as predictors for melanoma survival (Rendlemen *et al.,* 2009; Davies *et al*., 2012; Davies *et al*., 2014a; Davies *et al*., 2014b; Taylor *et al.,* 2015a; Orlow *et al.,* 2016). However, no published study so far has explored the joint effect of clinical factors, gene expression levels and genetic variants on melanoma survival. Hence, this chapter will explore the combined effect of clinical predictors, gene expressions, and genetic variants on MSS, and assess how well these variables predict melanoma survival both individually and jointly.

The next section describes the analyses for combining clinical predictors, gene expression levels and genetic variants to build prognostic models for melanoma survival. The data (clinical factors, whole-genome gene expression data, and genome-wide SNP data) used for analyses in this chapter were described in Chapter 2. Each type of data was analysed separately using a training set to identify the important predictors associated with MSS. Then, the selected predictors from the training set analyses were combined in an independent test set using three approaches (described in §6.2.2), and the predictive performance of models built in each approach was assessed. The splitting of data into a training set and a test set was as described in Chapter 3 (see §3.2.2.2). Four models were fitted, and this will be described further in §6.2.3. Section 6.3 presents the results from the four fitted models, and in Section 6.4 the results are discussed, comparing the predictive performance between the three approaches and the four models applied in this chapter.

6.2 Methods

6.2.1 Methods for integrating data

One strategy for building a prognostic model is to develop a model using a discovery dataset and to assess the model's predictive performance using an independent validation dataset, ideally using external data from a different source. When using the same data to build the prognostic model and to assess its performance, over-fitting occurs, and the model developed is not expected to perform as well when applied to new data. When external data are not available for validation, available datasets that are large enough can be randomly split into a 2/3 training set to develop the model and a 1/3 test set to assess the model (Harrell *et al.*, 1996; Harrell, 2001).

When using -omics data to build prognostic models, using ordinary regression to select important features is subject to over-fitting and unstable coefficients due to high-dimensionality and collinearity. Hence, the use of multiple -omics datasets to develop new prognostic models presents a challenge. The two main approaches to integrate different types of -omics data as discussed by Ritchie *et al.* (2015) in Chapter 1 (see §1.3) are multi-staged analysis (involves integrating information in a hierarchical manner) and meta-dimensional analysis (involves integrating multiple different types of data simultaneously).

Multi-staged analysis starts by finding associations between two or three different types of data, then uses results derived from the first level analysis to look for associations with another type of data in the second level analysis. Several studies have implemented this method to integrate genomic and transcriptomic data to identify eQTL patterns, which were then used in subsequent analysis. For example, Curtis *et al.* (2012) analysed the association of copy number variants, SNPs, and somatic copy number aberrations with gene expression levels to identify eQTL patterns (genome-wide eQTL, cis-eQTL, and trans-eQTL) in breast cancer patients. The authors then found that cis-copy number aberrations dominated the expression landscape, and used the top 1000 cis-associated genes as input for cluster analysis to identify novel subgroups of breast tumours. Burkhardt *et al.* (2015) also explored the association of SNPs and gene expression to identify eQTL patterns in whole blood, which were then used to identify novel regulatory mechanisms for amino acids and acylcarnites (compounds for the metabolism of fatty acids) in whole blood. Similarly, Gusev *et al.* (2016) integrated SNPs and gene expression data to identify eQTLs for height, and subsequently correlated these with the trait.

In meta-dimensional analysis, various studies used data reduction first before integrating different types of high-dimensional data. Data reduction methods that are commonly used are cluster analysis or penalized regression analysis. Several studies have implemented cluster analysis to reduce the dimension of gene expression data. In Pittman *et al.* (2004), cluster analysis was used to create clusters of genes, which were used to construct metagenes, an aggregate measure of expression of sets of genes. Then, the metagenes were integrated with clinical data to predict recurrence in breast cancer patients using classification tree models. In Verhaak *et al.* (2010), cluster analysis was used to identify gene-expression based subtypes of glioblastoma multiforme, a common brain tumour in adults. The tumour subtypes were then combined with genomic data (somatic mutations and DNA copy number) to identify the genomic patterns differentiating the subtypes. In Gentles *et al.* (2015), cluster analysis was performed before selecting gene expression signatures for data integration to identify different clusters of survival-associated genes representing different biological features. The top five most significant genes from each cluster were used to compute a molecular prognostic index, which was then combined with a clinical prognostic index into a composite risk model for patient survival in non-small cell lung cancer patients.

An example of a study that implemented penalized regression analysis for data reduction is by Mankoo *et al.* (2011), who used lasso penalized Cox regression to perform feature selection to derive molecular signatures from multiple genomic data (mRNA expression, microRNA expression, DNA methylation, and copy number aberration). The resulting molecular signatures for each type of data were then integrated to predict progression-free survival and overall survival in ovarian cancer patients.

6.2.2 Overview of the methods

This section presents an overview of the approaches taken here to combine the selected predictors, explained in more detail in §6.2.3 and §6.2.4.

The test set samples used to combine different types of predictor in this chapter was based on the split of gene expression data (n=699) into a training set (n=464) and a test set (n=235) as described in Chapter 3 §3.2.2.2. The final test samples consist of 190 samples with both gene expression and genotype data available. The analysis of each type of data in the training set excluded samples in the final test set samples and those with survival analysis exclusion criteria (patients with multiple melanomas, who were recruited into the study more than two years after diagnosis, or were missing cause of death).

In the training set, the important clinical predictors, gene expression levels, and genetic variants for MSS were identified. As the number of predictors for gene expression and genotype data is large (p >> N), a data reduction strategy is required to avoid over-fitting. Lasso penalized Cox regression was implemented to select the important predictors from -omics data as it not only reduces the dimensionality of data, but can also be used as a variable selection method. Unlike other prognostic studies in primary melanoma that use -omics data (Winnepenninckx *et al.*, 2006; Conway *et al.*, 2009; Davies *et al.,* 2012; Rendleman *et al.,* 2013; Davies *et al.,* 2014a; Davies *et al.,* 2014b; Gerami *et al.,* 2015), the feature that is relatively new in this analysis is the application of penalized Cox regression to identify -omics predictors associated with MSS. The analyses to identify important predictors from different types of data were conducted separately as follows:

- Clinical predictors: A training set consisting of 1,795 patients with clinical data available (excluding test set samples and those with survival analysis exclusion criteria) was used to explore the associations of five established clinical predictors with MSS using multivariable Cox regression
- Gene expression levels: A training set consisting of 424 patients with gene expression data available (excluding test set samples and those with survival analysis exclusion criteria) was used to explore the

associations of whole-genome gene expression levels (29,354 probes) with MSS using lasso penalized Cox regression

- SNPs: A training set consisting of 1,543 patients with genotype data available (excluding test set samples and those with survival analysis exclusion criteria) was used to explore the associations of selected SNPs across the genome (7,414 SNPs with P-values < 0.01 in univariable Cox models) with MSS using lasso penalized Cox regression

In the test set, combined data survival models were developed using the selected clinical predictors, gene expression levels and SNPs from the training set. Survival analysis exclusion criteria used in the training set were also applied in the test set. Three approaches were explored to combine the selected predictors from different types of data as follows:

- The first approach was based on **external estimation**. This approach assesses how well a prediction score developed in one dataset performs in independent data by using the estimates from the study that identified the predictor. Over-fitting does not occur as no re-estimation is done in the independent data. To apply this approach, a clinical score, gene expression score and SNP score were computed using estimates from the training set (estimates from multivariable Cox regression for clinical predictors and estimates from penalized Cox regression for both gene expression levels and SNPs). Then, the predictive performance of the scores was assessed individually and jointly using the *C*-index and AUC as further explained in §6.2.4.
- The second approach was based on **risk scores**. A clinical score, gene expression score and SNP score were also computed for this approach, but the effect of the scores on MSS was re-estimated in the test set using univariable and multivariable Cox regression. The scores were standardized before fitting the Cox regression as each score was on a different scale. The predictive performance of the scores was also assessed using the *C*-index and AUC individually and jointly. The scores were also calculated in a subset of 365 patients in the training set (samples with both gene expression and genotype data) to assess the predictive performance of the scores in the training set (these estimates

were expected to be subject to over-fitting, as the same data were used to build and assess the model; however, this was conducted as a baseline comparison).

- The third approach was based on **variable selection**. This approach was implemented to assess how well the individual predictors selected from the training set predict MSS in a new dataset; this was expected to give better estimates as the effect of each predictor was re-estimated in the new data. The selected predictors from the training set were combined using lasso penalized Cox regression analysis to deal with the high number of predictors and multicollinearity problem. The predictive performance of the full model, the model with selected predictors only, the model with selected -omic predictors only, and the model with Breslow thickness only was assessed using the *C*-index and AUC.

6.2.3 Models

Four models were fitted to combine the selected predictors and were compared in terms of predictive performance:

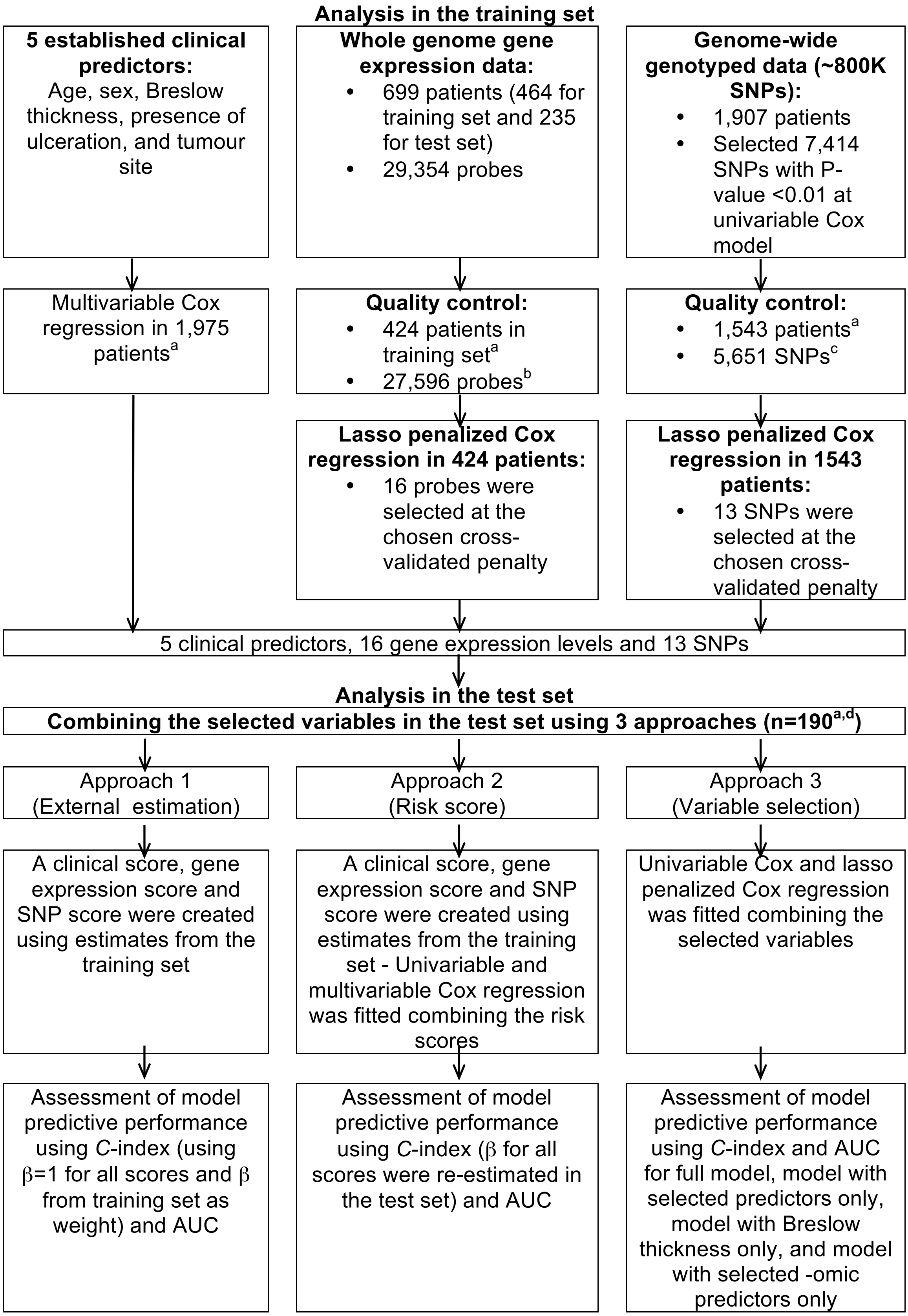
i. Model 1 Combined data survival models using variables selected from the training set in Chapter 3
ii. Model 2 Combined data survival models with prior cluster analysis
iii. Model 3 Combined data survival models using Lund clusters
iv. Model 4 Combined data survival models using clinical predictors and Lund classification

Model 1 (§6.2.3.1) was the main model to combine the selected predictors. However, as gene expression levels were found to be much more predictive of MSS than SNPs in the Chapter 3, alternative models were developed with the aim of improving predictive performance by investigating different methods of analysing the gene expression data as described further in Model 2 (§6.2.3.2) and Model 3 (§6.2.3.3).

Model 4 (§6.2.3.4) was not a new alternative model, but a classifier developed in another study conducted in Lund (Harbst *et al.*, 2012). This was analysed to determine how well it performed in the LMC to provide a comparison with a predictor that has been reported in the literature.

6.2.3.1 Model 1 Combined data survival models using variables selected from the training set in Chapter 3

Model 1 combines the selected five clinical predictors, 16 gene expression levels and 13 SNPs found in the training set (Chapter 3) using the **external estimation**, **risk score**, and **variable selection** approaches as described in §6.2.2. A flow chart of the analyses based on Model 1 is shown in Figure 6.1.


Analysis in the training set
5 established clinical predictors: Age, sex, Breslow thickness, presence of ulceration, and tumour site
Whole genome gene expression data: • 699 patients (464 for training set and 235 for test set) • 29,354 probes
Genome-wide genotyped data (~800K SNPs): • 1,907 patients • Selected 7,414 SNPs with P-value <0.01 at univariable Cox model
Multivariable Cox regression in 1,975 patients[a]
Quality control: • 424 patients in training set[a] • 27,596 probes[b]
Quality control: • 1,543 patients[a] • 5,651 SNPs[c]
Lasso penalized Cox regression in 424 patients: • 16 probes were selected at the chosen cross-validated penalty
Lasso penalized Cox regression in 1543 patients: • 13 SNPs were selected at the chosen cross-validated penalty
5 clinical predictors, 16 gene expression levels and 13 SNPs
Analysis in the test set
Combining the selected variables in the test set using 3 approaches (n=190[a,d])
Approach 1 (External estimation)
Approach 2 (Risk score)
Approach 3 (Variable selection)
A clinical score, gene expression score and SNP score were created using estimates from the training set
A clinical score, gene expression score and SNP score were created using estimates from the training set - Univariable and multivariable Cox regression was fitted combining the risk scores
Univariable Cox and lasso penalized Cox regression was fitted combining the selected variables
Assessment of model predictive performance using C-index (using β=1 for all scores and β from training set as weight) and AUC
Assessment of model predictive performance using C-index (β for all scores were re-estimated in the test set) and AUC
Assessment of model predictive performance using C-index and AUC for full model, model with selected predictors only, model with Breslow thickness only, and model with selected -omic predictors only

[a] Excluding patients with multiple melanomas, who were recruited into the study more than 2 years after diagnosis, or were missing cause of death

[b] Excluding probes with low proportion of samples detected and low variance

[c] Excluding SNPs with missing rate > 3%, MAF < 5% and P-value<10^{-4} for HWE test

[d] Patients with both gene expression and genotypes data in the test set (n=190)

Figure 6.1 Model 1 Combined data survival models

The idea of a risk score is explained further in §6.2.4.1, and below are **e**xample calculations to create the clinical score, gene expression score, and SNP score in the test set for Model 1. To calculate the clinical score in the test set, estimates from the multivariable Cox regression in Chapter 3 (Table 3.3) were used:

Clinical score = Age*0.03 + Sex(Male)*0.33 + Site†(Head/Neck)*0.11 + Breslow thickness*0.16 + Presence of ulceration (Yes)* 0.75

† If tumour site is trunk, the calculation is Site (Trunk)*0.44

† If tumour site is other, the calculation is Site (Other)*0.37

To calculate the gene expression score in the test set, estimates from the penalized Cox regression in Chapter 3 (Table 3.4) were used:

Gene expression score = ILMN_1701441*-0.05 + ILMN_3249501*0.17 + ILMN_1749829*0.10 + ILMN_1731206*-0.05 + ILMN_1764109*-0.03 + ILMN_2056167*0.03 + ILMN_3238435*-0.05 + ILMN_1695959*-0.06 + ILMN_1741648*-0.002 + ILMN_1784238*0.04 + ILMN_1778401*-0.02 + ILMN_1759729*0.03 + ILMN_2344221*0.04 + ILMN_2095633*-0.03 + ILMN_1700547*0.01 + ILMN_1735199*0.001

To calculate the SNP score in the test set, estimates from the penalized Cox regression in Chapter 3 (Table 3.8) were used:

SNP score = RS17837209*0.16 + RS9957831*0.02 + RS4768090*0.04 + RS2902554*0.05 + RS5770310*0.02 +RS10233832*-0.01 + RS17379771*0.02 + RS16956192*0.04 + RS2392477*-0.02 + RS6689263*-0.02 + RS11639902*-0.01 + RS12519276*0.01 + RS10941528*0.002

To calculate the *C*-index based on external estimation approach, the calculated clinical score, gene expression score and SNP score was used as the predictor in the model:

Predictor using β=1 for all scores = Clinical score*1 + Gene expression score*1 + SNP score*1

Predictor using β from training set (from Table 6.2) as weight = Clinical score*0.30 + Gene expression score*0.70 + SNP score*0.55

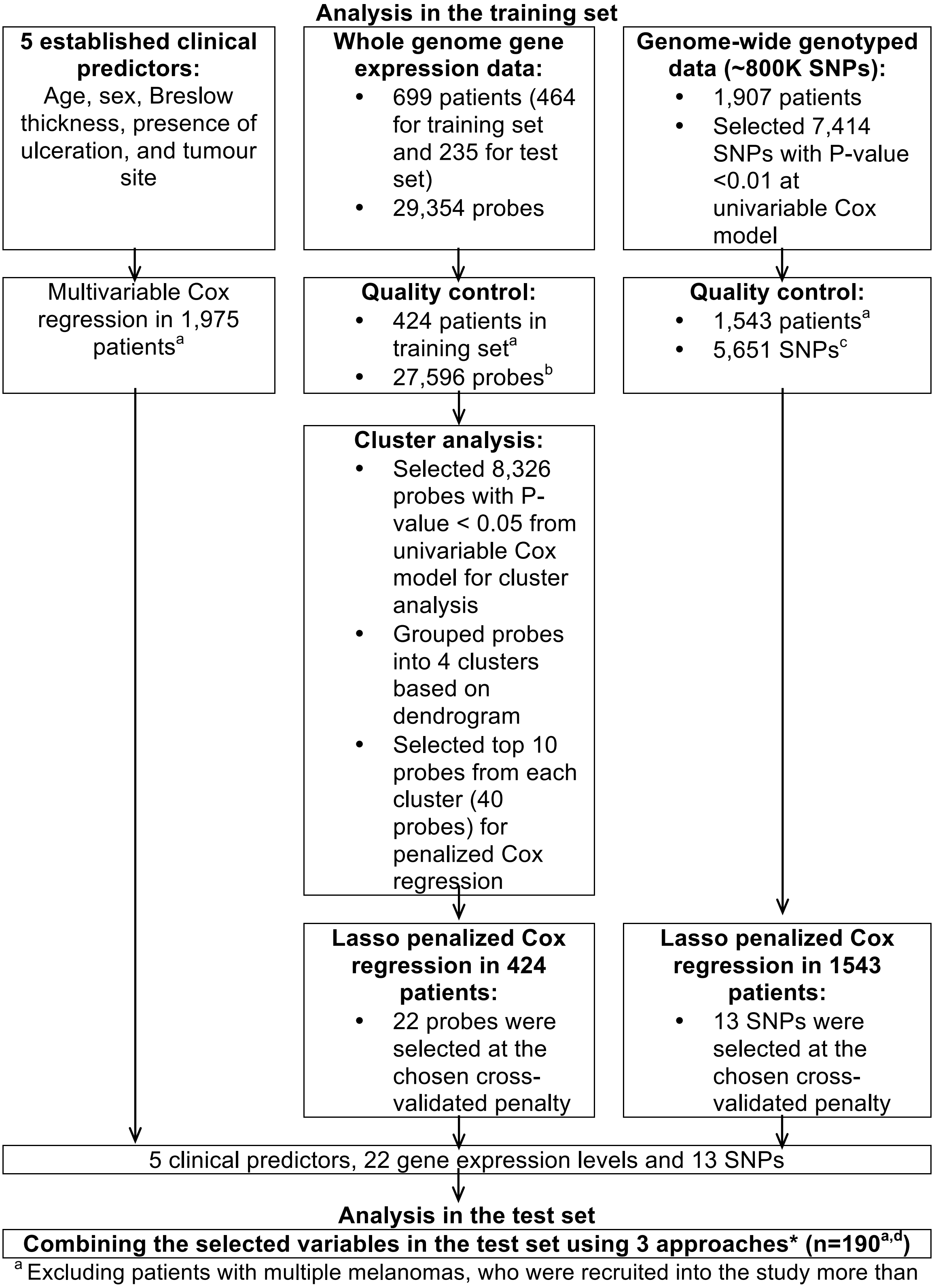
6.2.3.2 Model 2 Combined data survival models with prior cluster analysis

Model 2 is similar to Model 1 except in the analysis of gene expression probes, which were selected for inclusion in the model after a prior cluster analysis. A flow chart of the analyses based on Model 2 is shown in Figure 6.2.

The motivation for performing the cluster analysis was to give a better chance of including genes from different pathways among those selected in the penalized Cox regression analysis, as the previous analysis (Model 1) might select genes that are dominated by one pathway. This approach was based on Gentles *et al.* (2015) as discussed in 6.2.1.

Before cluster analysis, probes were filtered to select those associated with MSS at P-value less than 0.05 using univariable Cox regression. A total of 8,326 probes were selected for cluster analysis as further explained in §6.2.4.4. After deciding the number of clusters based on the dendrogram, the top 10 probes from each cluster were selected for penalized Cox regression of MSS. Only 10 probes from each cluster were selected for penalized regression as selecting too many probes from each cluster may cause the penalized model to select probes that were mainly or entirely from one cluster only, thus more likely to give similar results to Model 1.

After selecting probes from the penalized model, the selected gene expression levels were combined with the selected clinical predictors and SNPs using similar methods to Model 1.


Analysis in the training set
5 established clinical predictors: Age, sex, Breslow thickness, presence of ulceration, and tumour site
Whole genome gene expression data:
• 699 patients (464 for training set and 235 for test set)
• 29,354 probes
Genome-wide genotyped data (~800K SNPs):
• 1,907 patients
• Selected 7,414 SNPs with P-value <0.01 at univariable Cox model
Multivariable Cox regression in 1,975 patients^a
Quality control:
• 424 patients in training set^a
• 27,596 probes^b
Quality control:
• 1,543 patients^a
• 5,651 SNPs^c
Cluster analysis:
• Selected 8,326 probes with P-value < 0.05 from univariable Cox model for cluster analysis
• Grouped probes into 4 clusters based on dendrogram
• Selected top 10 probes from each cluster (40 probes) for penalized Cox regression
Lasso penalized Cox regression in 424 patients:
• 22 probes were selected at the chosen cross-validated penalty
Lasso penalized Cox regression in 1543 patients:
• 13 SNPs were selected at the chosen cross-validated penalty
5 clinical predictors, 22 gene expression levels and 13 SNPs
Analysis in the test set
Combining the selected variables in the test set using 3 approaches* (n=190^{a,d})

[a] Excluding patients with multiple melanomas, who were recruited into the study more than 2 years after diagnosis, or were missing cause of death

[b] Excluding probes with low proportion of samples detected and low variance

[c] Excluding SNPs with missing rate > 3%, MAF < 5% and P-value<10^{-4} for HWE test

[d] Patients with both gene expression and genotypes data in the test set (n=190)

*** Similar to approaches used in Model 1**

Figure 6.2 Model 2 Combined data survival models with prior cluster analysis

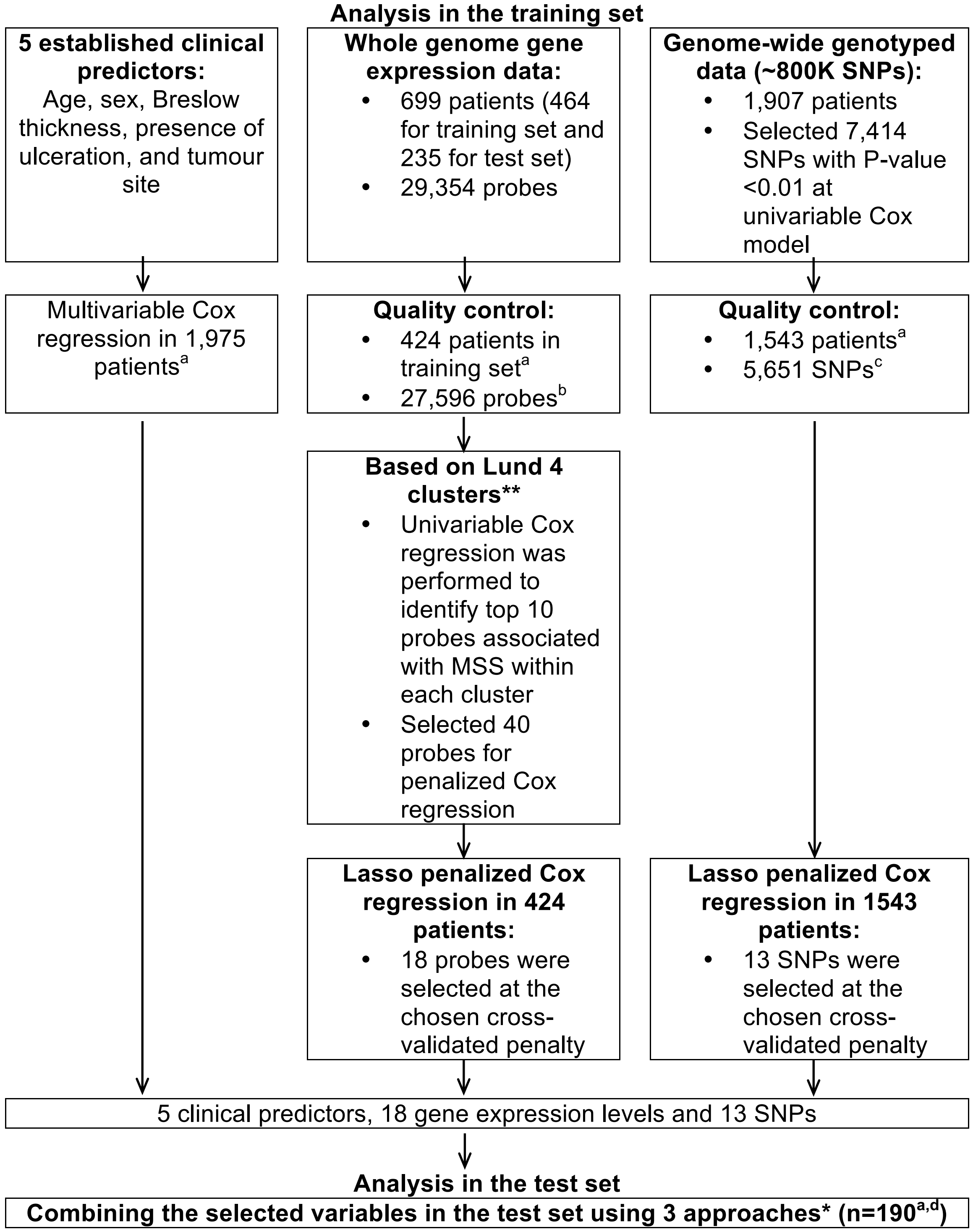
6.2.3.3 Model 3 Combined data survival models using Lund clusters

Model 3 differs from Model 2 in that the gene expression probes were selected from clusters based on a previously published cluster analysis (Harbst *et al*., 2012) as described in Chapter 1 §1.2.2. Figure 6.3 shows the flow chart of the analyses based on Model 3.

In Model 3, selection of the gene expression levels was performed using clusters of patients created by Harbst *et al.* (2012) in Lund to identify whether the model could be improved by using these clusters, which have a biological interpretation. In primary melanoma, Harbst *et al.* (2012) conducted an unsupervised hierarchical cluster analysis and found that patients can be classified into four molecular subgroups (*high-immune*, *normal-like*, *pigmentation* and *proliferative*) based on their gene expression signature. A list of 503 genes and their average gene expression values (across the four subgroups) that were used to classify patients was supplied by the Lund group. As the list contains gene names instead of probe names, there were some genes with more than one average expression values across the four subgroups. For the genes with several average expression values, only one value (the highest average expression value) was retained, leaving only 486 genes from which to create gene clusters based on Lund clusters.

Cluster membership for each gene in the list was defined by looking at the gene's average expression value across the four subgroups. When a cluster had the highest average expression for a gene compared to the other three clusters, the gene was assigned to that cluster. After identifying the gene's membership cluster in Lund's list, the gene names were matched to the whole genome DASL gene expression data in Leeds. There were 160 genes (228 probes) identified from the high-immune cluster, 136 genes (224 probes) from the normal-like cluster, 85 genes (142 probes) from the pigmentation cluster and 69 genes (100 probes) from the proliferative cluster. Of the 486 genes listed from Lund, only 450 genes were available in the LMC gene expression data (36 genes excluded).

In the training set, a univariable Cox regression was performed to identify the top 10 probes associated with MSS within each cluster. A total of 40 probes were then selected for penalized Cox regression. Only 10 probes were selected from each cluster to avoid selecting too many genes from the same cluster in the final model. After selecting probes from the penalized model, the selected gene expression levels were combined with the selected clinical predictors and SNPs using similar methods to Model 1.


Analysis in the training set
5 established clinical predictors:
Age, sex, Breslow thickness, presence of ulceration, and tumour site
Whole genome gene expression data:
• 699 patients (464 for training set and 235 for test set)
• 29,354 probes
Genome-wide genotyped data (~800K SNPs):
• 1,907 patients
• Selected 7,414 SNPs with P-value <0.01 at univariable Cox model
Multivariable Cox regression in 1,975 patients[a]
Quality control:
• 424 patients in training set[a]
• 27,596 probes[b]
Quality control:
• 1,543 patients[a]
• 5,651 SNPs[c]
Based on Lund 4 clusters**
• Univariable Cox regression was performed to identify top 10 probes associated with MSS within each cluster
• Selected 40 probes for penalized Cox regression
Lasso penalized Cox regression in 424 patients:
• 18 probes were selected at the chosen cross-validated penalty
Lasso penalized Cox regression in 1543 patients:
• 13 SNPs were selected at the chosen cross-validated penalty
5 clinical predictors, 18 gene expression levels and 13 SNPs
Analysis in the test set
Combining the selected variables in the test set using 3 approaches* (n=190[a,d])

[a] Excluding patients with multiple melanomas, who were recruited into the study more than 2 years after diagnosis, or were missing cause of death

[b] Excluding probes with low proportion of samples detected and low variance

[c] Excluding SNPs with missing rate > 3%, MAF < 5% and P-value<10^{-4} for HWE test

[d] Patients with both gene expression and genotypes data in the test set (n=190)

* Similar to approaches used in Model 1

**160 genes (228 probes) from high-immune cluster; 136 genes (224 probes) from normal cluster; 85 genes (142 probes) from pigmentation cluster; 69 genes (100 probes) from proliferative cluster

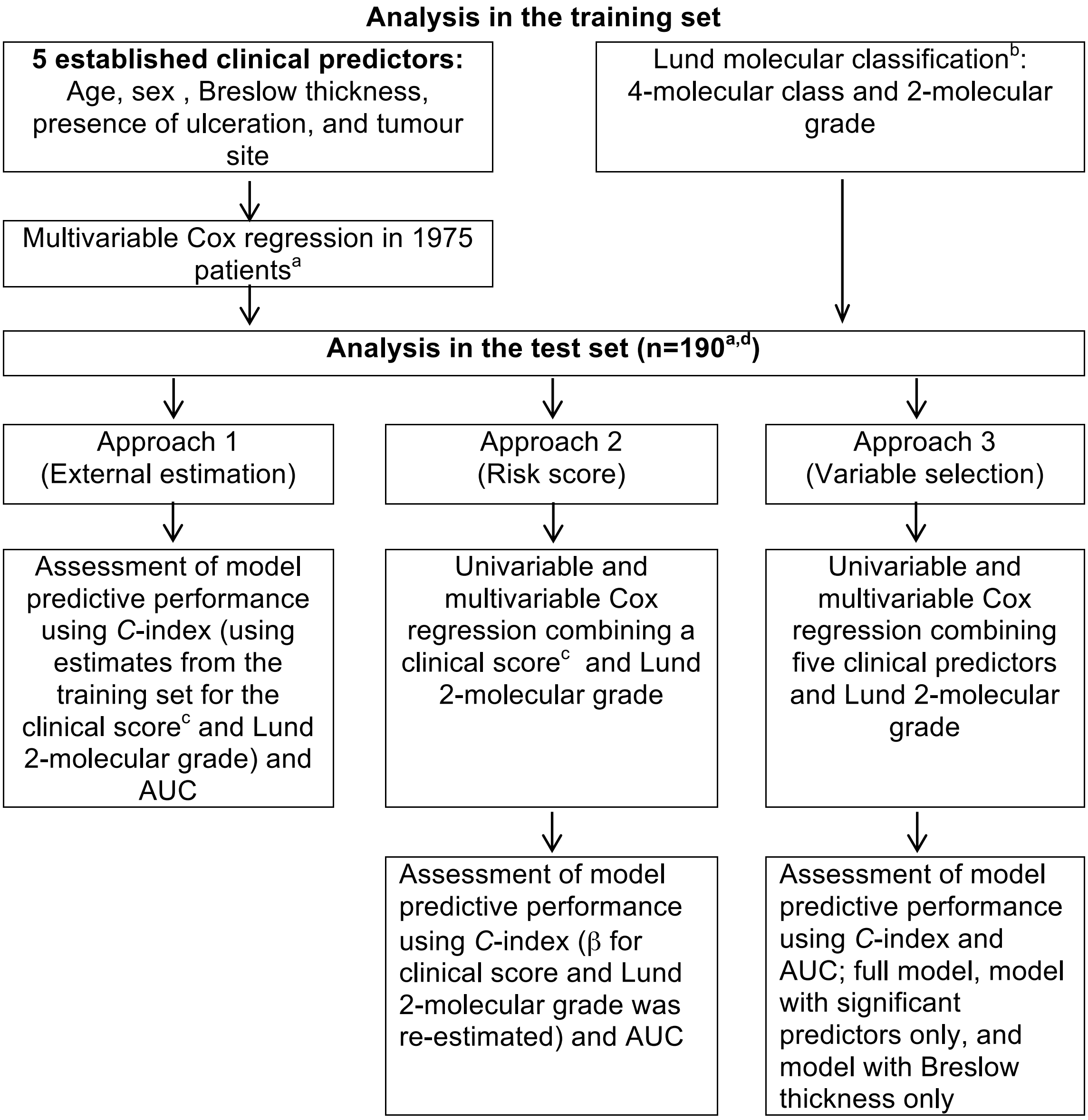
Figure 6.3 Model 3 Combined data survival models using Lund clusters

6.2.3.4 Model 4 Combined data survival models using clinical predictors and Lund classification

Model 4 combines the selected five clinical predictors with the Lund molecular classification of patients created by Harbst *et al.* (2012) to determine how well their model performs in the LMC.

The initial molecular classification developed by Harbst *et al.* (2012) classified patients into 4-molecular subgroups (high-immune, normal-like, pigmentation and proliferative). Later, they also found that patients can be classified into 2-molecular grades (low risk and high risk) which differed in terms of survival.

The *C*-index was calculated for each model combining the clinical score with the 2-molecular grades to determine whether the Lund classification could improve the clinical model. Patient classification in the test set was based on the 2-molecular grading (90 identified as low-grade and 96 as high-grade) as the test sample is too small when using the 4-molecular subgroups: 43 identified as high immune, 47 as normal-like, 74 as pigmentation, 22 as proliferative, and 4 unclassified. Figure 6.4 shows the Model 4 analyses flow chart.


Analysis in the training set
5 established clinical predictors: Age, sex , Breslow thickness, presence of ulceration, and tumour site
Lund molecular classification[b]: 4-molecular class and 2-molecular grade
Multivariable Cox regression in 1975 patients[a]
Analysis in the test set (n=190[a,d])
Approach 1 (External estimation)
Approach 2 (Risk score)
Approach 3 (Variable selection)
Assessment of model predictive performance using C-index (using estimates from the training set for the clinical score[c] and Lund 2-molecular grade) and AUC
Univariable and multivariable Cox regression combining a clinical score[c] and Lund 2-molecular grade
Univariable and multivariable Cox regression combining five clinical predictors and Lund 2-molecular grade
Assessment of model predictive performance using C-index (β for clinical score and Lund 2-molecular grade was re-estimated) and AUC
Assessment of model predictive performance using C-index and AUC; full model, model with significant predictors only, and model with Breslow thickness only

[a] Excluding patients with multiple melanomas, who were recruited into the study more than 2 years after diagnosis, or were missing cause of death

[b] Based on Harbst *et al.* (2012)

[c] Calculated using estimates from multivariable Cox regression of 5 clinical predictors in the training set

[d] Patients with both gene expression and genotypes data in the test set (n=190)

Figure 6.4 Model 4 Combined data survival models using clinical predictors and Lund classification

6.2.4 Statistical methods

6.2.4.1 Risk scores

"Risk score" in this study refers to a summary of the effect of several predictors weighted by their estimated effect size. Harrell (2001) suggested that in the presence of a large number of predictors, an alternative to including individual predictors in a model is to calculate a summary index of the related variables to reduce the number of predictors and to deal with correlation. Studies such as Gentles *et al.* (2015) have used summary scores instead of individual variables to develop their prognostic models. The clinical score in this chapter is the weighted sum over all selected clinical predictors (age, sex, Breslow thickness, presence of ulceration, and tumour site) and is calculated for each individual in the test set sample.

Clinical score = $\Sigma_i \beta_i c_i$,

where c_i is the value of variable *i*, and β_i is the coefficient obtained from the multivariable Cox regression in the training set.

The gene expression score in this chapter is the weighted sum over all gene expression levels selected by the penalized Cox model and is calculated for each individual in the test set sample.

Gene expression score = $\Sigma_i \beta_i g_i$,

where g_i is the gene expression level, and β_i is the estimate obtained from penalized Cox regression in the training set.

The SNP score in this chapter is the sum of genetic effects of trait-associated alleles weighted by their estimated effect sizes. The sum is over all trait-associated variants selected by the penalized Cox model and is calculated for each individual in the test set sample. Genetic risk scores (or polygenic risk scores) are now being used widely in genetic epidemiology (Dudbridge, 2013).

SNP score = $\Sigma_i \beta_i s_i$,

where s_i is the number of risk alleles carried by the individual at variant *i* (0, 1 or 2), and β_i is the estimate obtained from penalized Cox regression in the training set.

6.2.4.2 Univariable and multivariable Cox regression

Cox proportional hazards regression was performed to determine the significant predictors for MSS. This method has been described in Chapter 3. The Cox model was fitted using *coxph* function in the *survival* package in R software.

The Cox model is a semi-parametric model, it leaves the baseline hazard ($h_0(t)$) function unspecified, but assumes covariates enter the model linearly (Hosmer and Lemeshow, 1999). The hazard function for subject *j* at time is given by:

$$h_j(t) = h_0(t)\ \exp(ß_1 x_{j1} + ß_2 x_{j2} + ß_3 x_{j3} + \ldots + ß_p x_{jp})$$

where $ß_i$ is the coefficient and x_{ji} is the predictor, $i = 1, \ldots, p$.

The Cox model also assumes proportional hazards, where the effect of predictor X does not vary with time *t*. Residuals can be used to check the model assumptions. The most commonly used residuals for the Cox model are Schoenfeld residuals and Martingale residuals (Hosmer and Lemeshow, 1999).

Schoenfeld residuals represent the difference between the observed covariate and the expected covariate given the risk set at that time. The residuals are calculated for each covariate and can be used to check the proportional hazards assumption by plotting against time. Using a formal statistical test is a better method to assess the proportional hazards assumption than a graphical assessment. Therefore, *coxzph* function in the *survival* package was used to perform the Schoenfeld residuals test. The test provide P-value for individual predictor and a global test P-value for overall assumption of proportional hazards for all of the predictors. A P-value of less than 0.05 indicates the assumption is violated.

Martingale residuals represent the difference between the observed number of deaths (0 or 1) for subject *j*, and the expected number based on the fitted model. The residuals are defined for the j^{th} individual. Martingale residuals versus individual covariates can be plotted to check the linearity assumption of the covariate. A local linear regression curve that is parallel to the zero line in the plot indicates assumption of the linearity is fulfilled.

6.2.4.3 Penalized Cox regression

Penalized Cox regression was applied to select the important gene expression levels and SNPs in the training set. This method was described in Chapter 3 (see §3.2.3.3).

6.2.4.4 Cluster analysis

Gene expression clustering is used to group together genes based on similar patterns of gene expression so that genes within the same cluster have high similarity to each other, and genes in other clusters are less similar. The similarity between two expression patterns can be measured using proximity measures such as distance and correlation (Jiang *et al.*, 2004). The two most commonly used similarity measures for gene expression data are Euclidean distance and Pearson's correlation coefficient. In comparison of clustering methods for gene expression data, Gibbons and Roth (2002) shows that Euclidean distance performs better when applied to log ratio gene expression data, while Pearson's correlation performs better for non-ratio based data such as those from Affymetrix array technology.

In this analysis, hierarchical clustering was applied using the *hclust* package in R software using Euclidean distance as the similarity measure, and complete linkage (calculating the largest distance between any two members) as the method to calculate the distance between clusters. Other methods available include simple linkage (calculating the shortest distance between any two members) and average linkage (calculating the average distance between any two members), but the complete linkage method was chosen in this analysis as it has been found to outperform other methods (Gibbons and Roth, 2002).

Hierarchical clustering generates clusters that subdivide into a series of smaller clusters forming a tree-shaped data structure called a dendrogram. Through visual inspection for the most obvious cluster patterns, the dendrogram can be cut at some level to obtain a specified number of clusters. Then, the number of specified clusters can be used to reorder genes in the original dataset so that genes with similar expression patterns are grouped together.

6.2.4.5 Model predictive performance using *C*-index

Assessment of predictive accuracy for survival models include calibration and discrimination. Harrell *et al.* (1996) described calibration as a measure of agreement between the observed outcomes and predictions, and discrimination as a measure of the model's ability to distinguish individuals who experience the outcome from those who remained event free. Calibration in survival modelling can be assessed by creating risk groups (based on categorizing a prognostic score), and graphically comparing the Kaplan-Meier estimates of survival probabilities in these groups with the predicted survival from the prognostic model for the patients in each group. A model is well-calibrated if the Kaplan-Meier survival curves are close to the predicted survival curves. However, there is no consensus on the number of risk groups to be created and where to position the cut-off points. Categorizing a continuous score will also lead to loss of information and inaccurate prediction. Therefore, only discrimination was used to assess model performance in this analysis.

A commonly used measure of predictive discrimination is the *concordance* index or *C*-index. The *C*-index is defined by Harrell *et al.* (1996) as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. In survival modelling, the *C*-index is computed by identifying all possible pairs of subjects whose survival time can be ordered (where one has died at time t and the other survived up to at least that time), and then calculating the number of pairs that are concordant; the pair is concordant if the subject with higher predicted survival is the one who survived longer.

The *C*-index is obtained by summing the number of concordant pairs and dividing it with the number of all usable pairs. The *C*-index for censored data can be obtained by ignoring the pairs that cannot be used; if both subjects are censored at the same time, or if one subject has died and the follow-up time of the other is less than the failure time of the first. The *C*-index can measure predictive performance of a model derived from a set of covariates in a model. It estimates the probability of concordance between the predicted and observed survival, with a value that ranges from 0.5 (no predictive discrimination) to 1.0 (perfectly discriminating model) (Harrell *et al*., 1996).

The *C*-index in this chapter was calculated for all models, based on the **external estimation**, **risk score** and **variable selection** approaches. The calculation of *C*-index for external validation was performed using *rcorr.cens* function in the *Hmisc* package by Harrell (2001) implemented in R software. A linear predictor, lp (lp=coefficient*score) was obtained for each score (clinical score, gene expression score, and SNP score). Then, lp was used in the model to calculate the *C*-index for each score individually and jointly. The calculation of the *C*-index when re-estimating the coefficients in new data was performed using the *validate.cph* function in the *rms* package by Harrell (2001), implemented in R software for resampling validation of the Cox model accuracy indexes. Bootstrapping was used to correct for possible over-fitting (Harrell, 2001); this can be performed in the *validate.cph* function and the number of 500 resamples was chosen to obtained a more stable estimate of the *C*-index. The *C*-index was estimated by averaging the *C*-index calculated from the 500 bootstrap samples.

Below is shown an example of a *C*-index calculations using *rcorr.cens* function, for the clinical score, gene expression score, SNP score, and the combined score in the external estimation approach for Model 1. The number of all usable pairs that contribute to the *C*-index estimates were similar in all models.

$$C - \text{index for clinical score } = \frac{Number\ of\ concordant\ pairs}{Number\ of\ all\ usable\ pairs} = \frac{9946}{12946} = 0.7685$$

$$C - \text{index for gene expression score } = \frac{8258}{12946} = 0.6381$$

$$C - \text{index for SNP score } = \frac{5570}{12946} = 0.4304$$

$$C - \text{index for combined score } = \frac{8590}{12946} = 0.664$$

6.2.4.6 Model predictive performance using AUC

The AUC measures predictive performance for models with a binary outcome; however, this was estimated in this analysis for comparison with the literature, as previous studies (Rendleman *et al*., 2013; Nsengimana *et al*., 2015) used AUC to assess their models' performance. An AUC value closer to 1 indicates a model with good predictive ability, while AUC value of 0.05 indicates no predictive ability (Harrell, 2001).

The AUC was estimated following a logistic regression to predict death from melanoma (coded as 0 for No and 1 for Yes) for all models based on the external estimation, risk score, and variable selection approaches. The AUC estimated in this analysis does not take into account the variable length of follow-up in time-to-event data as it uses a binary outcome variable. Therefore, the estimates are not accurate to assess model performance for survival models.

For comparisons with the *C*-index, the AUC was calculated for each score individually and for the combined score, for both external estimation and the risk score approach. For the variable selection approach, the AUC was calculated for the full model, the model with the selected predictors only, the model with Breslow thickness only, and the model with the selected -omic predictors only.

6.3 Results

The sample characteristics in the test set are shown in Table 6.1. The test set samples initially comprised 216 patients with both gene expression and genotype data available, but reduced to 190 after excluding those with survival analysis exclusion criteria.

There were 6 and 11 patients with missing values for Breslow thickness and presence of ulceration variables, respectively. To avoid further reduction in the test set samples, missing values were replaced using a simple imputation method as the percentage of missing values was very small in both variables. For Breslow thickness, simple imputation was conducted by replacing the missing value with the mean value. For ulceration, it was assumed that ulceration was absent for those without pathology information for ulceration.

The median age for patients in the test set is 59 years. The majority of patients were female (53.2%), the most common tumour type was superficial spreading melanoma (49.5%), most had no ulceration (65.4%), the most common site of tumour was on the limbs (42.6%), and the most common disease stage was AJCC stage II (48.4%). The number of deaths in the test set is 45 (23.7%) with median survival time of 3.1 years among those who died. Median follow-up for survivors is 7.2 years.

Table 6.1 Sample characteristics in the test set (n=190)

Variables	n missing	n (%)
Age (years)	0	58.7 (21.1 – 78.2)*
Sex	0	
Female		101 (53.2)
Male		89 (46.8)
Tumour type	0	
Superficial spreading		94 (49.5)
Nodular		56 (29.5)
Lentigo maligna melanoma		2 (1.0)
Acral lentiginous		6 (3.2)
Unclassified		19 (10.0)
Other		13 (6.8)
Breslow thickness (mm)	6	2.5 (0.7 – 15)*
Presence of ulceration	11	
No		117 (65.4)
Yes		62 (34.6)
Tumour site	0	
Limbs		81 (42.6)
Head		25 (13.2)
Trunk		56 (29.5)
Other		28 (14.7)
AJCC stage	4	
Stage I		58 (31.2)
Stage II		90 (48.4)
Stage III		38 (20.4)
Follow-up time for patients who are still alive[†]	0	7.2 (1.6 – 13.6)*
Survival time for patients who died[†]	0	3.1 (0.9 – 9.2)*
Survival status	0	
Censored		145 (76.3)
Died of melanoma		45 (23.7)

* Median (Range) [†]MSS calculated in years

6.3.1 Model 1 Combined data survival models using variables selected from the training set in Chapter 3

Table 6.2 and Table 6.3 show the association of risk scores with MSS in the training set and test set, respectively. In the training set, all scores were significantly associated with MSS in both univariable and multivariable analysis (Table 6.2). However, these results are over-fitted as similar data were used to obtained the estimates used in the calculation of risk scores.

In the test set (Table 6.3), both clinical score and gene expression score were strongly associated with MSS in the univariable analysis, but only clinical score remained associated in the multivariable analysis. The clinical score and gene expression score show significant moderate correlation with each other (Table 6.4), thus could have affected the significance of the gene expression score in the multivariable analysis. When the correlation between the clinical predictors and the gene expression were explored further (Table 6.5), three clinical predictors (age, Breslow thickness, and presence of ulceration) were significantly correlated with the gene expression score. Age and Breslow thickness were are also highly correlated with each other.

Borderline significance was observed for the SNP score in both the univariable and multivariable analyses in the test set (Table 6.3). However, the direction of effect of the SNP score in the test set is opposite that observed in the training set. Therefore, the SNP score may not have a strong effect on MSS.

Table 6.2 Cox model combining risk scores* in the training set (n=365)**

Predictors	Univariable Cox model				Multivariable Cox model			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.64	1.90	0.08	4.6×10^{-14}	0.30	1.34	0.09	6.2×10^{-4}
Gene expression score[b]	0.86	2.35	0.08	2.0×10^{-16}	0.70	2.01	0.09	1.5×10^{-13}
SNP score[c]	0.69	2.00	0.09	8.9×10^{-16}	0.55	1.74	0.09	4.6×10^{-10}

Table 6.3 Cox model combining risk scores* in the test set (n=190**)

Predictors	Univariable Cox model				Multivariable Cox model[†]			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.86	2.36	0.14	4.8×10^{-10}	0.80	2.23	0.15	1.2×10^{-7}
Gene expression score[b]	0.46	1.58	0.15	1.8×10^{-3}	0.17	1.19	0.18	0.34
SNP score[c]	-0.33	0.72	0.17	0.05	-0.33	0.72	0.17	0.05

[*] calculated using estimates from multivariable Cox model for clinical predictors and from penalized Cox model for gene expression and SNPs data

**Subset of patients with both gene expression and genotype data

[a] created from 5 clinical predictors

[b] created from 16 gene expression levels selected by penalized Cox model

[c] created from 13 SNPs selected by penalized Cox model

[d] HR per 1 standard deviation; All scores were standardized prior to univariable and multivariable Cox regression as each score was on a different scale as shown in Figure 6.5

[†] Proportional hazards assumption was checked using Schoenfeld residuals test and not violated (results shown in Table 2 in Appendix II)

Table 6.4 Pairwise correlation between the scores in the test set (n=190)

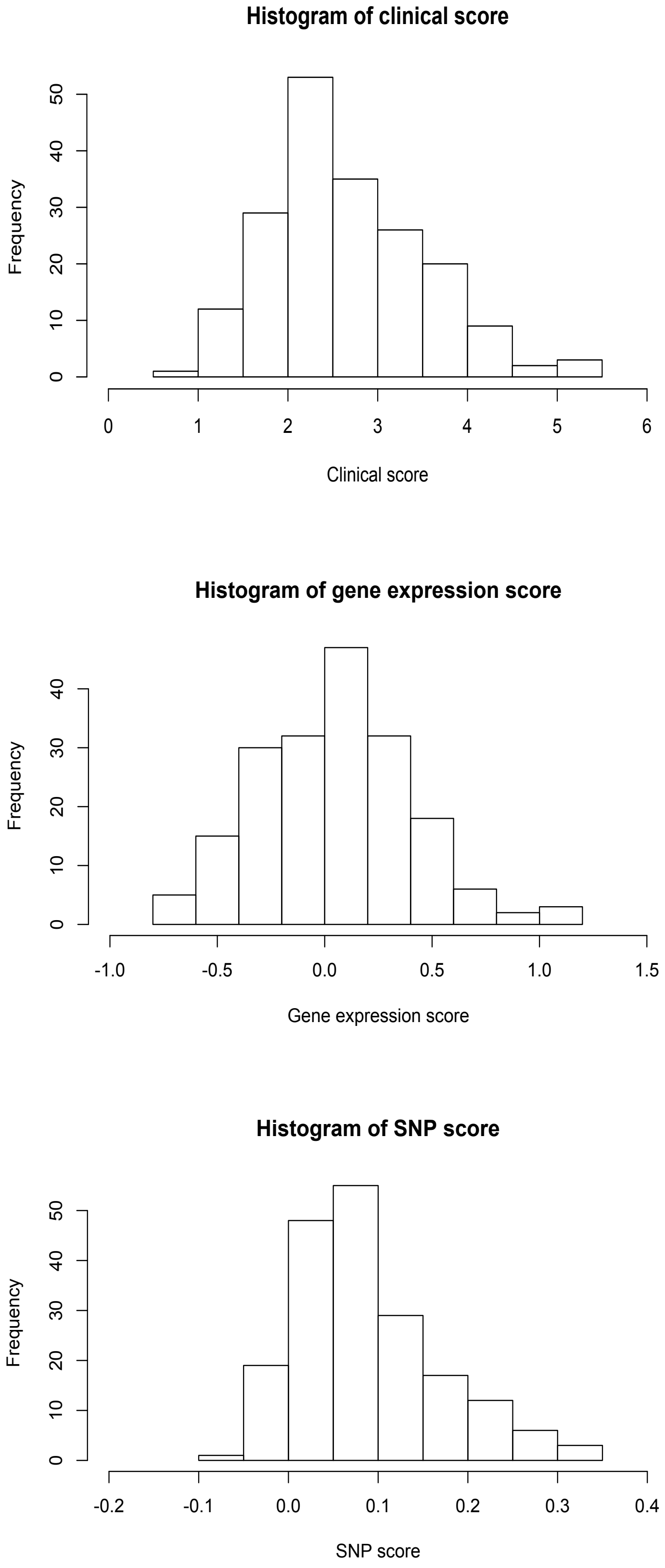
	Clinical score	Gene expression score	SNP score
Clinical score	1	0.40 (1.0×10^{-8})	-0.06 (0.45)
Gene expression score		1	0.01 (0.85)
SNP score			1

P-values are shown in brackets

Table 6.5 Pairwise correlation between clinical predictors and gene expression score in the test set (n=190)

	Age	Sex	Breslow	Ulceration	Site	Gene expression score
Age	1	0.17 (0.02)	0.23 (1.4×10^{-3})	0.13 (0.07)	0.08 (0.25)	0.28 (7.8×10^{-5})
Sex		1	0.04 (0.60)	0.10 (0.21)	0.22 (2.5×10^{-3})	0.11 (0.14)
Breslow thickness			1	0.36 (4.6×10^{-7})	0.22 (2.0×10^{-3})	0.32 (5.9×10^{-6})
Ulceration				1	0.21 (3.3×10^{-3})	0.25 (4.4×10^{-4})
Site					1	0.14 (0.06)
Gene expression score						1

P-values are shown in brackets



Histogram of clinical score

Histogram of gene expression score

Histogram of SNP score

Figure 6.5 Histograms of clinical score, gene expression score and SNP score

Table 6.6 presents the univariable analysis of the selected variables (5 clinical predictors, 16 gene expression levels and 13 SNPs) in the test set. Age at diagnosis, site of tumour (trunk, and other site compared with limbs as baseline), Breslow thickness, presence of ulceration, and expression level of eight genes (*LPAR1, NKD2, C1R, OSTC, HLA-DQB2, HLA-B, NDUFA8,* and *IGSF5*) show significant association with MSS in the univariable analysis. No significant association was observed for any of the selected SNPs.

Table 6.7 shows the variables chosen by penalized Cox model when combining the selected variables in the test set. Eight variables with non-zero coefficients were selected at the chosen penalty; Breslow thickness, presence of ulceration, expression level of *HLA-DQB2*, expression level of *OSTC*, expression level of *C1R*, age at diagnosis, SNP RS9957831, and site.

Table 6.6 Univariable analysis of the selected variables (5 clinical predictors, 16 gene expression levels and 13 SNPs) in Model 1 in the test set (n=190)

Predictors	Univariable Cox model			
	β	HR	SE	P-value
Age at diagnosis	0.04	1.04	0.01	3.1×10^{-3}
Sex (Male)	0.07	1.07	0.30	0.82
Breslow thickness	0.22	1.25	0.04	1.9×10^{-8}
Presence of ulceration	1.33	3.78	0.31	1.3×10^{-5}
Site (Head)	0.16	1.17	0.58	0.78
Site (Trunk)	0.75	2.13	0.39	0.05
Site (Other)	1.58	4.87	0.40	7.6×10^{-5}
ILMN_1701441 (*LPAR1*)	-0.45	0.64	0.16	0.01
ILMN_3249501 (*ZNF697*)	0.09	1.09	0.17	0.59
ILMN_1749829 (*DLG1*)	0.21	1.23	0.16	0.20
ILMN_1731206 (*NKD2*)	-0.33	0.72	0.13	0.01
ILMN_1764109 (*C1R*)	-0.48	0.62	0.14	4.2×10^{-4}
ILMN_2056167 (*OSTC*)	0.56	1.76	0.16	4.4×10^{-4}
ILMN_3238435 (*SNORA12*)	-0.21	0.81	0.16	0.20
ILMN_1695959 (*C21ORF63*)	-0.16	0.85	0.14	0.27
ILMN_1741648 (*HLA-DQB2*)	-0.71	0.49	0.15	4.1×10^{-6}
ILMN_1784238 (*SEC22B*)	0.19	1.21	0.15	0.21
ILMN_1778401 (*HLA-B*)	-0.36	0.70	0.14	0.01
ILMN_1759729 (*NDUFA8*)	0.38	1.46	0.16	0.02
ILMN_2344221 (*IGSF5*)	0.33	1.39	0.12	0.01
ILMN_2095633 (*FGF22*)	-0.22	0.80	0.17	0.19
ILMN_1700547 (*CHST9*)	0.08	1.08	0.15	0.60
ILMN_1735199 (*CIAPIN1*)	-0.04	0.96	0.13	0.77
RS17837209	-1.02	0.36	0.60	0.09
RS9957831	-0.60	0.55	0.36	0.09
RS4768090	-0.04	0.96	0.24	0.88
RS2902554	-0.47	0.63	0.35	0.18
RS5770310	-0.30	0.74	0.22	0.17
RS10233832	0.07	1.07	0.23	0.76
RS17379771	0.17	1.18	0.22	0.45
RS16956192	0.09	1.09	0.30	0.77
RS2392477	0.09	1.10	0.23	0.68
RS6689263	-0.05	0.95	0.31	0.88
RS11639902	-0.24	0.79	0.23	0.29
RS12519276	-0.10	0.90	0.22	0.64
RS10941528	0.25	1.28	0.22	0.27

The highlighted rows indicate the significant predictors at P-value < 0.05

Table 6.7 Variables chosen* by penalized Cox model of MSS when combining the selected variables (5 clinical predictors, 16 gene expression levels and 13 SNPs) in the test set (n=190)

Predictors	Penalized Cox model	Univariable Cox model			
	β	β	HR	SE	P-value
Breslow thickness	0.14	0.22	1.25	0.04	1.9×10^{-8}
Ulceration (absence vs presence)	0.41	1.33	3.78	0.31	1.3×10^{-5}
ILMN_1741648 (*HLA-DQB2*)	-0.05	-0.71	0.49	0.15	4.1×10^{-6}
ILMN _2056167 (*OSTC*)	0.47	0.56	1.76	0.16	4.4×10^{-4}
ILMN _1764109 (*C1R*)	-0.12	-0.48	0.62	0.14	4.2×10^{-4}
Age at diagnosis	0.01	0.04	1.04	0.01	3.0×10^{-3}
RS9957831	-0.17	-0.60	0.55	0.36	0.09
Site (limbs vs the rest of the body)	0.01	0.90	2.45	0.34	0.01

* In order of the variables that entered the model

The *C*-index and AUC estimates of the scores in the training set and test set are shown in Table 6.8. In the training set, the gene expression score (*C*-index=0.78) shows a higher predictive performance than the clinical score and SNP score, and jointly, the scores improved the predictive performance as might be expected (*C*-index=0.84). However, these results are over-fitted as the same data were used to developed and to assess the model. In the test set, the clinical score shows the highest predictive performance in both the external estimation and risk score approach (*C*-index=0.77). When combining the scores, no improvement was seen in the *C*-index in the external estimation (*C*-index=0.66 when using β=1 as weight; *C*-index=0.65 when using β from training set as weight), while similar *C*-index was seen in the risk score approach (*C*-index=0.76). The AUC estimates were comparable to the *C*-indices, where the clinical score has the highest predictive performance, and shows similar performance to the combined score.

The *C*-index and AUC estimates of the models based on the variable selection approach are shown in Table 6.9. The *C*-indices of the model with eight selected predictors, the model with Breslow thickness only, and the model with selected -omic predictors only are 0.77, 0.70, and 0.72 respectively. Whereas, the AUC estimates for these models are 0.80, 0.72, and 0.74 respectively. Both the *C*-index and AUC increased in the model with eight selected predictors compared to the model with Breslow thickness only, but were no better than the clinical score.

Table 6.8 *C*-index and AUC estimates of clinical score, gene expression score and SNP score in the training and test set

C-index	Clinical score	Gene expression score	SNP score	Combined score
Training set (n=365)[a]	0.704	0.784	0.716	0.836
Test set (n=190) (external estimation and using β=1 for all scores as weight)	0.769	0.638	0.43	0.664
Test set (n=190) (external estimation and using β from training set[b] as weight)	0.769	0.638	0.43	0.655
Test set (n=190)[a] (risk score approach)	0.770	0.639	0.566	0.760
AUC in the test set[c]	0.764	0.625	0.584	0.771

[a] *C*-index was estimated using bootstrapping method

[b] From Table 6.2

[c] AUC following logistic regression predicting of death from melanoma

Table 6.9 *C*-index and AUC estimates of the selected model in the test set (n=190)

	Full model (5 clinical predictors, 16 gene expression level, and 13 SNPs)	Model with 8 selected predictors (Breslow thickness, ulceration, *HLA-DQB2* expression level, *OSTC* expression level, *C1R* expression level, age, RS9957831, and site)	Model with Breslow thickness only	Model with 4 selected -omic predictors only
C-index [a]	0.716	0.765	0.70	0.720
AUC	0.885	0.799	0.716	0.737

[a] *C*-index was estimated using bootstrapping method

6.3.2 Model 2 Combined data survival models with prior cluster analysis

The gene expression levels selected using cluster analysis and penalized regression are presented in Table 6.10 in order of the probes that entered the model. Based on the cluster dendrogram from the cluster analysis (Figure 6.6), four clusters were chosen to group the probes. The top 10 most significant probes within each cluster were selected for penalized Cox regression, and the penalized Cox model selected 22 probes at the cross-validated penalty in the training set. There were seven probes (*C1R, C21orf63, SEC22B, HLA-DQB2, TMEM64, WDR3,* and *ANPEP*) selected from the first cluster, seven probes (*ZNF6970, DLGAP5, NKD2, OSTC, C1orf163, TBC1D7,* and *DGKA*) from the second cluster, five probes (*CENPM, SEMA4A, JMY, ALDH2,* and *BCL11B*) from the third cluster, and three probes (*NDUFA8, ORC6L,* and *TIPIN*) from the fourth cluster. All 22 probes were significantly associated with MSS in univariable analysis.

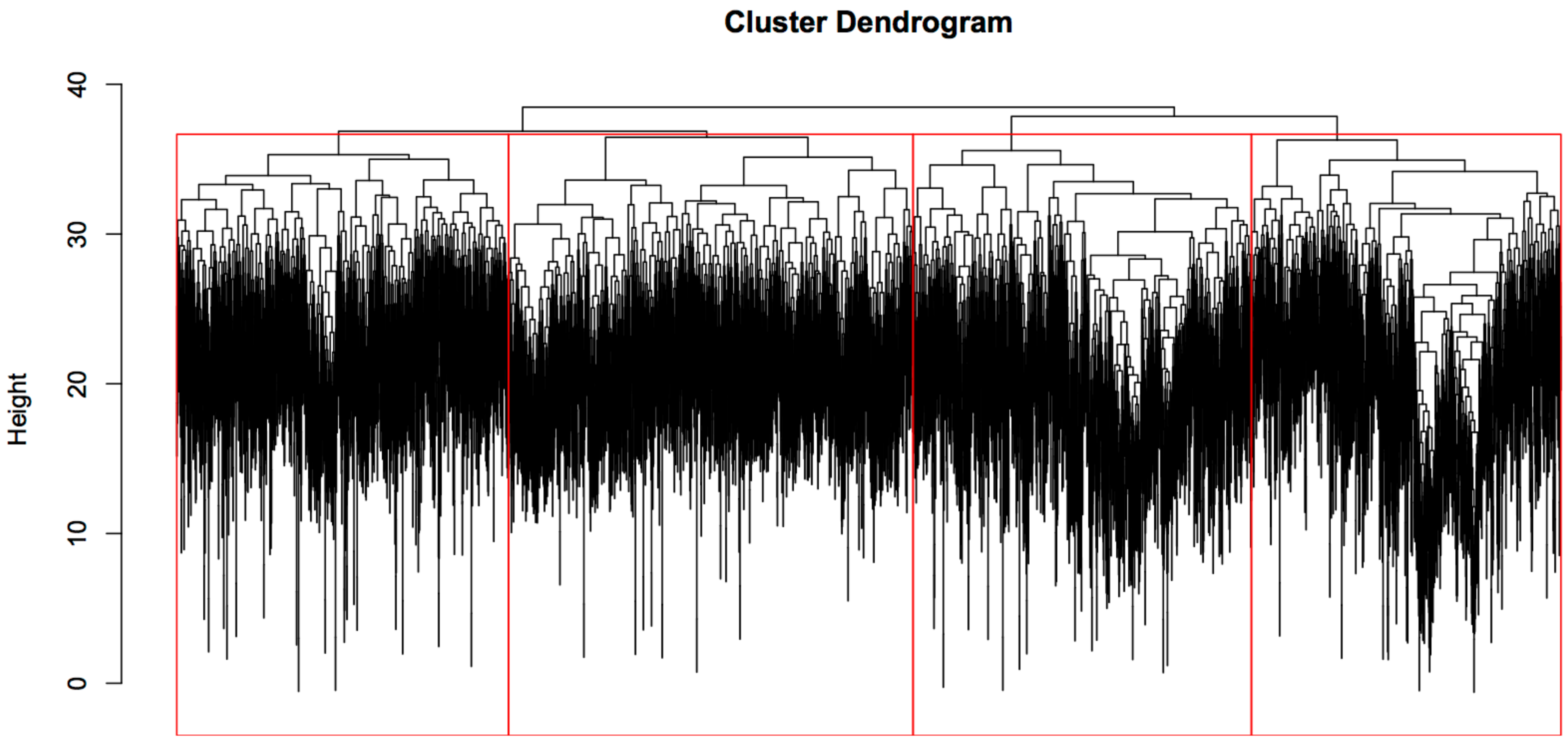


Figure 6.6 Cluster dengrogram for the gene expression data

Table 6.10 22 gene expression levels selected by penalized Cox model from 4 gene clusters in the training set (n=424)

Probe	Chr	Cluster	Penalized Cox model	Univariable Cox model			
			β	β	HR	SE	P-value
ILMN_3249501 (*ZNF6970*)	1	2	0.38	0.62	1.85	0.09	2.8 x 10^{-12}
ILMN_1749829 (*DLGAP5*)	14	2	0.05	0.52	1.68	0.10	1.5 x 10^{-7}
ILMN_1731206 (*NKD2*)	5	2	0.09	0.66	1.93	0.13	1.6 x 10^{-7}
ILMN_1764109 (*C1R*)	12	1	-0.05	-0.33	0.72	0.06	6.9 x 10^{-8}
ILMN_2056167 (*OSTC*)	4	2	0.02	0.50	1.64	0.09	6.4 x 10^{-8}
ILMN_1695959 (*C21orf63*)	21	1	0.18	-0.45	0.64	0.08	3.2 x 10^{-8}
ILMN_1784238 (*SEC22B*)	1	1	0.08	-0.43	0.65	0.08	1.1 x 10^{-7}
ILMN_1741648 (*HLA-DQB2*)	6	1	0.07	-0.41	0.66	0.08	1.4 x 10^{-7}
ILMN_1759729 (*NDUFA8*)	9	4	0.01	-0.47	0.62	0.08	1.6 x 10^{-9}
ILMN_1757415 (*C1orf163*)	1	2	0.13	0.53	1.70	0.09	1.3 x 10^{-9}
ILMN_2150402 (*TMEM64*)	8	1	-0.11	-0.44	0.65	0.07	2.3 x 10^{-10}
ILMN_1731070 (*ORC6L*)	16	4	-0.13	-0.57	0.56	0.08	1.1 x 10^{-11}
ILMN_2368721 (*CENPM*)	22	3	0.10	0.46	1.58	0.08	5.2 x 10^{-8}
ILMN_1711254 (*WDR3*)	1	1	-0.22	-0.44	0.65	0.07	1.7 x 10^{-9}
ILMN_1702787 (*SEMA4A*)	1	3	0.06	0.43	1.54	0.07	1.8 x 10^{-10}
ILMN_1761939 (*TIPIN*)	15	4	-0.14	-0.45	0.64	0.08	3.5 x 10^{-9}
ILMN_1762080 (*JMY*)	5	3	0.24	0.64	1.89	0.11	4.5 x 10^{-9}
ILMN_1661622 (*TBC1D7*)	6	2	0.08	0.71	2.04	0.12	3.2 x 10^{-9}
ILMN_1793859 (*ALDH2*)	12	3	0.03	0.51	1.67	0.09	5.2 x 10^{-8}
ILMN_2319910 (*DGKA*)	12	2	0.04	0.57	1.77	0.11	1.2 x 10^{-7}
ILMN_1667885 (*BCL11B*)	14	3	0.08	0.57	1.78	0.10	2.9 x 10^{-8}
ILMN_1763837 (*ANPEP*)	15	1	-0.10	-0.45	0.64	0.08	7.3 x 10^{-9}

Chr: Chromosome

In Table 6.11, all scores were significantly associated with MSS in both univariable and multivariable analysis, as expected, in the training set. In Table 6.12, similar results were observed as in the Model 1 (Table 6.3), where clinical score and gene expression score show significant association with MSS in the univariable analysis, but only clinical score remained significant in the multivariable analysis in the test set.

Table 6.11 Cox model combining risk scores* in the training set (n=365)**

Predictors	Univariable Cox model				Multivariable Cox model			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.64	1.90	0.08	4.6×10^{-14}	0.27	1.31	0.09	2.1×10^{-3}
Gene expression score[b]	1.06	2.90	0.10	2.0×10^{-16}	0.88	2.40	0.12	5.8×10^{-14}
SNP score[c]	0.69	2.00	0.09	8.9×10^{-16}	0.54	1.72	0.09	1.2×10^{-9}

Table 6.12 Cox model combining risk scores* in the test set (n=190)**

Predictors	Univariable Cox model				Multivariable Cox model[†]			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.86	2.36	0.14	4.8×10^{-10}	0.83	2.29	0.15	4.3×10^{-8}
Gene expression score[b]	0.39	1.48	0.14	0.01	0.09	1.10	0.17	0.58
SNP score[c]	-0.33	0.72	0.17	0.05	-0.33	0.72	0.17	0.05

[*] calculated using estimates from multivariable Cox model for clinical predictors and from penalized estimates for gene expression and SNPs data

**Subset of patients with both gene expression and genotype data

[a] created from 5 clinical predictors

[b] created from 22 gene expression levels selected by penalized Cox model

[c] created from 13 SNPs selected by penalized Cox model

[d] HR per 1 standard deviation; All scores were standardized prior to univariable and multivariable Cox regression

[†] Proportional hazards assumption was checked using Schoenfeld residuals test and not violated (results shown in Table 3 in Appendix II)

Table 6.13 shows the univariable analysis for the 22 gene expression levels in the test set. This table only shows univariable analysis for the gene expression levels as results for the five clinical predictors and 13 SNPs are similar to Model 1 in Table 6.6. In the test set, expression level of 12 genes (*TMEM64, ORC6L, CENPM, TIPIN, JMY, DLGAP5, ANPEP, OSTC, C21orf63, SEC22B, HLA-DQB2,* and *NDUFA8*) shows significant association with MSS in univariable analysis.

Table 6.14 shows the predictors selected by penalized Cox regression when combining five clinical predictors, 22 gene expression levels, and 13 SNPs in the test set. Nine predictors with non-zero coefficients were selected at the cross-validated penalty; Breslow thickness, presence of ulceration, expression level of *HLA-DQB2*, expression level of *OSTC*, expression level of *C1R*, age at diagnosis, SNP RS9957831, expression level of *ANPEP*, and site. Eight of the selected predictors were similar to the predictors selected in Model 1 except the expression level of *ANPEP*.

Table 6.13 Univariable analysis of the selected 22 gene expression levels in the test set (n=190)

Predictors	Univariable Cox model			
	β	HR	SE	P-value
ILMN_3249501 (*ZNF697*)	0.09	1.09	0.17	0.59
ILMN_1749829 (*DLGAP5*)	0.21	1.23	0.16	0.20
ILMN_1731206 (*NKD2*)	-0.33	0.72	0.13	0.01
ILMN_1764109 (*C1R*)	-0.48	0.62	0.14	4.2×10^{-4}
ILMN_2056167 (*OSTC*)	0.56	1.76	0.16	4.4×10^{-4}
ILMN_1695959 (*C21orf63*)	-0.16	0.85	0.14	0.27
ILMN_1784238 (*SEC22B*)	0.19	1.21	0.15	0.21
ILMN_1741648 (*HLA-DQB2*)	-0.71	0.49	0.15	4.1×10^{-6}
ILMN_1759729 (*NDUFA8*)	0.38	1.46	0.16	0.02
ILMN_1757415 (*C1orf163*)	0.14	1.15	0.18	0.45
ILMN_2150402 (*TMEM64*)	0.21	1.23	0.16	0.19
ILMN_1731070 (*ORC6L*)	0.34	1.4	0.17	0.04
ILMN_2368721 (*CENPM*)	0.18	1.2	0.16	0.25
ILMN_1711254 (*WDR3*)	0.19	1.21	0.15	0.20
ILMN_1702787 (*SEMA4A*)	-0.38	0.68	0.12	1.7×10^{-3}
ILMN_1761939 (*TIPIN*)	0.07	1.07	0.14	0.61
ILMN_1762080 (*JMY*)	-0.04	0.97	0.16	0.82
ILMN_1661622 (*TBC1D7*)	0.38	1.46	0.14	0.01
ILMN_1793859 (*ALDH2*)	-0.53	0.59	0.13	5.6×10^{-3}
ILMN_2319910 (*DGKA*)	-0.45	0.64	0.13	4.4×10^{-5}
ILMN_1667885 (*BCL11B*)	-0.44	0.64	0.13	7.9×10^{-4}
ILMN_1763837 (*ANPEP*)	-0.48	0.62	0.15	9.3×10^{-4}

The highlighted rows indicate the significant predictors at P-value < 0.05

Table 6.14 Variables chosen* by penalized Cox model of MSS when combining the selected variables (5 clinical predictors, 22 gene expression levels and 13 SNPs) in the test set (n=190)

Predictors	Penalized Cox model	Univariable Cox model			
	β	β	HR	SE	P-value
Breslow thickness	0.14	0.22	1.25	0.04	1.9×10^{-8}
Ulceration (absence vs presence)	0.41	1.33	3.78	0.31	1.3×10^{-5}
ILMN_1741648 (*HLA-DQB2*)	-0.04	-0.71	0.49	0.15	4.1×10^{-6}
ILMN_2056167 (*OSTC*)	0.46	0.56	1.76	0.16	4.4×10^{-4}
ILMN_1764109 (*C1R*)	-0.10	-0.48	0.62	0.14	4.2×10^{-4}
Age at diagnosis	0.01	0.04	1.04	0.01	3.1×10^{-3}
RS9957831	-0.18	-0.60	0.55	0.36	0.09
ILMN_1763837 (*ANPEP*)	-0.03	-0.48	0.62	0.15	9.3×10^{-4}
Site (limbs vs the rest of the body)	0.003	0.90	2.45	0.34	0.01

* In order of the variables that entered the model

Table 6.15 presents the *C*-index and AUC estimates for scores in the training set and test set for Model 2. Similar results were seen as in the Model 1 (Table 6.8), where the combined score improved the *C*-index in the training set, but not in the test set. The AUC estimate for the combined score also shows no improvement compared to the individual scores.

Table 6.16 presents the *C*-index and AUC estimates for models from the variable selection approach. The *C*-indices of the model with nine selected predictors, the model with Breslow thickness only, and the model with selected -omic predictors only are 0.76, 0.70, and 0.71 respectively, which are similar to the estimates in Model 1. The AUC for these models were also comparable to the Model 1 estimates.

Table 6.15 *C*-index and AUC estimates of clinical score, gene expression score and SNP score in the training and test set

C-index	Clinical score	Gene expression score	SNP score	Combined score
Training set (n=365)[a]	0.704	0.794	0.716	0.839
Test set (n=190) (external estimation and using β=1 for all scores as weight)	0.769	0.621	0.43	0.66
Test set (n=190) (external estimation and using β from training set[b] as weight)	0.769	0.621	0.43	0.609
Test set (n=190)[a] (risk score approach)	0.769	0.623	0.568	0.763
AUC in the test set[c]	0.764	0.611	0.584	0.772

[a] *C*-index was estimated using bootstrapping method

[b] From Table 6.11

[c] AUC following logistic regression predicting of death from melanoma

Table 6.16 *C*-index and AUC estimates of the selected model in the test set (n=190)

	Full model (5 clinical predictors, 22 expression level, and 13 SNPs)	Model with 9 selected predictors (Breslow thickness, ulceration, *HLA-DQB2* expression level, *OSTC* expression level, *C1R* expression level, age, RS9957831, *ANPEP* expression level, and site)	Model with Breslow thickness only	Model with 5 selected -omic predictors only
C-index [a]	0.725	0.761	0.70	0.714
AUC	0.918	0.800	0.716	0.739

[a] *C*-index were estimated using bootstrapping method

6.3.3 Model 3 Combined data survival models using Lund clusters to select gene expression levels for penalized Cox regression

The gene expression levels selected from four Lund clusters and penalized regression are shown in Table 6.17 in order of the probes that entered the model. After including the top 10 probes from each cluster for penalized Cox regression, the model selected 18 probes at cross-validated penalty in the training set. All 18 probes were significantly associated with MSS in univariable analysis.

Table 6.17 18 gene expression levels selected by penalized Cox model from 4 Lund clusters in the training set (n=424)

Probe	Cluster	Penalized Cox model	Univariable Cox model			
		β	β	HR	SE	P-value
ILMN_1741648 (*HLA-DQB2*)	Normal-like	-0.06	-0.45	0.64	0.08	3.49×10^{-9}
ILMN_1673721 (*EXO1*)	Proliferative	0.15	0.61	1.83	0.11	1.07×10^{-8}
ILMN_1664516 (*CENPF*)	Proliferative	0.03	0.51	1.66	0.10	1.70×10^{-7}
ILMN_1661622 (*TBC1D7*)	Pigmentation	0.10	0.50	1.64	0.09	6.40×10^{-8}
ILMN_1736096 (*DLL3*)	Pigmentation	0.08	0.51	1.66	0.10	1.05×10^{-6}
ILMN_1720373 (*SLC7A5*)	Pigmentation	0.14	0.44	1.55	0.09	1.90×10^{-6}
ILMN_1738832 (*SACS*)	Proliferative	0.11	0.46	1.58	0.10	1.38×10^{-6}
ILMN_1658143 (*RFC3*)	Proliferative	0.06	0.56	1.75	0.11	1.13×10^{-7}
ILMN_1751161 (*COL7A10*)	Normal-like	-0.10	-0.38	0.68	0.08	1.25×10^{-6}
ILMN_1795930 (*PTGER4*)	High-immune	-0.12	-0.43	0.65	0.08	3.2×10^{-7}
ILMN_1676191 (*DARS2*)	Pigmentation	0.06	0.58	1.79	0.12	7.43×10^{-7}
ILMN_1727087 (*GJA1*)	Normal-like	-0.03	-0.44	0.65	0.09	9.59×10^{-7}
ILMN_2043918 (*DLEU1*)	Pigmentation	0.02	0.47	1.61	0.11	1.42×10^{-5}
ILMN_1713088 (*MSI2*)	Pigmentation	0.02	0.57	1.77	0.12	2.53×10^{-6}
ILMN_1705477 (*CAMK1D*)	Normal-like	-0.01	-0.39	0.68	0.08	2.11×10^{-6}
ILMN_1770692 (*WDR12*)	Pigmentation	0.02	0.56	1.75	0.12	1.93×10^{-6}
ILMN_2051373 (*NEK2*)	Proliferative	0.03	0.42	1.52	0.09	8.29×10^{-6}
ILMN_1808071 (*KIF14*)	Proliferative	0.002	0.43	1.54	0.09	3.23×10^{-6}

Selected probes: 1 from high-immune cluster (*PTGER4*); 4 from normal-like cluster (*COL7A10, CAMK1D, GJA1, HLA-DQB2*); 7 from pigmentation cluster (*DARS2, WDR12, DLEU1, MSI2, TBC1D7, DLL3, SLC7A5*); 6 from proliferative cluster (*NEK2, EXO1, SACS, RFC3, CENPF, KIF14*)

Table 6.18 shows analysis in the training set to estimate the effect of risk scores on MSS. As expected, results in Table 6.18 also show that all scores were significantly associated with MSS in both univariable and multivariable analysis.

In Table 6.19, similar results were seen as in Model 1 and Model 2 (Table 6.3 and Table 6.12), where the clinical score and gene expression score show significant association with MSS in the univariable analysis, but only clinical score remained significant in the multivariable analysis in the test set.

Table 6.18 Cox model combining risk scores* in the training set (n=365)**

Predictors	Univariable Cox model				Multivariable Cox model			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.64	1.90	0.08	4.6×10^{-14}	0.33	1.40	0.09	2.8×10^{-4}
Gene expression score[b]	0.81	2.24	0.10	8.9×10^{-16}	0.54	1.72	0.12	2.7×10^{-6}
SNP score[c]	0.69	2.00	0.09	8.9×10^{-16}	0.55	1.74	0.09	1.8×10^{-10}

Table 6.19 Cox model combining risk scores* in the test set (n=190)**

Predictors	Univariable Cox model				Multivariable Cox model[†]			
	β	HR[d]	SE	P-value	β	HR[d]	SE	P-value
Clinical score[a]	0.86	2.36	0.14	4.8×10^{-10}	0.78	2.17	0.16	1.9×10^{-6}
Gene expression score[b]	0.55	1.74	0.16	3.8×10^{-4}	0.19	1.20	0.19	0.32
SNP score[c]	-0.33	0.72	0.17	0.05	-0.35	0.71	0.17	0.05

[*] calculated using estimates from multivariable Cox model for clinical predictors and from penalized estimates for gene expression and SNPs data

**Subset of patients with both gene expression and genotype data

[a] created from 5 clinical predictors

[b] created from 18 gene expression levels selected by penalized Cox model

[c] created from 13 SNPs selected by penalized Cox model

[d] HR per 1 standard deviation; All scores were standardized prior to univariable and multivariable Cox regression

[†] Proportional hazards assumption was checked using Schoenfeld residuals test and not violated (results shown in Table 4 in Appendix II)

Table 6.20 shows the univariable analysis of the 18 gene expression levels in the test set. Seven of the 18 gene expression levels (*HLA-DQB2, CENPF, TBC1D7, DLL3, SACS, GJA*, and *CAMK1D)* were significantly associated with MSS in univariable analysis. Univariable analysis for the five clinical predictors and 13 SNPs were not shown in this table as results were similar to Model 1 in Table 6.6.

When combining the five clinical predictors, 18 gene expression levels and 13 SNPs using penalized Cox regression analysis, the model selected nine predictors with non-zero coefficients at cross-validated penalty (Table 6.21). Six of the selected predictors (Breslow thickness, presence of ulceration, expression level of *HLA-DQB2*, age at diagnosis, SNP RS 9957831, and site) were similar to the predictors selected in Model 1 (Table 6.7). The other three selected predictors were expression level of *DLL3*, expression level of *CENPF* and SNP RS1737977.

Table 6.20 Univariable analysis of the selected 18 gene expression levels in the test set (n=190)

Predictors	Univariable Cox model			
	β	HR	SE	P-value
ILMN_1741648 (*HLA-DQB2*)	-0.71	0.49	0.15	4.1×10^{-6}
ILMN_1673721 (*EXO1*)	0.24	1.27	0.15	0.12
ILMN_1664516 (*CENPF*)	0.45	1.57	0.17	0.01
ILMN_1661622 (*TBC1D7*)	0.38	1.46	0.14	0.01
ILMN_1736096 (*DLL3*)	0.40	1.49	0.16	0.01
ILMN_1720373 (*SLC7A5*)	0.26	1.29	0.17	0.13
ILMN_1738832 (*SACS*)	0.32	1.38	0.13	0.01
ILMN_1658143 (*RFC3*)	0.27	1.31	0.15	0.07
ILMN_1751161 (*COL7A10*)	-0.16	0.85	0.13	0.20
ILMN_1795930 (*PTGER4*)	-0.24	0.79	0.13	0.06
ILMN_1676191 (*DARS2*)	0.36	1.43	0.2	0.07
ILMN_1727087 (*GJA1*)	-0.46	0.63	0.15	2.1×10^{-3}
ILMN_2043918 (*DLEU1*)	0.30	1.36	0.16	0.06
ILMN_1713088 (*MSI2*)	0.31	1.37	0.20	0.12
ILMN_1705477 (*CAMK1D*)	-0.36	0.70	0.12	3.4×10^{-3}
ILMN_1770692 (*WDR12*)	0.26	1.29	0.18	0.14
ILMN_2051373 (*NEK2*)	0.16	1.18	0.16	0.29
ILMN_1808071 (*KIF14*)	0.21	1.23	0.17	0.22

The highlighted rows indicate the significant predictors at P-value < 0.05

Table 6.21 Variables chosen by penalized Cox model of MSS when combining the selected variables (5 clinical predictors, 18 gene expression levels and 13 SNPs) in the test set (n=190)

Predictors	Penalized Cox model	Univariable Cox model			
	β	β	HR	SE	P-value
Breslow thickness	0.12	0.22	1.25	0.04	1.9×10^{-8}
Ulceration (absence vs presence)	0.46	1.33	3.78	0.31	1.3×10^{-5}
ILMN_1741648 (*HLA-DQB2*)	-0.10	-0.71	0.49	0.15	4.1×10^{-6}
Age at diagnosis	0.01	0.04	1.04	0.01	3.1×10^{-3}
RS9957831	-0.15	-0.6	0.55	0.36	0.09
ILMN_1736096 (*DLL3*)	0.003	0.40	1.49	0.16	0.01
Site (limbs vs the rest of the body)	0.0001	0.90	2.45	0.34	0.01
ILMN_1664516 (*CENPF*)	0.0002	0.45	1.57	0.17	0.01
RS17379771	0.0004	0.17	1.18	0.22	0.45

Table 6.22 shows the *C*-index and AUC estimates for scores in the training set and test set for Model 3. These results are similar to Model 1 (Table 6.8) and Model 2 (Table 6.15), where no improvements were seen in the combined score *C*-indices and AUC in the test set compared to the individual scores.

The *C*-index and AUC estimates for models based variable selection approach are shown in Table 6.23. Similar to results in Model 1 (Table 6.9) and Model 2 (Table 6.16), the model with the selected predictors by the penalized Cox analysis shows higher *C*-index and AUC estimates (*C*-index = 0.75 and AUC=0.81) than the model with Breslow thickness alone (*C*-index = 0.70 and AUC=0.72).

Table 6.22 *C*-index estimates of clinical score, gene expression score and SNP score in the training and test set

C-index	Clinical score	Gene expression score	SNP score	Combined score
Training set (n=365)[a]	0.704	0.732	0.716	0.819
Test set (n=190) (external estimation and using β=1 for all scores as weight)	0.769	0.657	0.43	0.680
Test set (n=190) (external estimation and using β from training set[b] as weight)	0.769	0.657	0.43	0.644
Test set (n=190)[a] (risk score approach)	0.771	0.658	0.570	0.762
AUC in the test set[c]	0.764	0.649	0.584	0.771

[a] *C*-index were estimated using bootstrapping method
[b] From Table 6.18
[c] AUC following logistic regression predicting of death from melanoma

Table 6.23 *C*-index and AUC estimates of the selected model in the test set (n=190)

	Full model	Model with 9 selected predictors (Breslow thickness, ulceration, *HLA-DQB2* expression level, age, RS9957831, *DLL3* expression level, site, *CENPF* expression level, and RS17379711)	Model with Breslow thickness only	Model with 5 selected -omic predictors only
C-index [a]	0.676	0.746	0.70	0.704
AUC	0.856	0.810	0.716	0.745

[a] *C*-index was estimated using bootstrapping method

6.3.4 Model 4 Combined data survival models using clinical predictors and Lund classification only

Of the 699 patients with gene expression data in the Leeds cohort, only 677 could be classified using the Lund 4-class molecular classification; 174 classified as high immune, 197 classified as normal-like, 222 classified as pigmentation, 84 classified as proliferative, and 22 unclassified. When using the 2-class molecular grading, 371 and 306 patients were identified as low-grade and high-grade, respectively. In the test set, 90 patients were identified as low-grade and 96 patients as high-grade. Four patients in the test set were unclassified.

The association of the 2-molecular grade and the selected clinical predictors with MSS were explored in the training set using 450 patients that can be classified with the Lund classification. In the univariable analysis, the 2-molecular grade shows significant association with MSS along with the five clinical predictors (Table 6.24). In the multivariable analysis, only age at diagnosis, Breslow thickness and the 2-molecular grade remained as the significant predictors for MSS.

In the test set, the 2-molecular grade, age at diagnosis, site, Breslow thickness, and presence of ulceration shows significant association with MSS in univariable analysis, but only Breslow thickness remained significant in multivariable analysis (Table 6.25).

Table 6.24 Cox model combining the selected clinical predictors and Lund 2-molecular grade in the training set* (n=450)

Predictors	Univariable Cox model				Multivariable Cox model			
	β	HR	SE	P-value	β	HR	SE	P-value
Age at diagnosis	0.04	1.04	0.01	3.3×10^{-6}	0.03	1.03	0.01	4.9×10^{-4}
Sex (female vs male)	0.55	1.73	0.18	2.0×10^{-3}	0.30	1.35	0.19	0.11
Site (limbs vs the rest of the body)	0.61	1.84	0.19	1.6×10^{-3}	0.26	1.30	0.21	0.21
Breslow thickness	0.14	1.15	0.02	5.0×10^{-10}	0.10	1.11	0.03	6.3×10^{-4}
Ulceration (absence vs presence)	0.72	2.05	0.18	4.9×10^{-5}	0.36	1.44	0.20	0.07
Lund 2-molecular grade (low-grade vs high-grade)	0.93	2.55	0.18	3.2×10^{-7}	0.58	1.79	0.19	2.7×10^{-3}

* Excluding the test set samples and those with survival analysis exclusion criteria

Table 6.25 Cox model combining the selected clinical predictors and Lund 2-molecular grade in the test set (n=190)

Predictors	Univariable Cox model				Multivariable Cox model			
	β	HR	SE	P-value	β	HR	SE	P-value
Age at diagnosis	0.04	1.04	0.01	3.1×10^{-3}	0.02	1.02	0.01	0.12
Sex (female vs male)	0.07	1.07	0.30	0.82	0.35	1.35	0.33	0.87
Site (limbs vs the rest of the body)	0.90	2.45	0.34	0.01	0.37	1.44	0.39	0.35
Breslow thickness	0.22	1.25	0.04	2.0×10^{-8}	0.13	1.14	0.05	0.01
Ulceration (absence vs presence)	1.33	3.78	0.31	1.6×10^{-5}	0.73	2.07	0.37	0.05
Lund 2-molecular grade (low-grade vs high-grade)	1.22	3.38	0.34	3.5×10^{-4}	0.49	1.63	0.39	0.21

Table 6.26 shows the *C*-index and AUC estimates for the clinical score and the 2-molecular grade in the test set. The *C*-index for the clinical score in both the external estimation and risk score approach was higher than the 2-molecular grade only. When combining the clinical score with the 2-molecular grade, no improvement was seen in the *C*-index for both approaches. For the AUC estimates, results were similar to the *C*-index estimates for both the individual predictor and the combined predictors.

Table 6.27 shows the *C*-index and AUC estimates for all six variables in the test set as only Breslow thickness shows significant association with MSS in multivariable analysis in Table 6.25. Both estimates were higher for the full model compared to Breslow thickness alone, but were similar to the clinical score.

Table 6.26 *C*-index and AUC estimates for clinical score* and Lund classification in the test set (n=190)

	Clinical score only	Lund 2-molecular grade only	Clinical score + Lund 2-molecular grade
C-index (external estimation approach)[a]	0.769	0.648	0.756
C-index (risk score approach)[b]	0.767	0.650	0.766
AUC[c]	0.764	0.638	0.765

* calculated using estimates from multivariable Cox regression of 5 clinical predictors in the training set (in Chapter 3)

[a] Estimates from the Lund study (Harbst *et al.*, 2012) for the 2-molecular grade were for overall survival and relapse-free survival only. Therefore, estimates from the training set (Table 6.24) were used in this analysis

[b] *C*-index was estimated using bootstrapping method (B=500)

[c] AUC following logistic regression predicting of death from melanoma

Table 6.27 *C*-index and AUC estimates for clinical predictors and Lund classification in the test set (n=190)

	Full model (Age, sex, site, Breslow thickness, presence of ulceration and Lund 2-molecular grade	Model with Breslow thickness only
C-index [a]	0.738	0.70
AUC	0.784	0.716

[a] *C*-index was estimated using bootstrapping method

6.3.5 Correlation between the gene expression scores from different approaches

Table 6.28 shows the pairwise correlation between the gene expression scores from Model 1 to Model 3. Gene scores from all models were strongly correlated with each other, with the strongest correlation from gene expression score in Model 1 and Model 2. Figure 6.7 shows a scatterplot of clinical score against gene expression score from Model 1 to identify the distribution of the scores for patients who were still alive and those who have died. For patients who are still alive, there is a moderate correlation between their clinical score and gene expression score. In Table 6.4, moderate correlation (r=0.4) was found between the clinical score and gene expression score.

Table 6.28 Pairwise correlation between the gene expression scores from different approaches in the test set (n=190)

	Gene expression score from Model 1	Gene expression score from Model 2	Gene expression score from Model 3
Gene expression score from Model 1	1	0.93	0.83
Gene expression score from Model 2		1	0.82
Gene expression score from Model 3			1

P-value for all pairwise correlations is $<2.2\text{x}10^{-16}$

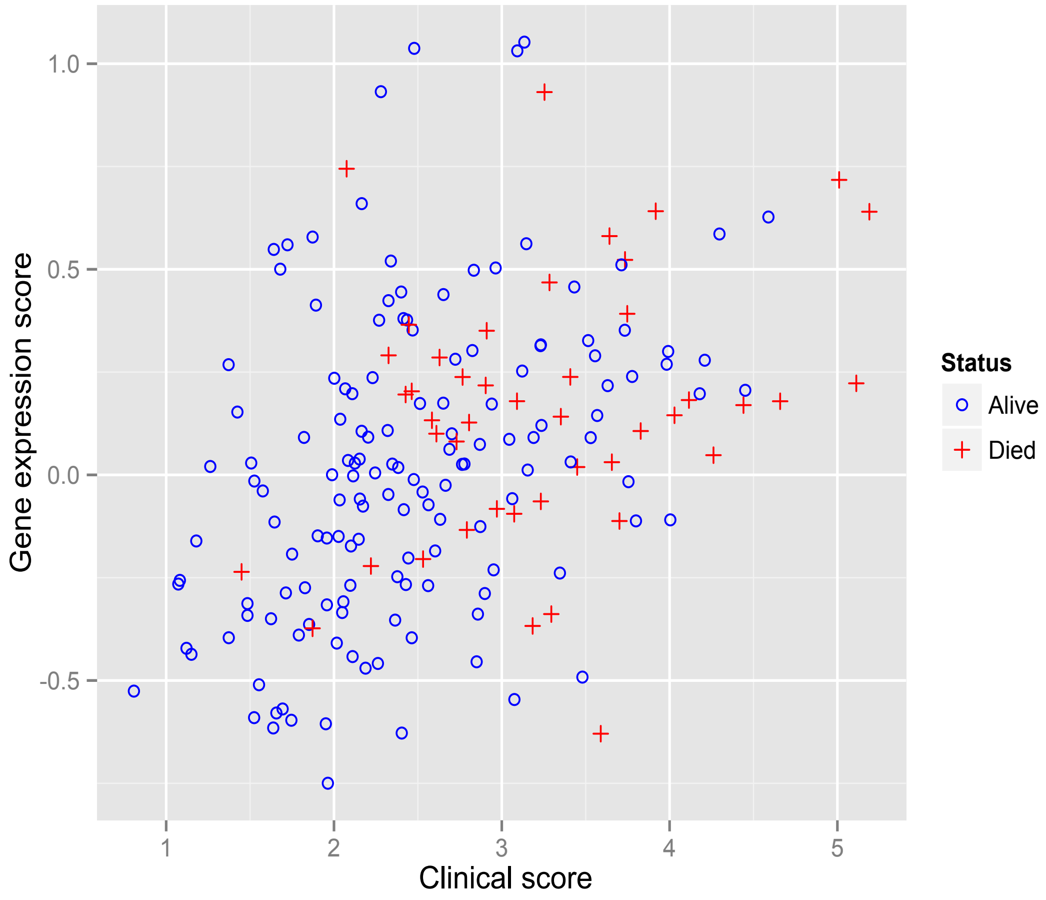


Figure 6.7 Scatterplot of clinical score versus gene expression score from Model 1 approach

6.4 Discussion

6.4.1 Predictive performance of each model

This chapter provides estimates of predictive ability for models developed from clinical predictors, gene expression levels, and genotype data. The predictive ability of these models was assessed in a test set of 190 samples with gene expression data and genotyped data available. As tumour sampling for gene expression profiling in the cohort was not random (samples were extracted from patients with thicker tumour for adequate extraction of RNA), this may explain the high proportion of deaths from melanoma among the 699 patients with gene expression data (23.7% in the test set and 17% in the training set). The split of the data into a training set and a test set among those with gene expression data however, was performed randomly.

Based on the risk score approach, the clinical score consistently shows the highest *C*-index compared to gene expression score and SNP score in all models in the test set (Model 1 *C*-indices: clinical score=0.77, gene expression score=0.64, and SNP score=0.57; Model 2 *C*-indices: clinical score=0.77, gene expression score=0.62, SNP score=0.58; Model 3 *C*-indices: clinical score=0.77, gene expression score=0.66, and SNP score=0.57). Similar results were observed in the three models when the external estimation approach was applied to compute the *C*-indices. Among the scores, the predictive ability of the SNP score was the lowest. This was not unexpected as the selected 13 SNPs from the training set did not meet a genome-wide significance level and none of the selected SNPs showed significant association with MSS in the test set in Table 6.6. The association of the SNP score with MSS in the test set was also not significant in any model (Table 6.3, Table 6.12, and Table 6.19), which may suggest that SNPs do not have a strong effect on MSS compared to clinical predictors and gene expression levels.

When the clinical score, gene expression score and SNP score were combined, no improvement in the *C*-indices was seen in all models in both the external estimation and risk score approach. The *C*-index (based on risk score approach) and AUC estimate for the combined score in Models 1 to 3 was 0.76, which was similar to the estimates for clinical score only. As the estimates of *C*-index in this analysis are quite variable, our results cannot conclude that combining the scores does not improve the predictive ability of the model. Results in this analysis are consistent with the literature (Lindholm *et al*., 2004; Buettner *et al*., 2005; Balch *et al*., 2009) which shows the established clinical predictors as strong predictors for melanoma survival. In the test set, Breslow thickness is the most significant predictor for melanoma survival as shown in the univariable analysis in Table 6.6.

In a recent melanoma prognostic survival model developed using clinical information only from Queensland Cancer Registry, Baade *et al*. (2015) reported that Breslow thickness explained most of the variation in their final prognostic model (including age, Breslow thickness, tumour site, presence of ulceration, presence of positive lymph nodes, and presence of metastasis). Their reported *C*-index for the melanoma severity index (calculated using variables in the final model) was 0.88, which indicates a good predictive ability. The high *C*-index in their study may be due to the extra variables used to calculate their index, and the study was based on a very large sample size (n=28,654 with 17,000 melanoma deaths) in comparison to the small test set used in this analysis.

Based on the variable selection approach, all models with the selected predictors only (*C*-index of Model 1 with 8 predictors=0.77; *C*-index of Model 2 with 9 predictors=0.76; *C*-index of Model 3 with 9 predictors=0.75) show higher predictive ability than the model with Breslow thickness only (*C*-index=0.70). The *C*-index for models with the selected -omic predictors only (Model 1=0.72; Model 2=0.71; Model 3=0.70) is comparable to the *C*-index of Breslow thickness alone in all models, which shows that the selected -omic predictors have similar predictive ability to Breslow thickness alone. Four clinical predictors (Breslow thickness, presence of ulceration, age at diagnosis and site) were selected in all models. However, since the *C*-indices for all models with the selected predictors were higher than the *C*-index for Breslow thickness alone, this shows the potential for improving the predictive ability when combining different types of predictors in the model.

When the clinical score was combined with the Lund 2-molecular grade, the *C*-index remained 0.77. The Lund 2-molecular grade had been previously assessed for its prognostic value in 300 patients in the LMC by Nsengimana *et al.* (2015), who reported an increase in the AUC for predicting death from melanoma from 0.68 when using AJCC stage alone to 0.72 when combining the AJCC stage and molecular grade. Similar AUC was found in Nsengimana *et al.* (2015) and in this analysis for the 2-molecular grade (0.65 *vs* 0.64). Some improvement was observed for the combined AJCC and molecular grade in Nsengimana *et al.* (2015) probably due to AJCC stage alone being less predictive compared with all the clinical predictors (age, sex, Breslow thickness, presence of ulceration and tumour site) used in this analysis.

In summary, results in this chapter indicate that clinical score is the best predictor of melanoma survival. Several gene expressions were also predictive of survival in the variable selection approach. However, as gene expressions seems to act through the clinical variables, it does not help in the final model.

6.4.2 Predictive performance between models

In addition to estimates of predictive ability from different types of data, these analyses also provide estimates of predictive ability for a gene expression score created using different approaches, and using different sets of probes. The *C*-indices for the gene expression score from Model 1 (16 probes selected using penalized Cox regression only), Model 2 (22 probes selected using clustering analysis and penalized Cox regression), and Model 3 (18 probes selected using Lund clustering and penalized Cox regression), were 0.64, 0.62, and 0.66, respectively. The *C*-indices for the three models were fairly similar. However, using only point estimates of the *C*-indices does not allow us to test whether the *C*-indices truly differ between the models. This could be shown using confidence intervals of the *C*-indices, but the current R package used to calculate the *C*-indices does not provide confidence interval values. Bootstrapping could be used in future analysis to calculate the confidence interval for *C*-index estimates.

The scores were highly correlated as shown in Table 6.27. As shown in Figure 6.8, there were 9 similar probes selected by Model 1 and Model 2 (r=0.93), while only 1 similar probes were chosen by Model 1 and Model 3 (r=0.83), and Model 2 and Model 3 (r=0.82), respectively. Between the three models, there was only one overlapped probe (*HLA-DQB2*).

In summary, although different set of probes were selected by different models, the gene expression scores were highly correlated with each other, and has only small variations in their predictive ability which was not adequate to improve the overall combined score predictive ability.

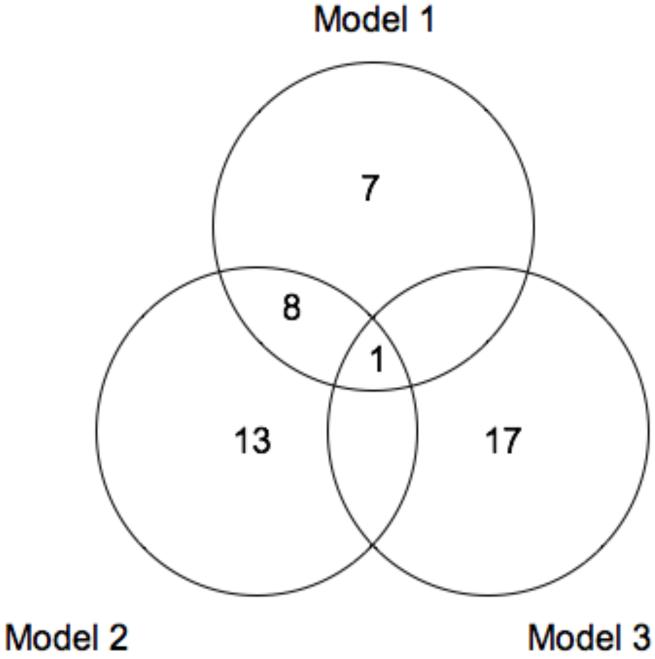


Figure 6.8 Number of overlapped probes between the three models

Chapter 7 Discussion and conclusions

7.1 Summary of main findings

This study uses clinical predictors, gene expression levels and genetic variants to build prognostic models for MSS, and explore the inter-relationships between these factors. In the survival analyses of single types of variable in Chapter 3, all the five established clinical predictors (age, sex, tumour site, Breslow thickness, and presence of ulceration) were significantly associated with MSS, consistent with previous studies. The penalized Cox model selected 16 gene expression levels (*LPAR1*, *ZNF697*, *DLG1*, *NKD2*, *C1R*, *OSTC*, *SNORA12*, *C21ORF63*, *HLA-DQB2*, *SEC22B*, *HLA-B*, *NDUFA8*, *IGSF5*, *FGF22*, *CHST,* and *CIAPIN1)*, and 13 SNPs (RS17837209, RS9957831, RS4768090, RS2902554, RS5770310, RS10233832, RS17379771, RS16956192, RS2392477, RS6689263, RS11639902, RS12519276, and RS10941528) that were significantly associated with MSS in the training set.

When the selected variables were combined in the test set in Chapter 6, using estimates from the training set, the results showed that combining the clinical score, gene expression score, and SNP score did not improve the predictive ability of the model compared to using the clinical score alone. The clinical score provides the highest predictive ability, and the SNP score the lowest. The clinical score and gene expression score was moderately correlated. Based on the variable selection approach, where the training set is only used to select variables, the predictive ability of the selected predictors using penalized Cox models were higher than the predictive ability of Breslow thickness alone, which may indicate the potential for improving the predictive ability when combining different types of predictors in the model.

When the associations of the selected 16 gene expression levels and 13 SNPs with the five established clinical predictors were explored in Chapter 5, most of the selected gene expression levels showed significant association with clinical predictors (10 significant associations with age at diagnosis, 3 with sex, 7 with tumour site, 15 with log-transformed Breslow thickness, and 11 with presence of ulceration), while only three selected SNPs showed significant association with clinical predictors, with only marginally significant associations.

Interestingly, the expression levels of just one of the selected genes (*CHST9*) showed no association with any of the clinical predictors.

GWAS of the 16 expression levels show that SNPs are associated with the expression level of two genes (*HLA-DQB2* and *NDUFA8*) at a genome-wide significance level, while others showed suggestive associations. Investigation of the melanoma susceptibility SNPs shows that one susceptibility SNP (RS1858550 in *PARP1*) was significantly associated with MSS. Of the 20 top susceptibility SNPs, two SNPs (RS498136 in *CCND1* and RS75570604 in *MC1R*) were significantly associated with age at diagnosis, three (RS1858550 in *PARP1,* RS2995264 in *OBFC1,* and RS73008229 in *ATM*) were significantly associated with sex, three (RS6750047 in *RMDN2*, RS250417 in *SLC45A2*, and RS10739221 in *TMEM38B*) were significantly associated with log-transformed Breslow thickness, and one (RS1393350 in *TYR*) was associated with presence of ulceration, mostly with only marginally significant associations. This suggests that genetic risk variants may not have strong effects on the important clinical predictors for melanoma survival. Evidence of eQTL associations (with the expression level of nearby genes) at the 5% significance level were seen in five of the susceptibility SNPs (RS12410869 with *ARNT* expression, RS6750047 with *RMDN2* expression, RS6914589 with *CDKAL1* expression, RS6088372 with *ASIP* expression, and RS408825 with *MX2* expression). When including other SNPs in the susceptibility loci and accounting for multiple testing, evidence of eQTLs were found in two regions only (*ARNT* and *MX2*).

When the heritability of survival in melanoma and Breslow thickness were estimated using GCTA in Chapter 4, no evidence of heritability was found for the 5-year and 10-year survival from melanoma. However, there was some evidence of heritability of Breslow thickness (h^2=0.21, P-value=0.01 when adjusting for centre only and h^2=0.18, P-value=0.03 when adjusting for centre, age and sex).

7.2 Final discussion

7.2.1 Heritability of survival from melanoma and Breslow thickness

The heritability analysis in this study did not find evidence of heritability of survival from melanoma, possibly due to limited sample size. However, a large Swedish study by Brandt *et al*. (2011) using the Swedish Family Cancer database reported a familial risk of dying from melanoma, with increased death from melanoma in offspring (standardized mortality rate of 2.13) and siblings (standardized mortality rate of 3.11), suggesting that there may be a genetic component in survival from melanoma. Therefore, future collaborative work between different research groups is needed to increase the sample size in order to identify genetic determinants for melanoma progression and survival.

Unlike cancer progression and survival, susceptibility to cancer is known to be heritable. Using genome-wide SNP data on a large sample and the GCTA tool, Zaitlen *et al*. (2013) reported significant heritability estimates for breast cancer and prostate cancer of 12% and 20%, respectively. For melanoma, all 20 susceptibility loci identified so far explain 19% of the familial relative risk (Law *et al*., 2015a). Evidence of heritability of susceptibility to melanoma is important as this provide insight into the biology of the disease. However, to date, no study has reported evidence of heritability of survival from any cancer using genome-wide SNP data.

This study found some evidence of heritability of Breslow thickness, suggesting the role of genetic factors, perhaps in the speed with which melanoma tumours grow. As Breslow thickness is the most important predictor for melanoma survival, future studies to identify genetic determinants of Breslow thickness are important to increase the understanding of how melanoma progresses.

7.2.2 Prognostic models for melanoma-specific survival

In agreement with other studies (Balch *et al.*, 2009; Baade *et al.,* 2015), our results found Breslow thickness to be the strongest prognostic factor for MSS in primary melanoma. The best predictor of melanoma survival in this study is the clinical score, consistent with other findings that found clinical variables to have higher predictive ability than individual -omic (Yuan *et al.*, 2014).

Yuan *et al.* (2014) used data from The Cancer Genome Atlas project to evaluate whether integrating clinical features with different types of -omic data (somatic copy-number alteration, DNA methylation, mRNA expression, microRNA expression, and protein expression) could improve prognostic models for four types of cancers (kidney, glioblastoma multiforme, ovarian and lung). They reported that model predictive power for three cancers (kidney, ovarian, and lung) improved only when integrating the clinical features with molecular subtypes derived from expression data, but not when using the gene-level features. However, the magnitude of gains in predictive power assessed by the *C*-index were small. Also, they point out that molecular subtypes are higher-level assemblies of individual gene features, thus may act as a more robust predictor than individual genes or a smaller set of genes.

In Harbst *et al*. (2012), their molecular classifiers for melanoma (4-molecular classification and 2-molecular grade) derived from gene expression data were associated with survival, but the importance of the classifier was assessed in terms of hazard ratios and P-values only, and not the added predictive ability. When assessed using AUC to predict death from melanoma in Nsengimana *et al*. (2015), who used a similar dataset to the one in this study, combining the Lund 2-molecular grade with AJCC stage improved the AUC by 4%. The discrepancy between results in this study and that of Nsengimana *et al.* (2015) may be due to AJCC stage being less predictive than the combined clinical predictors (age at diagnosis, sex, tumour site, Breslow thickness, and presence of ulceration) used in this study. Therefore, more studies in different populations should be performed to validate the predictive ability of the Lund classifiers.

In a recent study by Vazquez *et al*. (2016), it was found that adding whole-genome gene expression or methylation profiles to clinical covariates showed improvement in the AUC by up to 7% compared to clinical covariates alone in prediction of survival of breast cancer patients. An important difference between the study by Vazquez *et al*. (2016) and this study is their modelling approach, which used Bayesian generalized additive models, which are able to integrate data from multiple -omics, cope with high-dimensional inputs, accommodate interactions between different high-dimensional inputs, and assign different regularization parameters for different sets of inputs, allowing the model to weight different information from clinical covariates and from different -omics. A modelling approach that integrates -omic predictors with clinical covariates at the initial stage may perform better than an approach that analyzes the -omic data alone initially, as performed in this study, where the selection of gene expressions and SNPs was performed using penalized Cox regression separately in the training set. Although the association of the selected gene expression levels with MSS were highly significant in the test set as well as the training set, they seem to act through the clinical variables and do not improve the predictive ability of the final model. Therefore, future analysis should include clinical variables when performing the penalized Cox regression to find gene expression levels that predict MSS over and above the effect of the clinical predictors.

When using different methods to select the gene expression levels predictive of MSS, the expression level of *HLA-DQB2* was consistently selected by the three models in Chapter 6, suggesting that immune-related genes could be important for MSS. There is increasing evidence suggesting the predictive potential of immune-related genes in melanoma survival. In a review by Schramm *et al.* (2012), they found immune-related genes featured heavily in gene signatures identified from different studies. In Harbst *et al*. (2012), patients identified as having low-grade tumours using their molecular classifier have higher expression of immune-related genes. Further studies that looked into the association of immune genes with survival outcomes found significant association between the identified gene signatures and survival outcomes (Sivendran *et al*., 2014; Gerami *et al*., 2015). However, these studies rely only on hazard ratios and P-value to assess the effect of the gene signature on the survival outcomes, rather than assessing the predictive ability of the gene signatures. Future studies to identify prognostic biomarkers should also consider assessing the markers' predictive ability in addition to the clinical predictors important in melanoma survival.

7.3 Strengths and limitations

The strength of this study is the availability of detailed clinical, survival, and genomic data in a large cohort of incident melanoma cases, allowing data integration analysis to explore the combined effect of different types of data on melanoma survival. The LMC also has the largest number of melanoma patients with whole-genome gene expression data compared to other studies to date. Besides that, this study was able to focus on melanoma-specific survival, rather than overall survival, which is more relevant for studying disease-specific risk factors.

Several precautions were taken in this study to not introduce bias in survival analysis. Reliable sources of information (ONS and death certificate) were used for ascertainment of cases who died from melanoma. Patients who were recruited into the study more than two years after diagnosis were excluded, as recruiting cases into the study a long time after diagnosis can introduce bias. Methods that deal with left-censoring could be explored to allow inclusion of prevalent cases in future survival analysis (Azzato *et al*, 2009). The major advantage of using prevalent cases is gain in power, but this method will not make much different in the LMC as most of the cases are incident (only 1.5% (n=32) of cases in the cohort were recruited into the study more than two years after diagnosis). Potential issues of using prevalent cases in survival analysis are the possibility of missing individuals who died within a short period of time of their diagnosis and the effect of violation of the proportional hazards assumption, as shown in Azzato *et al.* (2009).

The main limitation of this study is the sample size. In Chapter 4, the heritability analysis does not have adequate power to detect heritability, hence the results do not provide clear evidence of heritability in survival from melanoma. This study was also underpowered to detect genome-wide significance for the survival GWAS in Chapter 3. Although the LMC has the largest number of melanoma patients with whole-genome gene expression data, splitting the patients into a training set (to develop the model) and a test set (to test the model) in Chapter 3 limit the number of patients in each dataset. Results in the test set were inconclusive due to small sample size (n=190).

In Chapter 5, the eQTL analysis to identify whether the 20 melanoma susceptibility SNPs have a possible role in regulation of gene expression levels limits the analysis to genes located near the SNP. This limits the multiple testing but may miss other genes associated with the SNPs, as regulatory variants could be distant from the gene in the genome (trans-eQTLs).

In addition, the prognostic models in this study focus very much on clinical predictors, gene expression and SNP data. Therefore, the prognostic models may miss other important predictors for melanoma such as tumour mutation status, TILs, and serum vitamin D level. Inclusion of these factors in the prediction model may improve the model's predictive ability. For example, inclusion of more predictors in Baade *et al.* (2015) contributes to a high *C*-index for the prognostic model. Epigenetics data could be included as potential predictors too as recent studies have shown association of DNA methylation (Roh *et al*., 2016) and microRNAs (Jayawardana *et al*., 2015; Saldanha *et al*., 2016) with melanoma outcomes.

Also, patients were not randomly selected for the microarray analyses, and hence may not be representative of patients with primary melanoma. There is a selection bias, as only samples from patients with thicker tumours (>0.75 mm) were selected for the microarray analysis, which was done to allow for adequate RNA extraction. Furthermore, thinner tumours might not be big enough to sample for research while leaving enough for clinical usefulness. The LMC is also enriched for cases with thicker tumours as patients with Breslow thickness < 0.75 mm were not recruited into the study after 2005, in order to maximise the value of the sample as a cohort looking at prognostic outcomes (Conway *et al*., 2009). As the gene expression levels from patients with thicker tumours could be different from patients with thinner tumour, results in this study may not be entirely generalizable.

7.4 Conclusions

In summary, the results in this study show that clinical predictors are the best predictors of survival, with Breslow thickness the strongest predictor, confirming current knowledge regarding the influence of clinical predictors on melanoma survival. The 16 gene expression levels associated with MSS in this study were also strongly predictive, but were highly associated with clinical predictors, especially Breslow thickness, suggesting gene expression influences MSS through clinical predictors. This study also shows there is a potential of combining different types of factors to improve the prognostic model for MSS based on the variable selection approach. In addition, results of the heritability analysis provide evidence that germline SNPs influence Breslow thickness.

7.5 Future recommendations

There are several recommendations for future work following the analyses in this study:

- Firstly, this study gives evidence that Breslow thickness is heritable. Future analysis should focus on increasing the sample size for this analysis. We are currently collaborating with a group in Australia to improve the heritability estimate for Breslow thickness using genome-wide SNP data.
- Secondly, GWAS of the 16 gene expression levels indicates that gene expression levels in the tumour may be partially predicted by SNPs. The use of genome-wide SNP data could be explored in the future to predict gene expression levels in patients who have not had gene expression levels measured but with available SNP data.
- Lastly, most of the gene expression probes selected in this analysis are highly correlated with the clinical predictors, especially Breslow thickness, presence of ulceration, and age at diagnosis. Therefore, for future analysis, penalized Cox regression of the whole-genome gene expression data should be performed including the clinical predictors in the model, in order to identify gene expression levels that predict MSS over and above the effect of the clinical predictors.

Appendix I

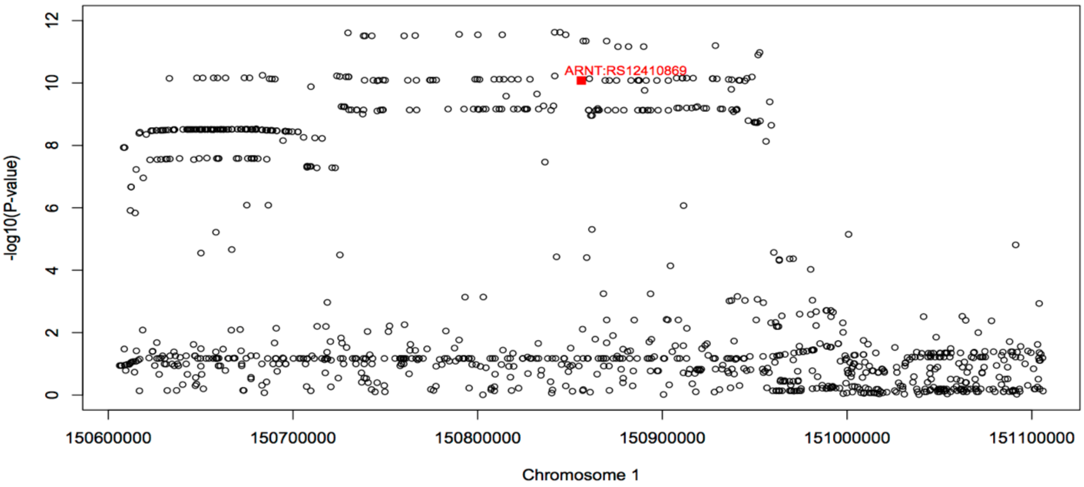


Figure 5.17 Manhattan plot for eQTL analysis for SNPs in ARNT region and ILMN 1762582 expression level

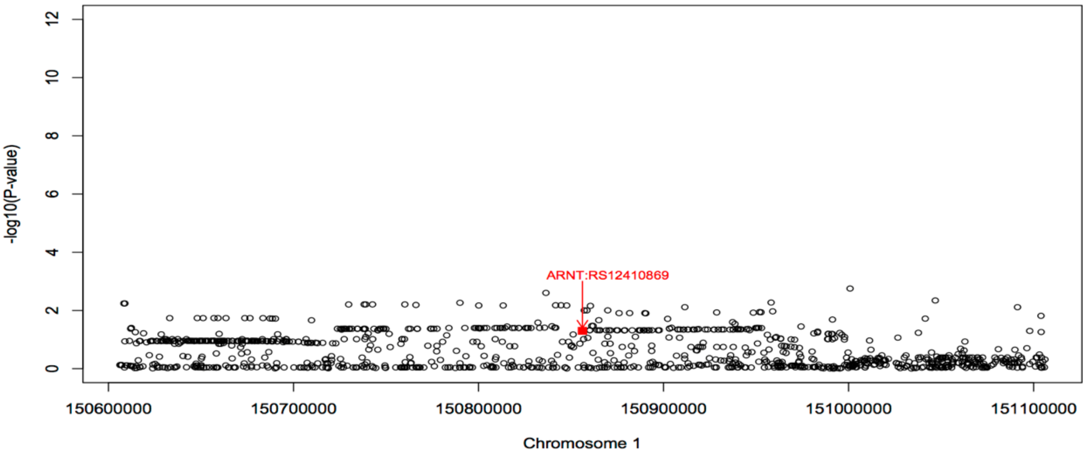


Figure 5.18 Manhattan plot for eQTL analysis for SNPs in ARNT region and ILMN 2347314 expression level

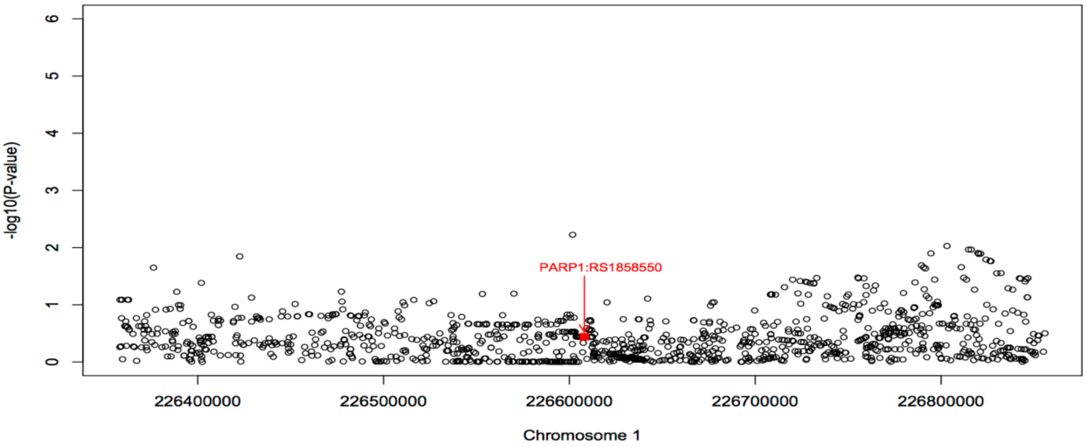


Figure 5.19 Manhattan plot for eQTL analysis for SNPs in PARP1 region and ILMN 1686871 expression level

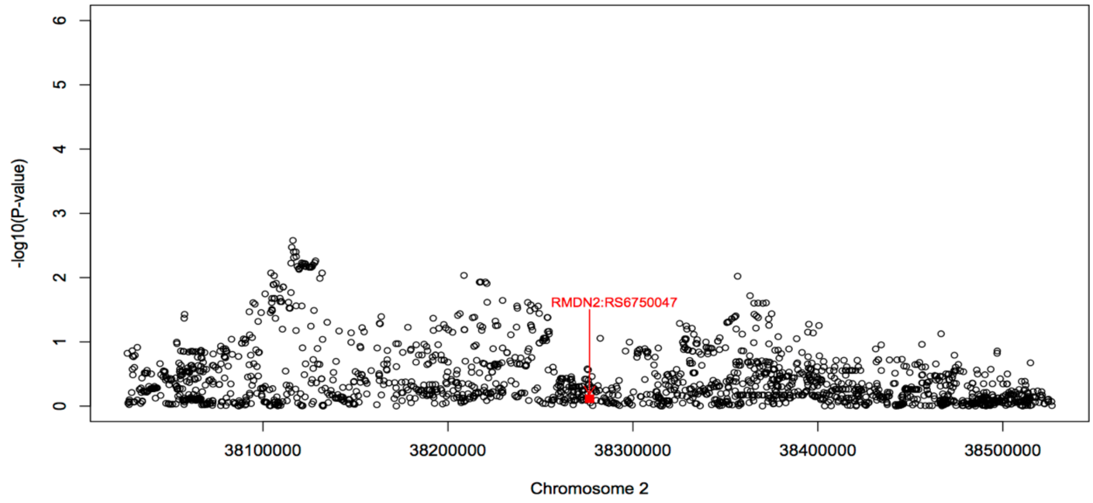


Figure 5.20 Manhattan plot for eQTL analysis for SNPs in RMDN2 region and ILMN 1812302 expression level

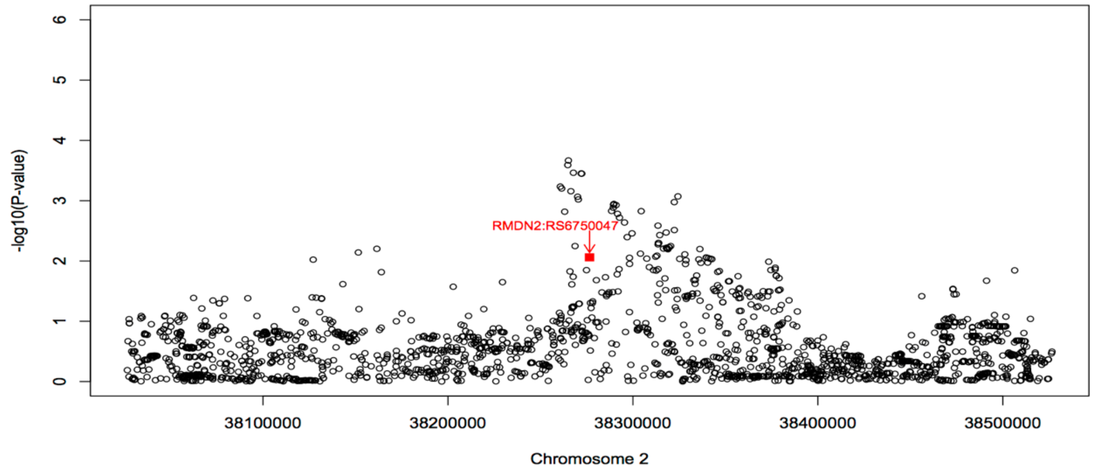


Figure 5.21 Manhattan plot for eQTL analysis for SNPs in RMDN2 region and ILMN 1693338 expression level

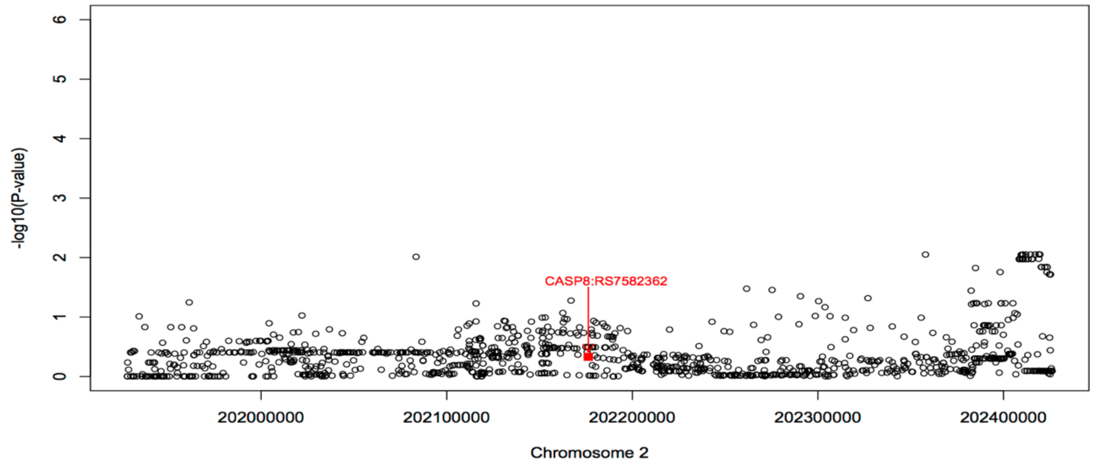


Figure 5.22 Manhattan plot for eQTL analysis for SNPs in CASP8 region and ILMN 1673757 expression level

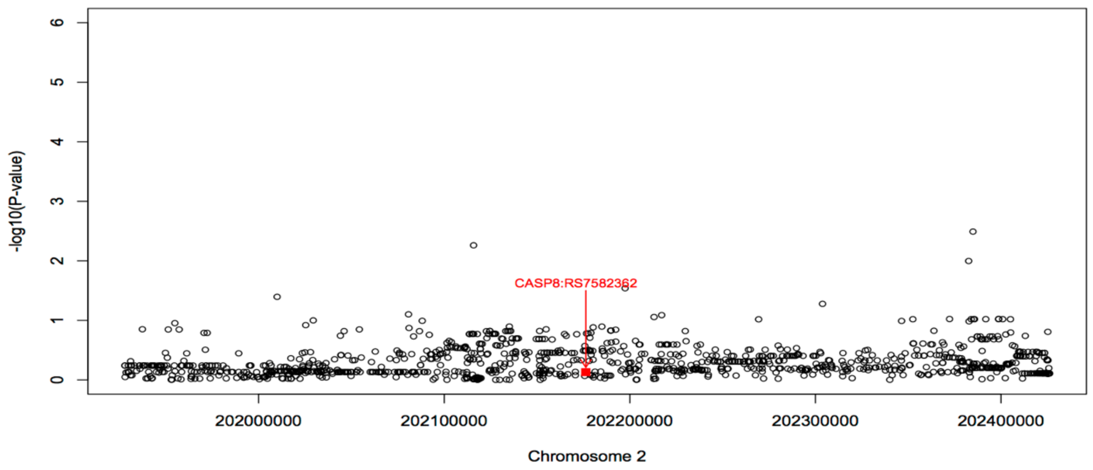


Figure 5.23 Manhattan plot for eQTL analysis for SNPs in CASP8 region and ILMN 1787749 expression level

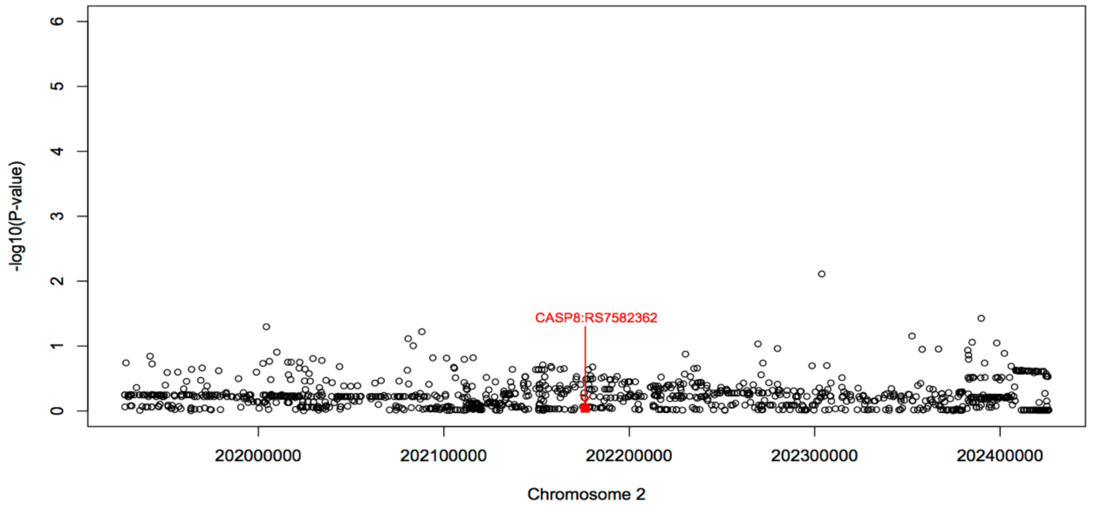


Figure 5.24 Manhattan plot for eQTL analysis for SNPs in CASP8 region and ILMN 1809313 expression level

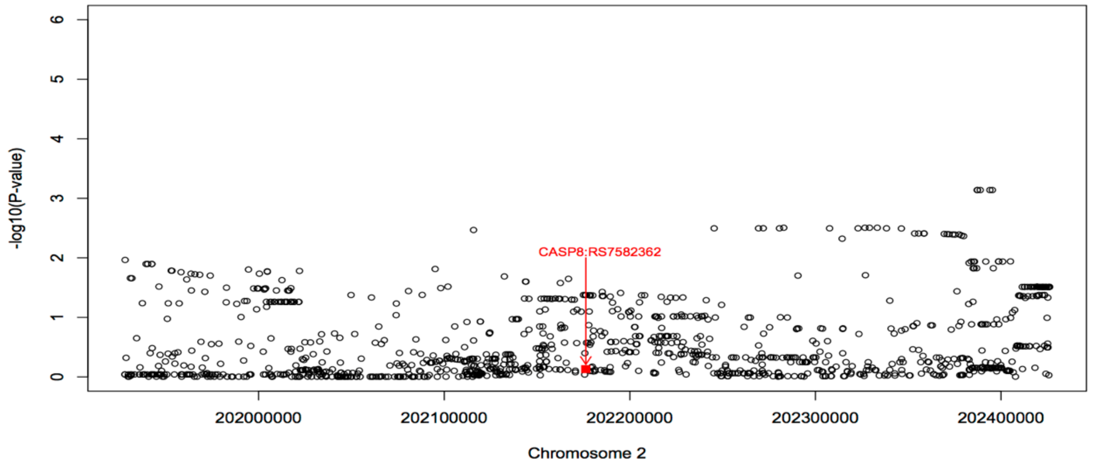


Figure 5.25 Manhattan plot for eQTL analysis for SNPs in CASP8 region and ILMN 2377733 expression level

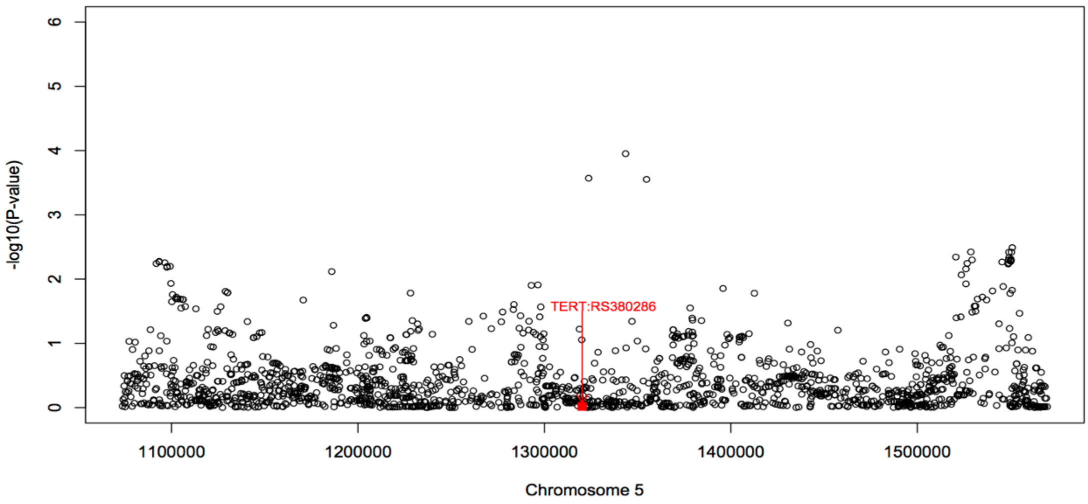


Figure 5.26 Manhattan plot for eQTL analysis for SNPs in TERT region and ILMN 1796005 expression level

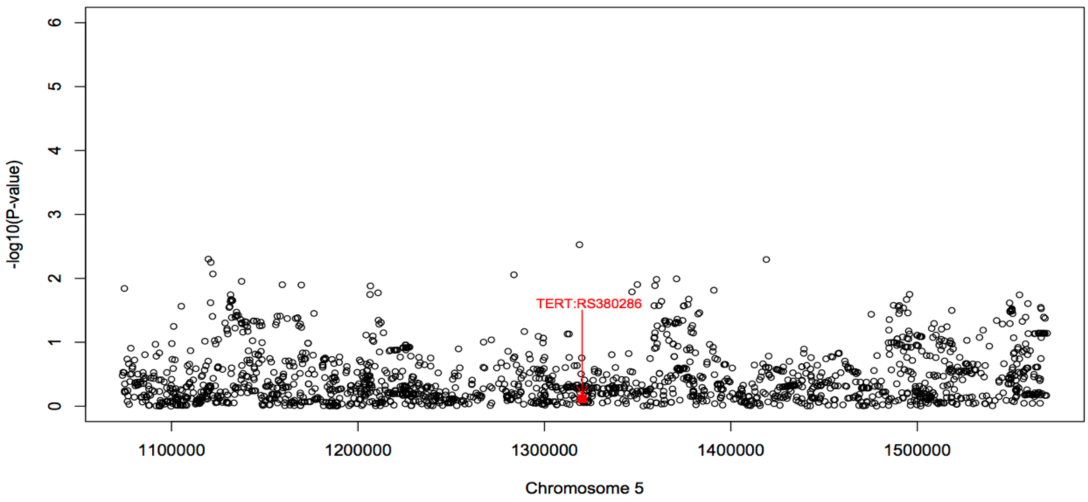


Figure 5.27 Manhattan plot for eQTL analysis for SNPs in TERT region and ILMN 2373119 expression level

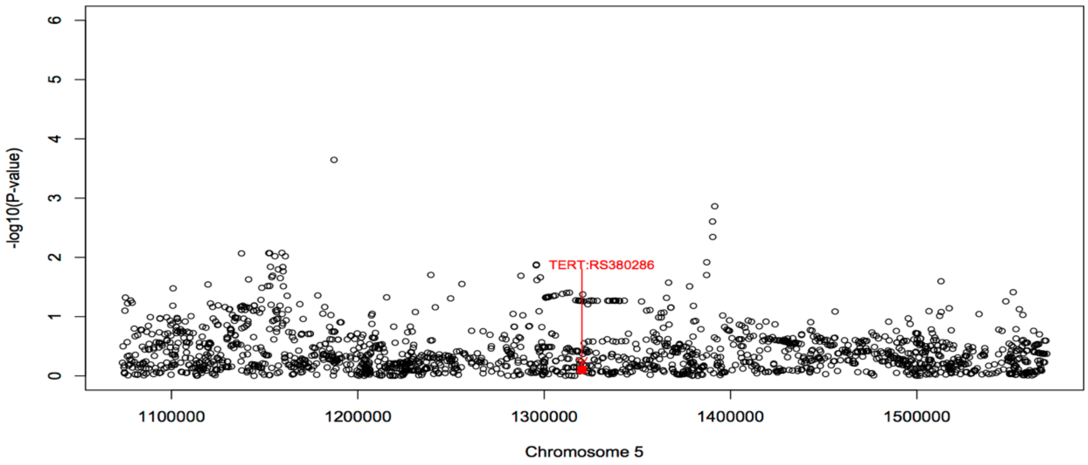


Figure 5.28 Manhattan plot for eQTL analysis for SNPs in TERT region and ILMN 1752802 expression level

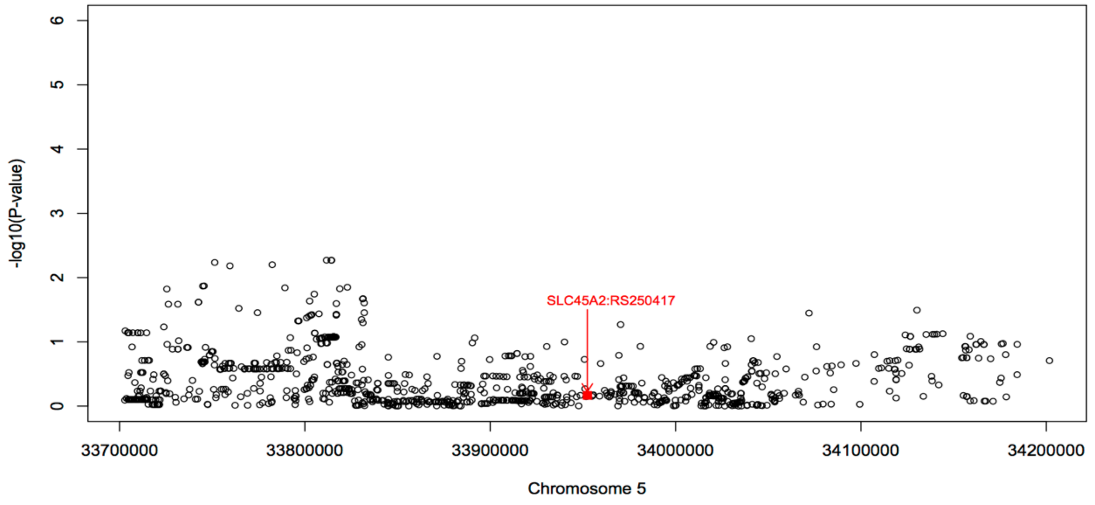


Figure 5.29 Manhattan plot for eQTL analysis for SNPs in SLC45A2 region and ILMN 1654165 expression level

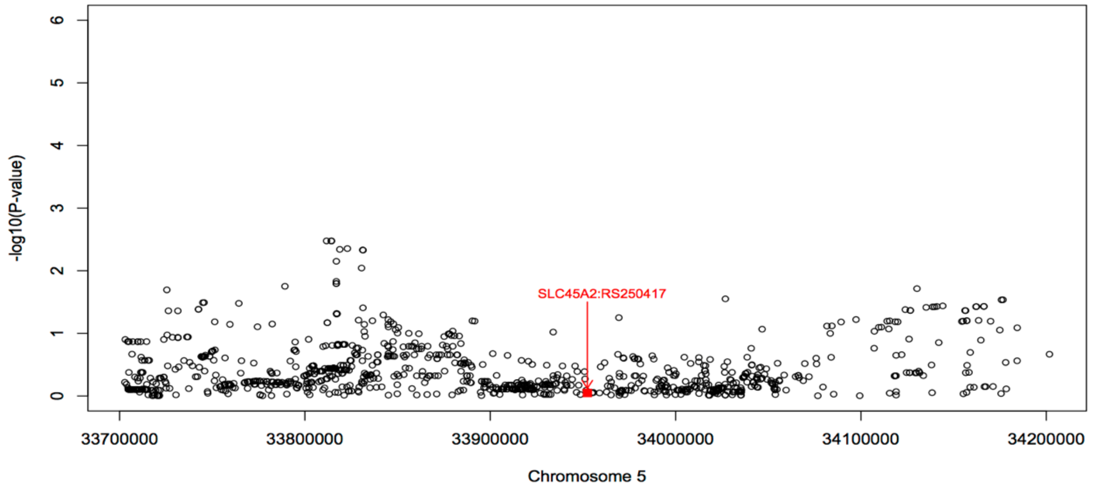


Figure 5.30 Manhattan plot for eQTL analysis for SNPs in SLC45A2 region and ILMN 1685259 expression level

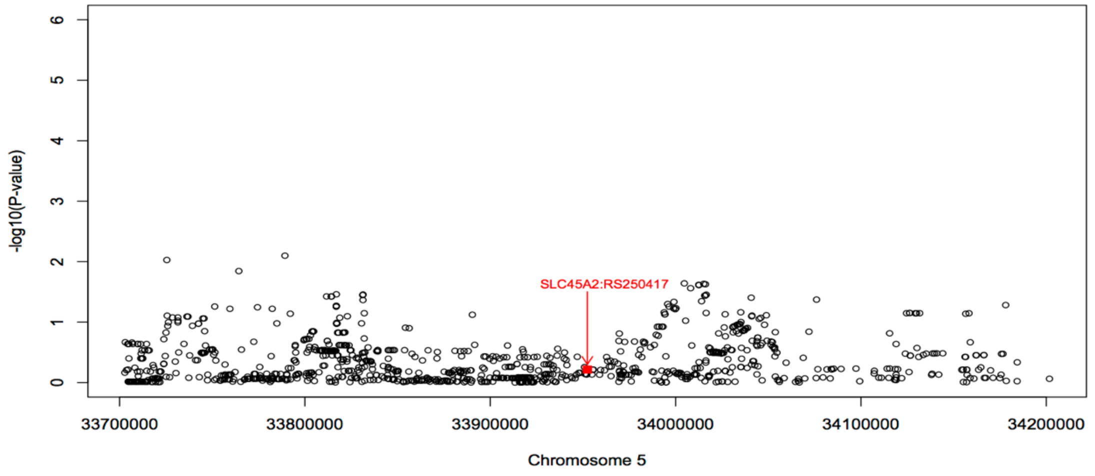


Figure 5.31 Manhattan plot for eQTL analysis for SNPs in SLC45A2 region and ILMN 2246188 expression level

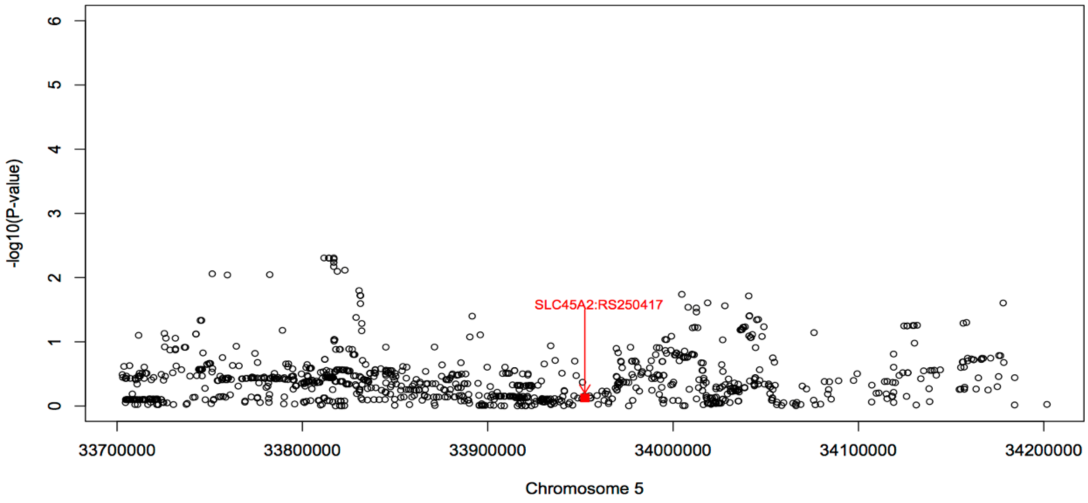


Figure 5.32 Manhattan plot for eQTL analysis for SNPs in SLC45A2 region and ILMN 2320391 expression level

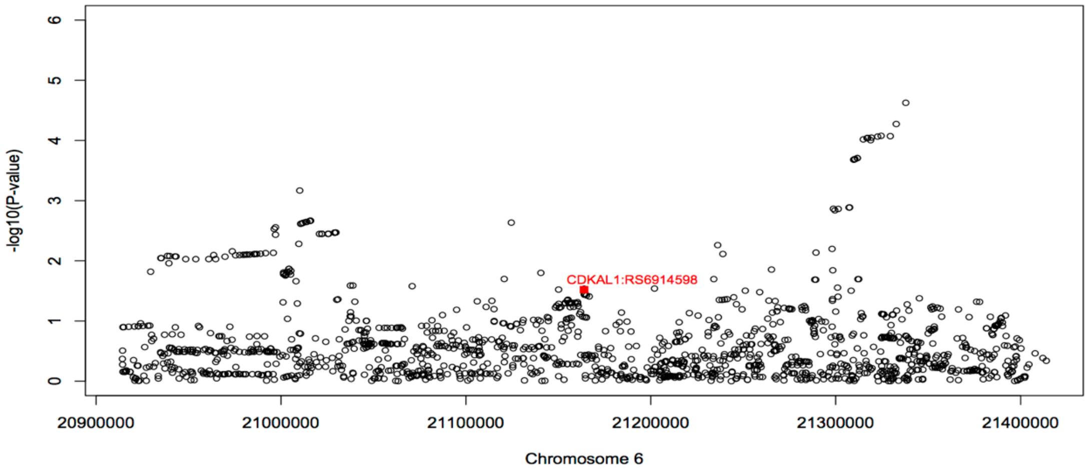


Figure 5.33 Manhattan plot for eQTL analysis for SNPs in CDKAL1 region and ILMN 1788022 expression level

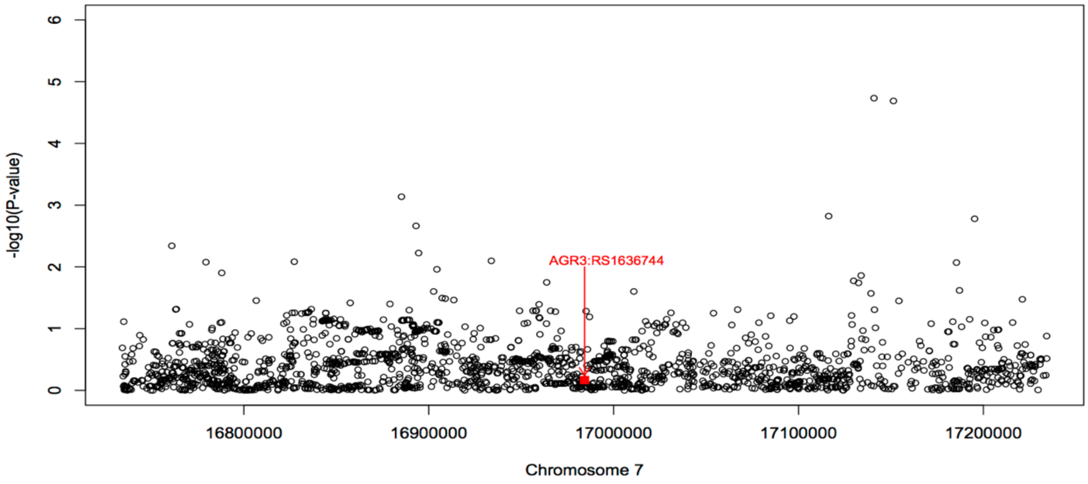


Figure 5.34 Manhattan plot for eQTL analysis for SNPs in AGR3 region and ILMN 1728787 expression level

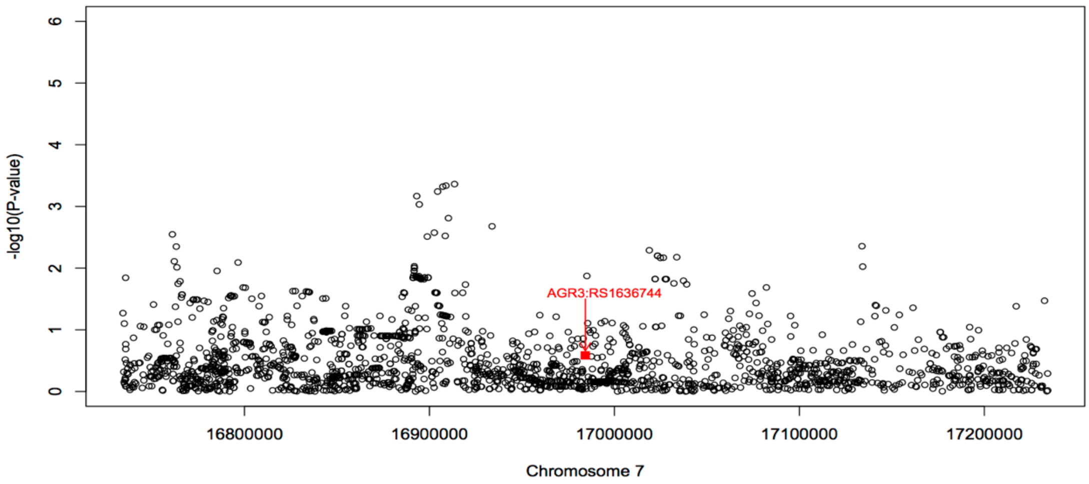


Figure 5.35 Manhattan plot for eQTL analysis for SNPs in AGR3 region and ILMN 2050246 expression level

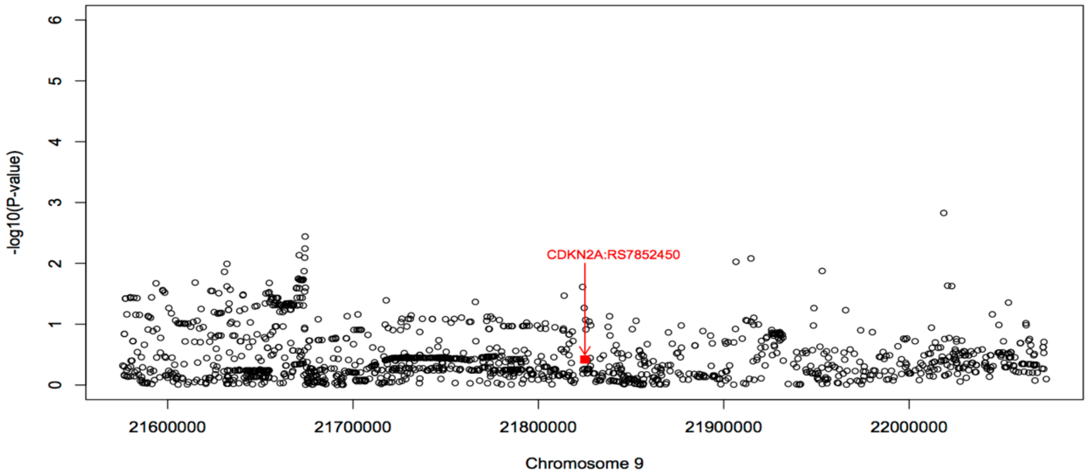


Figure 5.36 Manhattan plot for eQTL analysis for SNPs in CDKN2A region and ILMN 1717714 expression level

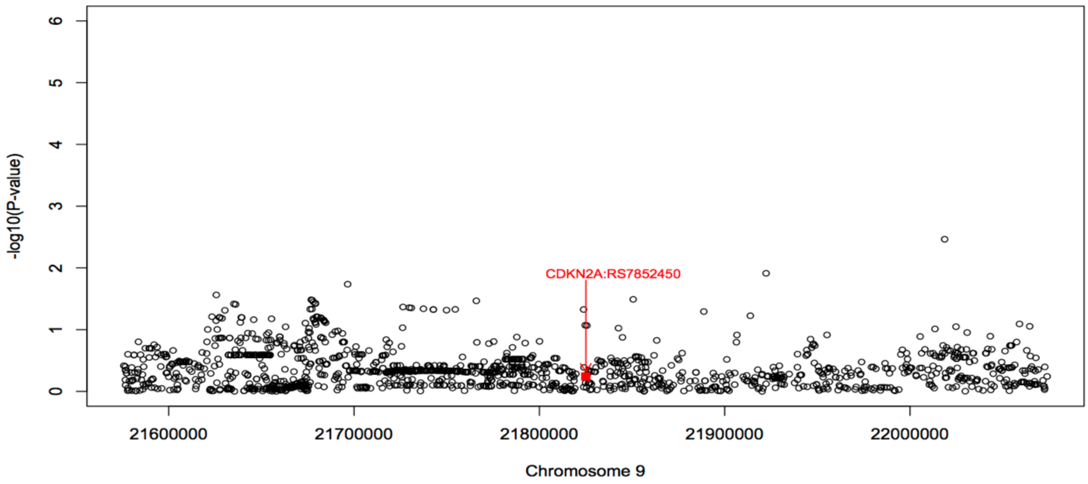


Figure 5.37 Manhattan plot for eQTL analysis for SNPs in CDKN2A region and ILMN 1744295 expression level

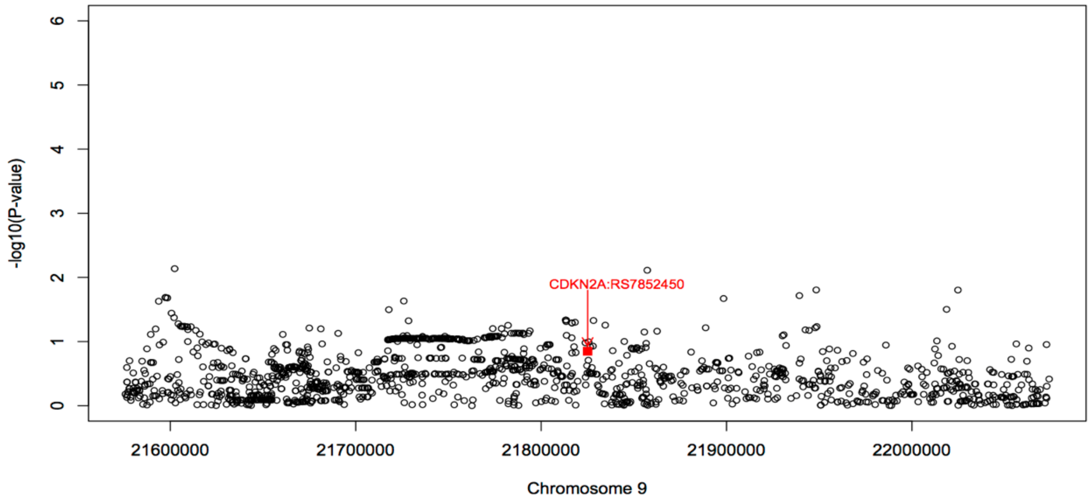


Figure 5.38 Manhattan plot for eQTL analysis for SNPs in CDKN2A region and ILMN 1757255 expression level

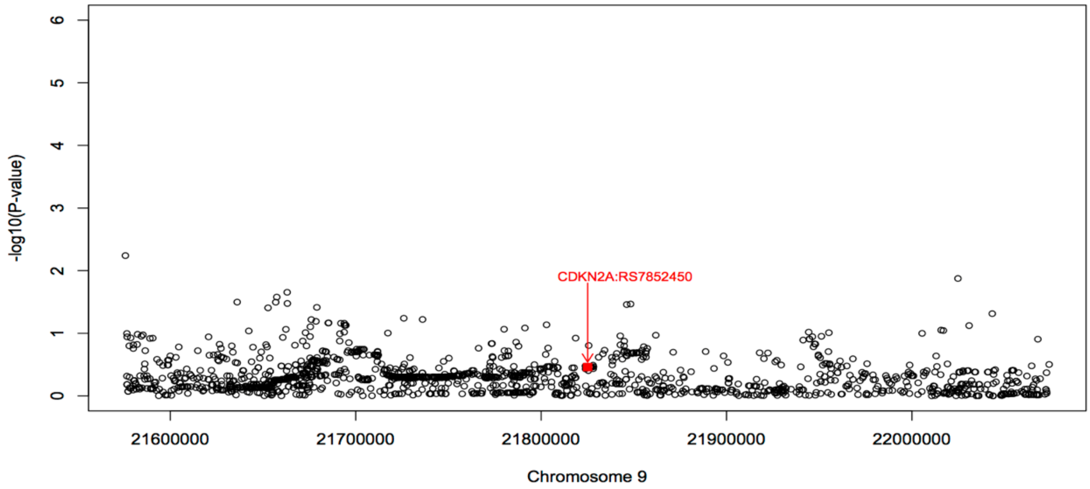


Figure 5.39 Manhattan plot for eQTL analysis for SNPs in CDKN2A region and ILMN 1753639 expression level

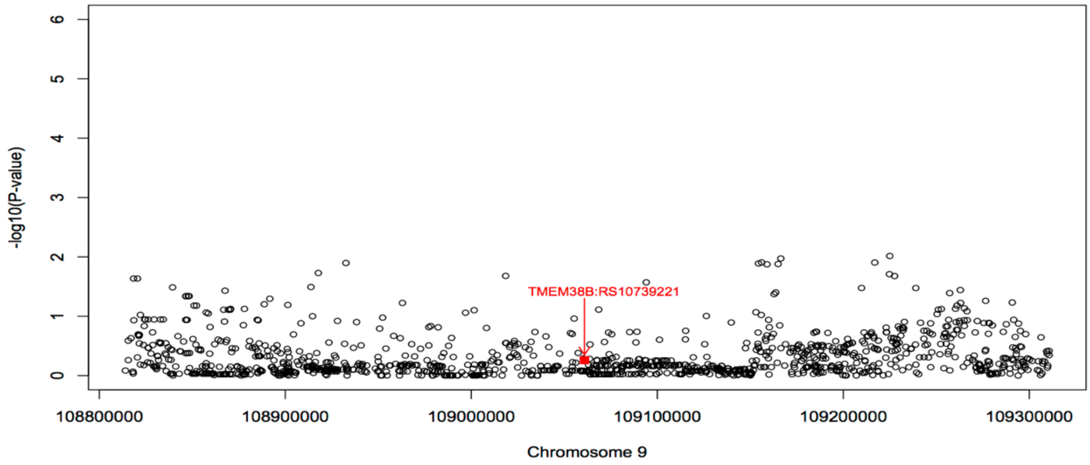


Figure 5.40 Manhattan plot for eQTL analysis for SNPs in TMEM38B region and ILMN 1669940 expression level

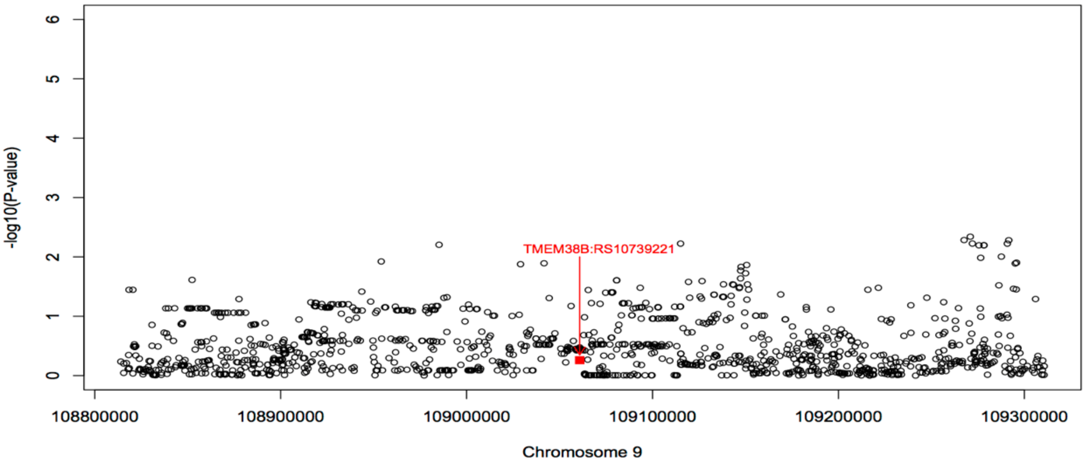


Figure 5.41 Manhattan plot for eQTL analysis for SNPs in TMEM38B region and ILMN 2093980 expression level

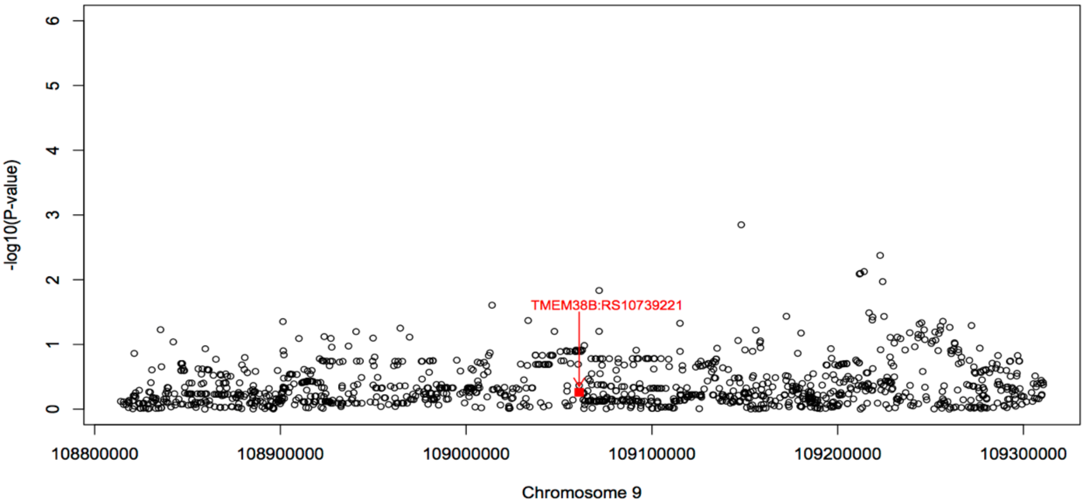


Figure 5.42 Manhattan plot for eQTL analysis for SNPs in TMEM38B region and ILMN 1722662 expression level

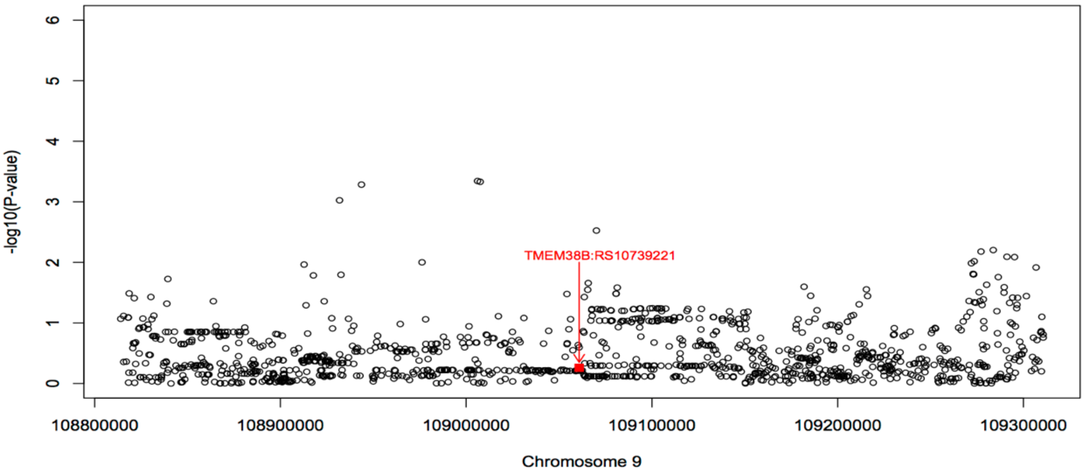


Figure 5.43 Manhattan plot for eQTL analysis for SNPs in TMEM38B region and ILMN 2135833 expression level

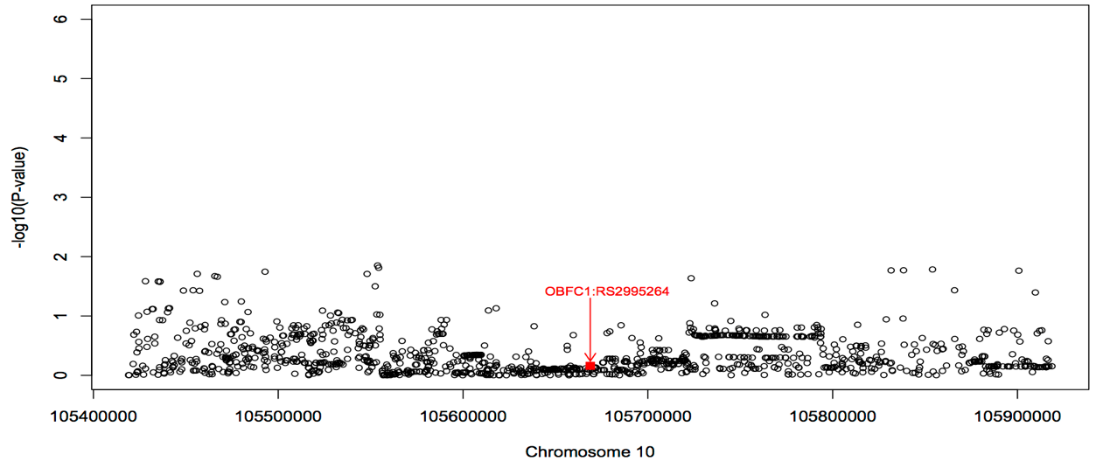


Figure 5.44 Manhattan plot for eQTL analysis for SNPs in OBCF1 region and ILMN 1789186 expression level

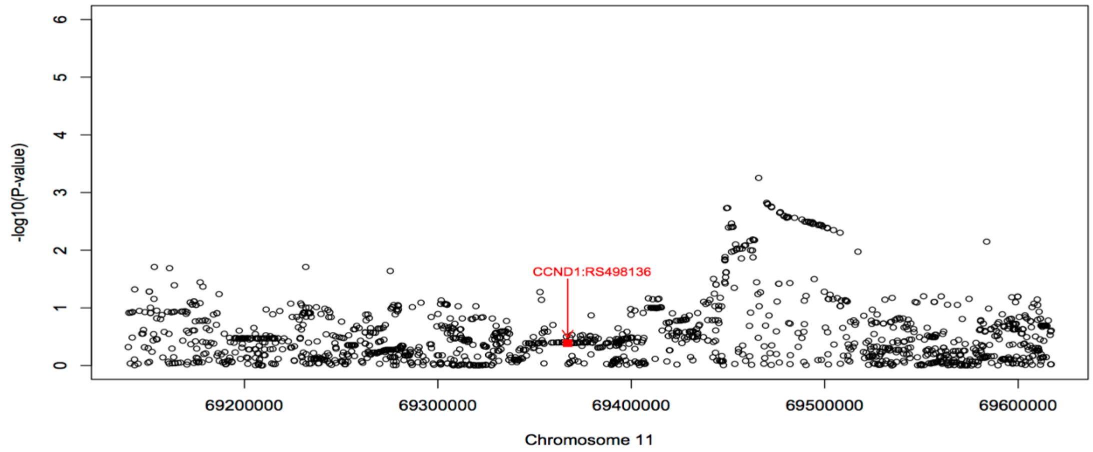


Figure 5.45 Manhattan plot for eQTL analysis for SNPs in CCND1 region and ILMN 1688480 expression level

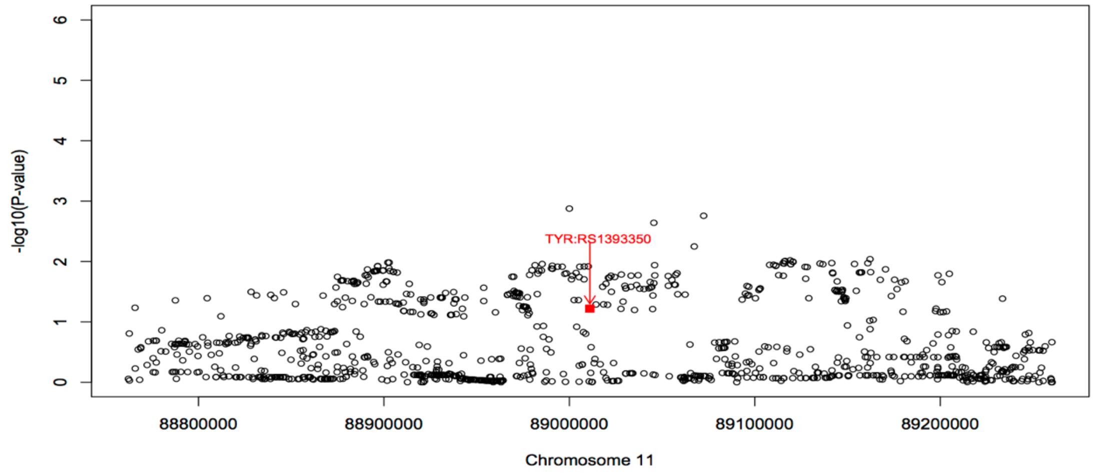


Figure 5.46 Manhattan plot for eQTL analysis for SNPs in TYR region and ILMN 1788774 expression level

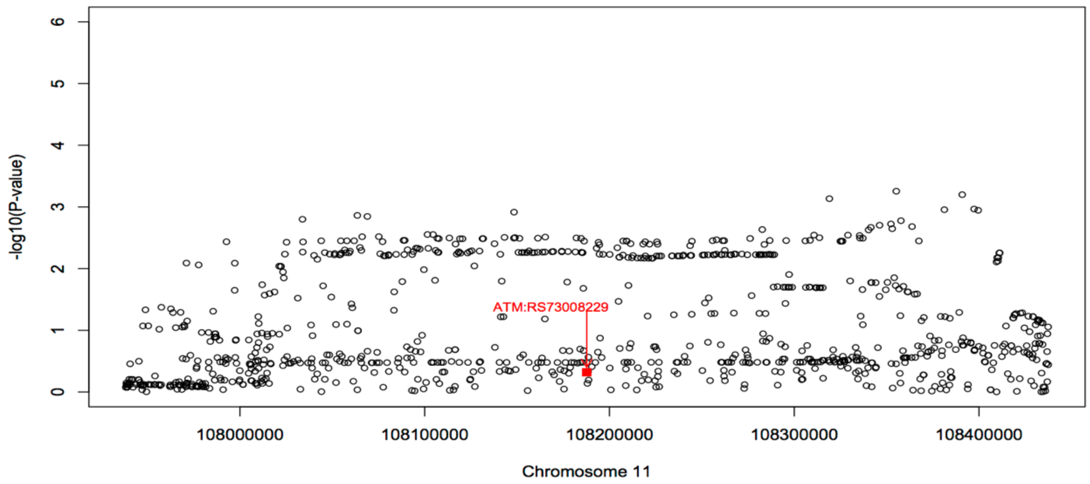


Figure 5.47 Manhattan plot for eQTL analysis for SNPs in ATM region and ILMN 1713630 expression level

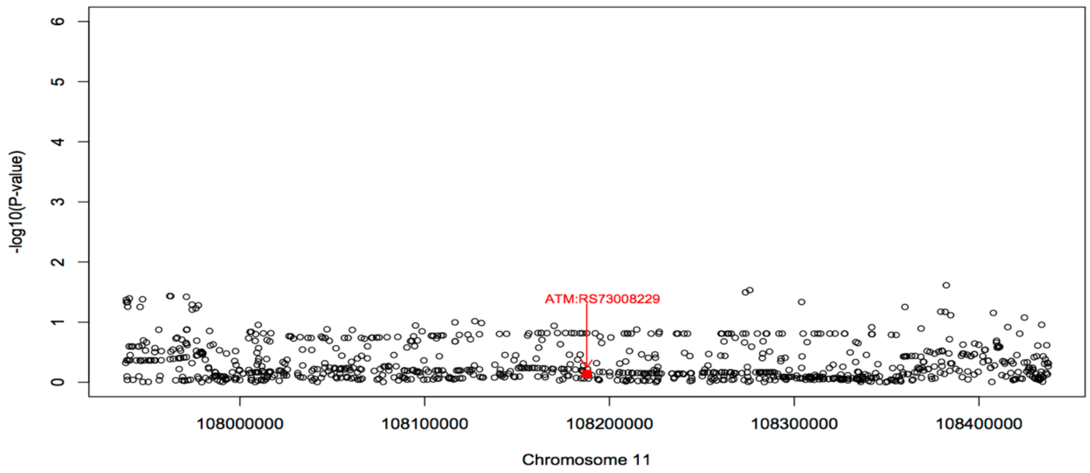


Figure 5.48 Manhattan plot for eQTL analysis for SNPs in ATM region and ILMN 1716231 expression level

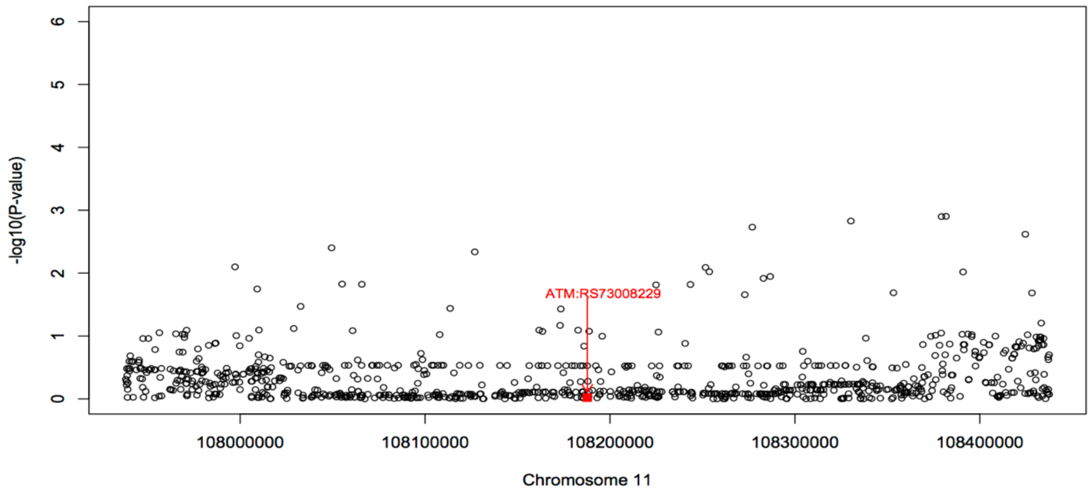


Figure 5.49 Manhattan plot for eQTL analysis for SNPs in ATM region and ILMN 1779214 expression level

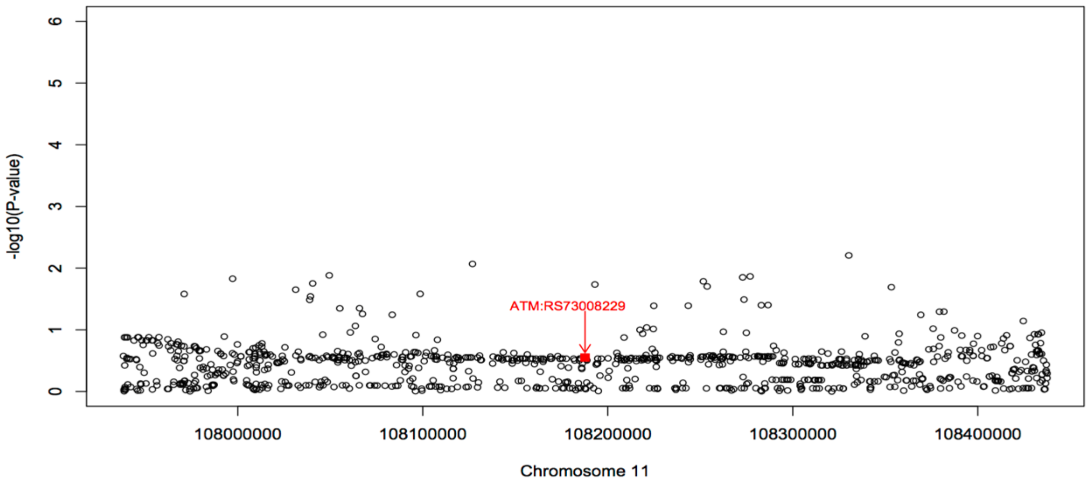


Figure 5.50 Manhattan plot for eQTL analysis for SNPs in ATM region and ILMN 2370825 expression level

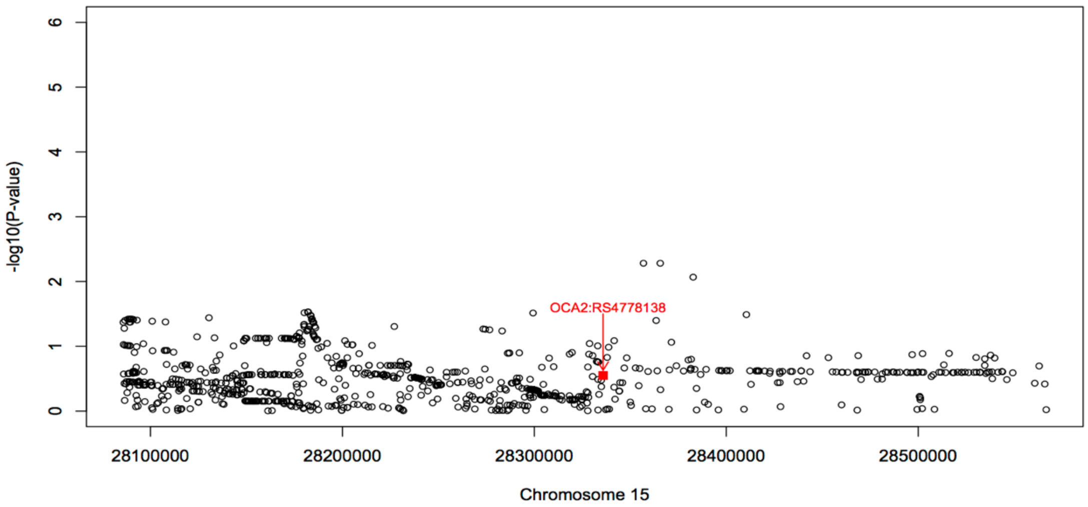


Figure 5.51 Manhattan plot for eQTL analysis for SNPs in OCA2 region and ILMN 1746116 expression level

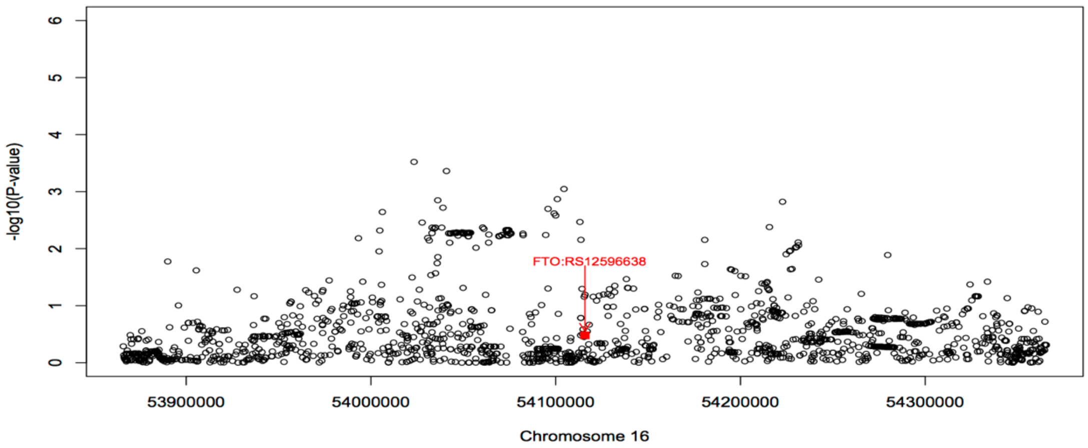


Figure 5.52 Manhattan plot for eQTL analysis for SNPs in FTO region and ILMN 2288070 expression level

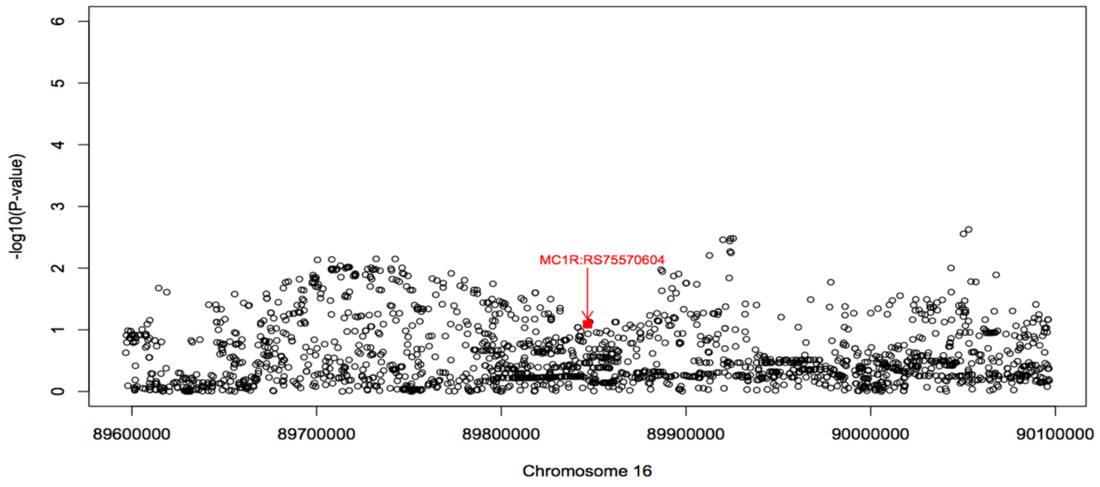


Figure 5.53 Manhattan plot for eQTL analysis for SNPs in MC1R region and ILMN 1653319 expression level

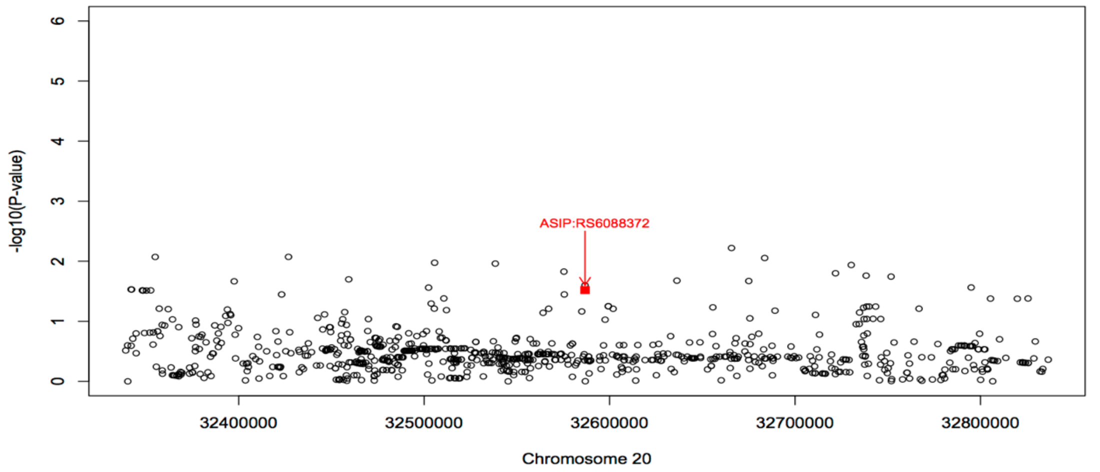


Figure 5.54 Manhattan plot for eQTL analysis for SNPs in ASIP region and ILMN 1791647 expression level

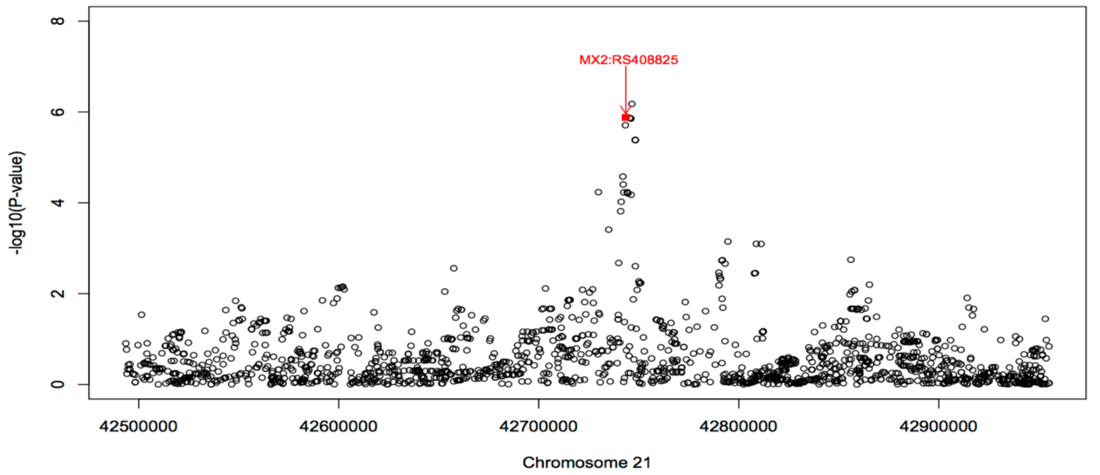


Figure 5.55 Manhattan plot for eQTL analysis for SNPs in MX2 region and ILMN 2231928 expression level

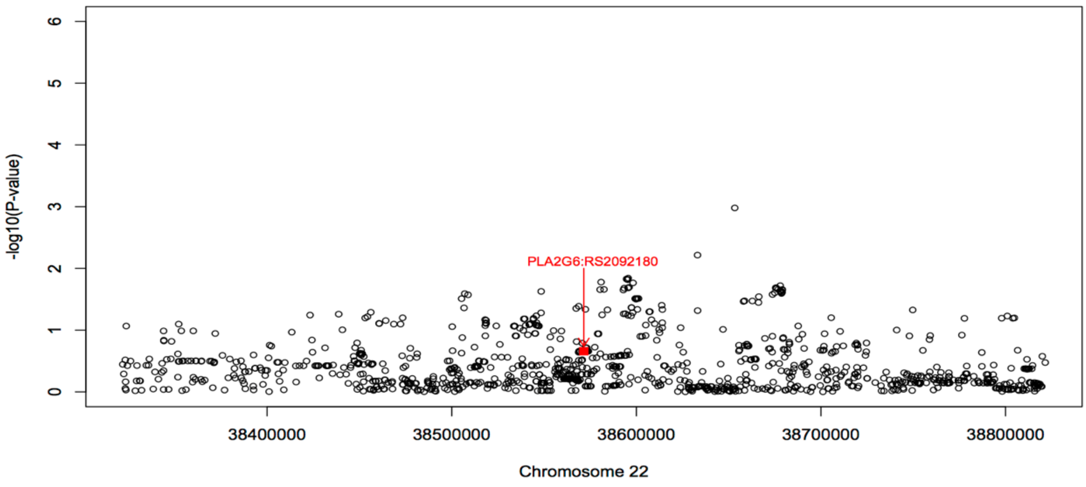


Figure 5.56 Manhattan plot for eQTL analysis for SNPs in PLA2G6 region and ILMN 1697654 expression level

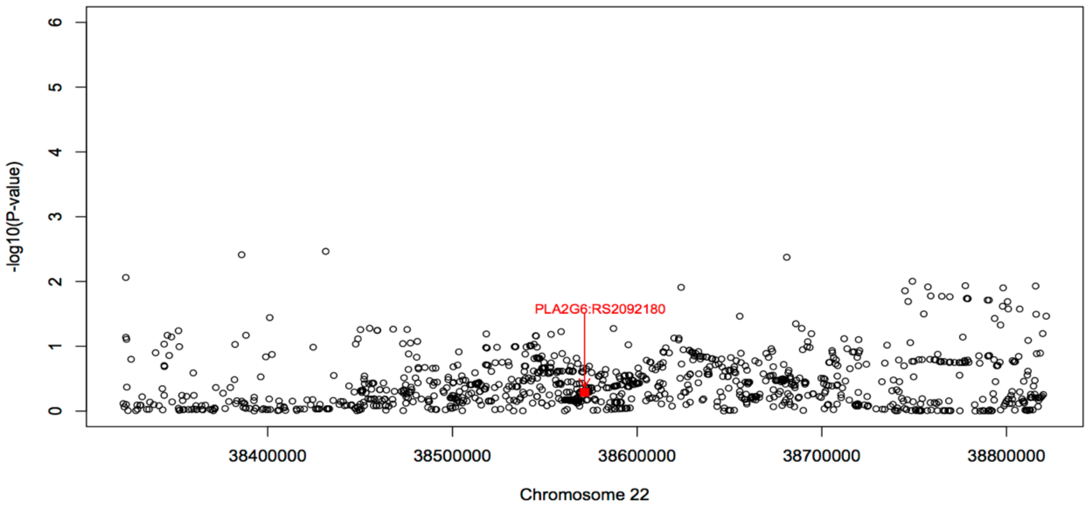


Figure 5.57 Manhattan plot for eQTL analysis for SNPs in PLA2G6 region and ILMN 1798955 expression level

Appendix II

Checking proportional hazards assumption for the multivariable Cox model in Table 3.3 in Chapter 3

Table 1 Schoenfeld residuals test for variables in the multivariable Cox model in Table 3.3

	rho	chisq	P-value
Age	0.02	0.07	0.79
Sex (Male)	-0.06	1.06	0.30
Tumour site (Head/Neck)	-0.03	0.26	0.61
Tumour site (Trunk)	0.04	0.41	0.52
Tumour site (Other)	0.07	1.50	0.22
Breslow thickness	-0.05	0.56	0.45
Presence of ulceration	-0.11	3.65	0.06
GLOBAL	NA	8.81	0.27

P-value for individual predictor and global test is not significant. Proportional hazards assumption is not violated

Model fitness assessment for the combined models in the test set (risk score approach)

1. Model 1 Combined data survival models (Model fitted in Table 6.3)

1.1 Checking linearity for continuous predictors using Martingale residuals

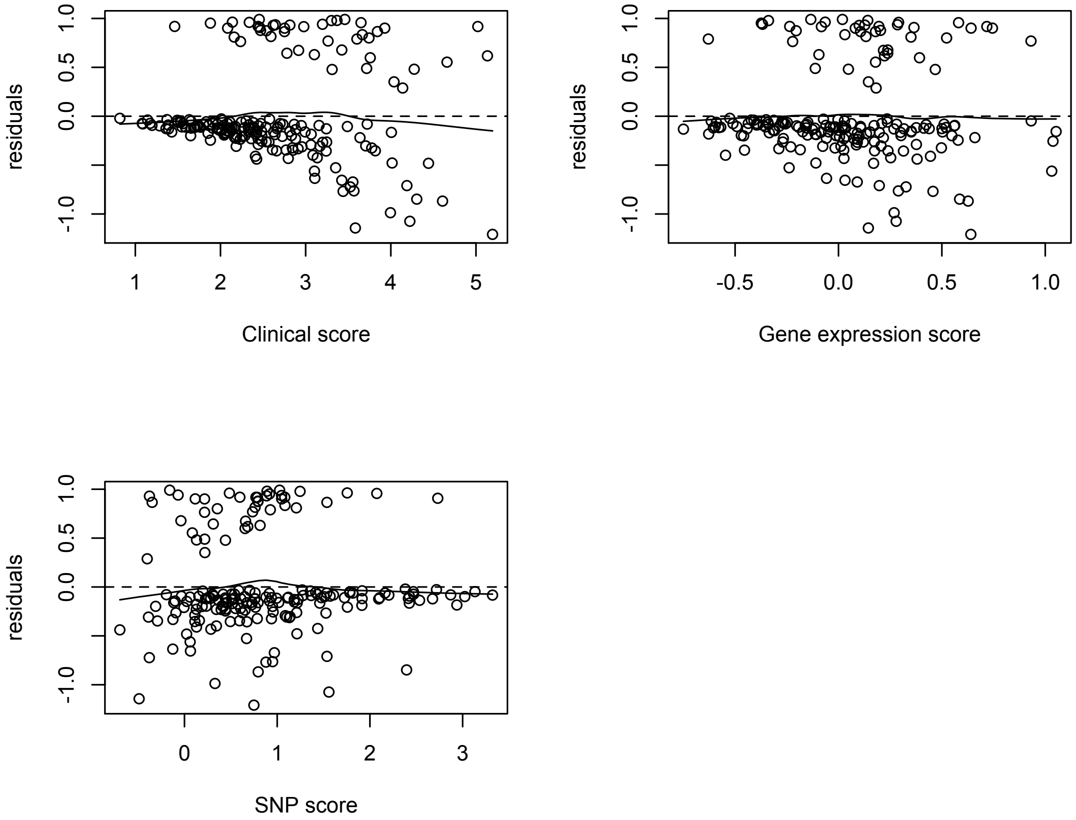


Figure 1. Martingale residuals versus covariate for clinical score, gene expression score and SNP score

Nonlinearity appear to be slight for clinical score and SNP score.

1.2 Checking proportional hazards assumption using Schoenfeld residuals test

Table 2 Schoenfeld residuals test for variables in the multivariable Cox model in Table 6.3

	rho	chisq	P-value
Clinical score	-0.09	0.34	0.56
Gene expression score	0.08	0.29	0.59
SNP score	-0.02	0.01	0.92
Global	NA	0.45	0.93

P-value for individual predictor and global test is not significant. Proportional hazards assumption is not violated

2. Model 2 Combined data survival models with prior cluster analysis (Model 1 fitted in Table 6.12)

2.1. Checking linearity for continuous predictors using Martingale residuals

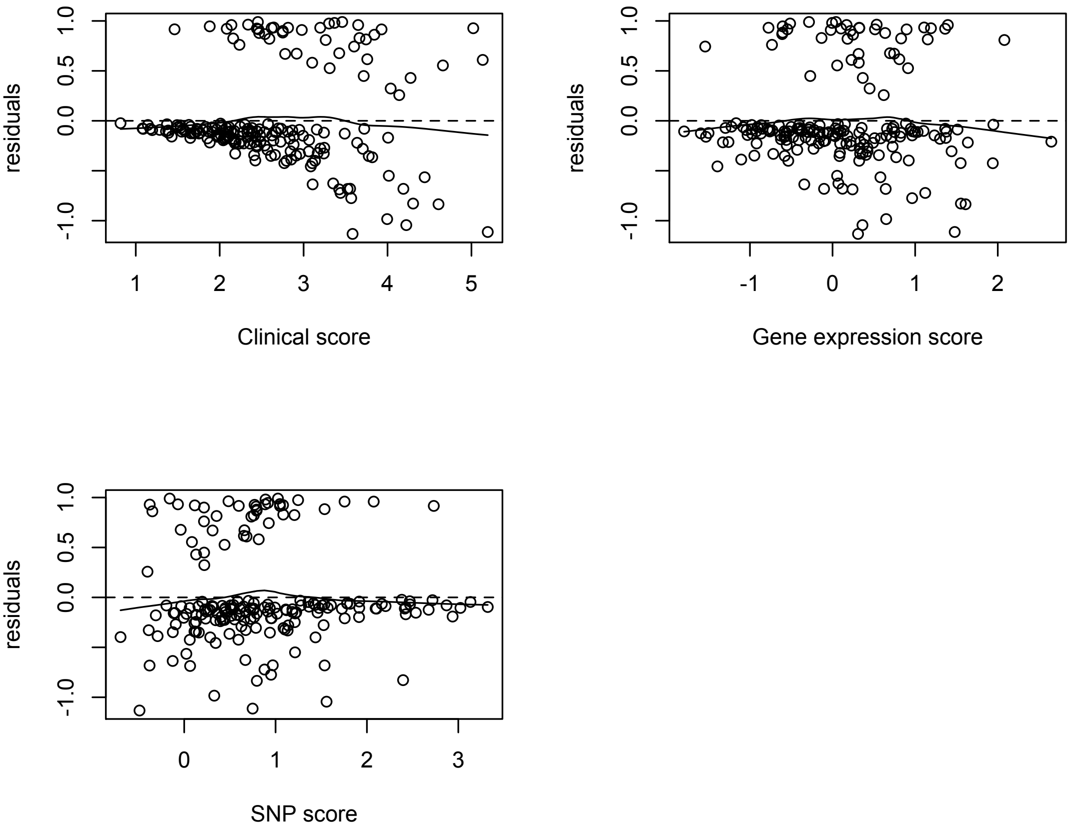


Figure 2. Martingale residuals versus covariate for clinical score, gene expression score and SNP score

Nonlinearity appear to be slight in all scores.

2.2 Checking proportional hazards assumption using Schoenfeld residuals test

Table 3 Schoenfeld residuals test for variables in the multivariable Cox model in Table 6.12

	rho	chisq	P-value
Clinical score	-0.09	0.36	0.55
Gene expression score	0.09	0.35	0.56
SNP score	-0.02	0.01	0.93
Global	NA	0.50	0.92

P-value for individual predictor and global test is not significant. Proportional hazards assumption is not violated.

3. Model 3 Combined data survival models using Lund cluster (Model fitted in Table 6.19)

3.1. Checking linearity for continuous predictors using Martingale residuals

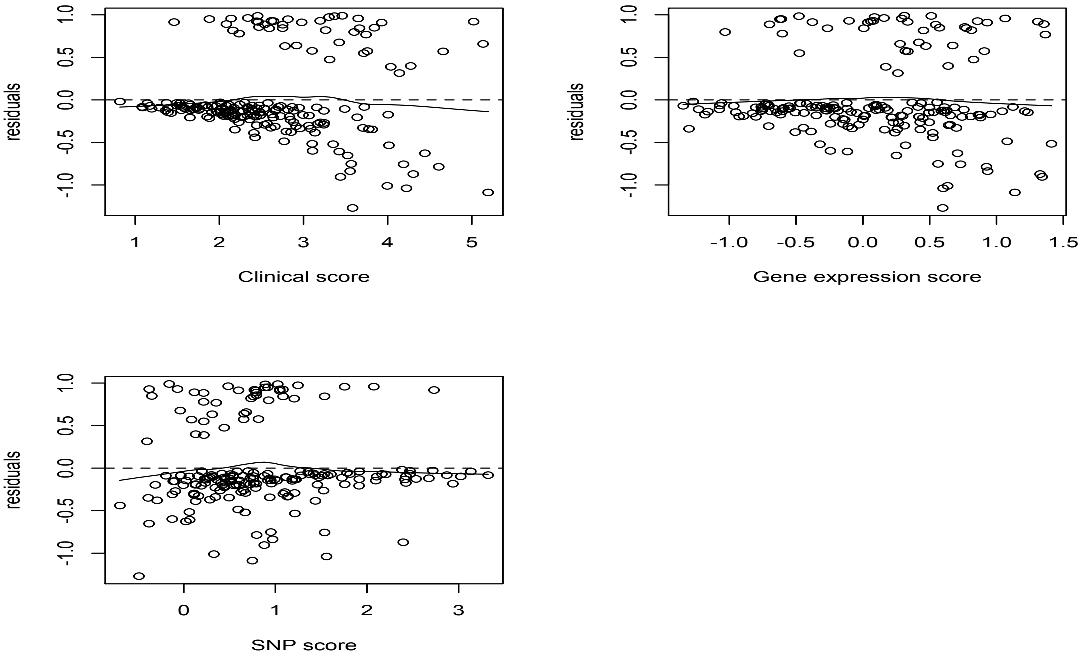


Figure 3. Martingale residuals versus covariate for clinical score, gene expression score and SNP score

Nonlinearity appear to be slight in clinical score and SNP score

3.2 Checking proportional hazards assumption using Schoenfeld residuals test

Table 4 Schoenfeld residuals test for variables in the multivariable Cox model in Table 6.19

	rho	chisq	P-value
Clinical score	-0.05	0.10	0.75
Gene expression score	-0.02	0.02	0.89
SNP score	-0.01	0.01	0.97
Global	NA	0.23	0.97

P-value for individual predictor and global test is not significant. Proportional hazards assumption is not violated.

References

Akslen, L. A., Angelini, S., Straume, O., Bachmann, I. M., Molven, A., Hemminki, K. & Kumar, R. 2005. BRAF and NRAS mutations are frequent in nodular melanoma but are not associated with tumor cell proliferation or patient survival. *J Invest Dermatol,* 125**,** 312-7.

Alonso, S. R., Tracey, L., Ortiz, P., Perez-Gomez, B., Palacios, J., Pollan, M., Linares, J., Serrano, S., Saez-Castillo, A. I., Sanchez, L., Pajares, R., Sanchez-Aguilera, A., Artiga, M. J., Piris, M. A. & Rodriguez-Peralto, J. L. 2007. A high-throughput study in melanoma identifies epithelial-mesenchymal transition as a major determinant of metastasis. *Cancer Res,* 67**,** 3450-60.

Azimi, F., Scolyer, R. A., Rumcheva, P., Moncrieff, M., Murali, R., Mccarthy, S. W., Saw, R. P. & Thompson, J. F. 2012. Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma. *J Clin Oncol,* 30**,** 2678-83.

Azzato, E. M., Greenberg, D., Shah, M., Blows, F., Driver, K. E., Caporaso, N. E. & Pharoah, P. D. P. 2009. Prevalent cases in observational studies of cancer survival: do they bias hazard ratio estimates? *British Journal of Cancer,* 100**,** 1806-11.

Baade, P., Meng, X., Youlden, D., Aitken, J. & Youl, P. 2012. Time trends and latitudinal differences in melanoma thickness distribution in Australia, 1990-2006. *Int J Cancer,* 130**,** 170-8.

Baade, P. D., Royston, P., Youl, P. H., Weinstock, M. A., Geller, A. & Aitken, J. F. 2015. Prognostic survival model for people diagnosed with invasive cutaneous melanoma. *BMC Cancer,* 15**,** 27.

Balch, C. M., Gershenwald, J. E., Soong, S. J., Thompson, J. F., Atkins, M. B., Byrd, D. R., Buzaid, A. C., Cochran, A. J., Coit, D. G., Ding, S., Eggermont, A. M., Flaherty, K. T., Gimotty, P. A., Kirkwood, J. M., Mcmasters, K. M., Mihm, M. C., Jr., Morton, D. L., Ross, M. I., Sober, A. J. & Sondak, V. K. 2009. Final version of 2009 AJCC melanoma staging and classification. *J Clin Oncol,* 27**,** 6199-206.

Barrett, J. H., Iles, M. M., Harland, M., Taylor, J. C., Aitken, J. F., Andresen, P. A., Akslen, L. A., Armstrong, B. K., Avril, M. F., Azizi, E., Bakker, B., Bergman, W., Bianchi-Scarra, G., Bressac-De Paillerets, B., Calista, D., Cannon-Albright, L. A., Corda, E., Cust, A. E., Debniak, T., Duffy, D., Dunning, A. M., Easton, D. F., Friedman, E., Galan, P., Ghiorzo, P., Giles, G. G., Hansson, J., Hocevar, M., Hoiom, V., Hopper, J. L., Ingvar, C., Janssen, B., Jenkins, M. A., Jonsson, G., Kefford, R. F., Landi, G., Landi, M. T., Lang, J., Lubinski, J., Mackie, R., Malvehy, J., Martin, N. G., Molven, A., Montgomery, G. W., Van Nieuwpoort, F. A., Novakovic, S., Olsson, H., Pastorino, L., Puig, S., Puig-Butille, J. A., Randerson-Moor, J., Snowden, H., Tuominen, R., Van Belle, P., Van Der Stoep, N., Whiteman, D. C., Zelenika, D., Han, J., Fang, S., Lee, J. E., Wei, Q., Lathrop, G. M., Gillanders, E. M., Brown, K. M., Goldstein, A. M., Kanetsky, P. A., Mann, G. J., Macgregor, S., Elder, D. E., Amos, C. I., Hayward, N. K., Gruis, N. A., Demenais, F., Bishop, J. A., Bishop, D. T.

& Geno, M. E. L. C. 2011. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat Genet,* 43**,** 1108-13.

Barrett, J. H., Taylor, J. C., Bright, C., Harland, M., Dunning, A. M., Akslen, L. A., Andresen, P. A., Avril, M. F., Azizi, E., Bianchi Scarra, G., Brossard, M., Brown, K. M., Debniak, T., Elder, D. E., Friedman, E., Ghiorzo, P., Gillanders, E. M., Gruis, N. A., Hansson, J., Helsing, P., Hocevar, M., Hoiom, V., Ingvar, C., Landi, M. T., Lang, J., Lathrop, G. M., Lubinski, J., Mackie, R. M., Molven, A., Novakovic, S., Olsson, H., Puig, S., Puig-Butille, J. A., Van Der Stoep, N., Van Doorn, R., Van Workum, W., Goldstein, A. M., Kanetsky, P. A., Pharoah, P. D., Demenais, F., Hayward, N. K., Newton Bishop, J. A., Bishop, D. T., Iles, M. M. & Geno, M. E. L. C. 2015. Fine mapping of genetic susceptibility loci for melanoma reveals a mixture of single variant and multiple variant regions. *Int J Cancer,* 136**,** 1351-60.

Benjamini, Y. & Yekutieli, D. 2001. The control of the flase discovery rate in multiple testing under dependency. *Annals of Statistics,* 29**,** 1165-1188.

Bishop, D. T., Demenais, F., Iles, M. M., Harland, M., Taylor, J. C., Corda, E., Randerson-Moor, J., Aitken, J. F., Avril, M. F., Azizi, E., Bakker, B., Bianchi-Scarra, G., Bressac-De Paillerets, B., Calista, D., Cannon-Albright, L. A., Chin, A. W. T., Debniak, T., Galore-Haskel, G., Ghiorzo, P., Gut, I., Hansson, J., Hocevar, M., Hoiom, V., Hopper, J. L., Ingvar, C., Kanetsky, P. A., Kefford, R. F., Landi, M. T., Lang, J., Lubinski, J., Mackie, R., Malvehy, J., Mann, G. J., Martin, N. G., Montgomery, G. W., Van Nieuwpoort, F. A., Novakovic, S., Olsson, H., Puig, S., Weiss, M., Van Workum, W., Zelenika, D., Brown, K. M., Goldstein, A. M., Gillanders, E. M., Boland, A., Galan, P., Elder, D. E., Gruis, N. A., Hayward, N. K., Lathrop, G. M., Barrett, J. H. & Bishop, J. A. 2009. Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet,* 41**,** 920-5.

Blom, D. J., Luyten, G. P., Mooy, C., Kerkvliet, S., Zwinderman, A. H. & Jager, M. J. 1997. Human leukocyte antigen class I expression. Marker of poor prognosis in uveal melanoma. *Invest Ophthalmol Vis Sci,* 38**,** 1865-72.

Bogunovic, D., O'neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y. L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A. C., Lonardi, S., Zavadil, J., Osman, I. & Bhardwaj, N. 2009. Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci U S A,* 106**,** 20429-34.

Boniol, M., Autier, P., Boyle, P. & Gandini, S. 2012. Cutaneous melanoma attributable to sunbed use: systematic review and meta-analysis. *BMJ,* 345**,** e4757.

Bovelstad, H. M., Nygard, S. & Borgan, O. 2009. Survival prediction from clinico-genomic models--a comparative study. *BMC Bioinformatics,* 10**,** 413.

Bovelstad, H. M., Nygard, S., Storvold, H. L., Aldrin, M., Borgan, O., Frigessi, A. & Lingjaerde, O. C. 2007. Predicting survival from microarray data--a comparative study. *Bioinformatics,* 23**,** 2080-7.

Brandt, A., Sundquist, J. & Hemminki, K. 2011. Risk of incident and fatal melanoma in individuals with a family history of incident or fatal melanoma or any cancer. *Br J Dermatol,* 165**,** 342-8.

Breslow, A. 1970. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Ann Surg,* 172**,** 902-8.

Brogelli, L., Reali, U. M., Moretti, S. & Urso, C. 1992. The prognostic significance of histologic regression in cutaneous melanoma. *Melanoma Res,* 2**,** 87-91.

Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., Ho, K. M., Ring, S., Hurles, M., Deloukas, P., Davey Smith, G. & Dermitzakis, E. T. 2014. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet,* 10**,** e1004461.

Buettner, P. G., Leiter, U., Eigentler, T. K. & Garbe, C. 2005. Development of prognostic factors and survival in cutaneous melanoma over 25 years: An analysis of the Central Malignant Melanoma Registry of the German Dermatological Society. *Cancer,* 103**,** 616-24.

Burkhardt, R., Kirsten, H., Beutner, F., Holdt, L. M., Gross, A., Teren, A., Tonjes, A., Becker, S., Krohn, K., Kovacs, P., Stumvoll, M., Teupser, D., Thiery, J., Ceglarek, U. & Scholz, M. 2015. Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genet,* 11**,** e1005510.

Burns, T., Breathnach, S., Cox, N. & Griffiths, C. 2004. *Rook's Textbook of Dermatology*, Blackwell Publishing.

Burton, A. L., Gilbert, J., Farmer, R. W., Stromberg, A. J., Hagendoorn, L., Ross, M. I., Martin, R. C., 2nd, Mcmasters, K. M., Scoggins, C. R. & Callender, G. G. 2011. Regression does not predict nodal metastasis or survival in patients with cutaneous melanoma. *Am Surg,* 77**,** 1009-13.

Chang, Y. M., Barrett, J. H., Bishop, D. T., Armstrong, B. K., Bataille, V., Bergman, W., Berwick, M., Bracci, P. M., Elwood, J. M., Ernstoff, M. S., Gallagher, R. P., Green, A. C., Gruis, N. A., Holly, E. A., Ingvar, C., Kanetsky, P. A., Karagas, M. R., Lee, T. K., Le Marchand, L., Mackie, R. M., Olsson, H., Osterlind, A., Rebbeck, T. R., Sasieni, P., Siskind, V., Swerdlow, A. J., Titus-Ernstoff, L., Zens, M. S. & Newton-Bishop, J. A. 2009a. Sun exposure and melanoma risk at different latitudes: a pooled analysis of 5700 cases and 7216 controls. *Int J Epidemiol,* 38**,** 814-30.

Chang, Y. M., Newton-Bishop, J. A., Bishop, D. T., Armstrong, B. K., Bataille, V., Bergman, W., Berwick, M., Bracci, P. M., Elwood, J. M., Ernstoff, M. S., Green, A. C., Gruis, N. A., Holly, E. A., Ingvar, C., Kanetsky, P. A., Karagas, M. R., Le Marchand, L., Mackie, R. M., Olsson, H., Osterlind, A., Rebbeck, T. R., Reich, K., Sasieni, P., Siskind, V., Swerdlow, A. J., Titus-Ernstoff, L., Zens, M. S., Ziegler, A. & Barrett, J. H. 2009b. A pooled analysis of melanocytic nevus phenotype and the risk of cutaneous melanoma at different latitudes. *Int J Cancer,* 124**,** 420-8.

Chen, X., Li, X., Chen, J., Zheng, P., Huang, S. & Ouyang, X. 2012. Overexpression of CIAPIN1 inhibited pancreatic cancer cell proliferation and was associated with good prognosis in pancreatic cancer. *Cancer Gene Ther,* 19**,** 538-44.

Cirenajwis, H., Ekedahl, H., Lauss, M., Harbst, K., Carneiro, A., Enoksson, J., Rosengren, F., Werner-Hartman, L., Torngren, T., Kvist, A., Fredlund, E., Bendahl, P. O., Jirstrom, K., Lundgren, L., Howlin, J., Borg, A., Gruvberger-Saal, S. K., Saal, L. H., Nielsen, K., Ringner, M., Tsao, H., Olsson, H., Ingvar, C., Staaf, J. & Jonsson, G. 2015. Molecular stratification of metastatic melanoma using gene expression profiling: Prediction of survival outcome and benefit from molecular targeted therapy. *Oncotarget,* 6**,** 12297-309.

Conway, C., Mitra, A., Jewell, R., Randerson-Moor, J., Lobo, S., Nsengimana, J., Edward, S., Sanders, D. S., Cook, M., Powell, B., Boon, A., Elliott, F., De Kort, F., Knowles, M. A., Bishop, D. T. & Newton-Bishop, J. 2009. Gene expression profiling of paraffin-embedded primary melanoma using the DASL assay identifies increased osteopontin expression as predictive of reduced relapse-free survival. *Clin Cancer Res,* 15**,** 6939-46.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., Mckinney, S., Group, M., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A. L., Brenton, J. D., Tavare, S., Caldas, C. & Aparicio, S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* 486**,** 346-52.

Davies, J. R., Field, S., Randerson-Moor, J., Harland, M., Kumar, R., Anic, G. M., Nagore, E., Hansson, J., Hoiom, V., Jonsson, G., Gruis, N. A., Park, J. Y., Guan, J., Sivaramakrishna Rachakonda, P., Wendt, J., Pjanova, D., Puig, S., Schadendorf, D., Okamoto, I., Olsson, H., Affleck, P., Garcia-Casado, Z., Puig-Butille, J. A., Stratigos, A. J., Kodela, E., Donina, S., Sucker, A., Hosen, I., Egan, K. M., Barrett, J. H., Van Doorn, R., Bishop, D. T. & Newton-Bishop, J. 2014a. An inherited variant in the gene coding for vitamin D-binding protein and survival from cutaneous melanoma: a BioGenoMEL study. *Pigment Cell Melanoma Res,* 27**,** 234-43.

Davies, J. R., Jewell, R., Affleck, P., Anic, G. M., Randerson-Moor, J., Ozola, A., Egan, K. M., Elliott, F., Garcia-Casado, Z., Hansson, J., Harland, M., Hoiom, V., Jian, G., Jonsson, G., Kumar, R., Nagore, E., Wendt, J., Olsson, H., Park, J. Y., Patel, P., Pjanova, D., Puig, S., Schadendorf, D., Sivaramakrishna Rachakonda, P., Snowden, H., Stratigos, A. J., Bafaloukos, D., Ogbah, Z., Sucker, A., Van Den Oord, J. J., Van Doorn, R., Walker, C., Okamoto, I., Wolter, P., Barrett, J. H., Timothy Bishop, D. & Newton-Bishop, J. 2014b. Inherited variation in the PARP1 gene and survival from melanoma. *Int J Cancer,* 135**,** 1625-33.

Davies, J. R., Randerson-Moor, J., Kukalizch, K., Harland, M., Kumar, R., Madhusudan, S., Nagore, E., Hansson, J., Hoiom, V., Ghiorzo, P., Gruis, N. A., Kanetsky, P. A., Wendt, J., Pjanova, D., Puig, S., Saiag, P., Schadendorf, D., Soufir, N., Okamoto, I., Affleck, P., Garcia-Casado, Z., Ogbah, Z., Ozola, A., Queirolo, P., Sucker, A., Barrett, J. H., Van Doorn, R., Bishop, D. T. & Newton-Bishop, J. 2012. Inherited variants in the MC1R gene and survival from cutaneous melanoma: a BioGenoMEL study. *Pigment Cell Melanoma Res,* 25**,** 384-94.

Deeb, K. K., Trump, D. L. & Johnson, C. S. 2007. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat Rev Cancer,* 7**,** 684-700.

Devitt, B., Liu, W., Salemi, R., Wolfe, R., Kelly, J., Tzen, C. Y., Dobrovic, A. & Mcarthur, G. 2011. Clinical outcome and pathological features associated with NRAS mutation in cutaneous melanoma. *Pigment Cell Melanoma Res,* 24**,** 666-72.

Dudbridge, F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet,* 9**,** e1003348.

Duffy, D. L., Iles, M. M., Glass, D., Zhu, G., Barrett, J. H., Hoiom, V., Zhao, Z. Z., Sturm, R. A., Soranzo, N., Hammond, C., Kvaskoff, M., Whiteman, D. C., Mangino, M., Hansson, J., Newton-Bishop, J. A., Genomel, Bataille, V., Hayward, N. K., Martin, N. G., Bishop, D. T., Spector, T. D. & Montgomery, G. W. 2010a. IRF4 variants have age-specific effects on nevus count and predispose to melanoma. *Am J Hum Genet,* 87**,** 6-16.

Duffy, D. L., Zhao, Z. Z., Sturm, R. A., Hayward, N. K., Martin, N. G. & Montgomery, G. W. 2010b. Multiple pigmentation gene polymorphisms account for a substantial proportion of risk of cutaneous malignant melanoma. *J Invest Dermatol,* 130**,** 520-8.

Ehret, G. B., Lamparter, D., Hoggart, C. J., Genetic Investigation Of Anthropometric Traits, C., Whittaker, J. C., Beckmann, J. S. & Kutalik, Z. 2012. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am J Hum Genet,* 91**,** 863-71.

Ellerhorst, J. A., Greene, V. R., Ekmekcioglu, S., Warneke, C. L., Johnson, M. M., Cooke, C. P., Wang, L. E., Prieto, V. G., Gershenwald, J. E., Wei, Q. & Grimm, E. A. 2011. Clinical correlates of NRAS and BRAF mutations in primary human melanoma. *Clin Cancer Res,* 17**,** 229-35.

Erdmann, F., Lortet-Tieulent, J., Schüz, J., Zeeb, H., Greinert, R., Breitbart, E. W. & Bray, F. 2013. International trends in the incidence of malignant melanoma 1953–2008—are recent generations at higher or lower risk? *International Journal of Cancer,* 132**,** 385-400.

Falchi, M., Bataille, V., Hayward, N. K., Duffy, D. L., Bishop, J. A., Pastinen, T., Cervino, A., Zhao, Z. Z., Deloukas, P., Soranzo, N., Elder, D. E., Barrett, J. H., Martin, N. G., Bishop, D. T., Montgomery, G. W. & Spector, T. D. 2009. Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. *Nat Genet,* 41**,** 915-9.

Ferlay, J., Steliarova-Foucher E Fau - Lortet-Tieulent, J., Lortet-Tieulent J Fau - Rosso, S., Rosso S Fau - Coebergh, J. W. W., Coebergh Jw Fau - Comber, H., Comber H Fau - Forman, D., Forman D Fau - Bray, F. & Bray, F. 2013. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer,* 49(6)**,** 1374-1403.

Fleet, J. C., Desmet, M., Johnson, R. & Li, Y. 2012. Vitamin D and cancer: a review of molecular mechanisms. *Biochem J,* 441**,** 61-76.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Abeni, D., Boyle, P. & Melchi, C. F. 2005a. Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi. *Eur J Cancer,* 41**,** 28-44.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Zanetti, R., Masini, C., Boyle, P. & Melchi, C. F. 2005b. Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors. *Eur J Cancer,* 41**,** 2040-59.

Garbe, C., Peris, K., Hauschild, A., Saiag, P., Middleton, M., Spatz, A., Grob, J. J., Malvehy, J., Newton-Bishop, J., Stratigos, A., Pehamberger, H. & Eggermont, A. 2010. Diagnosis and treatment of melanoma: European consensus-based interdisciplinary guideline. *Eur J Cancer,* 46**,** 270-83.

Gentles, A. J., Bratman, S. V., Lee, L. J., Harris, J. P., Feng, W., Nair, R. V., Shultz, D. B., Nair, V. S., Hoang, C. D., West, R. B., Plevritis, S. K., Alizadeh, A. A. & Diehn, M. 2015. Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non-Small Cell Lung Cancer. *J Natl Cancer Inst,* 107.

Gerami, P., Cook, R. W., Wilkinson, J., Russell, M. C., Dhillon, N., Amaria, R. N., Gonzalez, R., Lyle, S., Johnson, C. E., Oelschlager, K. M., Jackson, G. L., Greisinger, A. J., Maetzold, D., Delman, K. A., Lawson, D. H. & Stone, J. F. 2015. Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma. *Clin Cancer Res,* 21**,** 175-83.

Gibbons, F. D. & Roth, F. P. 2002. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res,* 12**,** 1574-81.

Gimotty, P. A., Guerry, D., Ming, M. E., Elenitsas, R., Xu, X., Czerniecki, B., Spitz, F., Schuchter, L. & Elder, D. 2004. Thin primary cutaneous malignant melanoma: a prognostic tree for 10-year metastasis is more accurate than American Joint Committee on Cancer staging. *J Clin Oncol,* 22**,** 3668-76.

Grundberg, E., Small, K. S., Hedman, A. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T. P., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, S. Y., Glass, D., Travers, M., Min, J. L., Ring, S., Ho, K., Thorleifsson, G., Kong, A., Thorsteindottir, U., Ainali, C., Dimas, A. S., Hassanali, N., Ingle, C., Knowles, D., Krestyaninova, M., Lowe, C. E., Di Meglio, P., Montgomery, S. B., Parts, L., Potter, S., Surdulescu, G., Tsaprouni, L., Tsoka, S., Bataille, V., Durbin, R., Nestle, F. O., O'rahilly, S., Soranzo, N., Lindgren, C. M., Zondervan, K. T., Ahmadi, K. R., Schadt, E. E., Stefansson, K., Smith, G. D., Mccarthy, M. I., Deloukas, P., Dermitzakis, E. T., Spector, T. D. & Multiple Tissue Human Expression Resource, C. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet,* 44**,** 1084-9.

Guitart, J., Lowe, L., Piepkorn, M., Prieto, V. G., Rabkin, M. S., Ronan, S. G., Shea, C. R., Tron, V. A., White, W. & Barnhill, R. L. 2002. Histological characteristics of metastasizing thin melanomas: a case-control study of 43 cases. *Arch Dermatol,* 138**,** 603-8.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppala, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P. & Pasaniuc, B. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet,* 48**,** 245-52.

Haqq, C., Nosrati, M., Sudilovsky, D., Crothers, J., Khodabakhsh, D., Pulliam, B. L., Federman, S., Miller, J. R., 3rd, Allen, R. E., Singer, M. I., Leong, S. P., Ljung, B. M., Sagebiel, R. W. & Kashani-Sabet, M. 2005. The gene expression signatures of melanoma progression. *Proc Natl Acad Sci U S A,* 102**,** 6092-7.

Harbst, K., Staaf, J., Lauss, M., Karlsson, A., Masback, A., Johansson, I., Bendahl, P. O., Vallon-Christersson, J., Torngren, T., Ekedahl, H., Geisler, J., Hoglund, M., Ringner, M., Lundgren, L., Jirstrom, K., Olsson, H., Ingvar, C., Borg, A., Tsao, H. & Jonsson, G. 2012. Molecular profiling reveals low- and high-grade forms of primary melanoma. *Clin Cancer Res,* 18**,** 4026-36.

Harrell, F. E., Jr. 2001. *Regression Modelling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer.

Harrell, F. E., Jr., Lee, K. L. & Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med,* 15**,** 361-87.

Hastie, T., Tibshirani, R. & Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

Hosmer, J. D. W. & Lemeshow, S. 1999. *Applied Survival Analysis: Regression Modelling of Time to Event Data*, Wiley.

Howie, B. N., Donnelly, P. & Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet,* 5**,** e1000529.

Huang, Y. T., Liang, L., Moffatt, M. F., Cookson, W. O. & Lin, X. 2015. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genet Epidemiol,* 39**,** 347-56.

Hussussian, C. J., Struewing, J. P., Goldstein, A. M., Higgins, P. A., Ally, D. S., Sheahan, M. D., Clark, W. H., Jr., Tucker, M. A. & Dracopoli, N. C. 1994. Germline p16 mutations in familial melanoma. *Nat Genet,* 8**,** 15-21.

Iles, M. M., Law, M. H., Stacey, S. N., Han, J., Fang, S., Pfeiffer, R., Harland, M., Macgregor, S., Taylor, J. C., Aben, K. K., Akslen, L. A., Avril, M. F., Azizi, E., Bakker, B., Benediktsdottir, K. R., Bergman, W., Scarra, G. B., Brown, K. M., Calista, D., Chaudru, V., Fargnoli, M. C., Cust, A. E., Demenais, F., De Waal, A. C., Debniak, T., Elder, D. E., Friedman, E., Galan, P., Ghiorzo, P., Gillanders, E. M., Goldstein, A. M., Gruis, N. A., Hansson, J., Helsing, P., Hocevar, M., Hoiom, V., Hopper, J. L., Ingvar, C., Janssen, M., Jenkins, M. A., Kanetsky, P. A., Kiemeney, L. A., Lang, J., Lathrop, G. M., Leachman, S., Lee, J. E., Lubinski, J., Mackie, R. M., Mann, G. J., Martin, N. G., Mayordomo, J. I., Molven, A., Mulder, S., Nagore, E., Novakovic, S., Okamoto, I., Olafsson, J. H., Olsson, H., Pehamberger, H., Peris, K., Grasa, M. P., Planelles, D., Puig, S., Puig-Butille, J. A., Randerson-Moor, J., Requena, C., Rivoltini, L., Rodolfo, M., Santinami, M., Sigurgeirsson, B., Snowden, H., Song, F., Sulem, P., Thorisdottir, K., Tuominen, R., Van Belle, P., Van Der Stoep, N., Van Rossum, M. M., Wei, Q., Wendt, J., Zelenika, D., Zhang, M., Landi, M. T., Thorleifsson, G., Bishop, D. T., Amos, C. I., Hayward, N. K., Stefansson, K., Bishop, J. A., Barrett, J. H., Geno, M. E. L. C., Q, M. & Investigators, A. 2013. A variant in FTO shows association with melanoma risk not due to BMI. *Nat Genet,* 45**,** 428-32, 432e1.

Jaeger, J., Koczan, D., Thiesen, H. J., Ibrahim, S. M., Gross, G., Spang, R. & Kunz, M. 2007. Gene expression signatures for tumor progression, tumor subtype, and tumor thickness in laser-microdissected melanoma tissues. *Clin Cancer Res,* 13**,** 806-15.

Jayawardana, K., Schramm, S. J., Tembe, V., Mueller, S., Thompson, J. F., Scolyer, R. A., Mann, G. J. & Yang, J. Y. 2015. Identification, Review and Systematic Cross-Validation of MicroRNA Prognostic Signatures in Metastatic Melanoma. *J Invest Dermatol*.

Jewell, R., Conway, C., Mitra, A., Randerson-Moor, J., Lobo, S., Nsengimana, J., Harland, M., Marples, M., Edward, S., Cook, M., Powell, B., Boon, A., De Kort, F., Parker, K. A., Cree, I. A., Barrett, J. H., Knowles, M. A., Bishop, D. T. & Newton-Bishop, J. 2010. Patterns of expression of DNA repair genes and relapse from melanoma. *Clin Cancer Res,* 16**,** 5211-21.

Jiang, D., Tang, C. & Zhang, A. 2004. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Trans. on Knowl. and Data Eng.,* 16**,** 1370-1386.

John, T., Black, M. A., Toro, T. T., Leader, D., Gedye, C. A., Davis, I. D., Guilford, P. J. & Cebon, J. S. 2008. Predicting clinical outcome through molecular profiling in stage III melanoma. *Clin Cancer Res,* 14**,** 5173-80.

Johnson, D. B., Lovly, C. M., Flavin, M., Panageas, K. S., Ayers, G. D., Zhao, Z., Iams, W. T., Colgan, M., Denoble, S., Terry, C. R., Berry, E. G., Iafrate, A. J., Sullivan, R. J., Carvajal, R. D. & Sosman, J. A. 2015. Impact of NRAS mutations for patients with advanced melanoma treated with immune therapies. *Cancer Immunol Res,* 3**,** 288-95.

Jonsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringner, M., Lillehaug, J. R., Borg, A. & Lonning, P. E. 2010. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin Cancer Res,* 16**,** 3356-67.

Kamb, A., Shattuck-Eidens, D., Eeles, R., Liu, Q., Gruis, N. A., Ding, W., Hussey, C., Tran, T., Miki, Y., Weaver-Feldhaus, J. & *et al*. 1994. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat Genet,* 8**,** 23-6.

Kashani-Sabet, M., Sagebiel, R. W., Ferreira, C. M., Nosrati, M. & Miller, J. R., 3RD 2002. Tumor vascularity in the prognostic assessment of primary cutaneous melanoma. *J Clin Oncol,* 20**,** 1826-31.

Law, M. H., Bishop, D. T., Lee, J. E., Brossard, M., Martin, N. G., Moses, E. K., Song, F., Barrett, J. H., Kumar, R., Easton, D. F., Pharoah, P. D., Swerdlow, A. J., Kypreou, K. P., Taylor, J. C., Harland, M., Randerson-Moor, J., Akslen, L. A., Andresen, P. A., Avril, M. F., Azizi, E., Scarra, G. B., Brown, K. M., Debniak, T., Duffy, D. L., Elder, D. E., Fang, S., Friedman, E., Galan, P., Ghiorzo, P., Gillanders, E. M., Goldstein, A. M., Gruis, N. A., Hansson, J., Helsing, P., Hocevar, M., Hoiom, V., Ingvar, C., Kanetsky, P. A., Chen, W. V., Geno, M. E. L. C., Essen-Heidelberg, I., Group, S. D. H. S., Q, M., Investigators, Q., Investigators, A., Group, A. M. S., Landi, M. T., Lang, J., Lathrop, G. M., Lubinski, J., Mackie, R. M., Mann, G. J., Molven, A., Montgomery, G. W., Novakovic, S., Olsson, H., Puig, S., Puig-Butille, J. A., Qureshi, A. A., Radford-Smith, G. L., Van Der Stoep, N., Van Doorn, R., Whiteman, D. C., Craig, J. E., Schadendorf, D., Simms, L. A., Burdon, K. P., Nyholt, D. R., Pooley, K. A., Orr, N., Stratigos, A. J., Cust, A. E., Ward, S. V., Hayward, N. K., Han, J., Schulze, H. J., Dunning, A. M., Bishop, J. A., Demenais, F., Amos, C. I., Macgregor, S. & Iles, M. M. 2015a. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet,* 47**,** 987-95.

Law, M. H., Rowe, C. J., Montgomery, G. W., Hayward, N. K., Macgregor, S. & Khosrotehrani, K. 2015b. PARP1 polymorphisms play opposing roles in melanoma occurrence and survival. *Int J Cancer,* 136**,** 2488-9.

Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. 2011. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet,* 88**,** 294-305.

Leiter, U., Buettner, P. G., Eigentler, T. K. & Garbe, C. 2004. Prognostic factors of thin cutaneous melanoma: an analysis of the central malignant melanoma registry of the german dermatological society. *J Clin Oncol,* 22**,** 3660-7.

Li, C., Yin, M., Wang, L. E., Amos, C. I., Zhu, D., Lee, J. E., Gershenwald, J. E., Grimm, E. A. & Wei, Q. 2013. Polymorphisms of nucleotide excision repair genes predict melanoma survival. *J Invest Dermatol,* 133**,** 1813-21.

Lindholm, C., Andersson, R., Dufmats, M., Hansson, J., Ingvar, C., Moller, T., Sjodin, H., Stierner, U., Wagenius, G. & Swedish Melanoma Study, G. 2004. Invasive cutaneous malignant melanoma in Sweden, 1990-1999. A prospective, population-based study of survival and prognostic factors. *Cancer,* 101**,** 2067-78.

Liu, W., Dowling, J. P., Murray, W. K., Mcarthur, G. A., Thompson, J. F., Wolfe, R. & Kelly, J. W. 2006. Rate of growth in melanomas: characteristics and associations of rapidly growing melanomas. *Arch Dermatol,* 142**,** 1551-8.

Llewellyn, C. H., Trzaskowski, M., Plomin, R. & Wardle, J. 2013. Finding the missing heritability in pediatric obesity: the contribution of genome-wide complex trait analysis. *Int J Obes (Lond),* 37**,** 1506-9.

Macgregor, S., Montgomery, G. W., Liu, J. Z., Zhao, Z. Z., Henders, A. K., Stark, M., Schmid, H., Holland, E. A., Duffy, D. L., Zhang, M., Painter, J. N., Nyholt, D. R., Maskiell, J. A., Jetann, J., Ferguson, M., Cust, A. E., Jenkins, M. A., Whiteman, D. C., Olsson, H., Puig, S., Bianchi-Scarra, G., Hansson, J., Demenais, F., Landi, M. T., Debniak, T., Mackie, R., Azizi, E., Bressac-De Paillerets, B., Goldstein, A. M., Kanetsky, P. A., Gruis, N. A., Elder, D. E., Newton-Bishop, J. A., Bishop, D. T., Iles, M. M., Helsing, P., Amos, C. I., Wei, Q., Wang, L. E., Lee, J. E., Qureshi, A. A., Kefford, R. F., Giles, G. G., Armstrong, B. K., Aitken, J. F., Han, J., Hopper, J. L., Trent, J. M., Brown, K. M., Martin, N. G., Mann, G. J. & Hayward, N. K. 2011. Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. *Nat Genet,* 43**,** 1114-8.

Maldonado, J. L., Fridlyand, J., Patel, H., Jain, A. N., Busam, K., Kageshita, T., Ono, T., Albertson, D. G., Pinkel, D. & Bastian, B. C. 2003. Determinants of BRAF mutations in primary melanomas. *J Natl Cancer Inst,* 95**,** 1878-90.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M. C., Chiarion-Sileni, V., Mocellin, S., Rossi, C. R., Bicciato, S., Wang, E., Marincola, F. M. & Zanovello, P. 2006. A gene expression signature associated with survival in metastatic melanoma. *J Transl Med,* 4**,** 50.

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A. & Sander, C. 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One,* 6**,** e24709.

Mann, G. J., Pupo, G. M., Campain, A. E., Carter, C. D., Schramm, S. J., Pianova, S., Gerega, S. K., De Silva, C., Lai, K., Wilmott, J. S., Synnott, M., Hersey, P., Kefford, R. F., Thompson, J. F., Yang, Y. H. & Scolyer, R. A. 2013. BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J Invest Dermatol,* 133**,** 509-17.

Marchini, J. & Howie, B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet,* 11**,** 499-511.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. & Cheung, V. G. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature,* 430**,** 743-7.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C. M., Muller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott, L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C. T., Rybin, D., Couper, D. J., Kao, W. H., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., Van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R., Platou, C. G., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stancakova, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burtt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutskov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Mannisto, S., *et al*. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet,* 44**,** 981-90.

Nagore, E., Oliver, V., Botella-Estrada, R., Moreno-Picot, S., Guillen, C. & Fortea, J. M. 2006. Clinicopathological analysis of 1571 cutaneous malignant melanomas in Valencia, Spain: factors related to tumour thickness. *Acta Derm Venereol,* 86**,** 50-6.

Newton-Bishop, J. A., Beswick, S., Randerson-Moor, J., Chang, Y. M., Affleck, P., Elliott, F., Chan, M., Leake, S., Karpavicius, B., Haynes, S., Kukalizch, K., Whitaker, L., Jackson, S., Gerry, E., Nolan, C., Bertram, C., Marsden, J., Elder, D. E., Barrett, J. H. & Bishop, D. T. 2009. Serum 25-hydroxyvitamin D3 levels are associated with breslow thickness at presentation and survival from melanoma. *J Clin Oncol,* 27**,** 5439-44.

Nguyen, D. V., Arpat, A. B., Wang, N. & Carroll, R. J. 2002. DNA microarray experiments: biological and technological aspects. *Biometrics,* 58**,** 701-17.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, A. K., Bataille, V., Tzenova Bell, J., Surdulescu, G., Dimas, A. S., Ingle, C., Nestle, F. O., Di Meglio, P., Min, J. L., Wilk, A., Hammond, C. J., Hassanali, N., Yang, T. P., Montgomery, S. B., O'rahilly, S., Lindgren, C. M., Zondervan, K. T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., Mccarthy, M. I., Dermitzakis, E. T., Spector, T. D. & Mu, T. C. 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet,* 7**,** e1002003.

Nsengimana, J., Laye, J., Filia, A., Walker, C., Jewell, R., Van Den Oord, J. J., Wolter, P., Patel, P., Sucker, A., Schadendorf, D., Jonsson, G. B., Bishop, D. T. & Newton-Bishop, J. 2015. Independent replication of a melanoma subtype gene signature and evaluation of its prognostic value and biological correlates in a population cohort. *Oncotarget,* 6**,** 11683-93.

Nurnberg, B., Graber, S., Gartner, B., Geisel, J., Pfohler, C., Schadendorf, D., Tilgen, W. & Reichrath, J. 2009. Reduced serum 25-hydroxyvitamin D levels in stage IV melanoma patients. *Anticancer Res,* 29**,** 3669-74.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S. & Aittokallio, T. 2014. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet,* 10**,** e1004754.

Olsen, C. M., Carroll, H. J. & Whiteman, D. C. 2010. Familial melanoma: a meta-analysis and estimates of attributable fraction. *Cancer Epidemiol Biomarkers Prev,* 19**,** 65-73.

Ongen, H., Andersen, C. L., Bramsen, J. B., Oster, B., Rasmussen, M. H., Ferreira, P. G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., Padioleau, I., Bielser, D., Romano, L., Tomlinson, I., Houlston, R. S., Esteller, M., Orntoft, T. F. & Dermitzakis, E. T. 2014. Putative cis-regulatory drivers in colorectal cancer. *Nature,* 512**,** 87-90.

Orlow, I., Reiner, A. S., Thomas, N. E., Roy, P., Kanetsky, P. A., Luo, L., Paine, S., Armstrong, B. K., Kricker, A., Marrett, L. D., Rosso, S., Zanetti, R., Gruber, S. B., Anton-Culver, H., Gallagher, R. P., Dwyer, T., Busam, K., Begg, C. B., Berwick, M. & Group, G. E. M. S. 2016. Vitamin D receptor polymorphisms and survival in patients with cutaneous melanoma: a population-based study. *Carcinogenesis,* 37**,** 30-8.

Park, J. Y., Amankwah, E. K., Anic, G. M., Lin, H. Y., Walls, B., Park, H., Krebs, K., Madden, M., Maddox, K., Marzban, S., Fang, S., Chen, W., Lee, J. E., Wei, Q., Amos, C. I., Messina, J. L., Sondak, V. K., Sellers, T. A. & Egan, K. M. 2013. Gene variants in angiogenesis and lymphangiogenesis and cutaneous melanoma progression. *Cancer Epidemiol Biomarkers Prev,* 22**,** 827-34.

Park, M. Y. & Hastie, T. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B,* 69**,** 659-677.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R. & West, M. 2004. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A,* 101**,** 8431-6.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. & Sham, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S., Aben, K. K., Blondal, T., Thorgeirsson, T. E., Thorleifsson, G., Kristjansson, K., Thorisdottir, K., Ragnarsson, R., Sigurgeirsson, B., Skuladottir, H., Gudbjartsson, T., Isaksson, H. J., Einarsson, G. V., Benediktsdottir, K. R., Agnarsson, B. A., Olafsson, K., Salvarsdottir, A., Bjarnason, H., Asgeirsdottir, M., Kristinsson, K. T., Matthiasdottir, S., Sveinsdottir, S. G., Polidoro, S., Hoiom, V., Botella-Estrada, R., Hemminki, K., Rudnai, P., Bishop, D. T., Campagna, M., Kellen, E., Zeegers, M. P., De Verdier, P., Ferrer, A., Isla, D., Vidal, M. J., Andres, R., Saez, B., Juberias, P., Banzo, J., Navarrete, S., Tres, A., Kan, D., Lindblom, A., Gurzau, E., Koppova, K., De Vegt, F., Schalken, J. A., Van Der Heijden, H. F., Smit, H. J., Termeer, R. A., Oosterwijk, E., Van Hooij, O., Nagore, E., Porru, S., Steineck, G., Hansson, J., Buntinx, F., Catalona, W. J., Matullo, G., Vineis, P., Kiltie, A. E., Mayordomo, J. I., Kumar, R., Kiemeney, L. A., Frigge, M. L., Jonsson, T., Saemundsson, H., Barkardottir, R. B., Jonsson, E., Jonsson, S., Olafsson, J. H., Gulcher, J. R., Masson, G., Gudbjartsson, D. F., Kong, A., Thorsteinsdottir, U. & Stefansson, K. 2009. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet,* 41**,** 221-7.

Rangel, J., Nosrati, M., Torabian, S., Shaikh, L., Leong, S. P., Haqq, C., Miller, J. R., 3rd, Sagebiel, R. W. & Kashani-Sabet, M. 2008. Osteopontin as a molecular prognostic marker for melanoma. *Cancer,* 112**,** 144-50.

Rendleman, J., Fau - Shang, S., Shang S Fau - Dominianni, C., Dominianni C Fau - Shields, J. F., Shields Jf Fau - Scanlon, P., Scanlon P Fau - Adaniel, C., Adaniel C Fau - Desrichard, A., Desrichard A Fau - Ma, M., Ma M Fau - Shapiro, R., Shapiro R Fau - Berman, R., Berman R Fau - Pavlick, A., Pavlick A Fau - Polsky, D., Polsky D Fau - Shao, Y., Shao Y Fau - Osman, I., Osman I Fau - Kirchhoff, T. & Kirchhoff, T. 2013. Melanoma risk loci as determinants of melanoma recurrence and survival.

Riker, A. I., Enkemann, S. A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y., Howell, P., Metge, B., Samant, R. S., Shevde, L. A., Li, W., Eschrich, S., Daud, A., Ju, J. & Matta, J. 2008. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med Genomics,* 1**,** 13.

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., Hemingway, H. & Group, P. 2013. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med,* 10**,** e1001380.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet,* 16**,** 85-97.

Rockman, M. V. & Kruglyak, L. 2006. Genetics of global gene expression. *Nat Rev Genet,* 7**,** 862-72.

Roh, M. R., Gupta, S., Park, K. H., Chung, K. Y., Lauss, M., Flaherty, K. T., Jonsson, G., Rha, S. Y. & Tsao, H. 2016. Promoter Methylation of PTEN Is a Significant Prognostic Factor in Melanoma Survival. *J Invest Dermatol,* 136**,** 1002-11.

Saldanha, G., Elshaw, S., Sachs, P., Alharbi, H., Shah, P., Jothi, A. & Pringle, J. H. 2016. microRNA-10b is a prognostic biomarker for melanoma. *Mod Pathol,* 29**,** 112-21.

Schramm, K., Marzi, C., Schurmann, C., Carstensen, M., Reinmaa, E., Biffar, R., Eckstein, G., Gieger, C., Grabe, H. J., Homuth, G., Kastenmuller, G., Magi, R., Metspalu, A., Mihailov, E., Peters, A., Petersmann, A., Roden, M., Strauch, K., Suhre, K., Teumer, A., Volker, U., Volzke, H., Wang-Sattler, R., Waldenberger, M., Meitinger, T., Illig, T., Herder, C., Grallert, H. & Prokisch, H. 2014. Mapping the genetic architecture of gene regulation in whole blood. *PLoS One,* 9**,** e93844.

Schramm, S. J., Campain, A. E., Scolyer, R. A., Yang, Y. H. & Mann, G. J. 2012. Review and cross-validation of gene expression signatures and melanoma prognosis. *J Invest Dermatol,* 132**,** 274-83.

Schumacher, M., Binder, H. & Gerds, T. 2007. Assessment of survival prediction models based on microarray data. *Bioinformatics,* 23**,** 1768-74.

Shaw, H. M., Rivers, J. K., Mccarthy, S. W. & Mccarthy, W. H. 1992. Cutaneous melanomas exhibiting unusual biologic behavior. *World J Surg,* 16**,** 196-202.

Shi, H., Zhou, Y., Liu, H., Chen, C., Li, S., Li, N., Li, X., Zhang, X., Zhang, H., Wang, W. & Zhao, Q. 2010. Expression of CIAPIN1 in human colorectal cancer and its correlation with prognosis. *BMC Cancer,* 10**,** 477.

Shinozaki, M., Fujimoto, A., Morton, D. L. & Hoon, D. S. 2004. Incidence of BRAF oncogene mutation and clinical relevance for primary cutaneous melanomas. *Clin Cancer Res,* 10**,** 1753-7.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., Luciano, M., Martin, N. G., Mortensen, J., Nistico, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G. & Kaprio, J. 2003. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res,* 6**,** 399-408.

Sivendran, S., Chang, R., Pham, L., Phelps, R. G., Harcharik, S. T., Hall, L. D., Bernardo, S. G., Moskalenko, M. M., Sivendran, M., Fu, Y., De Moll, E. H., Pan, M., Moon, J. Y., Arora, S., Cohain, A., Difeo, A., Ferringer, T. C., Tismenetsky, M., Tsui, C. L., Friedlander, P. A., Parides, M. K., Banchereau, J., Chaussabel, D., Lebwohl, M. G., Wolchok, J. D., Bhardwaj, N., Burakoff, S. J., Oh, W. K., Palucka, K., Merad, M., Schadt, E. E. & Saenger, Y. M. 2014. Dissection of immune gene networks in primary melanoma tumors critical for antitumor surveillance of patients with stage II-III resectable disease. *J Invest Dermatol,* 134**,** 2202-11.

Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. 2012. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet,* 91**,** 1011-21.

Steyerberg, E. W., Moons, K. G., Van Der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G. & Group, P. 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med,* 10**,** e1001381.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., Tavare, S., Deloukas, P. & Dermitzakis, E. T. 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet,* 1**,** e78.

Swick, J. M. & Maize, J. C., Sr. 2012. Molecular biology of melanoma. *J Am Acad Dermatol,* 67**,** 1049-54.

Taylor, N. J., Reiner, A. S., Begg, C. B., Cust, A. E., Busam, K. J., Anton-Culver, H., Dwyer, T., From, L., Gallagher, R. P., Gruber, S. B., Rosso, S., White, K. A., Zanetti, R., Orlow, I., Thomas, N. E., Rebbeck, T. R., Berwick, M., Kanetsky, P. A. & Group, G. E. M. S. 2015a. Inherited variation at MC1R and ASIP and association with melanoma-specific survival. *Int J Cancer,* 136**,** 2659-67.

Taylor, P. N., Porcu, E., Chew, S., Campbell, P. J., Traglia, M., Brown, S. J., Mullin, B. H., Shihab, H. A., Min, J., Walter, K., Memari, Y., Huang, J., Barnes, M. R., Beilby, J. P., Charoen, P., Danecek, P., Dudbridge, F., Forgetta, V., Greenwood, C., Grundberg, E., Johnson, A. D., Hui, J., Lim, E. M., Mccarthy, S., Muddyman, D., Panicker, V., Perry, J. R., Bell,

J. T., Yuan, W., Relton, C., Gaunt, T., Schlessinger, D., Abecasis, G., Cucca, F., Surdulescu, G. L., Woltersdorf, W., Zeggini, E., Zheng, H. F., Toniolo, D., Dayan, C. M., Naitza, S., Walsh, J. P., Spector, T., Davey Smith, G., Durbin, R., Richards, J. B., Sanna, S., Soranzo, N., Timpson, N. J., Wilson, S. G. & Consortium, U. K. 2015b. Whole-genome sequence-based analysis of thyroid function. *Nat Commun,* 6**,** 5681.

Thomas, D. C. 2004. *Statistical Methods in Genetic Epidemiology*, Oxford University Press.

Thomas, N. E., Busam, K. J., From, L., Kricker, A., Armstrong, B. K., Anton-Culver, H., Gruber, S. B., Gallagher, R. P., Zanetti, R., Rosso, S., Dwyer, T., Venn, A., Kanetsky, P. A., Groben, P. A., Hao, H., Orlow, I., Reiner, A. S., Luo, L., Paine, S., Ollila, D. W., Wilcox, H., Begg, C. B. & Berwick, M. 2013. Tumor-infiltrating lymphocyte grade in primary melanomas is independently associated with melanoma-specific survival in the population-based genes, environment and melanoma study. *J Clin Oncol,* 31**,** 4252-9.

Thomas, N. E., Edmiston, S. N., Alexander, A., Groben, P. A., Parrish, E., Kricker, A., Armstrong, B. K., Anton-Culver, H., Gruber, S. B., From, L., Busam, K. J., Hao, H., Orlow, I., Kanetsky, P. A., Luo, L., Reiner, A. S., Paine, S., Frank, J. S., Bramson, J. I., Marrett, L. D., Gallagher, R. P., Zanetti, R., Rosso, S., Dwyer, T., Cust, A. E., Ollila, D. W., Begg, C. B., Berwick, M., Conway, K. & Group, G. E. M. S. 2015. Association Between NRAS and BRAF Mutational Status and Melanoma-Specific Survival Among Patients With Higher-Risk Primary Melanoma. *JAMA Oncol,* 1**,** 359-68.

Thorn, M., Ponten, F., Bergstrom, R., Sparen, P. & Adami, H. O. 1994. Clinical and histopathologic predictors of survival in patients with malignant melanoma: a population-based study in Sweden. *J Natl Cancer Inst,* 86**,** 761-9.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B,* 58**,** 267-288.

Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. *Stat Med,* 16**,** 385-95.

Vazquez, A. I., Veturi, Y., Behring, M., Shrestha, S., Kirst, M., Resende, M. F., Jr. & De Los Campos, G. 2016. Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles. *Genetics,* 203**,** 1425-38.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., Hayes, D. N. & Cancer Genome Atlas Research, N. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* 17**,** 98-110.

Verweij, P. J. & Van Houwelingen, H. C. 1993. Cross-validation in survival analysis. *Stat Med,* 12**,** 2305-14.

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A., Chen, G. B., Lee, S. H., Wray, N. R., Goddard, M. E. & Yang, J. 2014. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet,* 10**,** e1004269.

Visscher, P. M., Hill, W. G. & Wray, N. R. 2008. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet,* 9**,** 255-66.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., Avril, M. F., Ortiz Romero, P. L., Robert, T., Balacescu, O., Eggermont, A. M., Lenoir, G., Sarasin, A., Tursz, T., Van Den Oord, J. J., Spatz, A., Melanoma Group Of The European Organization For, R. & Treatment Of, C. 2006. Gene expression profiling of primary cutaneous melanoma and clinical outcome. *J Natl Cancer Inst,* 98**,** 472-82.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T. H., D'ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., Huang, S., Kattenberg, M., Kochar, J., Middeldorp, C. M., Qu, A., Shabalin, A., Tischfield, J., Todd, L., Tzeng, J. Y., Van Grootheest, G., Vink, J. M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J. H., De Geus, E. J., Yin, Z., Penninx, B. W. & Boomsma, D. I. 2014. Heritability and genomics of gene expression in peripheral blood. *Nat Genet,* 46**,** 430-7.

Xu, X., Chen, L., Guerry, D., Dawson, P. R., Hwang, W. T., Vanbelle, P., Elder, D. E., Zhang, P. J., Ming, M. E., Schuchter, L. & Gimotty, P. A. 2012. Lymphatic invasion is independently prognostic of metastasis in primary cutaneous melanoma. *Clin Cancer Res,* 18**,** 229-37.

Yang, J., Benyamin, B., Mcevoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. & Visscher, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

Yin, J., Liu, H., Liu, Z., Wang, L. E., Chen, W. V., Zhu, D., Amos, C. I., Fang, S., Lee, J. E. & Wei, Q. 2015. Genetic variants in fanconi anemia pathway genes BRCA2 and FANCA predict melanoma survival. *J Invest Dermatol,* 135**,** 542-50.

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L. A., Xu, Y., Hess, K. R., Diao, L., Han, L., Huang, X., Lawrence, M. S., Weinstein, J. N., Stuart, J. M., Mills, G. B., Garraway, L. A., Margolin, A. A., Getz, G. & Liang, H. 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol,* 32**,** 644-52.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S. & Price, A. L. 2013. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet,* 9**,** e1003520.

Zhao, S., Kurenbekova, L., Gao, Y., Roos, A., Creighton, C. J., Rao, P., Hicks, J., Man, T. K., Lau, C., Brown, A. M., Jones, S. N., Lazar, A. J., Ingram, D., Lev, D., Donehower, L. A. & Yustein, J. T. 2015. NKD2, a negative regulator of Wnt signaling, suppresses tumor growth and metastasis in osteosarcoma. *Oncogene,* 34**,** 5069-79.

Zuo, L., Weger, J., Yang, Q., Goldstein, A. M., Tucker, M. A., Walker, G. J., Hayward, N. & Dracopoli, N. C. 1996. Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma. *Nat Genet,* 12**,** 97-9.