UNIVERSITY OF TORONTO

CHL 7001 Final Project Report

Professor Wei Xu

Student Name: Shu Guo

7/18/2013

Genome Wide Study on Early -Stage Head and Neck Squamous Cell Cancer Survival

Abstract

Objective: The purpose of this study is to conduct genome wide search on biomarkers that are related to symptoms and survival of early-stage head and neck squamous cell cancer.

Method: I fit a logistic and Cox proportional hazard models for each of 541,903 SNPs that covers all the autosomal chromosomes from 1 to 22, adjusting for the first three principal components of the SNPs data and treatment group. The same methods were used for the second step analysis on a few SNPs that discovered with strong relationship with phenotype data, adjusting for three PCs and some other covariates.

Results: Several SNPs have small p values in the first step of genome wide search (for example: p = 0.000003508 for rs3847141 and p = 0.000000200395 for rs1472080), and the second step statistical studies further confirmed the strong relationship with the phenotype data.

Conclusion: although none of the SNPs is significant in genome wide level, several of them might have strong relationship with overall survival, and further investigation of result is needed.

Introduction:

As a result of HapMap project, together with the advancements in the genotyping technology, it is now much easier to conduct population-based genetic association studies which presents an exciting opportunity to uncover the genetic contributors to the complex human diseases. These studies aim to discover the novel associations between genetic sequence information derived from unrelated individuals and a measure of disease progression or disease status that can be reproduced and replicated robustly. Head and neck cancer is a very complex human disease refers to a group of biologically similar cancers that start in the oral cavity, nasal cavity, paranasal sinuses, pharynx, and larynx. 90% of head and neck cancers are squamous cell carcinomas.[1] Head and neck cancer is believed to have strong associations with certain environment and life style risk factors such as cigarette and alcohol consumption, UV light and certain strains of viruses.[2] The number of new cases of head and neck cancers in the United States was 40,490 in 2006, accounting for about 3% of adult malignancies. 11,170 patients died of their disease in 2006.[3] In this study, I conducted a genome wide search for the genetic markers that are related to the larynx acute maxillary, which is a important symptom for head and neck cancer, and the overall survival for the patients. This study focused on autosomal GSVs and excluded all the sex GSVs. Genotyping assay of 531 samples was done using the Illumina HumanHap 610K BeadChips platform on the 540 randomized trial patients. After conducting systematic quality control on the raw genotyping data, a total of 497 patients and 541903 SNPs were used in the final analysis. I used the statistical methods to evaluate the relationship between all the SNPs and larynx acute maxillary(lam) and overall survival.

Data Description

Genetic data: Genotype information for 541,903 single-nucleotide polymorphisms (SNP) that cover all the autosomal choromosomes is recorded in binary files and PLINK was used to manipulated these data files. Minor allele frequencies are all greater than 1% and no SNP has strong departure from HWE ($p < 0.001$). There is no individual has missing genotypes greater than 5% and no SNP has more than 1% missing genetic data. Identity-by-state (IBS) was conducted for all pairs of individuals by using PLINK to identify excessively similar pairs of subjects, and no IBS values are greater than 0.8. The first three principal components for the SNP data are available for the analysis.

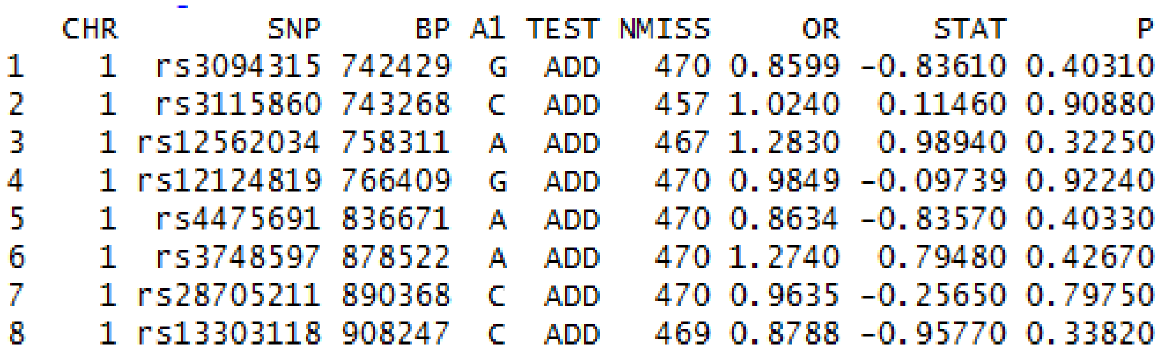
Phenotype trait and covariate variable: two phenotype traits were used in this study: larynx acute maxillary (lam) and overall survival, with individuals who has value of 0 are in the control group and those have value of 1 are in effected group. Covariate variables in the data set include treatment group, age, stage, gender, total RT dose, cancer stage and body mass index.

Statistical Methods:

Logistic regression models for lam: in the first step, logistic regression models were fit for each SNP, using lam as response phenotype variable. PLINK is an excellent software that can be used to conduct genome wide association studies, and I used PLINK to fit logistic regression models for all 541,903 SNPs, adjusted for three PCs and treatment group, and there are 475 individuals in this data set. The PLINK command that conduct this genome wide search is:

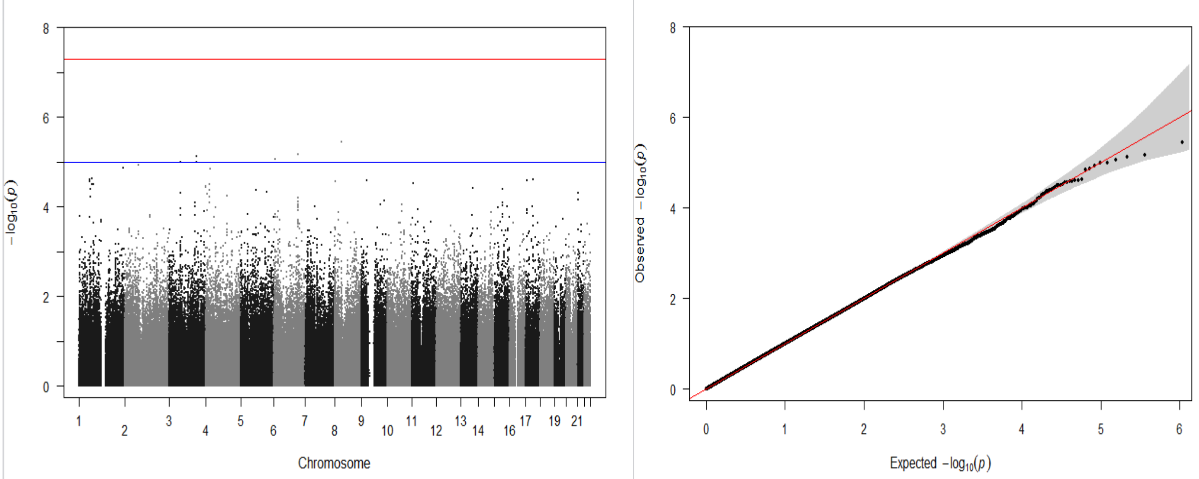
```
plink --bfile mydata --1 --logistic --pheno Phenolam.txt --covar CovarLam.txt --allow-no-sex --hide-covar
```

and name of the file that generated by this PLINK command is plink.assoc.logistic. We can use R to read this file and the first few lines of the data file is as follow:



```
    CHR         SNP     BP A1 TEST NMISS     OR     STAT       P
1     1   rs3094315 742429  G  ADD   470 0.8599 -0.83610 0.40310
2     1   rs3115860 743268  C  ADD   457 1.0240  0.11460 0.90880
3     1  rs12562034 758311  A  ADD   467 1.2830  0.98940 0.32250
4     1  rs12124819 766409  G  ADD   470 0.9849 -0.09739 0.92240
5     1   rs4475691 836671  A  ADD   470 0.8634 -0.83570 0.40330
6     1   rs3748597 878522  A  ADD   470 1.2740  0.79480 0.42670
7     1  rs28705211 890368  C  ADD   470 0.9635 -0.25650 0.79750
8     1  rs13303118 908247  C  ADD   469 0.8788 -0.95770 0.33820
```

we can see that in the above file, the first column is the choromosome numbers, followed by SNP names, BP, minor allele, SNP model that used in the tests, number of non-missing observations, odds ratio, test statistics, p-values. R also generated observation numbers for each SNPs that we can use them to extract SNPs. The Manhattan plot and QQ plot for the p-values are:



From the Manhattan plot we can see a few SNPs showing strong relationships with genotype trait, and the QQ plot fit reasonably well to the data.

In the second step, I picked eight SNPs that have smallest p-values in the result file generated by PLINK, and fit logistic regression models against phenotype variable lam, adjusted for three PCs, treatment group, total RT dose, gender, cancer stage and body mass index. After removing missing data, there are 471 observations and 17 variables in the data set. The result of the full model is as follow:

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.0926	1.1444	-0.0810	0.9355	
PC1	-2.8885	2.6291	-1.0990	0.2719	
PC2	-0.4307	2.4850	-0.1730	0.8624	
PC3	-0.4309	2.5607	-0.1680	0.8664	
rs6429215_G1	-0.8327	0.2631	-3.1650	0.0016	**
rs10865432_A1	-0.6146	0.2450	-2.5090	0.0121	*
rs216059_G1	-0.7086	0.2193	-3.2320	0.0012	**
rs10935647_A1	-1.0208	0.2642	-3.8640	0.0001	***
rs16873952_A1	0.8896	0.2710	3.2830	0.0010	**
rs2432755_G1	-0.7372	0.2276	-3.2400	0.0012	**
rs6901603_A1	-0.8014	0.2217	-3.6150	0.0003	***
rs3847141_G1	-0.7873	0.2218	-3.5490	0.0004	***
Groupe1	-0.1304	0.2192	-0.5950	0.5518	
DoseTot	0.0400	0.0150	2.6680	0.0076	**
SEX2	-0.0730	0.2754	-0.2650	0.7909	
STAD1200	-0.6331	0.2329	-2.7180	0.0066	**
BMI	0.0562	0.0248	2.2700	0.0232	*

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.95  on 470  degrees of freedom
Residual deviance: 506.17  on 454  degrees of freedom
AIC: 540.17

Number of Fisher Scoring iterations: 4
```

from the above table we can see all the SNPs are significant at 0.05 level, and several covariate variables have large p-values. Next I used Akaike Information Criterion (AIC) to conduct stepwise search through the space of possible models to find the best reduced model, and the final reduced model is:

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.0721	1.1286	-0.0640	0.9491	
rs6429215_G1	-0.8470	0.2610	-3.2450	0.0012	**
rs10865432_A1	-0.6390	0.2424	-2.6360	0.0084	**
rs216059_G1	-0.7377	0.2178	-3.3870	0.0007	***
rs10935647_A1	-0.9988	0.2624	-3.8060	0.0001	***
rs16873952_A1	0.8737	0.2688	3.2500	0.0012	**
rs2432755_G1	-0.7319	0.2268	-3.2270	0.0013	**
rs6901603_A1	-0.8072	0.2197	-3.6740	0.0002	***
rs3847141_G1	-0.7818	0.2201	-3.5510	0.0004	***

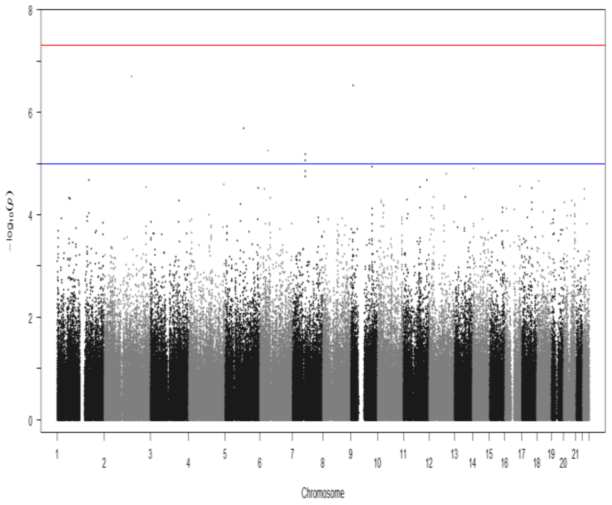
DoseTot	0.0401	0.0149	2.6880	0.0072	**
STAD1200	-0.6292	0.2313	-2.7200	0.0065	**
BMI	0.0528	0.0245	2.1560	0.0311	*

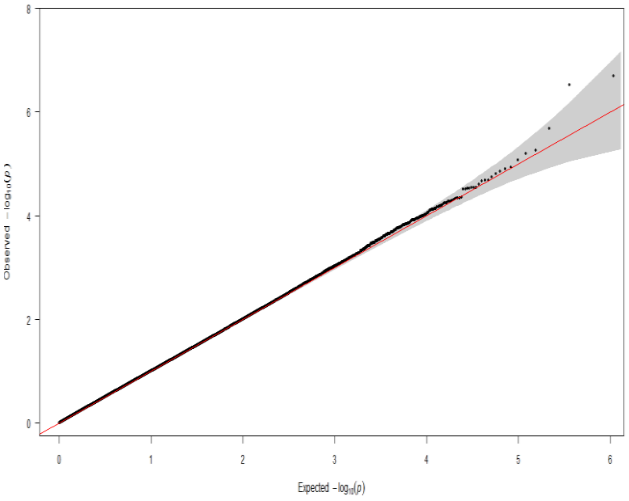
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null dviance: 638.95  on 470  degrees of freedom
Residual deviance: 507.85  on 459  degrees of freedom
AIC: 531.85
```

The p-values of the variables in the logistic model are all significant, and the residual deviance is close to the degree freedom. Therefore we conclude that the SNPs in the above table associated with phenotype trait lam, based on the information from the current data set.

Overall survival models: Cox PH models were used to conduct genome wide association study between SNPs and survival time of patients. With R-plugin, I was able to use PLINK to fit Cox PH models for all the SNPs, adjusted for three PCs and treatment group. The Manhattan plot and QQ plot for all the p-values in the result file are:





From the Manhattan plot we can see some SNPs are showing strong relationship with phenotype trait, and most of the points on QQ plot are close to the 45 degree line therefore we do not need to consider population stratification. Nine SNPs with small p-values in the result file were pick up to conduct the second step survival analysis, and the genetic data of the nine SNPs was extracted from the binary file using PLINK. There are 167 observations and 18 variables in the final data set, including 9 SNPs, 3 PCs, treatment group, cancer stage, gender, censor variable for overall survival and overall survival in days. R software was used to fit the Cox PH model and the result of the initial full model is:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
rs1472080_A1	0.8859	2.4252	0.2349	3.7720	0.0002	***
rs26505_A1	0.6325	1.8822	0.1559	4.0560	0.0000	***
rs11962379_C1	0.5862	1.7971	0.1764	3.3240	0.0009	***
rs2024237_G1	0.7176	2.0495	0.2647	2.7110	0.0067	**
rs10950167_G1	0.3166	1.3725	0.2054	1.5410	0.1232	
rs10491766_A1	0.8262	2.2847	0.2257	3.6600	0.0003	***
rs10759497_G1	0.6548	1.9247	0.1770	3.7000	0.0002	***
rs7300328_A1	0.3760	1.4564	0.1752	2.1460	0.0319	*
rs10145863_A1	0.5264	1.6928	0.1700	3.0970	0.0020	**
PC1	-0.9754	0.3770	1.6104	-0.6060	0.5447	
PC2	1.8513	6.3680	1.6516	1.1210	0.2623	
PC3	1.8185	6.1623	1.6627	1.0940	0.2741	
Age	0.0661	1.0683	0.0089	7.4390	0.0000	***
Group1	0.0604	1.0623	0.1504	0.4020	0.6879	
SEX2	-0.0855	0.9181	0.1959	-0.4360	0.6627	
STAD1200	0.7275	2.0698	0.1529	4.7560	0.0000	***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.756  (se = 0.022 )
Rsquare= 0.335   (max possible= 0.991 )
Likelihood ratio test= 190.3  on 16 df,   p=0
Wald test            = 182.1  on 16 df,   p=0
Score (logrank) test = 202.1  on 16 df,   p=0
```

Step wise AIC selection method was used again in search of best reduced model and the final reduced model is:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
rs1472080_A1	0.8768	2.4032	0.2315	3.7870	0.0002	***
rs26505_A1	0.5970	1.8167	0.1538	3.8810	0.0001	***
rs11962379_C1	0.6066	1.8341	0.1738	3.4900	0.0005	***
rs2024237_G1	0.9244	2.5203	0.2071	4.4640	0.0000	***
rs10491766_A1	0.8126	2.2537	0.2208	3.6810	0.0002	***
rs10759497_G1	0.6488	1.9133	0.1762	3.6830	0.0002	***
rs7300328_A1	0.4159	1.5157	0.1726	2.4090	0.0160	*
rs10145863_A1	0.5391	1.7145	0.1688	3.1940	0.0014	**
Age	0.0639	1.0660	0.0086	7.4280	1.11E-13	***
STAD1200	0.7019	2.0177	0.1511	4.6450	3.41E-06	***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.753  (se = 0.022 )
Rsquare= 0.328   (max possible= 0.991 )
Likelihood ratio test= 185.3  on 10 df,   p=0
Wald test            = 182.9  on 10 df,   p=0
Score (logrank) test = 199.8  on 10 df,   p=0
```

In the final model, all the p-values of the Wald z tests for predict variables are significant at 0.05 level, and all the global tests are significant. Graphic methods (plots are not shown here) were used to further evaluate this Cox PH model. The plot of H(t) against t on a log-log scale is very close to a straight line, which does not show any non-proportional hazard trend for the model. Most of the points on the Cox-Snell residual plot are close to the 45 degree line, which means the final model gives a reasonable fit to the data. The results of Grambsch and Therneau's test show that PH assumption holds for all the coefficients, and test values are listed below:

coefficients	rho	chisq	p
rs1472080_A1	-0.0789	1.2620	0.2610
rs26505_A1	-0.0577	0.6585	0.4170
rs11962379_C1	0.0734	1.0910	0.2960
rs2024237_G1	0.0766	1.1975	0.2740
rs10491766_A1	-0.0460	0.3994	0.5270
rs10759497_G1	0.0988	2.0286	0.1540
rs7300328_A1	-0.0040	0.0030	0.9560
rs10145863_A1	0.0703	0.9811	0.3220
Age	0.1261	3.0439	0.0810
STAD1200	-0.0700	0.9725	0.3240
GLOBAL	NA	12.4819	0.2540

The final model fits very well to the data set and we can conclude that the SNPs in the final model are associated with overall survival of patients, based on the information from the current data set.

Discussion

Although we can find some SNPs associated with lam and overall survival from the logistic models and Cox PH models mentioned previously, several aspect of the limitation of this study need to be considered before we can generalize the results. First, some other methods can be used to choose covariates for the genome wide search. For example, we can first find the best fit logistic and Cox PH models using the covariates in the original data set, and then use the covariates in the best fit models to conduct genome wide associate study using PLINK and R. In this way we may be able to find some other significant SNPs. Secondly, all the adjusted p-values for the 541,903 SNPs, both in logistic and Cox PH models, are not significant, which means none of the test statistics is significant in genome wide level. Also, the significant results in both the final logistic and Cox PH models are obtained from currently data set, and the power of the statistical tests are relatively low due to the sample size. Therefore before we can draw any conclusions from this study, further replication and investigation studies are needed.

References:

1. http://www.macmillan.org.uk/Cancerinformation/Cancertypes/Headneck/Aboutheadneckcancers/Typesofheadneckcancer.aspx

2. Ridge JA, Glisson BS, Lango MN, et al. "Head and Neck Tumors" in Pazdur R, Wagman LD, Camphausen KA, Hoskins WJ (Eds) Cancer Management: A Multidisciplinary Approach. 11 ed. 2008

3. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, Thun M (2006). "Cancer statistics, 2006". CA Cancer J Clin 56 (2): 106–30. doi:10.3322/canjclin.56.2.106. PMID 16514137.

4. Azad AK, Bairati I, Samson E, et al. Validation of genetic sequence variants as prognostic factors in early-stage head and neck squamous cell cancer survival. Clinical Cancer Research 2012;18(1):196-206.

5. Nan M. Laird, Christoph Lange (2010). The Fundamentals of Modern Statistical Genetics

6. Benjamin M. Neale, Manuel A.R. Ferreira, Sarah E. Medland, Danielle Posthuma (2008). Statistical Genetics: Gene Mapping Through Linkage and Association.

7. Andrea S. Foulkes (2009). Applied Statistical Genetics with R: For Population-based Association Studies.

8. Julina J. Faraway (2006). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models

9. W.N. Venables, B.D. Ripley (2003) Modern Applied Statistics with S

10. Peter H. Westfall, S. Stanley Young (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment