ANALYSIS OF AN IMMUNE FOCUSED TARGETED GENETIC ASSOCIATION STUDY IN INTERMEDIATE-RISK MELANOMA

by

Ying Qian

BS in Biological Science, Nanjing Agricultural University, China, 2013

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Ying Qian

It was defended on

July 6, 2015

and approved by

Ada Youk, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Ahmad Tarhini, MD, PhD, Associate Professor, Department of Medicine
School of Medicine, University of Pittsburgh

Thesis Advisor: Yan Lin, PhD, Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Yan Lin, PhD

ANALYSIS OF AN IMMUNE FOCUSED TARGETED GENETIC ASSOCIATION STUDY IN INTERMEDIATE-RISK MELANOMA

Ying Qian, MS

University of Pittsburgh, 2015

ABSTRACT

The Centers for Disease Control and Prevention (CDC) estimates that about 8,000 deaths in the United States are caused by melanoma skin cancer each year. Melanoma has become the most lethal skin cancer over the past three decades. Immunotherapies were introduced to Melanoma patients in the 60's, and Interferon Alpha (IFN α) is one of the mostly used drugs for immunotherapy. Previous studies showed that using IFN α-2b might increase the survival rate of patients with high-risk melanoma skin cancer. However, not all patients respond to immunotherapies. So ECOG 1697 (E1697) trial was performed to compare the effect of patients obtained four-week high-dose IFN-a2b and the control group. This project utilizes a subset of the E1697 patients to search for potential immune-related genes that are associated with the prognosis of patients with localized melanoma. Both SNP and gene level analysis were conducted. This study has important public health significance because it identifies genetic factors associated with prognosis of local melanoma, which may be used to guide the treatment of this subgroup of melanoma patients in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to give my sincere thanks to my thesis advisor, Dr. Yan Lin, for giving me the opportunity to participate in this project, and for patiently providing me guidance in learning new things and completing this thesis. I would like to thank Dr. Ahmad Tarhini for providing us this wonderful project, and all the support for my thesis. Also, a special thanks to Dr. Ada Youk who spent valuable time checking the thesis draft, and provided valuable comments and suggestions.

1.0 INTRODUCTION

1.1 MELANOMA

1.1.1 General Introduction

According to the American Cancer Society, of all cancers, skin cancer is by far the most common one. Melanoma is the deadliest type of skin cancer. Over the past three decades, melanoma has the fastest growth of incidence rate among all skin cancers. The Centers for Disease Control and Prevention (CDC) estimates that about 8,000 deaths in the United States are caused by melanoma skin cancer each year (Plescia, Protzel Berman, & White, 2011). The American Cancer Society estimates that in 2015, over 73,000 new melanoma cases will be diagnosed, and nearly 10,000 people are expected to die from it in the United States (American Cancer Society. Cancer Facts & Figures 2015).

Melanoma incidence is higher in whites than in blacks and Asians, and increases as people age. However, it is also one of the most common cancers in young adults (Bleyer, O'leary, Barr, & Ries, 2006), especially young women. Ultraviolet (UV) light exposure is a major risk factor for most melanomas (Parkin, Mesher, & Sasieni, 2011). Other known risk factors include large numbers of moles, fair skin, family or personal history of skin cancers, and a weakened immune system. Signs of melanoma typically seen include a new spot on the skin, a spot that is

changing in size, shape, or color, and a spot that looks different from all of the other spots on skin (known as the ugly duckling sign).

1.1.2 Immunotherapy

The treatments of melanoma include surgery, immunotherapy, targeted therapy, chemotherapy, and radiation therapy. Early-stage melanomas are often treated with surgery, but late-stage melanomas require advanced treatments after surgery. These advanced melanomas are difficult to treat with radiation and chemotherapy. Over the past few years, melanoma treatment is gradually transformed from the traditional chemotherapy and radiation therapy to immunotherapy and targeted therapy.

The human immune system is a collection of organs, special cells, and substances that play a protective role from infections and other diseases. Immune response has a strong impact on melanoma prognosis (Herrera-Gonzalez, 2013). Immunotherapies stimulate a patient's own immune system with medicines to recognize and destroy the melanoma cancer cells.

Immunotherapies were introduced to Melanoma patients in the 60's. One of the commonly used drugs for immunotherapy is Interferon Alpha (IFN α). Interferon is a man-made copy of human protein. It helps the immune system to fight viral infections. Interferon Alpha-2b (IFN α-2b) treatment is often given as a shot under the skin. Studies showed that using IFN α-2b might increase the survival rate of people with high-risk melanoma skin cancer (Kirkwood et al., 2004; Kirkwood et al., 1996).

1.1.3 Clinical Biomarker for Melanoma Patients on Immunotherapy

A biomarker usually refers to a measurable substance in the body that may be associated with the risk or prognosis of a certain disease. In melanoma immunotherapy, previous immune-based cancer therapies have found several serum biomarkers that may play potential prognostic or diagnostic roles for melanoma (Tartour et al., 1994; Wittke et al., 1999). However, these studies have not completely resolved the issue as how well the patients respond to immunotherapies. As a result, there is need to continue identifying immune biomarkers capable of predicting clinical responses (Disis, 2011).

1.2 E1697 STUDY

E1697 (ECOG 1697) is a randomized intergroup trial aimed to compare the effect of two treatment arms: (A) observations with no evidence of disease, (B) patients obtain four weeks high-dose IFN-a2b with no evidence of disease. The study was terminated for futility in Oct. 2010.

1.3 IMMUNOCHIP

Immunochip is a customized Illumina Infinium single-nucleotide polymorphism (SNP) microarray. It contains close to 200,000 genetic markers drawn from genomic regions possibly associated with one or more immune-mediated disease. Deep replication of meta-genome-wide

association studies (GWASs), and fine mapping of GWAS loci were the two major goals of Immunochip research (Parkes, Cortes, van Heel, & Brown, 2013).

Genetic association studies examine the association of genetic variants with a disease. Immunochip is a high-density SNP array that provides cost-effective genotyping of common and rare variants to fine-map the established immune-related loci. This is a powerful tool for immunogenetics gene mapping in identifying large numbers of genetic loci (Cortes & Brown, 2011).

1.4 GOAL OF THE STUDY

The effects of immunotherapies have been shown in previous studies on patients with melanoma skin cancer. However, not all patients respond to immunotherapies. This study utilizes a subset of the E1697 patients to search for potential immune-related genes that are associated with the prognosis of patients on either one-month high dose IFN α-2b arm or the observation arm. Our results will provide insights for the mechanism of how the patients' immune system affects the prognosis of melanoma and provide potential prognostic (and predictive) biomarkers for melanoma patients.

2.0 METHODS AND RESULTS

2.1 STUDY SAMPLE

This is a correlative study of E1697 (ECOG 1697), which is a phase III randomized trial to compare the efficacy of four weeks of treatment of high-dose IFN-a2b with the observation arm. The current analysis aimed to discover prognostic genetic markers of melanoma patients. The analysis set is a subset of data from E1697 trial, which contains 216 randomly selected subjects. Blood samples were obtained at the study entry, and Immunochip was used to genotype the patients.

2.2 DATA

2.2.1 Starting Files

The SAS file, *e1697_spore_29april15.sas7bdat*, is the clinical data I got for the subset of E1697 trail from the ECOG statistician, which contains the following variables:

Column1: case (case number: ranges from 15080 to 36000)

Column2: trtm (treatment: A=control group, B=4-week high-dose IFN-a2b group)

Column3: sex (1=male, 2=female)

Column4: BRSLW_THICKNESS (tumor Breslow's depth in millimeters)

Column5: CLARK_LVL (Clark's level)

Column6: LDH_RS (Lactate dehydrogenase value)

Column7: LDH_ULN (LDH upper limit of normal)

Column8: PIG (Pigmentation: 1= amelanotic, 2= melanotic, -1= unknown)

Column9: PS (ECOG Performance status)

Column10: ULCER_YN (Ulceration: 1=no, 2=yes, -1=unknown)

Column11: surv_y (survival years)

Column12: rfs (relapse free survival years)

Column13: rfs_ind (relapse free survival index: 1=event, 0=censored)

Column14: surv_s (survival index: 1=event, 0=censored)

Column15: age (age at diagnosis)

ImmunoChip_GeneAnnotation.csv, is a file with gene annotation information. It contains 197076 lines (SNPs) and 8 columns:

Column 1: Name (rs number for SNP identifier)

Column 2: Chr (Chromosome number)

Column 3: Coordinate

Column 4: GeneSymbol (abbreviation of gene name)

Column 5: GeneLocation

Column 6: ExonLocation

Column 7: CodingStatus

Column 8: AminoAcid1.AminoAcid2

2.2.2 PLINK

Plink was used to perform the Quality Control of the genotype data. Plink is an open-source command-line network connection tool written by Simon Tatham. It is a whole genome association analysis toolset for performing a range of basic, large-scale analyses (Purcell et al., 2007). The PLINK program and instructions can be found at http://pngu.mgh.harvard.edu/~purcell/plink/.

2.2.3 Binary Files

The original genotype data were in binary PED files. The BED file, *Mel_IC.bed*, held the actual genotype information. It was a compressed file, which cannot be viewed with a standard text editor as the FAM and BIM files. The FAM file, *Mel_IC.fam*, contained subject information. The first six columns of BED file are:

Column1: Family ID

Column2: Individual ID

Column3: Paternal ID

Column4: Maternal ID

Column5: Sex (1=male, 2=female)

Column6: Phenotype (-9=missing, 1=unaffected, 2=affected)

The BIM file, *Mel_IC.bim*, is an extended MAP file with two columns of allele names. The order of the columns are arranged as followed:

Column1: Chromosome

Column2: SNP Name

Column3: Cytogenetic Distance (in centimeter)

Column4: Physical Distance (bp)

Column5: Allele 1

Column6: Allele 2

2.2.4 Quality Control

In Genome-wide association studies (GWAS), the quality control (QC) procedure is a critical element to inspect and clean data by reducing both the number of individuals and the number of SNPs passed on to downstream analysis (Turner et al., 2011; Weale, 2010). Because hundreds of thousands of genotypes are generated in GWAS, the occurrence of unidentified genotyping error may lead to spurious results.

2.2.4.1 Relationship Check

Relationship check is used to identify and record discrepancies between pedigrees provided and relatedness inferred from the genotype data by estimating the coefficients of identity by descent (IBD) (Turner et al., 2011).

SNPs with minor allele frequency (MAF) < 0.05 (a total of 20658 SNPs) were removed given the very limited sample size of the study, because they tend to have poorly behaved test statistics.

```
./plink --bfile ../Mel_IC --maf 0.05 --genome --rel-check --genome-full
--min 0.05 --noweb --out Mel_IC_relationcheck
```

A list of heterozygous haploid genotypes was written to *Mel_IC_relationcheck.hh* file. Whole genome IBD information was written to *Mel_IC_relationcheck.genome* file.

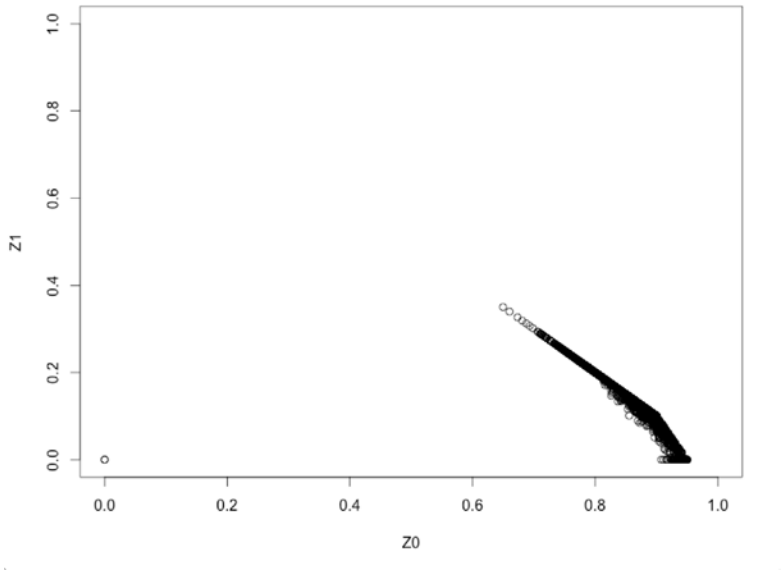


Figure 1. Relationship Check plot

Figure 1 is a plot showing the information of relative pairs of individuals. Z0 and Z1 denote the probability that individual1 and individual2 in a family share 0 or 1 allele at the marker locus. We expected to see all individual pairs on the diagonal. Figure 1 shows no specific pattern or weird points except for the unusual point near 0.00. This is consistent with the fact that all our subjects are not related to each other. The unusual point shares sample IDs as follows:

130624, 132789, 130777, 132879

2.2.4.2 Missing Data Check

We next checked the missing data by individual and by SNP.

```
./plink --bfile ../Mel_IC --missing --noweb --out Mel_IC_misscheck
./plink --bfile ../Mel_IC --het --noweb --out Mel_IC_misscheck
```

Through the first command line above, missing data information by individual was written to *Mel_IC_misscheck.imiss* file, and missing data information by locus was written to *Mel_IC_misscheck.lmiss* file. The second command line above wrote the individual

heterozygosity information to *Mel_IC_misscheck.het* file to check individuals with outlying heterozygosity rate.

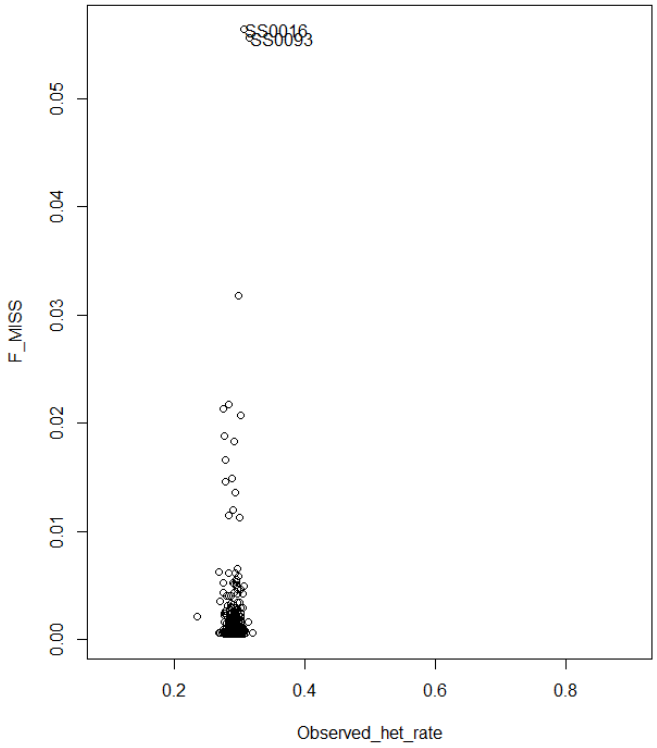


Figure 2. Missing data check by individual

The observed heterozygosity rate per individual is plotted on the x axis of Figure 2 and the proportion of missing SNPs per individuals is plotted on the y axis. Figure 2 indicated two samples (SS0016, SS0093) with high missing rate (proportion of sample missing > 0.05) at the top of the plot.

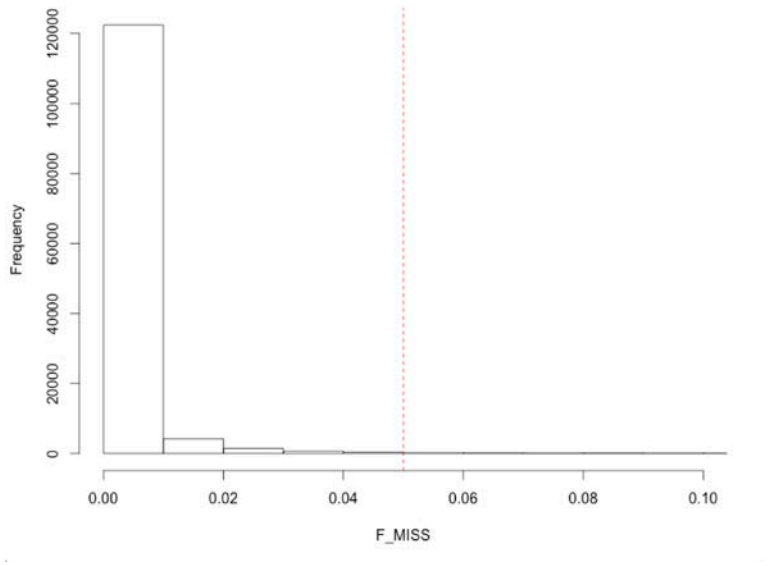


Figure 3. Missing data check by SNP

Figure 3 shows a histogram of the missing data rate. Most of the proportion of sample that is missing is close to 0.00.

2.2.4.3 Population Structure Check

Population stratification is the systematic difference in allele frequencies between subpopulations. Population stratification may introduce false positive results if not proper controlled. Population structure check is aimed to detect subpopulation structure of the study population using multidimensional scaling (MDS) on SNP genotype data (Turner et al., 2011). We chose the number of dimension to be 4.

```
./plink --bfile ../Mel_IC --remove rm.list.txt --make-bed --noweb --out
Mel_IC_removed
./plink --bfile Mel_IC_removed --noweb --indep 50 5 1.01
./plink --bfile Mel_IC_removed --extract plink.prune.in --make-bed --
noweb --out Mel_IC_pruned
```

```
./plink --bfile Mel_IC_pruned --maf 0.05 --noweb --out Mel_IC_popustra
–genome
./plink --bfile Mel_IC_pruned --maf 0.05 --noweb --out Mel_IC_mds --
read-genome Mel_IC_popustra.genome --cluster --mds-plot 4
```

The MDS plots of the 4 dimensions are shown in Figure 4. In our study, what we expected to see is that all the plots are almost like residue plots instead of any specific structure or pattern, so that we would not treat population structure as a confounder.

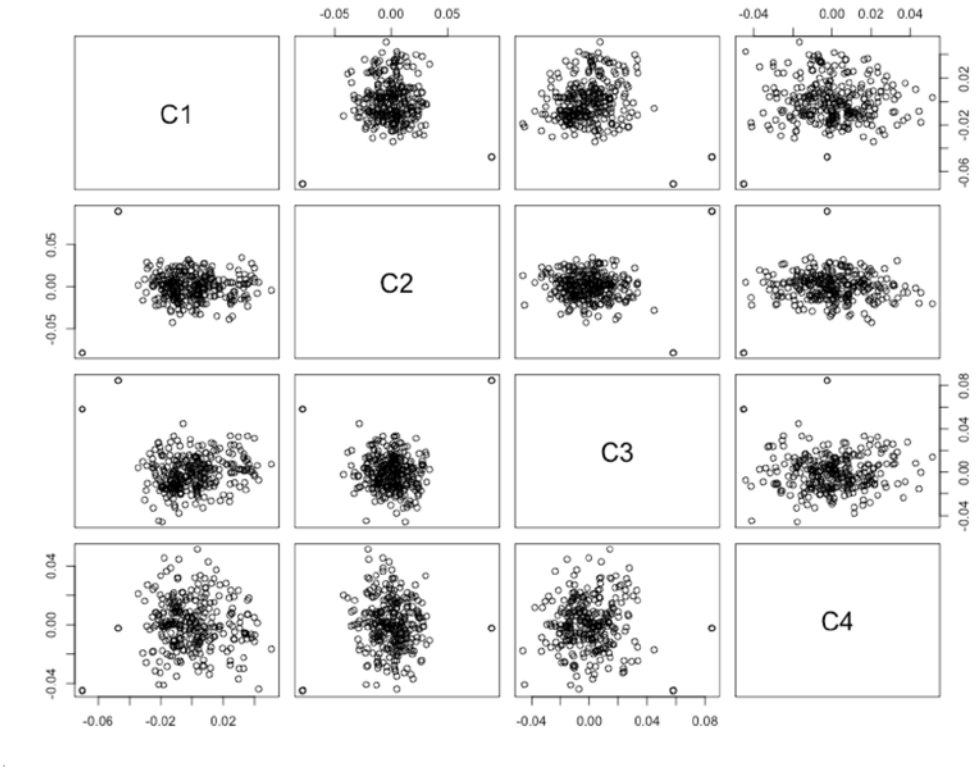


Figure 4. Population structure matrix

Figure 4 demonstrated that, overall, there is no obvious population structures, except for a few data points. Therefore, separate population structure plots with individual IDs were made to find those outliers.

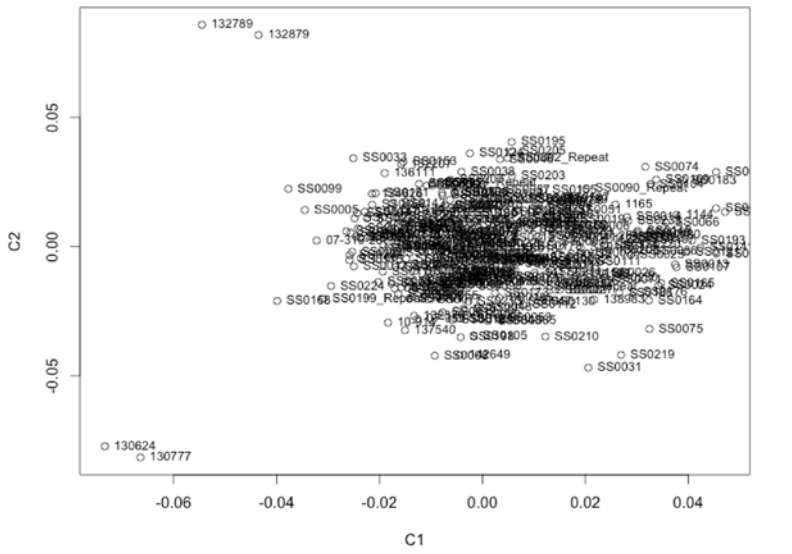


Figure 5. Population structure plot between C1 and C2

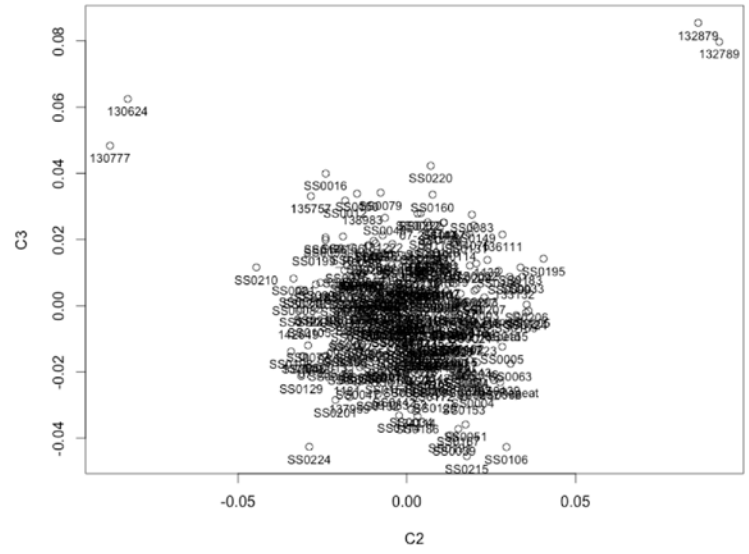


Figure 6. Population structure plot between C2 and C3

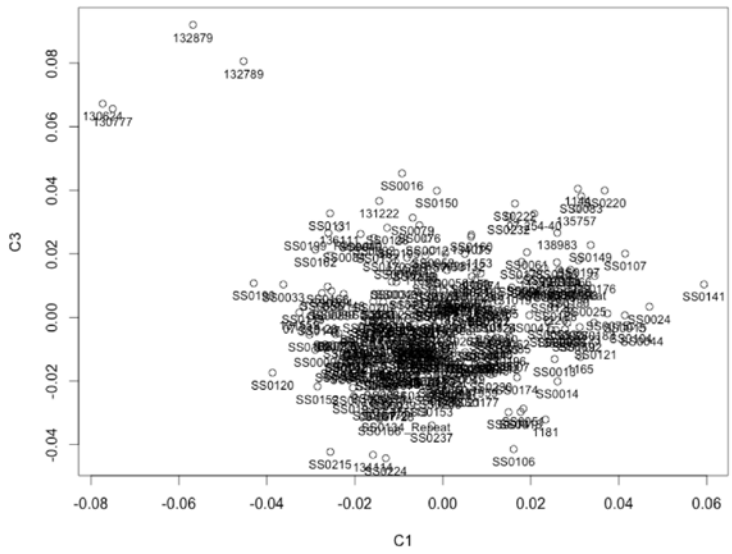


Figure 7. Population structure plot between C1 and C3

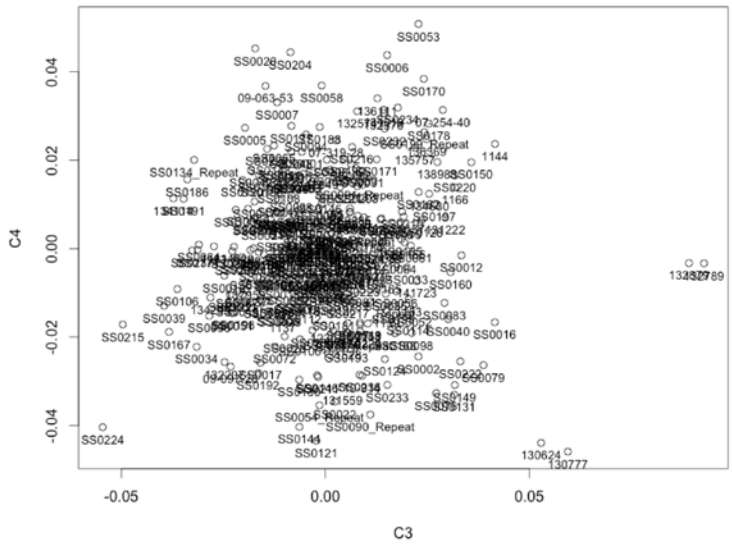


Figure 8. Population structure plot between C3 and C4

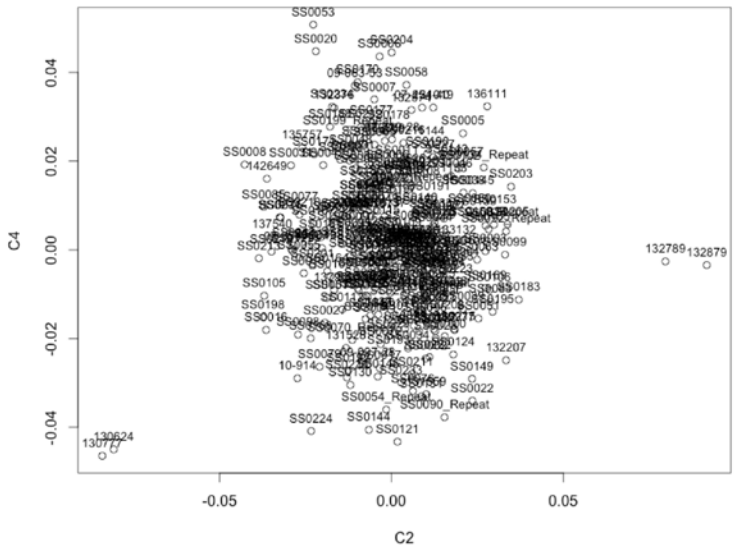


Figure 9. Population structure plot between C2 and C4

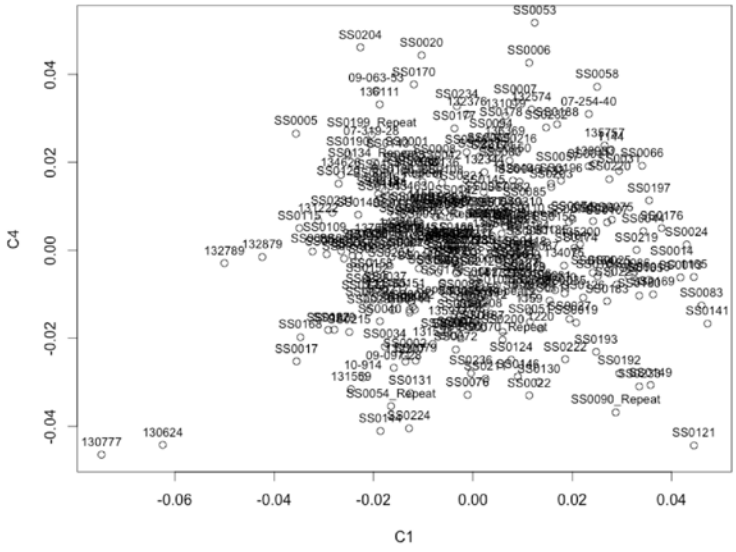


Figure 10. Population structure plot between C1 and C4

All of the five population structure plots above indicate the following four outliers: 130624, 130777, 132789, 132879

According to relationship check, missing data check and population structure check results, we finally decided to remove six samples (130624, 130777, 132789, 132879, SS0016, SS0093) from original data set after quality control process.

2.3 CLINICAL FACTORS

2.3.1 Model Selection

Cox proportional hazard regression was used to check the important clinical factors that affect the prognosis of the patients. The data file used in the analysis is *e1697_spore_29april15.csv*, which contains the final clinical data of these patients provided by the ECOG statistician. Clinical factors investigated in the analysis included trtm, sex, age, BRSLW_THICKNESS, CLARK_LVL, LDH_RS, LDH_ULN, PIG, PS and ULCER_YN (details listed in 2.2.1). Relapse-free survival (RFS) was used as the endpoint of the analysis.

Purposeful selection is a considerate method to select covariates in the regression model manually. It follows a slightly different logic to stepwise selection as proposed by Hosmer and Lemeshow (Hosmer Jr, Lemeshow, & Sturdivant, 2013). First, univariate analysis was performed for each covariate of interest and Wald test p-values are shown below in Table 1.

Table 1. Univariate model of RFS

Covariate	Wald Test p-value
Treatment (A/B)	0.71
Sex (Male/Female)	0.18*
Tumor Breslow's Thickness	0.01*
Clark's Level	0.15*
Lactate Dehydrogenase (LDH) Value	0.51
LDH Upper Limit of Normal	0.29
Pigmentation	0.20*
Performance Status	0.79
Ulceration (Yes/No)	0.26
Age at Diagnosis	0.01*

*significant at α=0.2 level

Five covariates had significant p-values at α=0.2. Following the steps of purposeful selection, a multivariable model with only two covariates, tumor Breslow's thickness and age at diagnosis, were included the final model. Table 2 lists the parameter estimates and Wald test p-values for covariates in the final model.

Table 2. Multivariable model of RFS

Covariate	Parameter Estimate	Wald Test p-value
Tumor Breslow's Thickness	0.801	0.036
Age at Diagnosis	0.025	0.018

2.4 TEST FOR ASSOCIATION AT SNP LEVEL

The GenABEL-package was used to conduct the SNP level analysis. This package performs an effective and powerful role in storing and handling GWAS data, as well as fast quality control procedures, testing of association, visualization of results, and easy interfaces to standard statistical and graphical procedures in R (Aulchenko, Ripke, Isaacs, & van Duijn, 2007).

Cox proportional hazards models were fit for RFS using the GenABEL package. Table 3 shows the results for the top 10 most significant associations, sorted by the Wald test p values. (Top 50 most significant association results are listed in appendix.)

Table 3. Summary for top 10 most significant association results at SNP level

SNP	Chr	Coordinate	Gene*	Location	P value
rs6944473	7	14326377	DGKB	INTRON	1.42E-06
rs10495124	1	217568816	LYPLAL1|LOC728510	INTERGENIC	1.82E-05
imm_12_2178130	12	2178130	CACNA1C	INTRON	3.96E-05
seq-rs2784110	1	197047009	PTPRC|LOC100131234	INTERGENIC	5.50E-05
rs17591522	1	217600391	LYPLAL1|LOC728510	INTERGENIC	5.98E-05
rs11942401	4	188052244	FAT|ZFP42	INTERGENIC	6.27E-05
rs6704463	1	217614448	LYPLAL1|LOC728510	INTERGENIC	7.18E-05
rs2095403	1	62632898	ANKRD38|USP1	INTERGENIC	8.27E-05
rs2839235	21	46625020	PCNT	INTRON	8.30E-05
rs3860187	10	49639139	WDFY4	INTRON	0.0001103

*Gene on which the SNP is located. When the SNP is located in between two genes, it is denoted as GENE1|GENE2.

A Manhattan plot of the SNP level results is shown in Figure 11. A Manhattan plot is a plot of the negative logarithm of the association p-value ($-\log_{10}$ P) for each single nucleotide polymorphism (SNP) against the genomic coordinates. We have one signal jumps above in chromosome 7. It seems to be very significant. But we are worried about this. This could be a sporadic positive or could be real because we don't have much information around it. So this signal needs to be checked out. Usually, a peak similar to chromosome 1 is expected to see for detecting the signals in genetic association study. Overall, we did not find many genome-wide significant results, which is expected for our sample size. Because the smallest p-value (the greatest negative logarithm) shown in the Manhattan plot is on chromosome 7, and chromosome 1 also has several small p-values, we also provided the chromosome level Manhattan plots for these two chromosomes (Figure 12) to see closely if they have some signals.

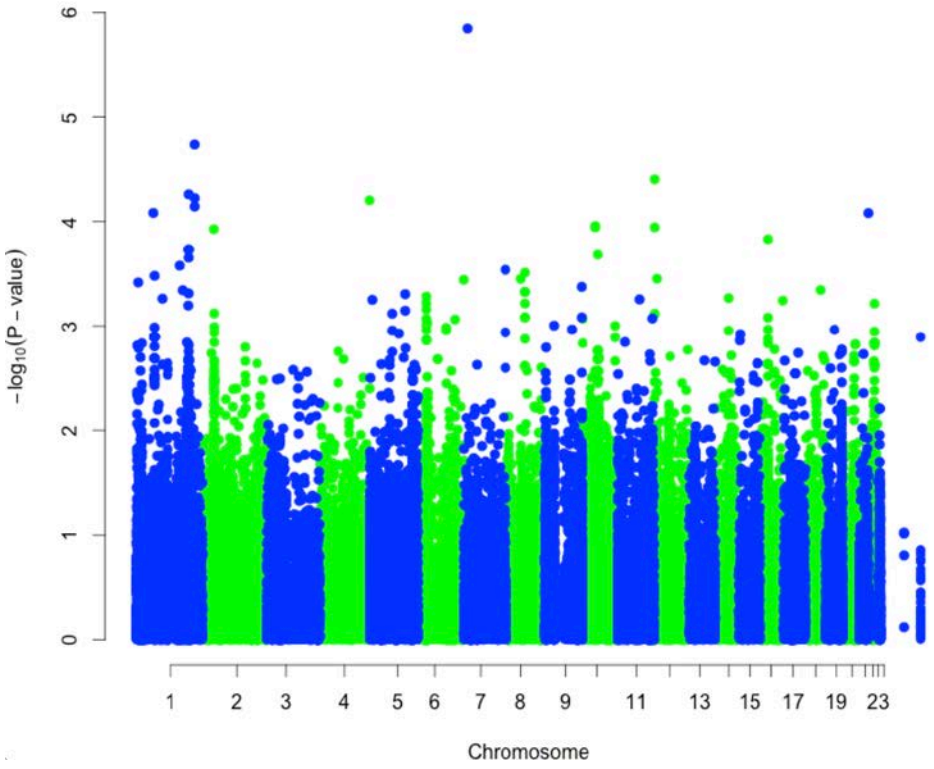


Figure 11. Manhattan plot

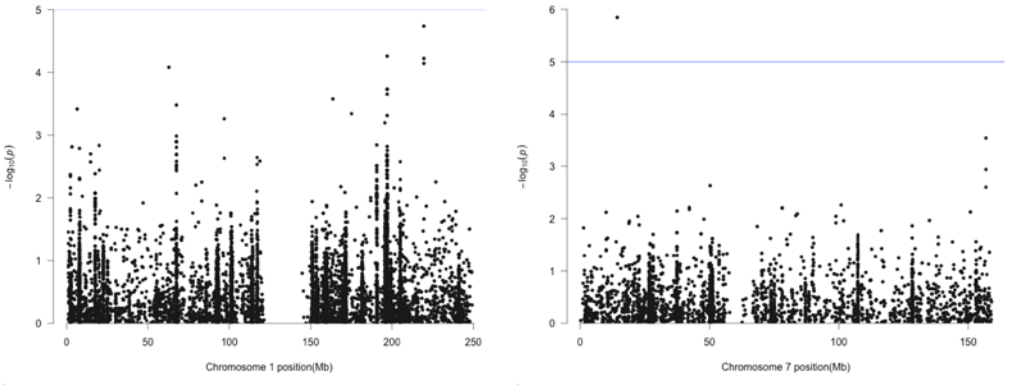


Figure 12. Manhattan plot for Chromosome 1 and 7

A Quantile-Quantile (QQ) plot of the SNP level analysis is shown in Figure 13. It plots the observed –log10 p-values against the expected –log10 p-values under the null model of no association. If all points fall on the diagonal line, then there is no association. It is expected that most of the SNPs, with the exception of a few, should be on the diagonal line. If most of the points deviate from the diagonal line, it is an indication that the observed association is spurious due to unknown underlying factors. In our case, no indication of inflated overall association was found.

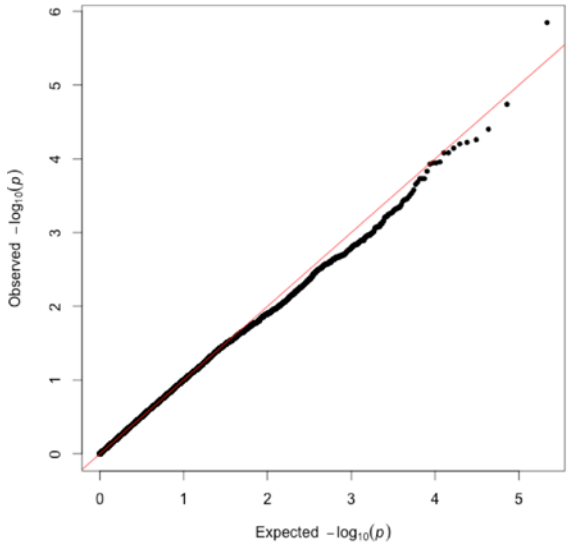
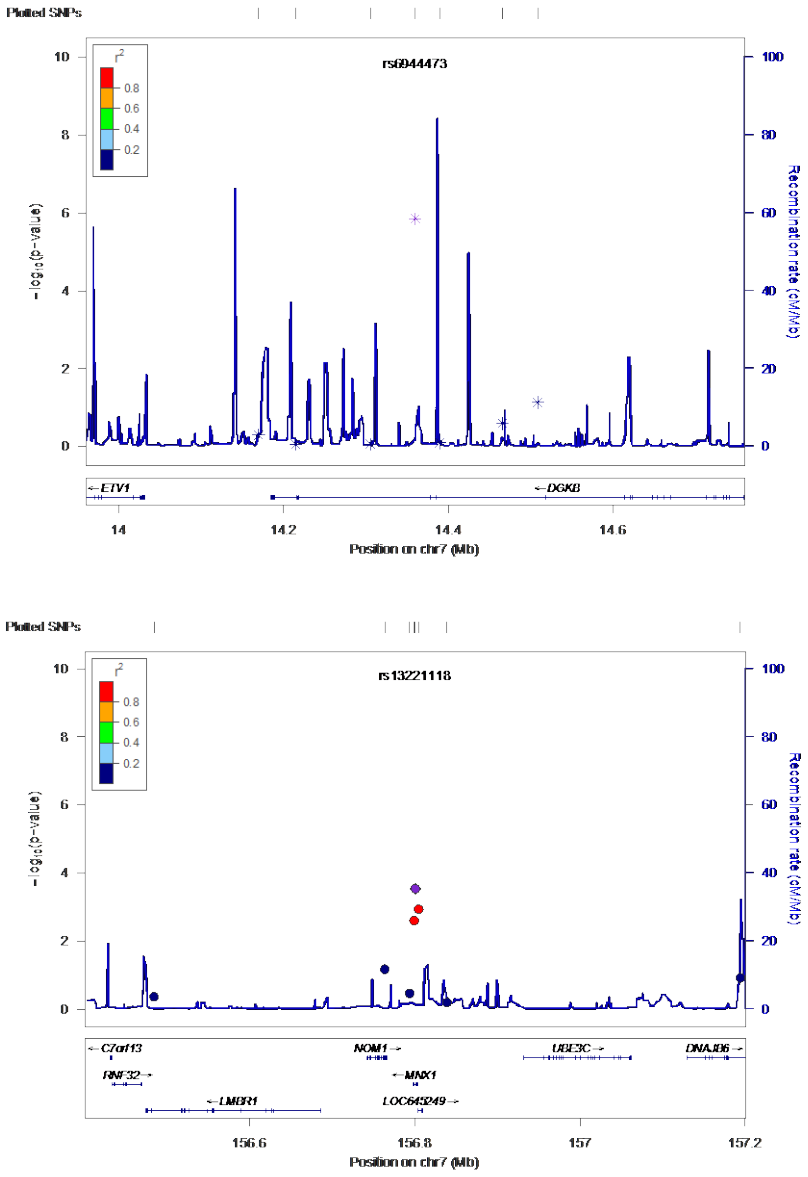


Figure 13. QQ plot

LocusZoom was used to plot the association results of the most significant SNPs. LocusZoom is a tool to plot the association results from GWAS, developed by Abecasis group. It is available at http://locuszoom.sph.umich.edu/locuszoom/. The purpose of these plots is to visualize nearby genes to infer the possible biological interpretation of the results.



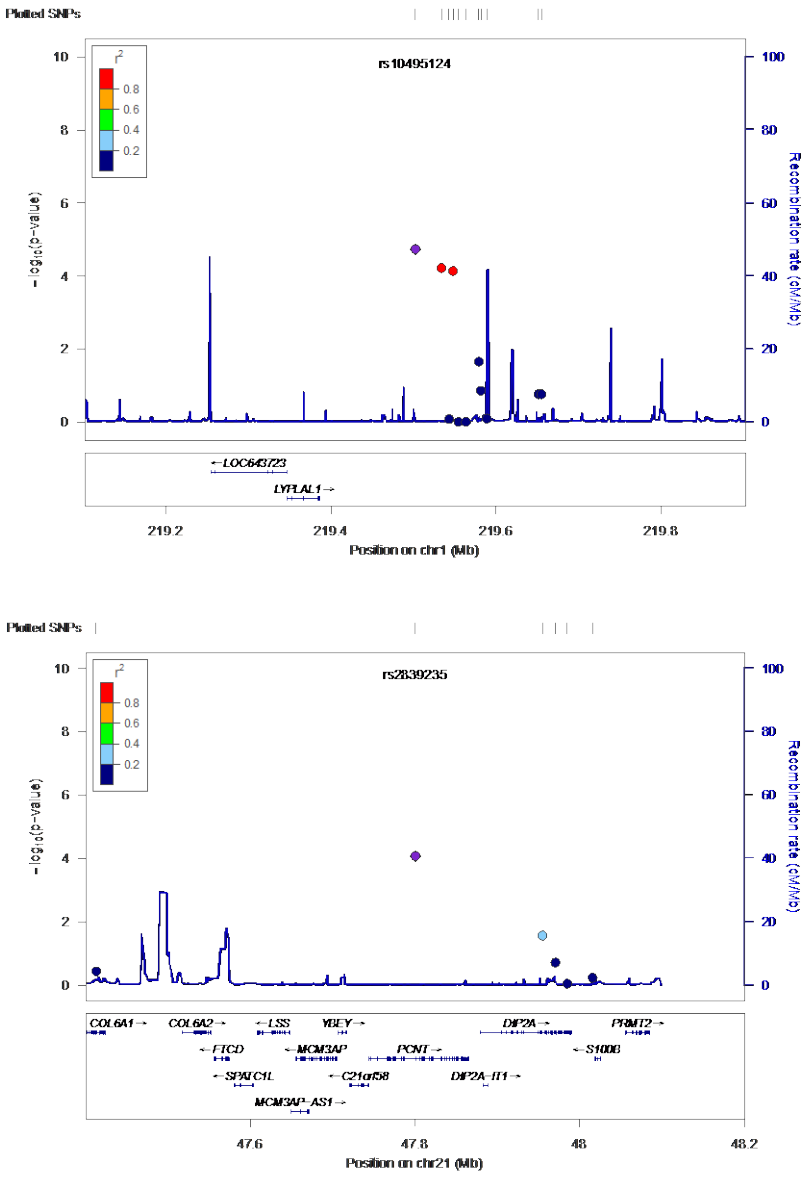


Figure 14. Regional plot of area surrounding interested SNPs of –log (P values) using LocusZoom

2.5 TEST FOR ASSOCIATION AT GENE LEVEL

Given the limited sample size, we also looked at the gene level analysis to improve power. Gene level analysis utilizes all the SNPs on (or near) the gene for the association test to improve the

power. Two methods were used for this analysis: the Sequence kernel association test (SKAT) (Lee et al., 2012; Lin et al., 2011) and the CoxKM (Cai, Tonini, & Lin, 2011; Lee et al., 2012). Both are kernel-based methods for gene set analysis. However, SKAT can only handle continuous and binary phenotype while CoxKM is designed for the time-to-event phenotype. For SKAT analysis RFS is dichotomized at 3 years.

2.5.1 SKAT

Sequence kernel association test (SKAT), is a kernel-based test method to look for the association between variants and phenotype (Lee et al., 2012). It utilizes a kernel matrix to aggregate individual SNP score statistics and computes p-values at gene level. Top 10 signals (based on the p values) of the SKAT data analysis are listed in Table 4.

Table 4. Top 10 SKAT results

GENE	Chr	Start	Stop	p value
DGKB	19	14136077	161585680	0.000147085
LOC340268	4	9834067	185223182	0.000332623
GABBR2	3	100274966	38086931	0.000336071
FBXL17	2	107045500	34732070	0.000684611
HTRA1	5	124216620	30136403	0.000800605
DUSP10	22	219470464	1629929	0.000835814
HLX	12	219074478	1635423	0.000836453
FBLN7	12	112615980	32848649	0.000907298
CPB2	23	45527945	84559380	0.000928026
CTCFL	10	55500243	68295641	0.000944519

2.5.2 CoxKM

CoxKM-package is an R package to perform Cox kernel machine SNP-set association test for association between SNP-set and a right-censored survival outcome. It uses the kernel machine Cox regression framework and performs a score test to assess the overall effect of the interested genetic markers (Lin et al., 2011). Two different kernels, the IBS and linear kernel, were used in this analysis, and the top 10 results are listed below. The IBS kernel is a kernel function that incorporates the IBS information. The results of these two kernels are very similar (Table 5).

Table 5. Top 10 coxKM results using IBS and linear kernel

GENE	n.marker.test	n.indiv	p.IBS	Q.IBS	df.IBS	p.linear	Q.linear	df.linear
HTRA1	2	205	1.00E-04	163.3571446	0.997232432	2.00E-04	318.2507069	1.0197953
FBXO32	4	205	6.00E-04	51.8164787	3.526320773	4.00E-04	208.1382071	3.552179132
FGF9	6	205	9.00E-04	53.7839697	3.490050315	0.0023	321.4549365	3.296177478
HLX	9	205	0.0012	49.1513785	4.810549398	NA	NA	NA
DUSP10	8	205	0.0012	55.69015018	4.034794788	3.00E-04	455.078092	4.084936
SOCS6	16	205	0.0017	21.30837138	10.92939678	NA	NA	NA
PEMT	4	205	0.0018	100.3035437	1.479564066	0.0021	401.2141747	1.492598419
RAI1	4	205	0.0018	100.3035437	1.479564066	0.0021	401.2141747	1.492598419
LOC642278	2	205	0.0019	133.8361564	1.066328026	0.001	267.6723128	1.034384141
KCNK1	2	205	0.0019	68.19834591	1.636928436	NA	NA	NA

2.6 CROSS REFERENCES OF DIFFERENT ANALYSIS RESULTS

2.6.1 Cross References of Gene Level Analysis

After getting the separated gene level analysis results by using SKAT and coxKM methods, comparisons of the top 50 significant gene results were made to search for the overlap between these two methods. As shown in Table 6, there are 10 overlapping genes between the SKAT and coxKM analysis results.

Table 6. Cross references of SKAT and coxKM analysis results

GENE	Chr	Start	Stop	n.marker.test	p.value.IBS	p.value.linear	p value.SKAT
HTRA1	5	124216620	30136403	2	1.00E-04	2.00E-04	0.000800605
DUSP10	22	219470464	1629929	8	0.0012	3.00E-04	0.000835814
HLX	12	219074478	1635423	9	0.0012	NA	0.000836453
FBLN7	12	112615980	32848649	2	0.0042	0.0087	0.000907298
ULK4	4	41834977	92844857	7	0.0089	0.0091	0.001026475
LOC642278	4	556195	241014568	2	0.0019	0.001	0.003086166
C20orf19	17	20735221	150859452	3	0.0023	0.0024	0.003348033
C20orf74	17	20735221	150859711	3	0.0023	0.0024	0.003348033
PEMT	20	17420920	159311566	4	0.0018	0.0021	0.007846131
RAI1	18	17478733	159295042	4	0.0018	0.0021	0.007846131

2.6.2 Cross References between Gene and SNP Level Analysis

We were also interested to see if some overlapping results would happen between the gene level and SNP level analysis results. Comparisons between the top 50 significant gene results and top 50 SNP level results were made. There are 3 overlapping genes between coxKM and SNP level analysis results (shown in Table 7), and only 1 overlapping gene between SKAT and SNP level analysis results (shown in Table 8).

Table 7. Cross references of SKAT and SNP level analysis results

GENE	Chr	SNP*	SNP.Coordinate	GeneLocation	p.value.SNP	p.value.SKAT
DGKB	7	rs6944473	14326377	INTRON	1.42E-06	0.000147085

*SNP is from the SNP level analysis

Table 8. Cross references of coxKM and SNP level analysis results

GENE	Chr	SNP*	SNP.Coordinate	GeneLocation	p.value.SNP	n.marker .test	p.value. IBS	p.value. linear
PLEKHG5	1	rs2986738	6470257	INTRON	0.000382766	2	0.0046	0.005
LOC100132924	10	rs9629920	49629651	INTRON	0.000113517	5	0.0055	0.0069
LOC100131234	1	seq-rs10800590	197042798	INTERGENIC	0.000221863	231	0.0087	0.0066

*SNPs are from the SNP level analysis

3.0 DISCUSSION AND FUTURE WORKS

3.1 GENERAL DISCUSSION

A total of 205 subjects passed QC and were included in the analysis. The total number of events in these 205 subjects is 61, which is rather small given the large number of SNPs (197076) tested. Thus, the statistical power for this analysis is extremely low. This exploratory analysis aims to generate a top rank list of genes to be followed up by larger studies. Thus, the p-values of the tests are not to be taken literally, rather, as a way of ranking the top hits. Different methods are used to confirm and complement each other.

In cancer research, OS is generally a more solid endpoint than RFS. The latter is subjected to the interval length of follow up. However, due to too few events in OS (32), we focused on the analysis using RFS as the phenotype.

To avoid bias and improve accuracy in our analysis, we first investigated the potential clinical factors that are associated with RFS in our study population. After model selection, two covariates of interest, Breslow's thickness and age at diagnosis, were left in the final. These two factors were controlled for in all of the following analyses.

3.2 SIGNIFICANT SIGNALS

Association tests were performed at two different levels. At SNP level, CoxPH models implemented in GenABEL-package were used. LocusZoom plots of 4 of the top hits, rs6944473, rs10495124, rs13221118 and rs2839235, were generated. We were not able to plot three other SNPs of interest because they are not assigned rs-numbers. SNP rs6944473 has a strong signal. However, the SNPs nearby do not seem to have strong association with the RFS thus we do not observe a typical "peak" as we usually see in a positive signal of a GWAS. Therefore, it could be a false positive. One possible cause of this could be genotype error. However, we are not able to check this because we do not have the raw data. We plotted the RFS plot by genotype of this SNP (Figure 15) to see if CoxPH is appropriate for the SNP. As shown in Figure 15 (I), the hazards between different genotype groups are proportional, although, the homozygous minor allele group has only 1 subject. We combined the subjects with minor allele together, as shown in Figure 15 (II), and applied log rank test. We obtained a significant p-value of 3.5E-7. Therefore, if this SNP is correctly genotyped, then it appears to be a significant predictor of RFS in our cohort. Further investigation of this result is needed.

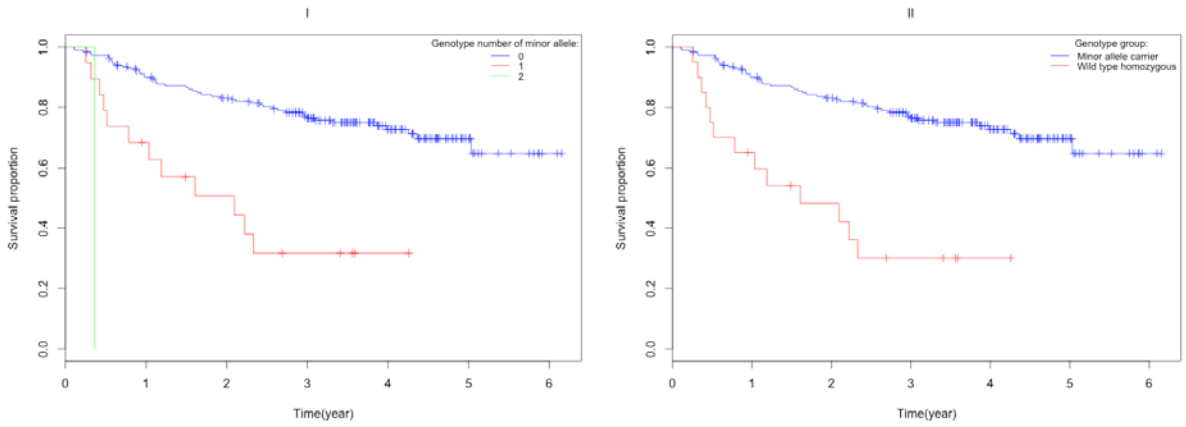


Figure 15. Kaplan-Meier curves of RFS by rs6944473 genotype
(I. by genotype number, II. by genotype group)

At the gene level, two different kernel-based methods were used to test for association. CoxKM was used to test the association at gene level using the RFS as the phenotype. The SKAT was applied to a dichotomized (at 3 year) RFS endpoint. Both are kernel-based tests, and both in theory require much larger sample size than our study cohort. It is reassuring that when we compared the top 50 lists of the two methods, 10 overlapped, which is what we expected, because the phenotypes are largely correlated.

When we cross-referenced the gene and SNP level analyses, the overlaps were very limited. DGKB showed up again in the SKAT analysis. The total number of SNPs on this gene included in the analysis for SKAT is 12. However, it is possible that the SKAT result is mostly driven by the one very significant SNP.

3.3 FUTURE WORKS

As discussed above, the results between the SNP and gene level analyses overlapped poorly. The analyses combined patients from both arms of the trial given that the trial is negative. However, at the molecular level, it is still possible that these two groups of patients responded differently. Thus, we plan to reanalyze the data stratified by patient treatment. This will further reduce the power of the study. However, we've experienced a similar situation where the subgroup analysis gave us more consistent results.

In consulting with a geneticist, we will work with the PI of the study and try to understand the biological function of the top hits. A validation of the genotyping of potential signals using a targeted platform, e.g. the Sequenome chip, will further strengthen the results of this analysis.

APPENDIX A: ADDITIONAL TABLES

Table 9. Summary for top 50 most significant association results at SNP level

SNP	Chr	Coordinate	GeneSymbol*	Location	P value
rs6944473	7	14326377	DGKB	INTRON	1.42E-06
rs10495124	1	217568816	LYPLAL1|LOC728510	INTERGENIC	1.82E-05
imm_12_2178130	12	2178130	CACNA1C	INTRON	3.96E-05
seq-rs2784110	1	197047009	PTPRC|LOC100131234	INTERGENIC	5.50E-05
rs17591522	1	217600391	LYPLAL1|LOC728510	INTERGENIC	5.98E-05
rs11942401	4	188052244	FAT|ZFP42	INTERGENIC	6.27E-05
rs6704463	1	217614448	LYPLAL1|LOC728510	INTERGENIC	7.18E-05
rs2095403	1	62632898	ANKRD38|USP1	INTERGENIC	8.27E-05
rs2839235	21	46625020	PCNT	INTRON	8.30E-05
rs3860187	10	49639139	WDFY4	INTRON	0.0001103
rs9629920	10	49629651	LOC100132924	INTRON	0.000113517
imm_12_2187865	12	2187865	CACNA1C	INTRON	0.000113531
rs9309074	2	41821802	SLC8A1|LDHAL3	INTERGENIC	0.000118057
imm_16_31279175	16	31279175	ITGAX	CODING	0.000147564
seq-rs10800591	1	197063314	PTPRC|LOC100131234	INTERGENIC	0.000185575

Table 9 Continued

seq-rs12406470	1	197062743	PTPRC|LOC100131234	INTERGENIC	0.000185575
seq-rs6427752	1	197062012	PTPRC|LOC100131234	INTERGENIC	0.000185575
1kg_10_59547220	10	59547220	ZWINT|IPMK	INTERGENIC	0.000207381
seq-rs10800590	1	197042798	PTPRC|LOC100131234	INTERGENIC	0.000221863
rs2226007	1	161860160	NUF2|LOC729952	INTERGENIC	0.00026407
rs13221118	7	156493383	MNX1	INTRON	0.000289407
rs7837005	8	72161586	XKR9|EYA1	INTERGENIC	0.000307944
imm_1_67442201	1	67442201	IL23R	INTRON	0.000330605
rs10845202	12	10726132	STYK1|CSDA	INTERGENIC	0.000352037
rs270793	8	56012168	LOC100128419|XKR4	INTERGENIC	0.000354308
rs6927768	6	170708910	TBP	INTRON	0.000361202
rs2986738	1	6470257	PLEKHG5	INTRON	0.000382766
imm_9_138290709	9	138290709	QSOX2|DKFZP434A062	INTERGENIC	0.000421789
rs1979302	18	55002156	SEC11C|GRP	INTERGENIC	0.000452031
rs12146041	1	173368784	TNN	INTRON	0.000454158
rs1156956	8	71917592	XKR9|EYA1	INTERGENIC	0.000471323
rs7845516	8	71848376	XKR9|EYA1	INTERGENIC	0.000471323
seq-rs2784114	1	197037793	PTPRC|LOC100131234	INTERGENIC	0.000486758
rs12153520	5	130446033	CHSY-2|HINT1	INTERGENIC	0.00049527
rs9262632	6	31132787	HCG22	UTR	0.000521266
rs2332094	14	69139425	LOC100130174|KIAA0247	INTERGENIC	0.000540284
rs545152	1	96659092	LOC729977|LOC440595	INTERGENIC	0.000547392
rs990648	11	79730135	ODZ4|MGC33846	INTERGENIC	0.000555701

Table 9 Continued

rs6864771	5	7850291	ADCY2	INTRON	0.000560977
rs9923190	16	85014134	LOC732275|LOC283904	INTERGENIC	0.000570215
rs11753208	6	31113411	LOC729792	INTRON	0.000602604
rs12155783	8	72161680	XKR9|EYA1	INTERGENIC	0.000608692
imm_22_38042260	22	38042260	RPL3	INTRON	0.000610902
imm_1_195577864	1	195577864	CRB1	INTRON	0.00063566
rs7768644	6	31110080	LOC729792	INTRON	0.000692246
rs7733977	5	130466000	CHSY-2|HINT1	INTERGENIC	0.000712495
1kg_2_43478582	2	43478582	THADA	INTRON	0.000758235
rs9283781	5	82103862	LOC92270|TMEM167	INTERGENIC	0.000761276
imm_12_2194668	12	2194668	CACNA1C	INTRON	0.000761736
imm_9_138290450	9	138290450	QSOX2|DKFZP434A062	INTERGENIC	0.0008274

*Gene on which the SNP is located. When the SNP is located in between two genes, it is denoted as GENE1|GENE2.

Table 10. Top 50 SKAT results

GENE	Chr	Start	Stop	p value
DGKB	19	14136077	161585680	0.000147085
LOC340268	4	9834067	185223182	0.000332623
GABBR2	3	100274966	38086931	0.000336071
FBXL17	2	107045500	34732070	0.000684611
HTRA1	5	124216620	30136403	0.000800605
DUSP10	22	219470464	1629929	0.000835814
HLX	12	219074478	1635423	0.000836453
FBLN7	12	112615980	32848649	0.000907298
CPB2	23	45527945	84559380	0.000928026
CTCFL	10	55500243	68295641	0.000944519
ULK4	4	41834977	92844857	0.001026475
ADAMTS2	12	178485124	10945296	0.001097248
GABRA5	14	24573461	138614368	0.001109697
GRB2	3	70823779	52366748	0.001656428
DDHD1	5	52524755	70796945	0.001727406
LOC645434	5	139880112	24206432	0.002133753
KIAA0195	3	70915763	52191086	0.00217213
BMP4	11	52693414	70783451	0.002432467
LOC100131472	19	10233678	181965614	0.002537652
LOC729112	17	122598338	30505819	0.002589555

Table 10 Continued

LOC642278	4	556195	241014568	0.003086166
FGF21	5	53951133	69333750	0.003108941
GRIK4	2	119870238	30923014	0.003142867
C20orf19	17	20735221	150859452	0.003348033
C20orf74	17	20735221	150859711	0.003348033
LOC100130010	4	12094704	167311824	0.003372483
TMEM89	23	48633471	79449021	0.003482006
VARS2	4	30990256	119147430	0.003965721
MTMR2	13	95208901	41013631	0.004078789
SPON1	22	13882056	162191705	0.00457111
GINS3	13	56971665	67130788	0.00472719
SH3MD4	2	108973688	34522053	0.004736383
C10orf67	14	23654346	141102100	0.0051071
SULT1A1	21	28517358	129249695	0.005224223
LOC100132354	23	43866851	87430929	0.005596336
ADRB2	5	148201190	22583965	0.006717942
SH3TC2	5	148201190	22583965	0.006717942
BCAT2	5	53953425	69312471	0.006810132
FLJ44815	4	29723177	127285007	0.006961246
COL7A1	23	48579063	79458866	0.00724393
AK5	9	77498829	49458078	0.007419735
CCL1	4	29710751	127300962	0.007768454
C9orf11	14	27223272	132113876	0.007777693

Table 10 Continued

PEMT	20	17420920	159311566	0.007846131
RAI1	18	17478733	159295042	0.007846131
SEMA3A	20	83166896	46044577	0.007899354
TBX15	2	118685496	31104111	0.008055778
ZNF516	3	71278968	51888761	0.008316658
C11orf49	23	47011024	80133854	0.008337512
LOC645000	4	60737084	61843768	0.008477575

Table 11. Top 50 coxKM results using IBS and linear kernel

GENE	n.marker.test	n.indiv	p.IBS	Q.IBS	df.IBS	p.linear	Q.linear	df.linear
HTRA1	2	205	1.00E-04	163.3571446	0.997232432	2.00E-04	318.2507069	1.0197953
FBXO32	4	205	6.00E-04	51.8164787	3.526320773	4.00E-04	208.1382071	3.552179132
FGF9	6	205	9.00E-04	53.7839697	3.490050315	0.0023	321.4549365	3.296177478
HLX	9	205	0.0012	49.1513785	4.810549398	NA	NA	NA
DUSP10	8	205	0.0012	55.69015018	4.034794788	3.00E-04	455.078092	4.084936
SOCS6	16	205	0.0017	21.30837138	10.92939678	NA	NA	NA
PEMT	4	205	0.0018	100.3035437	1.479564066	0.0021	401.2141747	1.492598419
RAI1	4	205	0.0018	100.3035437	1.479564066	0.0021	401.2141747	1.492598419
LOC642278	2	205	0.0019	133.8361564	1.066328026	0.001	267.6723128	1.034384141
KCNK1	2	205	0.0019	68.19834591	1.636928436	NA	NA	NA
C20orf74	3	205	0.0023	78.02869894	1.874982136	0.0024	234.0860968	1.868850054
C20orf19	3	205	0.0023	78.02869894	1.874982136	0.0024	234.0860968	1.868850054
IL1F7	5	205	0.0033	45.87091418	1.903022303	0.0039	233.7458099	1.898427645
LOC729668	22	205	0.0034	52.51872205	2.941225234	0.0039	1151.468446	2.844106193

Table 11 Continued

BACH2	207	205	0.0035	28.79968437	8.606043205	0.0038	5867.49402	9.217885776
CCDC88C	3	205	0.0036	85.93053123	1.205689231	0.003	251.9361541	1.239301028
RRM1	3	205	0.0042	43.02581801	2.757766733	0.004	129.077454	2.74399433
FBLN7	2	205	0.0042	67.98504669	2.016110267	0.0087	113.9488161	1.904888143
PLEKHG5	2	205	0.0046	52.66691091	1.779263457	0.005	105.3338218	1.786707106
C14orf181	70	205	0.005	34.29378735	5.221603951	0.004	2370.634724	5.12025002
UBLCP1	111	205	0.0051	23.36087174	7.906729142	0.0053	2628.336445	8.059648632
MAP3K8	121	205	0.0052	28.06843994	7.416439869	0.0039	1151.468446	2.844106193
CHODL	2	205	0.0053	69.87374076	1.844069232	0.0055	139.7474815	1.80580238
LOC100132924	5	205	0.0055	42.50840709	2.032615732	0.0069	212.5420355	2.034380308
IDE	10	205	0.0056	36.50485026	3.234346832	NA	NA	NA
ACTN1	62	205	0.0058	32.7955529	5.2595226	0.0053	2006.406783	5.162854725
LOC730134	6	205	0.0058	35.88389223	4.489042795	0.0065	215.3033534	4.380506614
LOC100131866	64	205	0.0059	34.90505968	4.542110238	0.0054	2223.256908	4.695085661
LOC100128781	56	205	0.006	40.62994815	3.530145757	0.0057	2269.892288	3.728449707
CBLN2	20	205	0.0062	16.17112482	13.3626406	0.0071	321.3913423	13.22349141

Table 11 Continued

PHLDB1	45	205	0.0063	37.48481044	4.380171538	0.0053	1715.572943	4.406236761
FTHL7	12	205	0.0063	26.17782462	6.714230682	0.0023	321.4549365	3.296177478
ADRA1B	78	205	0.0064	23.196342	9.132998512	0.0038	1875.790706	9.257734887
MRPL36	2	205	0.007	36.04350296	1.093684329	0.006	75.62701289	1.082086459
KIR3DL2	3	205	0.0071	56.49377593	2.033623651	0.0079	156.7243266	2.016101053
CLEC2B	140	205	0.0074	16.91395194	12.56710791	0.0075	2374.514151	12.61887258
LOC728727	11	205	0.0076	31.14237744	5.204061415	0.0087	345.1852147	5.254971518
FLJ41046	8	205	0.0076	27.17208389	4.079984235	0.0068	217.1868471	4.156190183
FLJ42418	8	205	0.0076	27.17208389	4.079984235	0.0068	217.1868471	4.156190183
DLC1	6	205	0.0082	33.05230852	4.132508543	0.0053	205.4312161	4.239507095
WFDC12	3	205	0.0082	61.92950098	1.099297828	0.0127	169.9468802	0.973506929
RUNX2	2	205	0.0083	50.71762835	2.021478056	0.0083	101.4352567	1.954351493
DSCAML1	3	205	0.0084	24.35274809	2.550270076	0.0081	73.05824427	2.579148
DYRK2	145	205	0.0087	26.0997574	6.63316068	NA	NA	NA
DIP2C	6	205	0.0087	37.83638328	3.841283055	0.0089	227.0182997	3.852243302
LOC100131234	231	205	0.0087	19.27530844	9.438183814	0.0066	4473.958268	9.945299748

Table 11 Continued

FZD8	5	205	0.0089	36.69497966	3.091913567	0.0074	183.4748983	3.199695772
ULK4	7	205	0.0089	70.15212838	1.488121134	0.0091	485.790286	1.467081239
HLA-E	482	205	0.0092	28.35820247	5.175844034	0.0087	13783.93721	5.180179863
TNR	2	205	0.0093	54.69729222	1.953768899	NA	NA	NA

APPENDIX B: R CODE

```
### Quality Control ###

###1.Relationship Check

# Relationship check
setwd("~/Dropbox/thesis/QC/relationcheck")

relation <- read.table("Mel_IC_relationcheck.genome", header = T)
head(relation)
attach(relation)

plot(Z0, Z1, col = RT, xlim = c(0, 1), ylim = c(0, 1),
        main = "Relationship Check")
with(relation, text(Z0[which(Z0 < 0.01)] + 0.05,
        Z1[which(Z0 < 0.01)], IID1[which(Z0 < 0.01)]), cex = 0.5)

s <- relation[relation$Z0 < 0.1, ]
t <- s$IID1 # 130624 132789
      u <- s$IID2 # 130777 132879

###2.Missing Data Check

# Missing data check by individual
setwd("~/Dropbox/thesis/QC/missingcheck")

het <- read.table("Mel_IC_misscheck.het", header = T)
miss <- read.table("Mel_IC_misscheck.imiss", header = T)
het_miss <- merge(het, miss, by = c("FID", "IID"))

# Calculate the observed heterozygosity rate
Observed_het_rate <- (het_miss$N.NM. - het_miss$O.HOM.) / het_miss$N.NM.
het.miss <- data.frame(het_miss, Observed_het_rate)
head(het.miss)

with(het.miss, plot(Observed_het_rate, F_MISS, xlim = c(0.1, 0.9), main =
 "Missing data check"))
with(het.miss, text(Observed_het_rate[which(F_MISS > 0.05)] + 0.05,
 F_MISS[which(F_MISS > 0.05)], IID[which(F_MISS > 0.05)]))

b <- het.miss[het.miss$F_MISS > 0.01, ]
a <- b$IID # 15 sample IDs
# SS0110, SS0137, SS0016, SS0025, SS0045_Repeat, SS0054_Repeat,
 SS0070_Repeat, SS0090_Repeat, SS0091_Repeat, SS0092_Repeat,
 SS0134_Repeat, SS0159_Repeat, SS0199_Repeat, SS0217_Repeat, SS0093
b1 <- het.miss[het.miss$F_MISS > 0.05, ]
a1 <- b1$IID # 2 sample IDs
      # SS0016, SS0093
```

```
# Missing data check by SNP
lmiss <- read.table("Mel_IC_misscheck.lmiss", header = T)
dim(lmiss) # 129903      5

with(lmiss, hist(F_MISS, breaks = 100, xlim = c(0, 0.1)))
abline(v = 0.05, col = "red", lty = 2)
      dim(lmiss[which(lmiss$F_MISS < 0.05), ]) # 129043      5

###3.Population Structure Check

# Population structure check
setwd("~/Dropbox/thesis/QC/popustra")

mds <- read.table("Mel_IC_mds.mds", header = T)
head(mds)
attach(mds)

# plot matrix
plot(mds[, c(-1, -2, -3)], main = "Population structure check")

# make separate population structure plots to find outliers
# C1-C2
# add noise to separate overlapped points
x = jitter(mds[, 4], factor = 500)
y = jitter(mds[, 5], factor = 500)

# make population structure check plot with individual IDs
plot(x, y, main = "Population structure check", xlab = "C1", ylab = "C2")
text(x, y, labels = IID, pos = 4, cex = 0.8)

# C2-C3
# add noise to separate overlapped points
x1 = jitter(mds[, 5], factor = 500)
y1 = jitter(mds[, 6], factor = 500)

# make population structure check plot with individual IDs
plot(x1, y1, main = "Population structure check", xlab = "C2", ylab =
 "C3")
text(x1, y1, labels = IID, pos = 1, cex = 0.8)

# C1-C3
# add noise to separate overlapped points
x2 = jitter(mds[, 4], factor = 500)
y2 = jitter(mds[, 6], factor = 500)

# make population structure check plot with individual IDs
plot(x2, y2, main = "Population structure check", xlab = "C1", ylab =
 "C3")
text(x2, y2, labels = IID, pos = 1, cex = 0.8)

# C3-C4
# add noise to separate overlapped points
x3 = jitter(mds[, 6], factor = 500)
y3 = jitter(mds[, 7], factor = 500)

# make population structure check plot with individual IDs
plot(x3, y3, main = "Population structure check", xlab = "C3", ylab =
 "C4")
text(x3, y3, labels = IID, pos = 1, cex = 0.8)
```

```
# C2-C4
# add noise to separate overlapped points
x4 = jitter(mds[, 5], factor = 500)
y4 = jitter(mds[, 7], factor = 500)

# make population structure check plot with individual IDs
plot(x4, y4, main = "Population structure check", xlab = "C2", ylab =
 "C4")
text(x4, y4, labels = IID, pos = 3, cex = 0.8)

# C1-C4
# add noise to separate overlapped points
x5 = jitter(mds[, 4], factor = 500)
y5 = jitter(mds[, 7], factor = 500)

# make population structure check plot with individual IDs
plot(x5, y5, main = "Population structure check", xlab = "C1", ylab =
 "C4")
text(x5, y5, labels = IID, pos = 3, cex = 0.8)



### Model Selection ###

setwd("~/Dropbox/thesis/4-27-15/Mel-GenABEL_Ying")

e1697 <- read.csv("e1697_spore_29april15.csv",header=T)   # 216
head(e1697)

# only include treatment A or B
ab <- e1697[e1697$trtm=="A"|e1697$trtm=="B",]   # 216 obs, since e1697
 only include treatment A and B

# change column names
names(ab)
colnames(ab)[13] <- "cens.RFS"
colnames(ab)[14] <- "cens.OS"
colnames(ab)[4] <- "BRSLW"
colnames(ab)[5] <- "CLARK"

# calculate OS and RFS in days
ab$OS.n <- ab$surv_y*365
ab$RFS.n <- ab$rfs*365

# dichotomize BRSLW using cutoff=2
sum(ab$BRSLW<=2)   # 52
sum(ab$BRSLW>2)   #164
ab[ab$BRSLW<=2,]$BRSLW <- 0
ab[ab$BRSLW>2,]$BRSLW <- 1



library(survival)

surv <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ strata(trtm) + sex
 + BRSLW + CLARK + LDH_RS + LDH_ULN + PIG + PS + ULCER_YN + age, data =
 pheno)
summary(surv)   # p=0.02, Rsquare= 0.085

# univariable models
```

```
surv.trtm <- survfit(Surv(time=pheno$RFS.n, event=cens.RFS) ~ trtm, data =
 pheno)
plot(surv.trtm)
surv.trtm.cox <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ trtm, data
 = pheno)
summary(surv.trtm.cox)   # p=0.71

surv.sex <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ sex, data =
 pheno)
summary(surv.sex)   # p=0.18* <0.2

surv.brslw <- survfit(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW, data
 = pheno)
plot(surv.brslw)
surv.brslw.cox <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW,
 data = pheno)
summary(surv.brslw.cox)   # p=0.01* <0.2

surv.clark <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ CLARK, data =
 pheno)
summary(surv.clark)   # p=0.15* <0.2

surv.ldh <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ LDH_RS, data =
 pheno)
summary(surv.ldh)   # p=0.51

surv.ldhu <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ LDH_ULN, data
 = pheno)
summary(surv.ldhu)   # p=0.29

surv.pig <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ PIG, data =
 pheno)
summary(surv.pig)   # p=0.195* <0.2

surv.ps <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ PS, data =
 pheno)
summary(surv.ps)   # p=0.79

surv.ulcer <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ ULCER_YN,
 data = pheno)
summary(surv.ulcer)   # p=0.26

surv.age <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ age, data =
 pheno)
summary(surv.age)   # p=0.006* <0.2

# fit a multivariable model containing all variables significant in the
 univariable analysis at p<0.2 level
# sex, BRSLW, CLARK, PIG, age
surv.multi <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ sex + BRSLW +
 CLARK + PIG + age, data = pheno)
summary(surv.multi)   # p=0.004
# Wald test p-values: sex=0.57, BRSLW=0.05*, CLARK=0.33, PIG=0.12,
 age=0.03*

# delete sex and refit the multivariable model
# BRSLW, CLARK, PIG, age
surv.multi1 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW +
 CLARK + PIG + age, data = pheno)
summary(surv.multi1)   # p=0.002
# Wald test p-values: BRSLW=0.05*, CLARK=0.34, PIG=0.13, age=0.017*
```

```
# estimates of coefficients are virtually unchanged

# delete CLARK and refit the multivariable model
# BRSLW, PIG, age
surv.multi2 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + PIG
 + age, data = pheno)
summary(surv.multi2)   # p=0.0013
# Wald test p-values: BRSLW=0.03*, PIG=0.12, age=0.012*
# estimates of coefficients are virtually unchanged

# delete PIG and refit the multivariable model
# BRSLW, age
surv.multi3 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age,
 data = pheno)
summary(surv.multi3)   # p=0.0014
# Wald test p-values: BRSLW=0.0355*, age=0.0176*
# estimates of coefficients are virtually unchanged


# add ULCER_YN and refit the multivariable model to see if ULCER_YN
 becomes significant
# BRSLW, age, ULCER_YN
surv.multi4 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age
 + ULCER_YN, data = pheno)
summary(surv.multi4)   # p=0.002
# Wald test p-values: BRSLW=0.03*, age=0.03*, ULCER_YN=0.26
# ULCER_YN is not significant, so no need to add it

# add LDH_ULN and refit the multivariable model to see if LDH_ULN becomes
 significant
# BRSLW, age, LDH_ULN
surv.multi5 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age
 + LDH_ULN, data = pheno)
summary(surv.multi5)   # p=0.003
# Wald test p-values: BRSLW=0.03*, age=0.02*, LDH_ULN=0.27
# LDH_ULN is not significant, so no need to add it

# add LDH_RS and refit the multivariable model to see if LDH_RS becomes
 significant
# BRSLW, age, LDH_RS
surv.multi6 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age
 + LDH_RS, data = pheno)
summary(surv.multi6)   # p=0.003
# Wald test p-values: BRSLW=0.035*, age=0.017*, LDH_RS=0.47
# LDH_RS is not significant, so no need to add it

# add trtm and refit the multivariable model to see if trtm becomes
 significant
# BRSLW, age, trtm
surv.multi7 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age
 + trtm, data = pheno)
summary(surv.multi7)   # p=0.004
# Wald test p-values: BRSLW=0.0364*, age=0.0175*, trtm=0.71
# trtm is not significant, so no need to add it

# add PS and refit the multivariable model to see if PS becomes
 significant
# BRSLW, age, PS
surv.multi8 <- coxph(Surv(time=pheno$RFS.n, event=cens.RFS) ~ BRSLW + age
 + PS, data = pheno)
summary(surv.multi8)   # p=0.004
```

```
# Wald test p-values: BRSLW=0.035*, age=0.0167*, PS=0.69
# PS is not significant, so no need to add it

### final model only contains BRSLW and age


write.csv(format(ab,digits=0), "model_brsage.csv", quote=F, row.names=F)



### Test for Association at SNP Level ###

#Prepare for creating genotype data
library(GenABEL)
library(gdata)
library(qqman)
library(survival)

setwd("~/Dropbox/thesis/5-26-15/Mel-GenABEL_Ying")

# prepare for creating genotype data
ab <- read.csv("model_brsage.csv", header=T)   # 216
colnames(ab)[1] <- "SEQ_NUMBER"

Mel2013 <-
 read.csv("10.28.13_genotype_SP_Melanoma_Tarhini_DataSheet1.csv",
 header=T)   # 299
names(Mel2013)
Mel2013 <- Mel2013[, c(1,5,39,57)]

data <- merge(ab, Mel2013, by="SEQ_NUMBER")   # 215
data <- data[order(data$X.3), ]
head(data)
names(data)

id <- data[, c(20,18)]

write.table(id, "id.list.txt", quote=F, sep="\t", row.names=F,
 col.names=F)


# after using Linux to generate .tped and .tfam files, use them to create
 genotype file
convert.snp.tped(tped="Clean-Mel_IC.tped", tfam="Clean-Mel_IC.tfam",
 out="Clean-Mel_IC.raw", strand="u")


# create phenotype data
head(data)
names(data)

pheno <- data[, c(18,16,14,17,13,2:10,15)]
head(pheno)   # 215 obs, 15 var
colnames(pheno)[1] <- "id"

write.table(format(pheno,digits=0), "pheno.txt", quote=F, sep="\t",
 row.names=F)


# since pheno.txt has 215 obs, but Clean-Mel_IC.raw only has 205 obs,
# pheno.txt needs to be checked line-by-line
```

```
pheno.clean <- read.table("pheno-clean.txt", header=T)   # 205 obs
# SS0045, SS0054, SS0070, SS0090, SS0091, SS0092, SS0134, SS0159, SS0199,
 SS0217 were removed

# count number of RFS events and OS events
event <- pheno.clean[pheno.clean$cens.RFS == 1, ]   # 61 event, 144 censor
event_OS <- pheno.clean[pheno.clean$cens.OS == 1, ]   # 61 event, 144
 censor

# change sex from 1/2 (1 = male, 2 = female) to 0/1 (0=female and 1=male)
pheno.clean$sex <- ifelse(pheno.clean$sex == 2, pheno.clean$sex <- 0,
 pheno.clean$sex <- 1)
write.table(pheno.clean, "pheno-clean.txt", quote=F, sep="\t",
 row.names=F, col.names=T)


# use GenABEL package
data <- load.gwaa.data(phe="pheno-clean.txt", gen="Clean-Mel_IC.raw",
 force=TRUE)

# run Cox proportional hazards models

cox.RFS <- mlreg(GASurv(RFS.n, cens.RFS) ~ 1, data)

CHR = data@gtdata@chromosome
BP = data@gtdata@map

RFS.7 <- cox.RFS@ results[which(cox.RFS@ results$ P1df < 10^(-7)),
 c("effB", "se_effB", "chi2.1df", "P1df" )]
RFS.4 <- cox.RFS@ results[which(cox.RFS@ results$ P1df < 10^(-4)),
 c("effB", "se_effB", "chi2.1df", "P1df" )]
dim(RFS.7)   # 0 obs, 4 var
dim(RFS.4)   # 4 obs, 4 var

RFS.all <- cox.RFS@ results[ , c("effB", "se_effB", "chi2.1df", "P1df" )]
SNP = row.names(RFS.all)
RFS.all = data.frame(SNP,CHR,BP,RFS.all)

write.table(RFS.all, sep="\t", file="RFS.all.txt", row.names=FALSE)


#--- QQ plot and manhattan plot ---#

#-- plot RFS --#
plot.RFS <- data.frame(CHR=data@gtdata@chromosome, BP=data@gtdata@map,
 P=cox.RFS@results$P1df)
plot.RFS$CHR <- as.numeric(as.character(drop.levels(plot.RFS$CHR)))
dim(plot.RFS)   # 108300 obs, 3 var
head(plot.RFS)

qq(plot.RFS$P, main="Q-Q plot for RFS")


### Manhattan plot of cox.RFS and find the most significant association
 SNPs ###

plot(cox.RFS, ylim=c(0,7), pch=19, main="Manhattan plot for RFS")

bestHits <- descriptives.scan(cox.RFS, top=50)
# Summary for top 50 most significant association results, sorted by P1df
```

```
# adjust for BRSLW and age

cox.RFS.brsage <- mlreg(GASurv(RFS.n, cens.RFS) ~ BRSLW + age, data)

RFS.7.brsage <- cox.RFS.brsage@ results[which(cox.RFS.brsage@ results$
 P1df < 10^(-7)), c("effB", "se_effB", "chi2.1df", "P1df" )]
RFS.4.brsage <- cox.RFS.brsage@ results[which(cox.RFS.brsage@ results$
 P1df < 10^(-4)), c("effB", "se_effB", "chi2.1df", "P1df" )]
dim(RFS.7.brsage)   # 0, 4
dim(RFS.4.brsage)   # 9, 4

RFS.brsage.all <- cox.RFS.brsage@ results[, c("effB", "se_effB",
 "chi2.1df", "P1df" )]
RFS.brsage.all = data.frame(SNP,CHR,BP,RFS.brsage.all)

write.table(RFS.brsage.all, sep="\t", file="RFS.brsage.all.txt",
 row.names=FALSE)

#--- QQ plot and manhattan plot ---#

#-- plot RFS --#
plot.RFS.brsage <- data.frame(CHR=data@gtdata@chromosome,
 BP=data@gtdata@map, P=cox.RFS.brsage@results$P1df)
plot.RFS.brsage$CHR <-
 as.numeric(as.character(drop.levels(plot.RFS.brsage$CHR)))
dim(plot.RFS.brsage)   # 108300 obs, 3 var
head(plot.RFS.brsage)

qq(plot.RFS.brsage$P, main="Q-Q plot for RFS.brsage")


### Manhattan plot of cox.RFS and find the most significant association
 SNPs ###
plot(cox.RFS.brsage, ylim=c(0,6), pch=19, main="Manhattan plot for
 RFS.brsage")

bestHits.brsage <- descriptives.scan(cox.RFS.brsage,top=50)
# Summary for top 50 most significant association results, sorted by P1df



### Manhattan plot of cox.RFS but only plot Chr1 and Chr7 ###
plot.RFS.brsage.1.7 <- plot.RFS.brsage[plot.RFS.brsage$CHR == 1 |
 plot.RFS.brsage$CHR == 7, ]   # 15799

### find gene annotation for bestHits.brsage
bestHits.brsage. <- data.frame(rownames(bestHits.brsage), bestHits.brsage)
 # make row.names as a new column
colnames(bestHits.brsage.)[1] <- "Name"

anno <- read.table("ImmunoChip_GeneAnnotation.txt", header=T, fill=T)   #
 197076

bestHits.brsage1 <- merge(anno, bestHits.brsage., by="Name")   # 50 obs,
 23 var
head(bestHits.brsage1)
names(bestHits.brsage1)
```

```
bestHits50 <- bestHits.brsage1[, c(1:5,10,18)]
head(bestHits50)
bestHits50 <- bestHits50[order(bestHits50$P1df), ]

write.csv(bestHits50, file="bestHits50.csv", row.names=FALSE)


### prepare file for LocusZoom
plot.RFS.brsage1 <- data.frame(rownames(plot.RFS.brsage), plot.RFS.brsage)
 # make row.names as a new column
colnames(plot.RFS.brsage1)[1] <- "SNP"

write.csv(plot.RFS.brsage1, file="~/Dropbox/thesis/5-11-
 15/LocusZoom/RFS_brsage.csv", row.names=FALSE)

write.table(plot.RFS.brsage1, file="~/Dropbox/thesis/5-11-
 15/LocusZoom/RFS_brsage.txt", row.names=FALSE)



### Test for Association at Gene Level ###

#setwd("~/Dropbox/thesis/5-19-15/coxKM_SKAT")

#-- Genotype data
tped <- read.table("Clean-Mel_IC.tped")   # 108300, 414
colnames(tped)[2] <- "Name"

anno <- read.table("ImmunoChip_GeneAnnotation.txt", header=T, fill=T)   #
 197076, 8

merge <- merge(anno, tped, by="Name")   # 108300, 421
head(merge)
names(merge)

Interested.Gene <- merge[, c(1:5)]   # 108300, 5
head(Interested.Gene)

### delete X, Y chromosomes
Interested.Gene <- Interested.Gene[Interested.Gene$Chr != "X" &
 Interested.Gene$Chr != "Y", ]   # 107816

### modify INTERGENIC GeneSymbol into separated rows
INTERGENIC <- Interested.Gene[Interested.Gene$GeneLocation ==
 "INTERGENIC", ]   # 59381
noINTERGENIC <- Interested.Gene[Interested.Gene$GeneLocation !=
 "INTERGENIC", ]   # 48435

INTERGENIC$GeneSymbol1 <- gsub("[|].+$", "", INTERGENIC$GeneSymbol)
INTERGENIC$GeneSymbol2 <- gsub("^.+[|]", "", INTERGENIC$GeneSymbol)
INTERGENIC$GeneSymbol <- NULL
head(INTERGENIC)
names(INTERGENIC)

Intergenic1 <- INTERGENIC[, c(1,2,3,5,4)]
Intergenic2 <- INTERGENIC[, c(1,2,3,6,4)]
colnames(Intergenic1)[4] <- "GeneSymbol"
colnames(Intergenic2)[4] <- "GeneSymbol"

Interested_Gene <- rbind(Intergenic1, Intergenic2, noINTERGENIC)
# 167197
```

```
Interested_Gene <- Interested_Gene[order(Interested_Gene$Coordinate), ]
Interested_Gene2 <- Interested_Gene[order(-Interested_Gene$Coordinate), ]

#write.csv(Interested_Gene, file="Interested_Gene.csv", row.names=FALSE)
#write.csv(Interested_Gene2, file="Interested_Gene2.csv", row.names=FALSE)

# Dichotomize RFS
Dicho.RFS <- pheno$cens.RFS
Dicho.RFS[pheno$RFS.n < 3*365 & pheno$cens.RFS == 1] <- 1
Dicho.RFS[pheno$RFS.n > 3*365 & pheno$cens.RFS == 1] <- 0
Dicho.RFS[pheno$RFS.n > 3*365 & pheno$cens.RFS == 0] <- 0
Dicho.RFS[is.na(pheno$RFS.n)] <- NA
table(Dicho.RFS)

#-- Phenotype data
pheno <- read.table("pheno-clean.txt", header = TRUE)

#-- Code genotype as (0, 1, 2)
hwe <- read.table("plink.hwe", header=T)   # 324900, 9
hwe <- hwe[which(hwe$TEST == "ALL"), ]   # 108300, 9

Minor.allele <- hwe$A1
Major.allele <- hwe$A2

Code0 <- paste(Major.allele, Major.allele, sep="_")
Code1a <- paste(Major.allele, Minor.allele, sep="_")
Code1b <- paste(Minor.allele, Major.allele, sep="_")
Code2 <- paste(Minor.allele, Minor.allele, sep="_")

library(coxKM)
library(SKAT)

time=proc.time()

Result.OS.IBS <- NULL
Result.OS.linear <- NULL
Result.RFS.IBS <- NULL
Result.RFS.linear <- NULL

unique1 <- unique(Interested_Gene$GeneSymbol)
length(unique1)   #12384
unique2 <- unique(Interested_Gene2$GeneSymbol)
length(unique2)   #12384

Gene.name <- NULL
Chrom <- NULL
Chr.pos <- NULL
Number.SNP <- NULL
Skat.Dicho.RFS <- NULL
Skat.Dicho.OS <- NULL


for(s in 1:length(unique1)){
  set.seed(1)

  Chr <- Interested_Gene$Chr[which.max(Interested_Gene$GeneSymbol ==
 unique1[s])]
  Start <- Interested_Gene$Coordinate[which.max(Interested_Gene$GeneSymbol
 == unique1[s])]
  Stop <-
 Interested_Gene2$Coordinate[which.max(Interested_Gene2$GeneSymbol ==
```

```
unique1[s])]

  Geno <- tped[which(tped$V1 == Chr & tped$V4 > Start & tped$V4 < Stop), ]

  if(dim(Geno)[1] == 0){
    Result.OS.IBS <- rbind(Result.OS.IBS, c(NA, NA, 0, NA, NA))
    Result.OS.linear <- rbind(Result.OS.linear, c(NA, NA, 0, NA, NA))
    Result.RFS.IBS <- rbind(Result.RFS.IBS, c(NA, NA, 0, NA, NA))
    Result.RFS.linear <- rbind(Result.RFS.linear, c(NA, NA, 0, NA, NA))
  }

  if(dim(Geno)[1] == 1){
    Result.OS.IBS <- rbind(Result.OS.IBS, c(NA, NA, 1, NA, NA))
    Result.OS.linear <- rbind(Result.OS.linear, c(NA, NA, 1, NA, NA))
    Result.RFS.IBS <- rbind(Result.RFS.IBS, c(NA, NA, 1, NA, NA))
    Result.RFS.linear <- rbind(Result.RFS.linear, c(NA, NA, 1, NA, NA))
  }

  if(dim(Geno)[1] > 1){
    Gene.name <- c(Gene.name,
as.character(unique(Interested_Gene$Name)[s]))
    Chrom <- c(Chrom, Interested_Gene$Chr[s])
    Chr.pos <- c(Chr.pos, Start)
    Number.SNP <- c(Number.SNP, dim(Geno)[1])

    n = ((dim(Geno)[2]-4)/2)
    genotype <- matrix(NA, nrow=dim(Geno)[1], ncol=n)
    ind <- 5

    for(i in 1:n){
      genotype[,i] <- paste(Geno[, ind], Geno[,ind+1], sep="_")
      ind <- ind+2
    }

    genotype.012 <- genotype

    for(i in 1:dim(genotype)[2]){
      genotype.012[which(genotype[,i] %in% Code0), i] <- 0
      genotype.012[which(genotype[,i] %in% Code1a), i] <- 1
      genotype.012[which(genotype[,i] %in% Code1b), i] <- 1
      genotype.012[which(genotype[,i] %in% Code2), i] <- 2
    }

    table(genotype.012)

    genotype.012 <- matrix(as.numeric(genotype.012),
ncol=dim(genotype.012)[2])
    dim(genotype.012)
    Z = t(genotype.012)


    ### Dichotomize RFS with SKAT-O
    obj <- SKAT_Null_Model(Dicho.RFS ~ 1, out_type="D")
    Skat.Dicho.RFS <- c(Skat.Dicho.RFS, SKAT(Z, obj,
method="optimal.adj")$p.value)

    ### coxKM
    fit <- coxKM(Z=t(genotype.012), U=pheno$OS.n, Delta=pheno$cens.OS,
kernel="IBS")
    Result.OS.IBS <- rbind(Result.OS.IBS, unlist(fit))
```

```
    fit <- coxKM(Z=t(genotype.012), U=pheno$OS.n, Delta=pheno$cens.OS,
 kernel="linear")
    Result.OS.linear <- rbind(Result.OS.linear, unlist(fit))

    Z <- matrix(t(genotype.012), ncol=dim(genotype.012)[1])

    fit <- coxKM(Z=Z, U=pheno$RFS.n, Delta=pheno$cens.RFS, kernel="IBS")
    Result.RFS.IBS <- rbind(Result.RFS.IBS, unlist(fit))

    fit <- coxKM(Z=Z, U=pheno$RFS.n, Delta=pheno$cens.RFS,
 kernel="linear")
    Result.RFS.linear <- rbind(Result.RFS.linear, unlist(fit))

  }
}

rownames(Result.OS.IBS) <- unique1
rownames(Result.OS.linear) <- unique1
rownames(Result.RFS.IBS) <- unique1
rownames(Result.RFS.linear) <- unique1

Gene <- Interested_Gene[!duplicated(Interested_Gene$GeneSymbol), ]
Result.OS.IBS<- cbind(Result.OS.IBS, Gene)
Result.OS.linear<- cbind(Result.OS.linear, Gene)
Result.RFS.IBS<- cbind(Result.RFS.IBS, Gene)
Result.RFS.linear<- cbind(Result.RFS.linear, Gene)

Result.Dicho.RFS<- cbind(Gene.name, Chrom, Chr.pos, Number.SNP,
 Skat.Dicho.RFS)


write.table(Result.OS.IBS, "ALL_Result.OS.IBS.txt")
write.table(Result.OS.linear, "ALL_Result.OS.linear.txt")
write.table(Result.RFS.IBS, "ALL_Result.RFS.IBS.txt")
write.table(Result.RFS.linear, "ALL_Result.RFS.linear.txt")
write.table(Result.Dicho.RFS, "Result.Dicho.RFS.txt")

total_time=(proc.time()-time)/60
total_time



### Plot survival curves for significant signal rs6944473 (DGKB) ###

setwd("~/Dropbox/thesis/6-16-15")

noDGKB <- read.table("Clean-Mel_IC_noDGKB.raw", header=T)

pheno.clean <- read.table("pheno-clean.txt", header=T)   # 205 obs
colnames(pheno.clean)[1] <- "IID"

pheno.clean$RFS.year <- pheno.clean$RFS.n / 365

DGKB_pheno <- merge(pheno.clean, noDGKB, by="IID")
names(DGKB_pheno)

# draw survival curve for the most significant SNP rs6944473
library(survival)

surv <- survfit(Surv(RFS.year, cens.RFS) ~ rs6944473_G, data = DGKB_pheno)
```

```r
plot(surv, col=c("blue","red","green"), lty=1, xlab="Time(year)",
 ylab="Survival proportion")
legend("topright", legend=c("0","1","2"), col=c("blue","red","green"),
 horiz=FALSE, y.intersp=0.9, bty='n', cex=0.8, lty=1, pt.cex = 1,
 title="Genotype number of minor allele:")
mtext("I", side=3, line=1, at=3)

table(DGKB_pheno$rs6944473_G)   # 0:185, 1:19, 2:1


### dichotomize rs6944473_G as 0 and 1/2 because we only have 1 subject
 has 2 minor alleles
Dicho <- DGKB_pheno$rs6944473_G
Dicho[DGKB_pheno$rs6944473_G == 0] <- 0
Dicho[DGKB_pheno$rs6944473_G == 1 | DGKB_pheno$rs6944473_G == 2] <- 1
table(Dicho)

DGKB_pheno1 <- cbind(DGKB_pheno, Dicho)   # 0:185, 1:20

surv1 <- survfit(Surv(RFS.year, cens.RFS) ~ Dicho, data = DGKB_pheno1)

plot(surv1, col=c("blue","red"), lty=1, xlab="Time(year)", ylab="Survival
 proportion")
legend("topright", legend=c("Minor allele carrier", "Wild type
 homozygous"), col=c("blue","red"), horiz=FALSE, y.intersp=0.9, bty='n',
 cex=0.8, lty=1, pt.cex = 1, title="Genotype group:")
mtext("II", side=3, line=1, at=3)

# log-rank test
survdiff<-survdiff(Surv(RFS.year, cens.RFS) ~ Dicho, data=DGKB_pheno1)
survdiff   # p = 3.49e-07
```

BIBLIOGRAPHY

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296. doi: 10.1093/bioinformatics/btm108

Bleyer, A., O'leary, M., Barr, R., & Ries, L. (2006). Cancer epidemiology in older adolescents and young adults 15 to 29 years of age, including SEER incidence and survival: 1975-2000. *Cancer epidemiology in older adolescents and young adults 15 to 29 years of age, including SEER incidence and survival: 1975-2000*.

Cai, T., Tonini, G., & Lin, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics, 67*(3), 975-986. doi: 10.1111/j.1541-0420.2010.01544.x

Cortes, A., & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res Ther, 13*(1), 101.

Disis, M. L. (2011). Immunologic biomarkers as correlates of clinical response to cancer immunotherapy. *Cancer Immunol Immunother, 60*(3), 433-442. doi: 10.1007/s00262-010-0960-8

Herrera-Gonzalez, N. E. (2013). *Interaction Between the Immune System and Melanoma*: INTECH Open Access Publisher.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Kirkwood, J. M., Manola, J., Ibrahim, J., Sondak, V., Ernstoff, M. S., & Rao, U. (2004). A pooled analysis of eastern cooperative oncology group and intergroup trials of adjuvant high-dose interferon for melanoma. *Clin Cancer Res, 10*(5), 1670-1677.

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C., & Blum, R. H. (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. *J Clin Oncol, 14*(1), 7-17.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., . . . Lin, X. (2012). Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet, 91*(2), 224-237. doi: 10.1016/j.ajhg.2012.06.007

Lin, X., Cai, T., Wu, M. C., Zhou, Q., Liu, G., Christiani, D. C., & Lin, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol, 35*(7), 620-631. doi: 10.1002/gepi.20610

Parkes, M., Cortes, A., van Heel, D. A., & Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics, 14*(9), 661-673.

Parkin, D., Mesher, D., & Sasieni, P. (2011). 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010. *British journal of cancer, 105*, S66-S69.

Plescia, M., Protzel Berman, P., & White, M. C. (2011). Melanoma surveillance in the United States. *J Am Acad Dermatol, 65*(5 Suppl 1), S1-2. doi: 10.1016/j.jaad.2011.05.031

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Tartour, E., Dorval, T., Mosseri, V., Deneux, L., Mathiot, C., Brailly, H., . . . Fridman, W. H. (1994). Serum interleukin 6 and C-reactive protein levels correlate with resistance to IL-2 therapy and poor survival in melanoma patients. *Br J Cancer, 69*(5), 911-913.

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., . . . Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet, Chapter 1*, Unit1.19. doi: 10.1002/0471142905.hg0119s68

Weale, M. E. (2010). Quality control for genome-wide association studies *Genetic Variation* (pp. 341-372): Springer.

Wittke, F., Hoffmann, R., Buer, J., Dallmann, I., Oevermann, K., Sel, S., . . . Atzpodien, J. (1999). Interleukin 10 (IL-10): an immunosuppressive factor and independent predictor in patients with metastatic renal cell carcinoma. *Br J Cancer, 79*(7-8), 1182-1184. doi: 10.1038/sj.bjc.6690189