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
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

**PREFACE** .......................................................................................................................... IX

1.0 **INTRODUCTION** ........................................................................................................ 1

  1.1 **MELANOMA** ....................................................................................................... 1

    1.1.1 General Introduction ....................................................................................... 1

    1.1.2 Immunotherapy ................................................................................................ 2

    1.1.3 Clinical Biomarker for Melanoma Patients on Immunotherapy .............. 3

  1.2 **E1697 STUDY** ................................................................................................... 3

  1.3 **IMMUNOCHIP** ................................................................................................... 3

  1.4 **GOAL OF THE STUDY** ..................................................................................... 4

2.0 **METHODS AND RESULTS** ...................................................................................... 5

  2.1 **STUDY SAMPLE** ............................................................................................... 5

  2.2 **DATA** .................................................................................................................... 5

    2.2.1 Starting Files ..................................................................................................... 5

    2.2.2 PLINK ............................................................................................................... 7

    2.2.3 Binary Files ....................................................................................................... 7

    2.2.4 Quality Control ................................................................................................ 8

      2.2.4.1 Relationship Check ............................................................................... 8

      2.2.4.2 Missing Data Check ............................................................................. 9
2.2.4.3 Population Structure Check ............................................................... 11

2.3 CLINICAL FACTORS ..................................................................................... 16

2.3.1 Model Selection ............................................................................................... 16

2.4 TEST FOR ASSOCIATION AT SNP LEVEL ............................................... 18

2.5 TEST FOR ASSOCIATION AT GENE LEVEL ........................................... 22

2.5.1 SKAT ............................................................................................................... 23

2.5.2 CoxKM ............................................................................................................ 24

2.6 CROSS REFERENCES OF DIFFERENT ANALYSIS RESULTS .......... 26

2.6.1 Cross References of Gene Level Analysis .................................................... 26

2.6.2 Cross References between Gene and SNP Level Analysis ...................... 28

3.0 DISCUSSION AND FUTURE WORKS .............................................................. 30

3.1 GENERAL DISCUSSION ................................................................................ 30

3.2 SIGNIFICANT SIGNALS ................................................................................ 31

3.3 FUTURE WORKS ............................................................................................. 32

APPENDIX A: ADDITIONAL TABLES ................................................................................. 33

APPENDIX B: R CODE ............................................................................................................ 43

BIBLIOGRAPHY ....................................................................................................................... 56
**LIST OF TABLES**

Table 1. Univariate model of RFS ................................................................................................................................................. 17
Table 2. Multivariable model of RFS .......................................................................................................................................... 17
Table 3. Summary for top 10 most significant association results at SNP level ................................................................. 18
Table 4. Top 10 SKAT results ....................................................................................................................................................... 24
Table 5. Top 10 coxKM results using IBS and linear kernel ........................................................................................................... 25
Table 6. Cross references of SKAT and coxKM analysis results ................................................................................................. 27
Table 7. Cross references of SKAT and SNP level analysis results ............................................................................................ 28
Table 8. Cross references of coxKM and SNP level analysis results .......................................................................................... 29
Table 9. Summary for top 50 most significant association results at SNP level ................................................................. 33
Table 10. Top 50 SKAT results ...................................................................................................................................................... 36
Table 11. Top 50 coxKM results using IBS and linear kernel ......................................................................................................... 39
LIST OF FIGURES

Figure 1. Relationship Check plot .................................................................................................. 9
Figure 2. Missing data check by individual .................................................................................. 10
Figure 3. Missing data check by SNP ........................................................................................... 11
Figure 4. Population structure matrix ........................................................................................... 12
Figure 5. Population structure plot between C1 and C2 ............................................................... 13
Figure 6. Population structure plot between C2 and C3 ............................................................... 13
Figure 7. Population structure plot between C1 and C3 ............................................................... 14
Figure 8. Population structure plot between C3 and C4 ............................................................... 14
Figure 9. Population structure plot between C2 and C4 ............................................................... 15
Figure 10. Population structure plot between C1 and C4 ............................................................. 15
Figure 11. Manhattan plot ............................................................................................................. 19
Figure 12. Manhattan plot for Chromosome 1 and 7 .................................................................... 20
Figure 13. QQ plot ........................................................................................................................ 20
Figure 14. Regional plot of area surrounding interested SNPs of –log (P values) using LocusZoom ................................................................................................................................... 22
Figure 15. Kaplan-Meier curves of RFS by rs6944473 genotype ................................................ 31
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 $\alpha$). Interferon is a man-made copy of human protein. It helps the immune system to fight viral infections. Interferon Alpha-2b (IFN $\alpha$-2b) treatment is often given as a shot under the skin. Studies showed that using IFN $\alpha$-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)
Column 4: BRSLW_THICKNESS (tumor Breslow’s depth in millimeters)

Column 5: CLARK_LVL (Clark’s level)

Column 6: LDH_RS (Lactate dehydrogenase value)

Column 7: LDH_ULN (LDH upper limit of normal)

Column 8: PIG (Pigmentation: 1= amelanotic, 2= melanotic, -1= unknown)

Column 9: PS (ECOG Performance status)

Column 10: ULCER_YN (Ulceration: 1=no, 2=yes, -1=unknown)

Column 11: surv_y (survival years)

Column 12: rfs (relapse free survival years)

Column 13: rfs_ind (relapse free survival index: 1=event, 0=censored)

Column 14: surv_s (survival index: 1=event, 0=censored)

Column 15: age (age at diagnosis)

*Immu*no*Chip* _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
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 is a plot showing the information of relative pairs of individuals. Z0 and Z1 denote the probability that individual 1 and individual 2 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.
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 properly 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
```
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 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
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

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

*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

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Parameter Estimate</th>
<th>Wald Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor Breslow’s Thickness</td>
<td>0.801</td>
<td>0.036</td>
</tr>
<tr>
<td>Age at Diagnosis</td>
<td>0.025</td>
<td>0.018</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Coordinate</th>
<th>Gene*</th>
<th>Location</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6944473</td>
<td>7</td>
<td>14326377</td>
<td>DGKB</td>
<td>INTRON</td>
<td>1.42E-06</td>
</tr>
<tr>
<td>rs10495124</td>
<td>1</td>
<td>217568816</td>
<td>LYPLAL1</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm_12_2178130</td>
<td>12</td>
<td>2178130</td>
<td>CACNA1C</td>
<td>INTRON</td>
<td>3.96E-05</td>
</tr>
<tr>
<td>seq-rs2784110</td>
<td>1</td>
<td>197047009</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs17591522</td>
<td>1</td>
<td>217600391</td>
<td>LYPLAL1</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs11942401</td>
<td>4</td>
<td>188052244</td>
<td>FAT</td>
<td>ZFP42</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs6704463</td>
<td>1</td>
<td>217614448</td>
<td>LYPLAL1</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs2095403</td>
<td>1</td>
<td>62632898</td>
<td>ANKRD38</td>
<td>USP1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs2839235</td>
<td>21</td>
<td>46625020</td>
<td>PCNT</td>
<td>INTRON</td>
<td>8.30E-05</td>
</tr>
<tr>
<td>rs3860187</td>
<td>10</td>
<td>49639139</td>
<td>WDFY4</td>
<td>INTRON</td>
<td>0.0001103</td>
</tr>
</tbody>
</table>

*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 (-log₁₀ 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.
A Quantile-Quantile (QQ) plot of the SNP level analysis is shown in Figure 13. It plots the observed $-\log_{10} p$-values against the expected $-\log_{10} p$-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.
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/](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

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

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

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

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

<table>
<thead>
<tr>
<th>GENE</th>
<th>Chr</th>
<th>SNP*</th>
<th>SNP.Coordinate</th>
<th>GeneLocation</th>
<th>p.value.SNP</th>
<th>p.value.SKAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGKB</td>
<td>7</td>
<td>rs6944473</td>
<td>14326377</td>
<td>INTRON</td>
<td>1.42E-06</td>
<td>0.000147085</td>
</tr>
</tbody>
</table>

*SNP is from the SNP level analysis
Table 8. Cross references of coxKM and SNP level analysis results

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

*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](image)

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

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Coordinate</th>
<th>GeneSymbol*</th>
<th>Location</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6944473</td>
<td>7</td>
<td>14326377</td>
<td>DGKB</td>
<td>INTRON</td>
<td>1.42E-06</td>
</tr>
<tr>
<td>rs10495124</td>
<td>1</td>
<td>217568816</td>
<td>LYP\textsubscript{L}AL\textsubscript{1}</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm\textsubscript{12}_2178130</td>
<td>12</td>
<td>2178130</td>
<td>CACNA1C</td>
<td>INTRON</td>
<td>3.96E-05</td>
</tr>
<tr>
<td>seq-rs2784110</td>
<td>1</td>
<td>197047009</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs17591522</td>
<td>1</td>
<td>217600391</td>
<td>LYP\textsubscript{L}AL\textsubscript{1}</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs11942401</td>
<td>4</td>
<td>188052244</td>
<td>FAT</td>
<td>ZFP42</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs6704463</td>
<td>1</td>
<td>217614448</td>
<td>LYP\textsubscript{L}AL\textsubscript{1}</td>
<td>LOC728510</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs2095403</td>
<td>1</td>
<td>62632898</td>
<td>ANKRD38</td>
<td>USP1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs2839235</td>
<td>21</td>
<td>46625020</td>
<td>PCNT</td>
<td>INTRON</td>
<td>8.30E-05</td>
</tr>
<tr>
<td>rs3860187</td>
<td>10</td>
<td>49639139</td>
<td>WDFY4</td>
<td>INTRON</td>
<td>0.0001103</td>
</tr>
<tr>
<td>rs9629920</td>
<td>10</td>
<td>49629651</td>
<td>LOC100132924</td>
<td>INTRON</td>
<td>0.000113517</td>
</tr>
<tr>
<td>imm\textsubscript{12}_2187865</td>
<td>12</td>
<td>2187865</td>
<td>CACNA1C</td>
<td>INTRON</td>
<td>0.000113531</td>
</tr>
<tr>
<td>rs9309074</td>
<td>2</td>
<td>41821802</td>
<td>SLC8\textsubscript{A}1</td>
<td>LDHAL3</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm\textsubscript{16}_31279175</td>
<td>16</td>
<td>31279175</td>
<td>ITGAX</td>
<td>CODING</td>
<td>0.000147564</td>
</tr>
<tr>
<td>seq-rs10800591</td>
<td>1</td>
<td>197063314</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Chromosome</td>
<td>Position</td>
<td>Feature</td>
<td>Distance from Reference</td>
<td>P-Value</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>----------</td>
<td>---------</td>
<td>------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>seq-rs12406470</td>
<td>1</td>
<td>197062743</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>seq-rs6427752</td>
<td>1</td>
<td>197062012</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>1kg_10_59547220</td>
<td>10</td>
<td>59547220</td>
<td>ZWINT</td>
<td>IPMK</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>seq-rs10800590</td>
<td>1</td>
<td>197042798</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs2226007</td>
<td>1</td>
<td>161860160</td>
<td>NUF2</td>
<td>LOC729952</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs13221118</td>
<td>7</td>
<td>156493383</td>
<td>MNX1</td>
<td>INTRON</td>
<td>0.000289407</td>
</tr>
<tr>
<td>rs7837005</td>
<td>8</td>
<td>72161586</td>
<td>XKR9</td>
<td>EYA1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm_1_67442201</td>
<td>1</td>
<td>67442201</td>
<td>IL23R</td>
<td>INTRON</td>
<td>0.000330605</td>
</tr>
<tr>
<td>rs10845202</td>
<td>12</td>
<td>10726132</td>
<td>STYK1</td>
<td>CSDA</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs270793</td>
<td>8</td>
<td>56012168</td>
<td>LOC100128419</td>
<td>XKR4</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs6927768</td>
<td>6</td>
<td>170708910</td>
<td>TBP</td>
<td>INTRON</td>
<td>0.000361202</td>
</tr>
<tr>
<td>rs2986738</td>
<td>1</td>
<td>6470257</td>
<td>PLEKHG5</td>
<td>INTRON</td>
<td>0.000382766</td>
</tr>
<tr>
<td>imm_9_138290709</td>
<td>9</td>
<td>138290709</td>
<td>QSOX2</td>
<td>DKFZP434A062</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs1979302</td>
<td>18</td>
<td>55002156</td>
<td>SEC11C</td>
<td>GRP</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs12146041</td>
<td>1</td>
<td>173368784</td>
<td>TNN</td>
<td>INTRON</td>
<td>0.000454158</td>
</tr>
<tr>
<td>rs1156956</td>
<td>8</td>
<td>71917592</td>
<td>XKR9</td>
<td>EYA1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs7845516</td>
<td>8</td>
<td>71848376</td>
<td>XKR9</td>
<td>EYA1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>seq-rs2784114</td>
<td>1</td>
<td>197037793</td>
<td>PTPRC</td>
<td>LOC100131234</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs12153520</td>
<td>5</td>
<td>130446033</td>
<td>CHSY-2</td>
<td>HINT1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs9262632</td>
<td>6</td>
<td>31132787</td>
<td>HCG22</td>
<td>UTR</td>
<td>0.000521266</td>
</tr>
<tr>
<td>rs2332094</td>
<td>14</td>
<td>69139425</td>
<td>LOC100130174</td>
<td>KIAA0247</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs545152</td>
<td>1</td>
<td>96659092</td>
<td>LOC729977</td>
<td>LOC440595</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs990648</td>
<td>11</td>
<td>79730135</td>
<td>ODZ4</td>
<td>MGC33846</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>SNP</td>
<td>Chrom.</td>
<td>Position</td>
<td>Gene(s)</td>
<td>Feature</td>
<td>p-value</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>rs6864771</td>
<td>5</td>
<td>7850291</td>
<td>ADCY2</td>
<td>INTRON</td>
<td>0.000560977</td>
</tr>
<tr>
<td>rs9923190</td>
<td>16</td>
<td>85014134</td>
<td>LOC732275</td>
<td>LOC283904</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>rs11753208</td>
<td>6</td>
<td>31113411</td>
<td>LOC729792</td>
<td>INTRON</td>
<td>0.000602604</td>
</tr>
<tr>
<td>rs12155783</td>
<td>8</td>
<td>72161680</td>
<td>XKR9</td>
<td>EYA1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm_22_38042260</td>
<td>22</td>
<td>38042260</td>
<td>RPL3</td>
<td>INTRON</td>
<td>0.000610902</td>
</tr>
<tr>
<td>imm_1_195577864</td>
<td>1</td>
<td>195577864</td>
<td>CRB1</td>
<td>INTRON</td>
<td>0.00063566</td>
</tr>
<tr>
<td>rs7768644</td>
<td>6</td>
<td>31110080</td>
<td>LOC729792</td>
<td>INTRON</td>
<td>0.000692246</td>
</tr>
<tr>
<td>rs7733977</td>
<td>5</td>
<td>130466000</td>
<td>CHSY-2</td>
<td>HINT1</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>1kg_2_43478582</td>
<td>2</td>
<td>43478582</td>
<td>THADA</td>
<td>INTRON</td>
<td>0.000758235</td>
</tr>
<tr>
<td>rs9283781</td>
<td>5</td>
<td>82103862</td>
<td>LOC92270</td>
<td>TMEM167</td>
<td>INTERGENIC</td>
</tr>
<tr>
<td>imm_12_2194668</td>
<td>12</td>
<td>2194668</td>
<td>CACNA1C</td>
<td>INTRON</td>
<td>0.000761736</td>
</tr>
<tr>
<td>imm_9_138290450</td>
<td>9</td>
<td>138290450</td>
<td>QSOX2</td>
<td>DKFZP434A062</td>
<td>INTERGENIC</td>
</tr>
</tbody>
</table>

*Gene on which the SNP is located. When the SNP is located in between two genes, it is denoted as GENE1|GENE2.*
<table>
<thead>
<tr>
<th>GENE</th>
<th>Chr</th>
<th>Start</th>
<th>Stop</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGKB</td>
<td>19</td>
<td>14136077</td>
<td>161585680</td>
<td>0.000147085</td>
</tr>
<tr>
<td>LOC340268</td>
<td>4</td>
<td>9834067</td>
<td>185223182</td>
<td>0.000332623</td>
</tr>
<tr>
<td>GABBR2</td>
<td>3</td>
<td>100274966</td>
<td>38086931</td>
<td>0.000336071</td>
</tr>
<tr>
<td>FBXL17</td>
<td>2</td>
<td>107045500</td>
<td>34732070</td>
<td>0.000684611</td>
</tr>
<tr>
<td>HTRA1</td>
<td>5</td>
<td>124216620</td>
<td>30136403</td>
<td>0.000800605</td>
</tr>
<tr>
<td>DUSP10</td>
<td>22</td>
<td>219470464</td>
<td>1629929</td>
<td>0.000835814</td>
</tr>
<tr>
<td>HLX</td>
<td>12</td>
<td>219074478</td>
<td>1635423</td>
<td>0.000836453</td>
</tr>
<tr>
<td>FBLN7</td>
<td>12</td>
<td>112615980</td>
<td>32848649</td>
<td>0.000907298</td>
</tr>
<tr>
<td>CPB2</td>
<td>23</td>
<td>45527945</td>
<td>84559380</td>
<td>0.000928026</td>
</tr>
<tr>
<td>CTCFL</td>
<td>10</td>
<td>55500243</td>
<td>68295641</td>
<td>0.000944519</td>
</tr>
<tr>
<td>ULK4</td>
<td>4</td>
<td>41834977</td>
<td>92844857</td>
<td>0.001026475</td>
</tr>
<tr>
<td>ADAMTS2</td>
<td>12</td>
<td>178485124</td>
<td>10945296</td>
<td>0.001097248</td>
</tr>
<tr>
<td>GABRA5</td>
<td>14</td>
<td>24573461</td>
<td>138614368</td>
<td>0.001109697</td>
</tr>
<tr>
<td>GRB2</td>
<td>3</td>
<td>70823779</td>
<td>52366748</td>
<td>0.001656428</td>
</tr>
<tr>
<td>DDHD1</td>
<td>5</td>
<td>52524755</td>
<td>70796945</td>
<td>0.001727406</td>
</tr>
<tr>
<td>LOC645434</td>
<td>5</td>
<td>139880112</td>
<td>24206432</td>
<td>0.002133753</td>
</tr>
<tr>
<td>KIAA0195</td>
<td>3</td>
<td>70915763</td>
<td>52191086</td>
<td>0.002172136</td>
</tr>
<tr>
<td>BMP4</td>
<td>11</td>
<td>52693414</td>
<td>70783451</td>
<td>0.002432467</td>
</tr>
<tr>
<td>LOC100131472</td>
<td>19</td>
<td>10233678</td>
<td>181965614</td>
<td>0.002537652</td>
</tr>
<tr>
<td>LOC729112</td>
<td>17</td>
<td>122598338</td>
<td>30505819</td>
<td>0.002589555</td>
</tr>
</tbody>
</table>

**Table 10.** Top 50 SKAT results
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Start Position</th>
<th>End Position</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC642278</td>
<td>4</td>
<td>556195</td>
<td>241014568</td>
<td>0.003086166</td>
</tr>
<tr>
<td>FGF21</td>
<td>5</td>
<td>53951133</td>
<td>69333750</td>
<td>0.003108941</td>
</tr>
<tr>
<td>GRIK4</td>
<td>2</td>
<td>119870238</td>
<td>30923014</td>
<td>0.003142867</td>
</tr>
<tr>
<td>C20orf19</td>
<td>17</td>
<td>20735221</td>
<td>150859452</td>
<td>0.003348033</td>
</tr>
<tr>
<td>C20orf74</td>
<td>17</td>
<td>20735221</td>
<td>150859711</td>
<td>0.003348033</td>
</tr>
<tr>
<td>LOC100130010</td>
<td>4</td>
<td>12094704</td>
<td>167311824</td>
<td>0.003372483</td>
</tr>
<tr>
<td>TMEM89</td>
<td>23</td>
<td>48633471</td>
<td>79449021</td>
<td>0.003482006</td>
</tr>
<tr>
<td>VARS2</td>
<td>4</td>
<td>30990256</td>
<td>119147430</td>
<td>0.003965721</td>
</tr>
<tr>
<td>MTMR2</td>
<td>13</td>
<td>95208901</td>
<td>41013631</td>
<td>0.004078789</td>
</tr>
<tr>
<td>SPON1</td>
<td>22</td>
<td>13882056</td>
<td>162191705</td>
<td>0.00457111</td>
</tr>
<tr>
<td>GINS3</td>
<td>13</td>
<td>56971665</td>
<td>67130788</td>
<td>0.00472719</td>
</tr>
<tr>
<td>SH3MD4</td>
<td>2</td>
<td>108973688</td>
<td>34522053</td>
<td>0.004736383</td>
</tr>
<tr>
<td>C10orf67</td>
<td>14</td>
<td>23654346</td>
<td>141102100</td>
<td>0.0051071</td>
</tr>
<tr>
<td>SULT1A1</td>
<td>21</td>
<td>28517358</td>
<td>129249695</td>
<td>0.005224223</td>
</tr>
<tr>
<td>LOC100132354</td>
<td>23</td>
<td>43866851</td>
<td>87430929</td>
<td>0.005596336</td>
</tr>
<tr>
<td>ADRB2</td>
<td>5</td>
<td>148201190</td>
<td>22583965</td>
<td>0.006717942</td>
</tr>
<tr>
<td>SH3TC2</td>
<td>5</td>
<td>148201190</td>
<td>22583965</td>
<td>0.006717942</td>
</tr>
<tr>
<td>BCAT2</td>
<td>5</td>
<td>53953425</td>
<td>69312471</td>
<td>0.006810132</td>
</tr>
<tr>
<td>FLJ44815</td>
<td>4</td>
<td>29723177</td>
<td>127285007</td>
<td>0.006961246</td>
</tr>
<tr>
<td>COL7A1</td>
<td>23</td>
<td>48579063</td>
<td>79458866</td>
<td>0.00724393</td>
</tr>
<tr>
<td>AK5</td>
<td>9</td>
<td>77498829</td>
<td>49458078</td>
<td>0.007419735</td>
</tr>
<tr>
<td>CCL1</td>
<td>4</td>
<td>29710751</td>
<td>127300962</td>
<td>0.007768454</td>
</tr>
<tr>
<td>C9orf11</td>
<td>14</td>
<td>27223272</td>
<td>132113876</td>
<td>0.007777693</td>
</tr>
<tr>
<td>Gene</td>
<td>ID</td>
<td>Value1</td>
<td>Value2</td>
<td>Value3</td>
</tr>
<tr>
<td>----------</td>
<td>----</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>PEMT</td>
<td>20</td>
<td>17420920</td>
<td>159311566</td>
<td>0.007846131</td>
</tr>
<tr>
<td>RAI1</td>
<td>18</td>
<td>17478733</td>
<td>159295042</td>
<td>0.007846131</td>
</tr>
<tr>
<td>SEMA3A</td>
<td>20</td>
<td>83166896</td>
<td>46044577</td>
<td>0.007899354</td>
</tr>
<tr>
<td>TBX15</td>
<td>2</td>
<td>118685496</td>
<td>31104111</td>
<td>0.008055778</td>
</tr>
<tr>
<td>ZNF516</td>
<td>3</td>
<td>71278968</td>
<td>51888761</td>
<td>0.008316658</td>
</tr>
<tr>
<td>C11orf49</td>
<td>23</td>
<td>47011024</td>
<td>80133854</td>
<td>0.008337512</td>
</tr>
<tr>
<td>LOC645000</td>
<td>4</td>
<td>60737084</td>
<td>61843768</td>
<td>0.008477575</td>
</tr>
</tbody>
</table>

Table 10 Continued
Table 11. Top 50 coxKM results using IBS and linear kernel

<table>
<thead>
<tr>
<th>GENE</th>
<th>n.marker.test</th>
<th>n.indiv</th>
<th>p.IBS</th>
<th>Q.IBS</th>
<th>df.IBS</th>
<th>p.linear</th>
<th>Q.linear</th>
<th>df.linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTRA1</td>
<td>2</td>
<td>205</td>
<td>1.00E-04</td>
<td>163.3571446</td>
<td>0.997232432</td>
<td>2.00E-04</td>
<td>318.2507069</td>
<td>1.0197953</td>
</tr>
<tr>
<td>FBXO32</td>
<td>4</td>
<td>205</td>
<td>6.00E-04</td>
<td>51.8164787</td>
<td>3.526320773</td>
<td>4.00E-04</td>
<td>208.1382071</td>
<td>3.552179132</td>
</tr>
<tr>
<td>FGF9</td>
<td>6</td>
<td>205</td>
<td>9.00E-04</td>
<td>53.7839697</td>
<td>3.490050315</td>
<td>0.0023</td>
<td>321.4549365</td>
<td>3.296177478</td>
</tr>
<tr>
<td>HLX</td>
<td>9</td>
<td>205</td>
<td>0.0012</td>
<td>49.1513785</td>
<td>4.810549398</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>DUSP10</td>
<td>8</td>
<td>205</td>
<td>0.0012</td>
<td>55.69015018</td>
<td>4.034794788</td>
<td>3.00E-04</td>
<td>455.078092</td>
<td>4.084936</td>
</tr>
<tr>
<td>SOCS6</td>
<td>16</td>
<td>205</td>
<td>0.0017</td>
<td>21.30837138</td>
<td>10.92939678</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>PEMT</td>
<td>4</td>
<td>205</td>
<td>0.0018</td>
<td>100.3035437</td>
<td>1.479564066</td>
<td>0.0021</td>
<td>401.2141747</td>
<td>1.492598419</td>
</tr>
<tr>
<td>RAI1</td>
<td>4</td>
<td>205</td>
<td>0.0018</td>
<td>100.3035437</td>
<td>1.479564066</td>
<td>0.0021</td>
<td>401.2141747</td>
<td>1.492598419</td>
</tr>
<tr>
<td>LOC642278</td>
<td>2</td>
<td>205</td>
<td>0.0019</td>
<td>133.8361564</td>
<td>1.066328026</td>
<td>0.001</td>
<td>267.6723128</td>
<td>1.034384141</td>
</tr>
<tr>
<td>KCNK1</td>
<td>2</td>
<td>205</td>
<td>0.0019</td>
<td>68.19834591</td>
<td>1.636928436</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>C20orf74</td>
<td>3</td>
<td>205</td>
<td>0.0023</td>
<td>78.02869894</td>
<td>1.874982136</td>
<td>0.0024</td>
<td>234.0860968</td>
<td>1.868850054</td>
</tr>
<tr>
<td>C20orf19</td>
<td>3</td>
<td>205</td>
<td>0.0023</td>
<td>78.02869894</td>
<td>1.874982136</td>
<td>0.0024</td>
<td>234.0860968</td>
<td>1.868850054</td>
</tr>
<tr>
<td>IL1F7</td>
<td>5</td>
<td>205</td>
<td>0.0033</td>
<td>45.87091418</td>
<td>1.903022303</td>
<td>0.0039</td>
<td>233.7458099</td>
<td>1.898427645</td>
</tr>
<tr>
<td>LOC729668</td>
<td>22</td>
<td>205</td>
<td>0.0034</td>
<td>52.51872205</td>
<td>2.941225234</td>
<td>0.0039</td>
<td>1151.468446</td>
<td>2.844106193</td>
</tr>
<tr>
<td>Gene</td>
<td>CD</td>
<td>FC</td>
<td>p-value</td>
<td>Log2FC</td>
<td>p-value</td>
<td>Log2FC</td>
<td>Log2FC</td>
<td>p-value</td>
</tr>
<tr>
<td>------------</td>
<td>----</td>
<td>----</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>BACH2</td>
<td>207</td>
<td>205</td>
<td>0.0035</td>
<td>28.79968437</td>
<td>8.606043205</td>
<td>0.0038</td>
<td>5867.49402</td>
<td>9.217885776</td>
</tr>
<tr>
<td>CCDC88C</td>
<td>3</td>
<td>205</td>
<td>0.0036</td>
<td>85.93053123</td>
<td>1.205689231</td>
<td>0.003</td>
<td>251.9361541</td>
<td>1.239301028</td>
</tr>
<tr>
<td>RRM1</td>
<td>3</td>
<td>205</td>
<td>0.0042</td>
<td>43.02581801</td>
<td>2.757766733</td>
<td>0.004</td>
<td>129.077454</td>
<td>2.74399433</td>
</tr>
<tr>
<td>FBLN7</td>
<td>2</td>
<td>205</td>
<td>0.0042</td>
<td>67.98504669</td>
<td>2.016110267</td>
<td>0.0087</td>
<td>113.9488161</td>
<td>1.904888143</td>
</tr>
<tr>
<td>PLEKKG5</td>
<td>2</td>
<td>205</td>
<td>0.0046</td>
<td>52.66691091</td>
<td>1.779263457</td>
<td>0.005</td>
<td>105.3338218</td>
<td>1.786707106</td>
</tr>
<tr>
<td>C14orf181</td>
<td>70</td>
<td>205</td>
<td>0.005</td>
<td>34.29378735</td>
<td>5.221603951</td>
<td>0.004</td>
<td>2370.634724</td>
<td>5.12025002</td>
</tr>
<tr>
<td>UBLCP1</td>
<td>111</td>
<td>205</td>
<td>0.0051</td>
<td>23.36087174</td>
<td>7.906729142</td>
<td>0.0053</td>
<td>2628.336445</td>
<td>8.059648632</td>
</tr>
<tr>
<td>MAP3K8</td>
<td>121</td>
<td>205</td>
<td>0.0052</td>
<td>28.06843994</td>
<td>7.416439869</td>
<td>0.0039</td>
<td>1151.468446</td>
<td>2.844106193</td>
</tr>
<tr>
<td>CHODL</td>
<td>2</td>
<td>205</td>
<td>0.0053</td>
<td>69.87374076</td>
<td>1.844069232</td>
<td>0.0055</td>
<td>139.7474815</td>
<td>1.80580238</td>
</tr>
<tr>
<td>LOC100132924</td>
<td>5</td>
<td>205</td>
<td>0.0055</td>
<td>42.50840709</td>
<td>2.032615732</td>
<td>0.0069</td>
<td>212.5420355</td>
<td>2.034380308</td>
</tr>
<tr>
<td>IDE</td>
<td>10</td>
<td>205</td>
<td>0.0056</td>
<td>36.50485026</td>
<td>3.234346832</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>ACTN1</td>
<td>62</td>
<td>205</td>
<td>0.0058</td>
<td>32.7955529</td>
<td>5.2595226</td>
<td>0.0053</td>
<td>2006.406783</td>
<td>5.162854725</td>
</tr>
<tr>
<td>LOC730134</td>
<td>6</td>
<td>205</td>
<td>0.0058</td>
<td>35.88389223</td>
<td>4.489042795</td>
<td>0.0065</td>
<td>215.3033534</td>
<td>4.380506614</td>
</tr>
<tr>
<td>LOC100131866</td>
<td>64</td>
<td>205</td>
<td>0.0059</td>
<td>34.90505968</td>
<td>4.542110238</td>
<td>0.0054</td>
<td>2223.256908</td>
<td>4.695085661</td>
</tr>
<tr>
<td>LOC100128781</td>
<td>56</td>
<td>205</td>
<td>0.006</td>
<td>40.62994815</td>
<td>3.530145757</td>
<td>0.0057</td>
<td>2269.892288</td>
<td>3.728449707</td>
</tr>
<tr>
<td>CBLN2</td>
<td>20</td>
<td>205</td>
<td>0.0062</td>
<td>16.17112482</td>
<td>13.3626406</td>
<td>0.0071</td>
<td>321.3913423</td>
<td>13.22349141</td>
</tr>
<tr>
<td>Gene</td>
<td>ID</td>
<td>Sample</td>
<td>p-value</td>
<td>Log2 Fold Change</td>
<td>q-value</td>
<td>Ave Log2FC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----</td>
<td>--------</td>
<td>---------</td>
<td>------------------</td>
<td>---------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHLDB1</td>
<td>45</td>
<td>205</td>
<td>0.0063</td>
<td>37.48481044</td>
<td>4.380171538</td>
<td>0.0053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTHL7</td>
<td>12</td>
<td>205</td>
<td>0.0063</td>
<td>26.17782462</td>
<td>6.714230682</td>
<td>0.0023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADRA1B</td>
<td>78</td>
<td>205</td>
<td>0.0064</td>
<td>23.196342</td>
<td>9.132998512</td>
<td>0.0038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRPL36</td>
<td>2</td>
<td>205</td>
<td>0.007</td>
<td>36.04350296</td>
<td>1.093684329</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIR3DL2</td>
<td>3</td>
<td>205</td>
<td>0.0071</td>
<td>56.49377593</td>
<td>2.033623651</td>
<td>0.0079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLEC2B</td>
<td>140</td>
<td>205</td>
<td>0.0074</td>
<td>16.91395194</td>
<td>12.56710791</td>
<td>0.0075</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC728727</td>
<td>11</td>
<td>205</td>
<td>0.0076</td>
<td>31.14237744</td>
<td>5.204061415</td>
<td>0.0087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ41046</td>
<td>8</td>
<td>205</td>
<td>0.0076</td>
<td>27.17208389</td>
<td>4.079984235</td>
<td>0.0068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLJ42418</td>
<td>8</td>
<td>205</td>
<td>0.0076</td>
<td>27.17208389</td>
<td>4.079984235</td>
<td>0.0068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DLC1</td>
<td>6</td>
<td>205</td>
<td>0.0082</td>
<td>33.05230852</td>
<td>4.132508543</td>
<td>0.0053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFDC12</td>
<td>3</td>
<td>205</td>
<td>0.0082</td>
<td>61.92950099</td>
<td>1.099297828</td>
<td>0.0127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX2</td>
<td>2</td>
<td>205</td>
<td>0.0083</td>
<td>50.71762835</td>
<td>2.021478056</td>
<td>0.0083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSCAML1</td>
<td>3</td>
<td>205</td>
<td>0.0084</td>
<td>24.35274809</td>
<td>2.550270076</td>
<td>0.0081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DYRK2</td>
<td>145</td>
<td>205</td>
<td>0.0087</td>
<td>26.0997574</td>
<td>6.63316068</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIP2C</td>
<td>6</td>
<td>205</td>
<td>0.0087</td>
<td>37.83638328</td>
<td>3.841283055</td>
<td>0.0089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC100131234</td>
<td>231</td>
<td>205</td>
<td>0.0087</td>
<td>19.27530844</td>
<td>9.438183814</td>
<td>0.0066</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11 Continued
<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FZD8</td>
<td>5</td>
<td>205</td>
<td>0.0089</td>
<td>36.69497966</td>
<td>3.091913567</td>
<td>0.0074</td>
<td>183.4748983</td>
</tr>
<tr>
<td>ULK4</td>
<td>7</td>
<td>205</td>
<td>0.0089</td>
<td>70.15212838</td>
<td>1.488121134</td>
<td>0.0091</td>
<td>485.790286</td>
</tr>
<tr>
<td>HLA-E</td>
<td>482</td>
<td>205</td>
<td>0.0092</td>
<td>28.35820247</td>
<td>5.175844034</td>
<td>0.0087</td>
<td>13783.93721</td>
</tr>
<tr>
<td>TNR</td>
<td>2</td>
<td>205</td>
<td>0.0093</td>
<td>54.69729222</td>
<td>1.953768899</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 11 Continued
### Quality Control ###

#### 1. Relationship Check ####

```r
# 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 ####

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

het <- read.table("Mel IC misscheck.het", header = T)
mis <- read.table("Mel IC misscheck.imiss", header = T)
het_miss <- merge(het, mis, 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$F_MISS > 0.05, ]
a1 <- b$IID  # 2 sample IDs
# SS0016, SS0093
```

---

43
# Missing data check by SNP

```r
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 + FS + 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)
surv.multi2
# 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)
surv.multi3
# 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)
surv.multi4
# 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)
surv.multi5
# 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)
surv.multi6
# 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)
surv.multi7
# 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)
surv.multi8
# Wald test p-values: BRSLW=0.0364*, age=0.0175*, PS=0.13
# PS is not significant, so no need to add it
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="	", 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="	", file="RFS.all.txt", row.names=FALSE)

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

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.brsage[, 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.brsage <- data.frame(rownames(plot.RFS.brsage), plot.RFS.brsage)
# make row.names as a new column
colnames(plot.RFS.brsage)[1] <- "SNP"
write.csv(plot.RFS.brsage, file="~/Dropbox/thesis/5-11-15/LocusZoom/RFS_brsage.csv", row.names=FALSE)
write.table(plot.RFS.brsage, 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
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 ==
unique2[s])]
    }

52
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))

53
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)
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

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


