


8-2012

## Genetic Predictors of Clinical Outcomes in Non-Small Cell Lung Cancer Patients

Xia Pu

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)

 Part of the [Epidemiology Commons](#)

### Recommended Citation

Pu, Xia, "Genetic Predictors of Clinical Outcomes in Non-Small Cell Lung Cancer Patients" (2012). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 264.

[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/264](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/264)

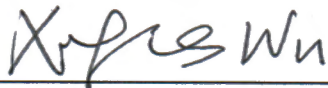
This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

# Genetic Predictors of Clinical Outcomes in Non-Small Cell Lung Cancer Patients

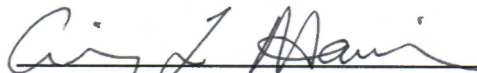
by

*Xia Pu, M.S.*

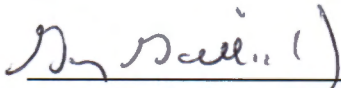
APPROVED:



\_\_\_\_\_  
Xifeng Wu, M.D., Ph.D., Advisor



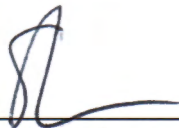
\_\_\_\_\_  
Craig L. Hanis, Ph.D.



\_\_\_\_\_  
Gary E. Gallick, Ph.D.



\_\_\_\_\_  
Jian Gu, Ph.D.



\_\_\_\_\_  
Scott M. Lippman, M.D.

APPROVED:

\_\_\_\_\_  
Dean, The University of Texas  
Graduate School of Biomedical Sciences at Houston

# **Genetic Predictors of Clinical Outcomes in Non-Small Cell Lung Cancer Patients**

A

## **DISSERTATION**

Presented to the Faculty of  
The University of Texas  
Health Science Center at Houston  
and  
The University of Texas  
M. D. Anderson Cancer Center  
Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

## **DOCTOR OF PHILOSOPHY**

by

Xia Pu, M.S.  
Houston, Texas

August 2012

*Dedicated*  
*To*  
*My Families*

## Acknowledgments

Foremost, I would like to express my deepest gratitude to my advisor Dr. Xifeng Wu for her mentorship and continuous support. I am extremely grateful for her patience, kindness, and willingness to share her knowledge. Her enthusiasm for science has greatly inspired and encouraged me. I consider myself extremely fortunate to have her as my mentor.

I am also grateful for my committee, Drs. Scott M. Lippman, Gary E. Gallick, Craig L. Hanis, and Jian Gu. They have been generous with their knowledge, experiences and precious time, and been a source of support and encouragement.

I also want to thank the faculty, staff and my colleagues at the University Of Texas Graduate School Of Biomedical Science at Houston as well as the University of Texas MD Anderson Cancer Center. I am especially grateful to our team members who have treated me like family by giving me great support in both scientific and daily life, especially Drs. Jian Gu, Jie Lin, Yuanqing Ye, Michelle A.T. Hildebrandt and David W. Chang, who have generously shared with me their expertise, knowledge and experience.

I would also like to thank my friends and classmates both in the US and in China for their friendship, help and supports. Lastly but most importantly, I want to thank my parents and my husband or their deep love and caring.

## ABSTRACT

Lung cancer is the leading cause of cancer-related mortality in the US. Emerging evidence has shown that host genetic factors can interact with environmental exposures to influence patient susceptibility to the diseases as well as clinical outcomes, such as survival and recurrence. We aimed to identify genetic prognostic markers for non-small cell lung cancer (NSCLC), a major (85%) subtype of lung cancer, and also in other subgroups. With the fast evolution of genotyping technology, genetic association studies have went through candidate gene approach, to pathway-based approach, to the genome wide association study (GWAS). Even in the era of GWAS, pathway-based approach has its own advantages on studying cancer clinical outcomes: it is cost-effective, requiring a smaller sample size than GWAS easier to identify a validation population and explore gene-gene interactions. In the current study, we adopted pathway-based approach focusing on two critical pathways - miRNA and inflammation pathways. MicroRNAs (miRNA) post-transcriptionally regulate around 30% of human genes. Polymorphisms within miRNA processing pathways and binding sites may influence patients' prognosis through altered gene regulation. Inflammation plays an important role in cancer initiation and progression, and also has shown to impact patients' clinical outcomes.

We first evaluated 240 single nucleotide polymorphisms (SNPs) in miRNA biogenesis genes and predicted binding sites in NSCLC patients to determine associations with clinical outcomes in early-stage (stage I and II) and late-stage (stage III and IV) lung cancer patients, respectively. First, in 535 early-stage patients, after correcting multiple comparisons, *FZD4*:rs713065 (hazard ratio [HR]:0.46, 95% confidence interval [CI]:0.32-0.65) showed a

significant inverse association with survival in early stage surgery-only patients. *SPI*:rs17695156 (HR:2.22, 95% CI:1.44-3.41) and *DROSHA*:rs6886834 (HR:6.38, 95% CI:2.49-16.31) conferred increased risk of progression in the all patients and surgery-only populations, respectively. *FAS*:rs2234978 was significantly associated with improved survival in all patients (HR:0.59, 95% CI:0.44-0.77) and in the surgery plus chemotherapy populations (HR:0.19, 95% CI:0.07-0.46).. Functional genomics analysis demonstrated that this variant creates a miR-651 binding site resulting in altered miRNA regulation of *FAS*, providing biological plausibility for the observed association. We then analyzed these associations in 598 late-stage patients. After multiple comparison corrections, no SNPs remained significant in the late stage group, while the top SNP *NAT1*:rs15561 (HR=1.98, 96%CI=1.32-2.94) conferred a significantly increased risk of death in the chemotherapy subgroup.

To test the hypothesis that genetic variants in the inflammation-related pathways may be associated with survival in NSCLC patients, we first conducted a three-stage study. In the discovery phase, we investigated a comprehensive panel of 11,930 inflammation-related SNPs in three independent lung cancer populations. A missense SNP (rs2071554) in *HLA-DOB* was significantly associated with poor survival in the discovery population (HR: 1.46, 95% CI: 1.02-2.09), internal validation population (HR: 1.51, 95% CI: 1.02-2.25), and external validation (HR: 1.52, 95% CI: 1.01-2.29) population. Rs2900420 in *KLRK1* was significantly associated with a reduced risk for death in the discovery (HR: 0.76, 95% CI: 0.60-0.96) and internal validation (HR: 0.77, 95% CI: 0.61-0.99) populations, and the association reached borderline significance in the external validation population (HR: 0.80, 95% CI: 0.63-1.02). We also evaluated these inflammation-related SNPs in NSCLC patients

in never smokers. Lung cancer in never smokers has been increasingly recognized as distinct disease from that in ever-smokers. A two-stage study was performed using a discovery population from MD Anderson (411 patients) and a validation population from Mayo Clinic (311 patients). Three SNPs (*IL17RA*:rs879576, *BMP8A*:rs698141, and *STK*:rs290229) that were significantly associated with survival were validated ( $p < 0.05$ ), and two more SNPs (*CD74*:rs1056400 and *CD38*:rs10805347) were borderline significant ( $p = 0.08$ ) in the Mayo Clinic population. In the combined analysis, *IL17RA*:rs879576 resulted in a 40% reduction in the risk for death ( $p = 4.1 \times 10^{-5}$  [ $p = 0.61$ , heterogeneity test]). We also validated a survival tree created in MD Anderson population in the Mayo Clinic population.

In conclusion, our results provided strong evidence that genetic variations in specific pathways that examined (miRNA and inflammation pathways) influenced clinical outcomes in NSCLC patients, and with further functional studies, the novel loci have potential to be translated into clinical use.



## TABLE OF CONTENTS

<b>Approval Page</b> .....	i
<b>Tital Page</b> .....	ii
<b>Dedications</b> .....	iii
<b>Acknowledgments</b> .....	iv
<b>Abstract</b> .....	v
<b>Table of Contents</b> .....	viii
<b>List of Illustrations</b> .....	xii
<b>List of Tables</b> .....	xiv
<b>Chapter 1: General Introduction</b> .....	1
1.1 Epidemiology.....	2
1.1.1 Incidence and mortality.....	2
1.1.2 Risk Factors.....	3
1.1.2.1 Tobacco smoking.....	3
1.1.2.2 Environmental and occupational exposure.....	4
1.1.2.3 Other risk factors.....	5
1.2 Clinical Aspects.....	6
1.2.1 General overview.....	6
1.2.2 Prognosis and treatment by stage.....	7
1.3 Genetics of lung cancer.....	12
1.3.1 Somatic alterations.....	12
1.3.2 Genetic susceptibility.....	13

1.3.2.1	Linkage analysis .....	14
1.3.2.2	Genetic association studies .....	14
1.4	MicroRNA .....	18
1.4.1	MiRNA biogenesis .....	18
1.4.2	miRNA binding site polymorphism.....	21
1.4.3	miRNA and lung cancer .....	21
1.5	Inflammation pathways .....	22
1.6	Hypothesis and rationale.....	25
1.6.1	Hypothesis 1: miRNA-related genetic variations and survival and recurrence in NSCLC patients .....	25
1.6.2	Hypothesis 2: The inflammation pathway and survival in late stage NSCLC patients .....	25
1.6.3	Hypothesis 3: The inflammation pathway and survival in NSCLC patients in never smokers .....	26
<b>Chapter 2: Material and Methods.....</b>		<b>27</b>
2.1	Study populations and data collection .....	28
2.2	SNP selection, genotyping and quality control.....	29
2.2.1	miRNA related SNPs.....	29
2.2.2	inflammation related SNPs .....	31
2.3	Statistical analyses .....	34
2.4	Luciferase reporter assay .....	35

<b>Chapter 3: Results and Discussion</b> .....	37
3.1 miRNA-related genetic variations and clinical outcomes in NSCLC patients.....	38
3.1.1 miRNA-related genetic variations and survival and recurrence in early stage NSCLC patients.....	38
3.1.1.1 Patients characteristics.....	38
3.1.1.2 Associations between individual SNPs and NSCLC clinical outcomes.....	40
3.1.1.3 Survival tree analysis of SNPs associated with NSCLC survival .....	53
3.1.1.4 Internal validation using bootstrap re-sampling method .....	55
3.1.1.5 The effect of selected miRNA binding site variants on miRNA-regulation .....	55
3.1.1.6 Associations between individual SNPs and late stage patients survival .....	60
3.1.1.7 Associations between individual SNPs and survival in late stage patients treated with chemotherapy.....	61
3.1.1.8 Internal validation using bootstrap re-sampling method .....	65
3.1.2 Discussion.....	65
3.2 Genetic variations in inflammation related genes and survival in late stage NSCLC patients.....	72
3.2.1 Patient characteristics .....	72
3.2.2 Effects of inflammation-related SNPs on overall survival.....	75
3.2.3 Stratified analyses.....	82
3.2.4 Cumulative effects of the top two SNPs.....	83
3.2.5 Discussion.....	84
3.3 Genetic variations in inflammation pathway and survival in NSCLC patients in never smokers.....	87

3.3.1	Patient characteristics .....	87
3.3.2	Main effect of individual SNP on survival in the discovery, replication, and combined analysis.....	89
3.3.3	Main effects of individual SNPs on survival stratified by histology and stage...	94
3.3.4	Main effects of individual SNPs on survival in ever-smokers .....	96
3.3.5	Survival tree analysis .....	98
3.3.6	Discussion.....	100
<b>Chapter 4: Conclusions .....</b>		<b>104</b>
<b>Chapter 5: Strength and Limitations.....</b>		<b>106</b>
<b>Chapter 6: Future Directions .....</b>		<b>109</b>
<b>References.....</b>		<b>112</b>
<b>Appendix A: Other Peer-reviewed Publications during Ph.D. Study.....</b>		<b>144</b>
<b>Appendix B: supplementary table 1 miRNA related SNPs selected .....</b>		<b>147</b>
<b>VITA .....</b>		<b>153</b>

## List of Illustrations

Figure 1: Overall survival, median survival time and five-year survival by TNM stage .....	8
Figure 2: The scheme of miRNA biogenesis and regulation.....	20
Figure 3: inflammation pathways in response to a danger signal.....	23
Figure 4: Kaplan-Meier estimates of FAS:rs2234978 on overall survival:.....	41
Figure 5: Kaplan-Meier estimates on effect of SP1:rs17695156 on time to progression among early stage patients.....	42
Figure 6: Kaplan-Meier estimates of effect of FZD4:rs713065 on overall survival .....	44
Figure 7: Kaplan-Meier estimates of overall survival and time to progression in early stage NSCLC patents grouped by the number of unfavorable genotypes (UFG).....	47
Figure 8: Kaplan-Meier estimates on effect of DROSHA:rs6886834 on early stage progression among surgery-only patients .....	50
Figure 9: Kaplan-Meier estimates for the effect of RRM2B:rs5005121 genotypes on NSCLC progression in two treatment subgroups based on the dominant model: .....	52
Figure 10: Potential gene-gene interactions among SNPs identified in the survival analysis in early stage NSCLC patients .....	54
Figure 11: Effect of the FAS variant allele on miR-561 targeting and luciferase reporter expression: .....	57
Figure 12: Kaplan-Meier estimates for the effect of selected SNPs on NSCLC survival in patients treated chemotherapy: .....	64
Figure 13: Study design and workflow.....	76
Figure 14: Forest plot for meta-analysis of the association of single nucleotide polymorphisms.....	79

Figure 15: Kaplan-Meier estimates of HLA-DOB:rs2071554 genotypes and risk of death in late-stage patients treated with chemotherapy .....	80
Figure 16: Kaplan-Meier estimates of the effect of selected SNPs on survival probability in never-smokers with lung cancer .....	92
Figure 17: Potential gene-gene interactions among SNPs validated in the survival tree analysis.....	99

## List of Tables

Table 1: Prognostic Factors in Patients With Surgically Resected NSCLC.....	10
Table 2: Prognostic Factors in Patients With Advanced NSCLC .....	11
Table 3: miRNA Processing and predicted targets genes.....	30
Table 4: Inflammation-related pathways selected .....	32
Table 5: Host characteristics of early stage NSCLCs.....	39
Table 6: Effect of selected SNPs on survival in early stage NSCLC patients.....	46
Table 7: Effect of selected SNPs on progression in early stage NSCLC patients.....	49
Table 8: Host characteristics of late stage NSCLCs .....	59
Table 9: Selected SNPs with survival in late stage NSCLC patients .....	63
Table 10: Characteristics of the study populations at the time of analysis.....	74
Table 11: Inflammation-related single nucleotide polymorphisms (SNPs) that were found to affect overall survival in patients with late-stage non-small cell lung cancer .....	77
Table 12: Characteristics of the never-smokers with lung cancer .....	88
Table 13: SNPs with the same trend in both the MD Anderson and Mayo Clinic.....	90
Table 14: Effect of selected SNPs on survival in adenocarcinoma patients.....	95
Table 15: Effect of selected SNPs on survival according to smoking status in the MD Anderson population.....	97

## *Chapter 1: General Introduction*



## 1.1 Epidemiology

### 1.1.1 Incidence and mortality

Lung cancer incidence increases dramatically since last century, and it is now the second most common cancer in both sexes and the leading causes of cancer death worldwide (1). Although the incidence rate has reduced in men (1.9% per year) and started to decline in women (0.3% per year), it is estimated that 226,160 new lung cancer cases will be diagnosed in 2012, accounting for 14% of all new cancer diagnoses (2, 3).

Around 95% of all lung cancer cases are classified into two major histological types – non-small cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) and. NSCLCs develop from epithelial cells, while SCLCs originate from neuroendocrine-cells. Around 85% of lung cancer cases are NSCLCs. NSCLCs can be classified into several subtypes according to physical and chemical characteristics of tumor cells. Adenocarcinoma is slow-growing tumors that account for around 40% of all NSCLCs. Patients are usually diagnosed before the tumors reach 4 cm in diameter and have a better prognosis compared to other subtypes (4, 5). Around 25% to 30% of NSCLC cases are squamous cell carcinoma, usually observed with a tumor larger than 4cm (4). 10-15% of NSCLC patients have large cell carcinoma, where the tumors are often poorly differentiated, grow rapidly, and metastasize early (6).

The highest lung cancer incidence in male observed in the central and Eastern Europe, while African and Asian men have the lowest incident rate. In female, the highest incidence was found in North America, and in some part of Europe. Women in Spain is the least likely to develop lung cancer, where the percentage of women smoker just begin to increase (7).

Within the same geographic location, lung cancer incidences are usually different between

ethnic groups, for example, in the US, black male has the highest incident rate of lung cancer, which is 40% higher in black men compared to white men (Cancer Facts & Figures 2012).

Despite of the declining mortality rate in both men and women, lung cancer is still the leading cause of cancer-related mortality in the US. Emerging evidence shows that early stage lung cancer patients received surgery have good survival; however, most of lung cancer patients diagnosed with un-resectable tumor, thus the overall mortality for lung cancer remain very high. It is expected that over a quarter of all cancer-related deaths will be attributed to lung cancer this year, more than colon, breast and prostate cancer combined (2).

### 1.1.2 Risk Factors

#### 1.1.2.1 Tobacco smoking

It is estimated that tobacco smoking is related to around 80% of all cancer deaths. Ever since 1950s, tobacco smoking has been recognized as the leading risk factor for lung cancer. Smokers have at least ten-time higher risk of developing lung cancer compared to those who never smoked, and the excess risk was equally observed in both male and female (8).

Factors of smoking influence lung cancer risk include duration of smoking, smoking intensity (i.e. number of cigarettes per day)age of initiation, inhaling habit, types of tobacco products and time since quitting. Data has shown that more than 90% of lung cancer cases are caused by tobacco smoking (9, 10).

For current smokers, lung cancer risk is proportional to the pack-year of smoking. The cumulative risk for continuous smokers is 19%, compared to 1% in never smokers at age of 75 (11). Former smokers have lower risk of lung cancer, however, excess risk of lung

cancer was still observed compared to never smokers. Other type of smoking, such as pipes, or cigars also showed related to increased lung cancer risk.

Lung cancer in never smokers - Although smoking is the predominant risk factors for lung cancer, lung cancer develops in 15% of male and 53% of female never-smokers (12, 13). Over the past few decades, the proportion of never-smokers with lung cancer has increased strikingly (13). Lung cancer in never smokers has emerged as a major public health problem in studies tracking smoking and smoking cessation rate. Previous studies have reported differing tumor etiology and clinicopathological presentation according to smoking status in lung cancer patients with never-smokers as being more likely to be women, having adenocarcinomas, and having less-differentiated tumors (12-15). Genetic and epigenetic alterations also differ with fewer changes overall. Tumors from never-smokers also have a unique and predominant profile compared to those from smokers, such as chromosomal gains at 16p, promoter hypermethylation of hMLH1 and hMSH2, and distinct mutations of major oncogenes and tumor suppressor genes. For example, compared to smokers, never smokers have fewer mutations in *K-ras* and *tp53* genes, while with a higher rate of mutation in *EGFR*, results in better response rate of EGFR tyrosine kinase inhibitors in never smokers (12-14). These findings suggest different paths of carcinogenesis in ever- and never-smokers with lung cancer.

#### 1.1.2.2 Environmental and occupational exposure

In the US, around 40% of non-smokers are currently exposed to environmental second hand smoke (16). Second-hand smoking leads to up to 30% increased risk for those who do not smoke, which is estimated to contribute to three thousands lung cancer deaths each year

(17). Evidence shows that second-hand smoking is comparable to smoking in the excess risk of lung cancer.

Radon is another environmental risk factor for lung cancer, which commonly released from construction material concentrated in buildings, and is hard to be detected without specialized equipment. Based on the data from U.S. Environmental Protection Agency, radon is responsible for over 20,000 lung cancer cases each year, which makes it a major risk factor for lung cancer, especially in never smokers.

Occupational asbestos exposure is another important risk factor for lung cancer. People who worked in an environment with asbestos fibers have much higher risk to develop and die from lung cancer. Moreover, joint effect was observed for asbestos exposure with tobacco smoking.

Other environmental risk factors include air pollutions, diesel exhaust, and arsenic. Certain forms of silica and chromium also show associations with lung cancer.

#### 1.1.2.3 Other risk factors

People with family or personal history of cancer are at higher risk of developing lung cancer. First-degree relatives of lung cancer patients have higher risk and/or early onset of the disease. Evidence shows that persons who received radiation therapy to the chest, or to other cancers, have an increased risk of developing lung cancer.

## 1.2 Clinical Aspects

### 1.2.1 General overview

The most common lung cancer symptoms include: a persistent cough, chest pain, hoarseness, shortness of breath, wheezing, infections, bone pain, neurologic changes, and jaundice. Lung cancer patients are usually diagnosed with advanced stage diseases, due to the late presentation of symptoms or lack of symptoms (18-22).

TNM staging system, which maintained by the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC), is the widely accepted standard for NSCLC staging. Clinical stage is defined by physical exam, biopsies, imaging tests prior to treatment. Patients who undergo surgical resection may also have pathologic stage based on histological tests of resected tumor, which can provide more accurate information about the extent of disease(23).

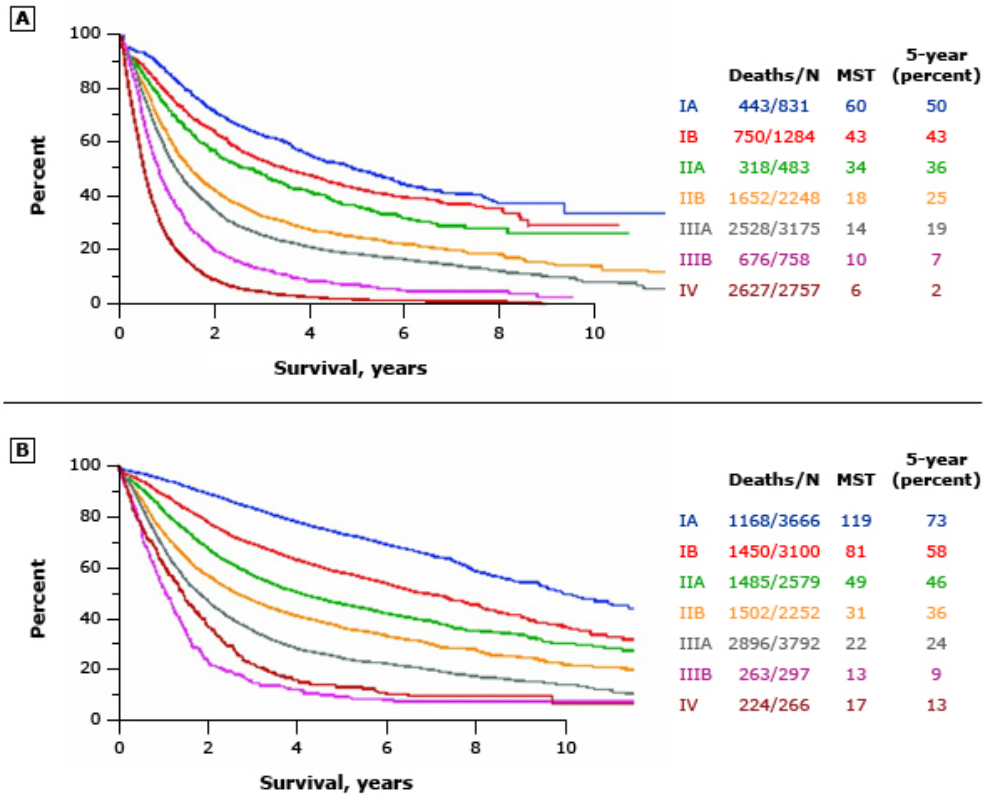
The treatment options for NSCLC includes surgery, radiation therapy, chemotherapy, target therapy and other local treatments, and treatments are usually used in combinations (24-26). Even with advances in treatment, the overall prognosis is still poor for NSCLC, with an overall 5-year survival rate (percentage of patients survival at least 5 years after diagnosis) of around 16% (1). In general, surgery is considered the most potentially curative treatment for early stage NSCLC and offers the best prognosis. Evidence shows that complete resection of localized tumor and associated lymph node largely benefit patients prognosis, and a post-operative chemotherapy could add significant benefit to patients' survival (27).

The clinicopathologic factors most often associated with prognosis include stage of cancer (tumor size and spread of disease), type of cancer (non-squamous histology),

presence of pulmonary symptoms, lymph nodes metastases, vascular invasion. Effort has been devoted to developing prediction modalities for patient prognosis based on these clinicopathologic factors; however, variations still exist within patients with same above mentioned characteristics (28-32).

### 1.2.2 Prognosis and treatment by stage

The TNM stage at diagnosis is the primary parameter used to estimate patients' prognosis and treatments. Early stage patients have the most promising 5-year survival rate of 30-49%, while survival for late stage patients is as low as 1% for stage IV patients (the Surveillance, Epidemiology and End Results [SEER] databases). A validation-study series using more than 31,000 lung cancer cases has provided the most extensive data of survival for each stage of patients (Figure 1), and a 59 months survival disparity was observed between stage IA and stage IV patients (33). Because the large difference in prognosis between patients with resectable and advanced diseases (34), the prognostic factors are commonly identified separately (Tables 1 and 2).



**Figure 1 Overall survival, median survival time and five-year survival by TNM stage**

**(A) clinical stage and (B) pathologic stage. (33) Reprinted by permission from J**

**Thorac Oncol, copyright (2007)**

About 30% of NSCLCs are diagnosed with early stage disease (stage I and II), which is considered as a localized disease and expected to have a generally good survival. Stage I patients are defined as those having a tumor limited to the lung without any invasion to the parietal pleura or main bronchi (35). Stage II NSCLC tumors are still in lung with or without invasion into local lymph nodes, and have not spread to distant sites. Surgical resection is the principle treatment for early stage NSCLCs. To obtain better outcome, chemotherapy (neoadjuvant/ adjuvant) and radiation therapy were performed to facilitate surgical resection and prevent recurrence when necessary (28, 36, 37). However, it is estimated that 20-25% of stage I or II patients will eventually develop recurrent or metastatic disease. Prevention of recurrence is the major concern for this group of patients (28).

A majority of NSCLC cases are diagnosed with late (stage III or IV) stage diseases, which are usually incurable. Typically, these patients present with metastatic disease, and only a few stage IIIA patients are eligible for surgery. Overall, late stage patients have a dismal 5-year survival rate of 5% (38). Standard treatment for these patients is platinum-based chemotherapy, which is reported to moderately prolong patients' survival (39, 40). Radiation therapy is usually performed in combination with chemotherapy either concurrently or sequentially. However, the response rate to chemotherapy is only 30% with a duration limited to 4-6 months. Furthermore, patients are at high risk for developing severe toxic effects (41). Therefore, patients with late stage disease are usually administered palliative treatments to reduce symptoms and improve quality of life (42, 43). As curative options are limited for these patients, a subset of this group may benefit from targeted therapies and may be candidates for clinical trials.



**Table 1: Prognostic Factors in Patients With Surgically Resected NSCLC**

Prognostic Factors	Tumor-Related Factors	Host-Related Factors
<b>Essential factors</b>	Stage Hypercalcemia <sup>54</sup>	Weight loss Performance status
<b>Additional factors</b>	<b>Anatomic</b> “T” factor Nodal level Intrapulmonary metastasis <b>Histologic</b> Grade Vessel invasion	Sex Age
<b>New or promising factors</b>	<b>Histologic</b> Cells in mitosis Lymphoid infiltration <b>Clinical chemistry</b> Blood group Ag  NSE CA-125 TPA <b>Proliferation markers</b> DNA ploidy and/or % S-phase PCNA Thymidine labeling <b>Cellular adhesion markers</b> CD44 <b>Other molecular biological markers</b> kRAS, RB gene, bcl-2, c-jun, MRP-1, EGFr (c-erbB-1), HGF, TPA, Cyclin D-1, P53, P21, c-fos, CYFRA-21-1, KAI-1, c-erbB-2, VEGF, sIL-2R, Cathepsin B	Smoking habit Quality of life Marital status Depressed mood CYPIA-1
		Coagulation factors Proteinuria CEA  Ki67 AgNOR  Plankoglobin

\* NSE = neuron-specific enolase; CEA = carcinoembryonic antigen; AgNOR = argyrophilic nucleolar organizer region; PCNA = proliferating cell nuclear antigen; RB = retinoblastoma; CYFRA-21 = serum assay for detection of cytokeratin 19 fragment; MRP = motility-related protein; kRAS = *ras* oncogene or protein; EGFr = epidermal growth factor receptor; HGF = hepatocyte growth factor; VEGF = vascular endothelial growth factor. Reprinted by permission from American College of Chest Physicians. (34)

**Table 2: Prognostic Factors in Patients With Advanced NSCLC**

<b>Prognostic Factors</b>	<b>Tumor-Related Factors</b>	<b>Host-Related Factors</b>
<b>Essential factors</b>	Stage (III vs IV) Hypercalcemia SVCO	Weight loss Performance status
<b>Additional factors</b>	<b>Anatomic</b> "T" factor "N" factor Clinical stage IIIA vs IIIB Number of sites involved Pleural effusion Liver metastases <b>Clinical chemistry/hematology</b> Hemoglobin LDH Albumin	Sex Symptoms Age
<b>New or promising factors</b>	<b>Clinical chemistry/hematology</b> Coagulation factors Proteinuria <b>Proliferation markers</b> DNA ploidy and/or % S-phase Ki-67 <b>Other molecular biologic markers</b> Replication errors 2p/3p K ras P53 c-erbB-1 TPA NSE <b>Other radiology</b> Thalium-201 uptake	Quality of life Marital status Depressed mood CYPIA1

\* SVCO = superior vena caval obstruction; NSE = neuron-specific enolase.  
Reprinted by permission from American College of Chest Physicians. (34)

### 1.3 Genetics of lung cancer

#### 1.3.1 Somatic alterations

Numerous molecular genetic abnormalities have been identified in lung cancer, such as chromosomal aberrations, alterations in major tumor suppressor gene (TSG) or oncogenes, many of the alterations are of great clinical importance (44, 45).

The most common identified mutation of lung cancers were in *KRAS* and the epidermal growth factor receptor (*EGFR*) tyrosine kinase gene. *KRAS* mutations, common form of *RAS* mutations, have been identified in lung tumors for two decades. Studies have shown that around 23% of all lung cancer cases carrying *KRAS* mutations, mostly in codons 12/13. Considerable efforts have been devoted to evaluate the predicting value of *KRAS* mutation on cancer drug response, which provided some evidence that *KRAS* might predict a poor response to adjuvant chemotherapy and kinase inhibitors (46-48). *EGFR* mutations are found in 15%-30% of NSCLC tumors (49). Although just recently identified in NSCLCs, *EGFR* mutations have attracted considerable attentions from clinics. *EGFR* mutations, especially in kinase domain, have been used as predictors for treatment response of *EGFR* kinase inhibitors, such as gefitinib. Other than *KRAS* and *EGFR*, other somatic mutations, such as *BRAF*, *ERBB2* and *TP53*, are also frequently identified in lung tumors.

Other genetic alterations, such as chromosomal alterations, somatic copy-number alteration, and loss-of heterogeneity (LOH), are also frequently found in lung cancer (50-52). For example, the loss of chromosome 3p has been identified in nearly half of non-small cell tumors (53), and LOH was observed in 90% of lung squamous cell carcinoma and in 67% of lung adenocarcinoma (52).

Somatic alterations have been investigated for their associations with prognosis (44, 45). For example, down-regulation of 3p genes (RASSF1A, FHIT,  $\beta$ -catenin) were found related with a poorer survival in NSCLC (54-56). In addition, many studies have found the role of several major oncogenes and TSGs in NSCLC prognosis. NSCLC tumors harboring *KRAS* mutations are smaller and poorly differentiated, patients have a higher mortality rate (57). In a study of advanced stage patients, it was found that compared to patients with *KRAS* mutated tumor, patients carrying *BRAF* mutations experienced a better prognosis (58, 59). Some other genes, such as growth factors (60), apoptosis genes (61), DNA repair gene (62-65), telomerase activity (66), inflammatory factors (67-70), plasminogen activator (71, 72), and matrix metalloproteinases (73) have also been described for their prognostic value.

### 1.3.2 Genetic susceptibility

Evidence of familial aggregation of lung cancer suggested a role of genetic components to lung cancer (74-77). For example, in a family-based study, a 2.4-fold increased risk of lung cancer was observed for the individuals whose relatives had developed lung cancer, the effect remained significant even after controlling for other risk factors (75). And in a recent large scale family linkage study of lung cancer, it was found that among 26,000 lung cancer patients screened in the study, 13.7% had at least one first-degree relative also developed lung cancer. The excess risk of lung cancer patients' relatives suggested the potential heritability of lung cancer. Association studies also provided supporting evidence of genetic component in the initiation and progression of lung cancer. Knowledge of lung cancer susceptibility genetic loci is the key for the understanding of the underlying mechanism of disease initiations and progressions. In general, lung cancer susceptibility genes were

categorized into high or moderate/ low risk (or penetrance) genes, for which family-based linkage analysis or genetic association studies were performed.

#### 1.3.2.1 Linkage analysis

The traditional strategy to identify high penetrance gene is the family-based linkage analysis followed by positional cloning. Family-based analysis can avoid potential bias caused by environmental factors, and has successfully mapped lots of genes associated with monogenic disorder including common cancers(78). High risk gene has a great impact on cancer risk for people carrying the variant allele, however, the frequency of variant allele of high-risk gene is very low in population, and thus the population attributable risk is low. Most of the high-risk genes are tumor suppressor genes (TSGs) discovered in the study of cancer syndromes, and show an autosomal dominant inherited fashion (Mendelian pattern) (79).

There are several gene mutation identified as potential high risk for lung cancer. For example, *TP53* mutations identified in family members with Li-Fraumeni syndrome were significantly associated with higher lung cancer risk and earlier age at onset. A family linkage study mapped a higher risk region to chromosome 6q23-25 (80), fine mapping of sequential studies further narrow it to *RGS17* gene (81, 82).

#### 1.3.2.2 Genetic association studies

Despite the great impact of high-penetrance genes on cancer development of individuals carrying mutated genes, it only accounts for 10% of cancers, and the remaining 90% of cancer were considered developing in a polygenic fashion with a complex interaction of

both environmental factors and multiple small and subtle genetic changes. Although only small proportion of people carrying low penetrance genes will develop cancer, and the effect of these low-penetrance genes usually cannot be distinguished clearly from environmental effect, the high prevalence of these low-penetrance genes in general population makes their identification of great impact in public health. Traditional family-based linkage analysis failed to identify this type of genes due to population heterogeneity and environmental confounders.(79) During the past decades, population-based association study has proved its value in discovering low/moderate penetrance loci. Based on “common disease common variant” hypothesis, association study identifies cancer susceptibility loci by comparing the frequency of the genetic variants between cancer patients and healthy controls. Numerous studies, particularly recent genome-wide association studies (GWAS), have unequivocally identified many low penetrance genetic loci for a variety of cancers(83).

Single nucleotide polymorphisms (SNP) are most commonly investigated form of genetic variations in cancer association studies. Evidence shows that SNPs would affect host gene either in terms of gene expression or protein activities, and have impact on lung cancer susceptibility and outcomes (84-86). Association studies can be either family- or population-based. By comparing the allele frequency of candidate loci between cases and healthy controls, population-based association studies are more widely used in cancer gene identification than family-based association study, in which elderly relatives of cancer patients are hard to recruit. Population-based association study has gone through a fast evolvement in the past decades, from candidate gene approach to pathway-based approach to genome-wide association approach (87, 88), and have been widely adopted in the

identification of low penetrance common alleles responsible for cancer susceptibility as well as patients' prognostic markers.

Candidate gene and pathway-based approach - Candidate gene approach is the earliest approach used to identify cancer susceptibility genes. This hypothesis driven approach is largely depending on a priori knowledge of SNPs and gene function. Most of the genes selected as candidate are genes encoding proteins within major known functional pathways and the SNPs are functional SNPs. Since the number of SNPs is limited, the genotyping cost is relatively low, and the sample size requirement is small. Pathway-based approach is an extension for candidate gene approach. Instead of analyzing a single gene or single variant, this approach focuses on gene variants of a whole biological or functional pathway. Pathway-based approach increases the coverage of analyzed region, but is still hypothesis-driven and based on existing knowledge. Because of the increasing number of variant genotyped, the cost of genotyping increases and chance of false discovery also increases. (86)

With its own strength of being based on prior knowledge of disease biology, candidate gene and pathway based approaches have been widely adopted to identify genetic predictors for lung cancer susceptibility loci (86). DNA repair pathway gene polymorphisms are most commonly identified to be associated with susceptibility of lung cancer (89-92). For example, polymorphisms in XRCC1 have repeatedly been identified to associate with lung cancer susceptibility (93-96). In a meta-analysis of 28 published epidemiological studies on nucleotide excision repair pathway gene, it was found that *ERCC2751Gln/Gln* and *XPA 23G/G* genotype were significantly associated with altered lung cancer risk (97). Besides the above mentioned studies, numerous studies on other pathways, such as cell

cycle (98-100), growth signaling (101-103), and apoptosis (85) pathways, have been identified as lung cancer susceptibility loci. Over past a few decades, studies have started to use candidate gene/ pathway-based approaches to investigate lung cancer outcomes, such as polymorphisms in DNA repair pathway (104-107), *AKT/mTOR* pathway (108-110), miRNA pathway (111-113), have showed evidence to related to survival in lung cancer patients.

Genome-wide association studies (GWAS) - not depending on any current knowledge, GWAS is a discovery-driven approach, providing a thorough screening of whole genome(114). Due to the large number of association tested, the requirement for statistical significance is very stringent ( $P\text{-value} < 10^{-8}$ ) and a multi-stage study design is usually performed to control for false discovery through successive validation steps (115).

Recent reports have clearly demonstrated the power of GWAS in identifying novel genetic loci of common diseases (83). Till date, fifteen GWAS studies have been reported on lung cancer (116-131). Compared to the identification of cancer susceptibility loci, only four studies have performed on lung cancer outcomes (124, 126, 127, 130). GWAS on outcomes studies has its limitations. The bottleneck is the requirement of large populations identified from multi-institutions to provide sufficient statistical power for GWAS analysis. And for outcome analysis, to obtain adequate clinical characteristics from all populations, such as histology, treatment regimens, and following-up information, is the pre-requisite for conducting such studies. Due to the heterogeneity nature of treatment regimens for lung cancer patients as well as the lack of comparable clinical/ follow-up data, to identify a comparable validation populations is usually a challenge, which largely hindered the



progress of GWAS on lung cancer outcomes (83). In this scenario, to initiate multi-institutional collaborations for a well-designed GWAS of stage or treatment-specific analysis of patients' outcomes is warranted.

Meanwhile, pathway-based approaches have its unique advantage as a powerful tool for outcome study. With limited number of candidate loci, the pathway-based approach required a much smaller sample size, and therefore is cost-effective and much easier to identify a validation population (132-137). Moreover, since pathway-based approaches are developed based on prior established knowledge of disease, it provides more coverage on the specific interested functional pathways relevant to disease, and is easier to discover gene-gene network interactions and study complex underlying biological network (132-137). In this context, a large scale pathway-based genetic variation study focusing on interesting biological pathways is both necessary and desirable for outcomes studies.

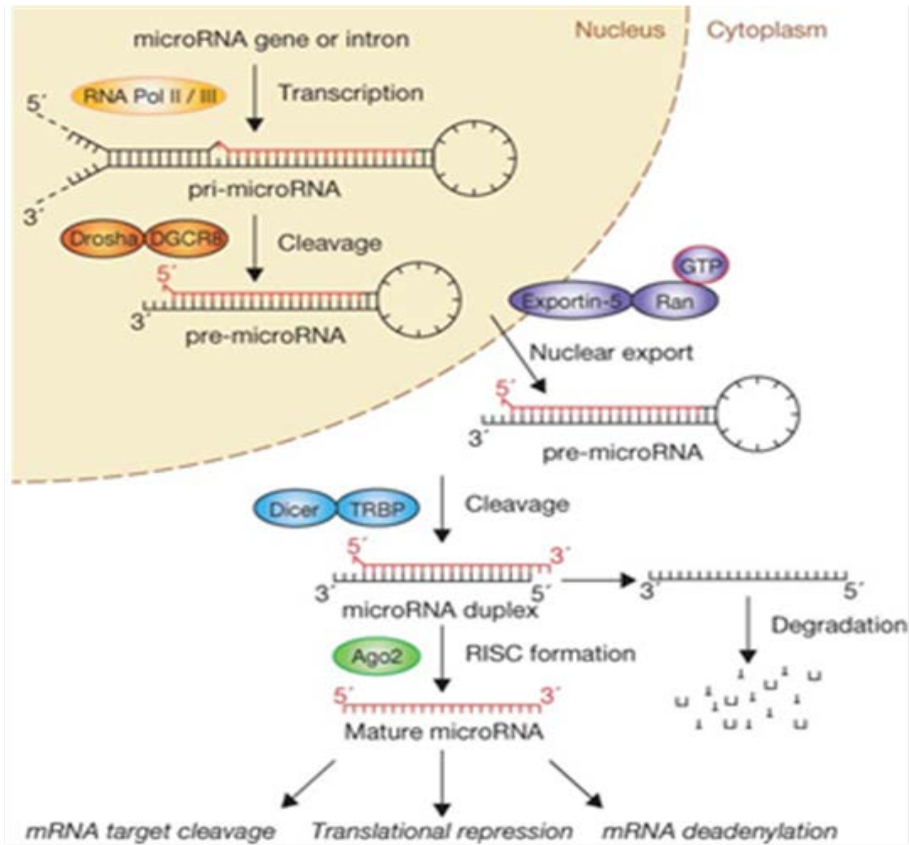
## **1.4 MicroRNA**

MicroRNAs (miRNAs) are a class of small non-coding RNAs approximately 22 nucleotides in length. Emerging evidence has shown that miRNAs function as oncogenes or tumor suppressor genes depending on the context (138-140) and have been shown to be potential biomarkers for cancer risk assessment, clinical treatment response, and prognosis (141).

### **1.4.1 MiRNA biogenesis**

MiRNAs undergo a complex processing procedure to produce the mature, functional unit. The initial step is the generation of pri-miRNA from the miRNA gene transcript through a

series of RNases. These pri-miRNA transcripts are then cleaved by Drosha, an RNase III endonuclease, producing an 85-nucleotide hairpin structure termed pre-miRNA. After exportation into the cytoplasm by Exportin-5/Ran-GTP complex, pre-miRNAs are further processed by DICER into an 18-25 nucleotide intermediate duplex. A single strand of this mature miRNA then becomes part of the RNA-inducing silencing complex (RISC) together with various other proteins, such as TARBP2, AGO2, GEMIN3, and GEMIN4. This complex then binds to the target mRNA to regulate gene function either through cleavage of the transcript by the RISC complex or induction of translational silencing through RNA-RNA interactions(142). Impaired miRNA processing has been reported to reduce stable miRNA levels and promote tumorigenesis (143), and genetic variations in several miRNA processing genes have been reported to influence the risk of several cancers, including bladder, esophageal and kidney cancer(144-146).



**Figure 2** The scheme of miRNA biogenesis and regulation

(147) Reprinted by permission from Macmillan Publishers Ltd: Nat Cell Biol, copyright (2009)

#### 1.4.2 miRNA binding site polymorphisms

Although miRNA genes are highly conserved with very few known genetic variations in the mature miRNA regions, the frequency of variations within miRNA target sites, which are located in less conserved 3' untranslated regions (UTRs), is much greater(148). Genetic variations within these sites are of interest because single nucleotide polymorphisms (SNPs) in the miRNA binding site may either disrupt the binding ability on an existing binding site or create a previous non-existing binding site, thus altering normal gene expression.

Recently, there has been an increasing interest in exploring miRNA binding site polymorphisms and their association with human diseases, ranging from mental disorders to cancers (149). Given the significant role of miRNA regulation, the fast growth of this field might revolutionize the way of cancer risk and prognosis prediction, and also help clinician to tailor personalized cancer therapy.

#### 1.4.3 miRNA and cancer

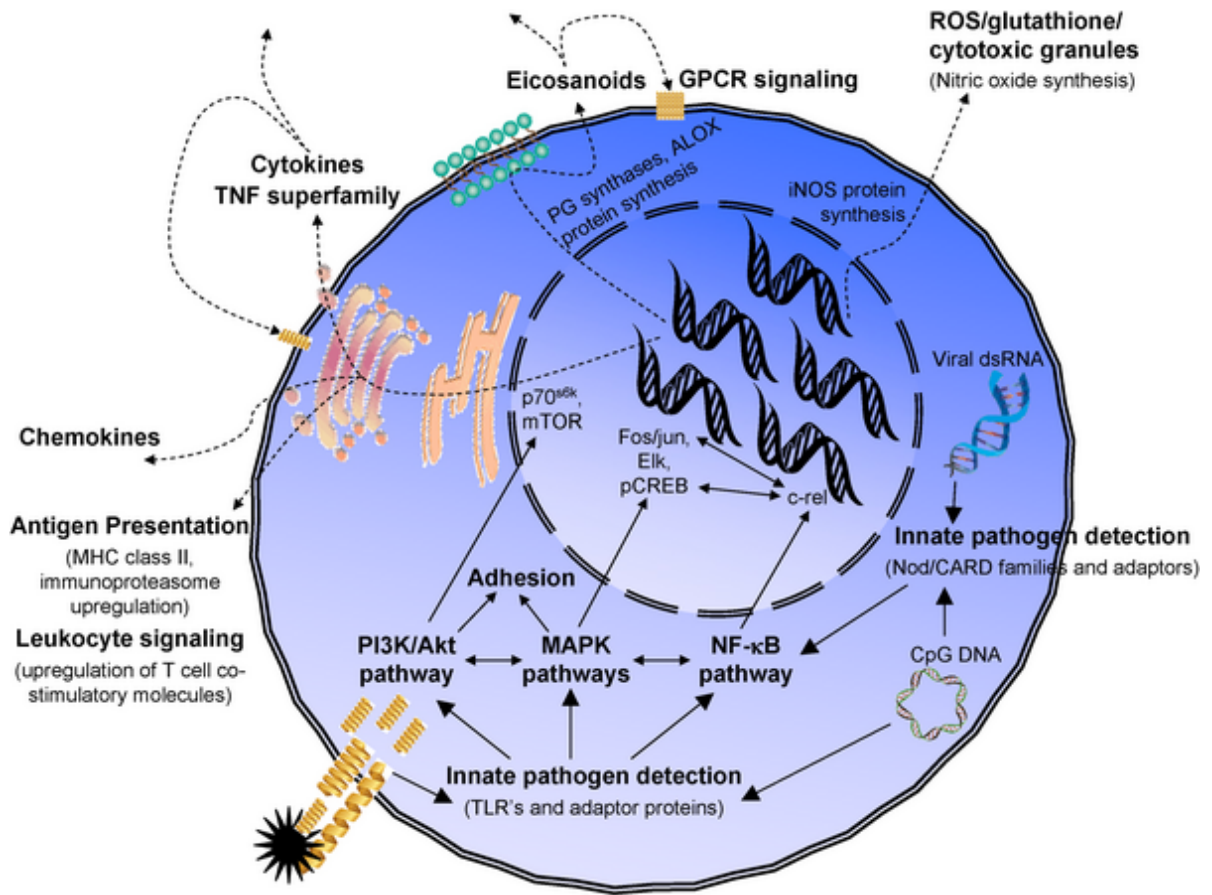
Impaired miRNA processing has been reported to reduce stable miRNA levels and promote tumorigenesis (143). Genetic variations in several miRNA processing genes have been reported to influence the risk of several cancers, including bladder, esophageal, kidney, and ovarian cancers (144, 146, 150, 151). In addition, variation in miRNA binding sites within 3' untranslated regions (3'UTR) of target genes may also affect miRNA-mRNA interaction and target gene expression, leading to altered cancer risk (152-157).

Evidence has shown that miRNAs are related to cancer prognosis including lung cancer. For examples, Yanaihara et al have reported that high hsa-miR-155 and low hsa-let-7a-2 expression correlated with poor survival (141), and distinct miRNA expression profile was

repeatedly observed between normal and tumor tissue of lung cancer patients (141, 158, 159).

### **1.5 Inflammation pathways**

Inflammation is an important cellular process that is activated in response to tissue damage, infections and other cellular processes (figure 2). However, a growing body of evidence supported a relationship between inflammation and cancer, with many cancers initiated at the site of inflammation. Products of the inflammatory response, such as free oxygen radicals, may induce harmful DNA alterations resulting in carcinogenesis and formation of invasive and/or metastatic phenotypes (160-165). Inflammatory cells and related signaling molecules could also be utilized by tumor to facilitate its progression and metastasis by generating a favorable micro-environment as well as promoting genetic instability and angiogenesis (161).



**Figure 3**inflammation pathways in response to a danger signal

(135)

The lung is a frequent site of infection and occasional site of chronic inflammation owing to environmental exposures. Furthermore, accumulating evidence shows that inflammation is associated with prognosis of various cancers, including lung cancer (166-169).

Poorer survival was found in cancer patients with elevated inflammatory markers. For example, high regulated Cox-2, which is a major enzyme involved in inflammatory response, is found in lung cancer, and associated with immune suppression, VEGF over expression, and also promotes angiogenesis and tumor invasions (67-70). Many studies have provided evidence that Cox-2 expression is a prognostic factor for NSCLC (68, 170, 171). In a study of 162 resected NSCLCs, more than 7 years difference in the median survival time was observed between patients with highest and lowest Cox-2 expression level (170, 171). Also, It is found that elevated circulating levels of C-reactive protein, an acute-phase reactant in inflammatory response, were associated with poor survival in NSCLC patients (168, 169). A few studies have explored associations between selected inflammation gene polymorphisms and lung cancer prognosis, with inconsistent results because of small sample sizes (172, 173).

## 1.6 Hypothesis and rationale

### 1.6.1 Hypothesis 1: miRNA-related genetic variations are associated with survival and recurrence in NSCLC patients

MicroRNAs (miRNA) post-transcriptionally regulate over 30% of human genes, miRNA was found de-regulated in most of human tumors. Evidence also showed miRNA are related to lung cancer prognosis. Given its important role, in this study, we hypothesized that miRNA-related polymorphisms, including polymorphism in miRNA processing genes, and miRNA binding sites in major cancer-related genes, could influence various cellular processes, such as tumor cell survival and drug response, thus have an impact on the clinical outcomes in NSCLC patients.

### 1.6.2 Hypothesis 2: Genetic variations in the inflammation pathways are associated with survival in late stage NSCLC patients

Lung cancer patients are usually diagnosed with advanced stage disease, which commonly treated with chemotherapy combination regimens. Inflammation has a well-established role with carcinogenesis, and it is estimated that inflammation contributes to 15% of cancer deaths. Evidence showed that inflammatory molecules and effectors not only increases the risk of developing cancer, but promotes tumor progression and mediate cancer patients' response to treatment and prognosis . Thus, we hypothesized that polymorphisms in major inflammation-related genes may affect inflammatory responses thus influence survival in late stage NSCLC patients



1.6.3 Hypothesis 3: Genetic variants in the inflammation pathway are associated with survival in never smokers among NSCLC patients

Lung cancer in never-smokers (LCINS) is increasingly recognized as a distinct disease from that in ever-smokers owing to substantial differences in etiology, clinical characteristics, and prognosis. Identification of specific prognostic and predictive markers for lung cancer in never-smokers beyond the general markers for lung cancer is warranted. Inflammation plays an important role in cancer initiation and progression, as well as influence clinical outcomes. In the present study, we hypothesized that inflammation-related genetic variants could influence host gene function and inflammatory responses, thus would have impact on NSCLC patients' prognosis through smoking independent mechanisms.

## *Chapter 2: Material and Methods*

## 2.1 Study populations and data collection

MD Anderson discovery population: Patients from The University of Texas MD Anderson Cancer Center included in this study were part of an ongoing lung study that has been recruiting since 1995. All patients were non-Hispanic white, had histologically confirmed (AJCC v6.) NSCLC. A structured questionnaire was used to collect epidemiologic and demographic data during an in-person interview with each patient. In addition, genomic DNA was extracted from peripheral blood samples obtained from each patient using the QIAamp DNA extraction kit (Qiagen, Valencia, CA), following standard protocol. Clinical and follow-up data were obtained from medical records. Each patient signed an informed consent form, and this study was approved by the MD Anderson Institutional Review Board.

Harvard University population: The details of the Harvard population have been described in detail previously (174). In brief, this lung cancer study was initiated in 1992; patients were recruited at the Massachusetts General Hospital. All participants in that study were at least 18 years old white patients with a confirmed primary lung cancer. An interviewer-administered questionnaire was used to collect epidemiologic data (demographics, occupational exposures, smoking history) for each patient. Peripheral blood was drawn from each patient for DNA extraction.

Mayo Clinic population: Patients at Mayo Clinic had newly diagnosed, histopathological confirmed primary NSCLC. A structured questionnaire was used to collect detailed epidemiological data on the patients. These patients participated in a long-term follow-up study from 1997 to 2008 described in detail previously (175, 176). Medical records were

reviewed for clinical and epidemiological data abstraction. All analyses were restricted to Caucasian patients to minimize effects of population structure.

## **2.2 SNP selection, genotyping and quality control**

### **2.2.1 miRNA related SNPs**

Gene and SNP selection: We had previously constructed a custom Illumina iSelect chip containing a comprehensive panel of approximately 10,000 SNPs from 998 cancer-related genes. The detailed description of this chip, including the SNP and gene selection schema, has been described previously (177). Eight miRNA processing genes (DDX20, DGCR8, DICER1, RNASEN, EIF2C1, GEMIN4, RAN, and XPO5) were among the 998 genes on this chip with 77 tagging (10 kb flanking and within each gene, linkage coefficient  $r^2 > 0.8$ ) and potential functional SNPs genotyped. We used the PolymiRTS database (Bao, Zhou et al. 2007) to identify SNPs in predicted binding sites for the 998 genes included on the chip and identified a total 163 SNPs from 133 genes with these criteria. All the selected SNPs had a minor allele frequency (MAF) greater than 0.01 in the Caucasian population. Table 3 listed the genes we selected in our study, and supplementary table 1 provides the entire list of all SNP analyzed.

**Table 3: miRNA Processing and predicted targets genes**

Processing Gene	Target genes					
DICER1	ACVR1B	C8orf49	E2F7	IGF2BP1	POLH	SPP1
DDX20	ADH5	C9orf9	EPHX2	IGFBP2	PON1	SST
DGCR8	ADH6	CASP2	EPS8L3	IGFBP5	RAD51L3	SSTR1
EI2FC1	ALDH18A1	CASP7	ERN1	IL1R1	RET	SSTR2
GEMIN4	ALDH5A1	CASP8	FANCD02	KRAS	RICTOR	ST7L
RAN	ANGPT4	CAV1	FAS	MBD1	RNF175	STK6
RNASEN (Drosha)	ANGPTL1	CAV2	FGF2	MDM4	RPA1	SUFU
XPO5	ATG4A	CD34	FGF5	MLL	RPS6KA3	SULT4A1
	ATG9A	CD4	FGF9	MTHFR	RPS6KB1	TLR4
	ATP5A1	CD44	FLJ35220	MTR	RPS6KB2	TNFRSF10D
	ATP5L	CDC7	FLJ38991	NAT1	RRM1	TNFRSF21
	ATP6V1C1	CDKN1B	FZD3	NDUFA6	RRM2B	TNFSF10
	BAG3	CDKN2A	FZD4	NEIL2	RXRA	TPM2
	BAG5	CDKN2C	GHITM	NFAT5	SETD1A	TXN2
	BAX	COL18A1	GHRHR	NFKBIB	SIRT3	UGT2A3
	BCL2L11	COX4NB	GPR30	NODAL	SMAD1	VDR
	BCL2L2	DCTN5	GPX3	NOTCH1	SMAD3	VEGF
	BIRC4	DDB2	GPX7	NR1I2	SMAD7	WNT11
	BIRC5	DGCR8	GSTM3	NUDT6	SMC1L2	WNT2B
	BIRC6	DICER1	HIP1	OGG1	SMO	XRCC5
	BNIP3L	DNMT3B	HSPB8	PDGFC	SNAI1	
	BTBD10	E2F2	IGF2AS	PGRMC2	SP1	

Genotyping and quality control: Genomic DNA was extracted from peripheral blood samples using the QIAamp DNA extraction kit (Qiagen, Valencia, CA) following manufacturer's protocol. SNPs genotyping was performed using iSelect Infinium II genotyping platform (Illumina, San Diego, CA, USA) according to the standard Infinium II assay protocol. Only SNPs with a cluster call rate  $>0.95$  were included in the analysis. DAVID gene ontology database (<http://david.abcc.ncifcrf.gov/home.jsp>) was used to analyze gene function clustering.

### 2.2.2 inflammation related SNPs

Gene and SNP selection: Compilation of the genes involved in the inflammatory response was performed based on a published panel of inflammation-associated genes (135) and a database of diabetes and inflammation genes (T1DBase [<http://www.t1dbase.org>]; University of Cambridge, Cambridge, UK). Tagging SNPs for candidate genes based on data from an European population were identified using data from the International HapMap Project, based on National Center for Biotechnology Information B36 assembly and dbSNP b126. For each gene, sequences 10 kb before the transcription start site and 10 kb after the transcription end site were included in the tag SNP selection using the Tagger pairwise method (Broad Institute, Cambridge, MA, USA) with an  $r^2$  threshold of 0.8 and minor allele frequency of at least 0.05 (178). The compiled SNP list was sent to Illumina (San Diego, CA, USA) for designability analysis using their array design tool. Only SNPs that exceeded the threshold score ( $>0.4$ ) were considered designable. In total, 11,930 SNPs (supplementary table 1) were included for construction of an Infinium II iSelect Custom Genotyping BeadChip (Illumina).

**Table 4: Inflammation-related pathways selected**

Pathway	No. of genes	No. of SNPs
Adhesion-extravasation-migration	12	108
Apoptosis signaling	67	834
Complement cascade	3	8
Cytokine signaling	266	3139
Glucocorticoid/PPAR signaling	24	258
Innate pathogen detection	53	542
Leukocyte signaling	132	2023
MAPK signaling	156	2854
Natural killer cell signaling	31	296
Phagocytosis-Ag presentation	41	488
PI3K/AKT signaling	45	580
ROS/glutathione/cytotoxic granules	25	231
TNF superfamily signaling	49	569
Total	904	11930

PPAR=peroxisome proliferator-activated receptor; MAPK=mitogen-activated protein kinase; PI3K=phosphatidylinositol 3-kinase; ROS=reactive oxygen species; TNF=tumor necrosis factor.

Genotyping and quality control: Genotyping for inflammation SNPs were performed at MD Anderson Cancer Center, Mayo Clinics and Harvard University using different platforms for discovery and validation purpose:

1) MD Anderson: detailed genotyping and quality control methods used in the discovery phase have been previously described (179). Briefly, genotyping was performed according to the standard Infinium II assay protocol for the iSelect HD BeadChips (Illumina). Quality control measures were applied to the datasets, excluding any DNA samples or SNPs with a call rate (percentage of data available for all SNPs or samples) <95%. For patients with direct relatives also enrolled in the study, only 1 patient within the relationship, the one whose DNA sample had a higher SNP call rate, was included in the final analysis. SNPs with MAF <0.01 were excluded. For validation purpose, genotyping for SNPs selected in the discovery phase was done either through the design of a custom Illumina Infinium iSelect BeadChip or using existing Illumina HumanHap300/ HumanHap317/ HumanHap660 genotyping data. Quality control for the Illumina Infinium iSelect BeadChip was performed on the basis of sample and SNP call rates; we removed any samples or SNPs with a call rate <95%. Detailed quality control measures for the Illumina HumanHap300/HumanHap317/HumanHap660 BeadChip have been described previously; these were also based on genotyping call rate (call rate >95% for all samples and SNPs included). SNPs with MAF <0.01 were also excluded (180).

2) Mayo clinics: SNPs selected for validation at Mayo Clinic were genotyped at Mayo Clinic's Genotyping Core Facility using a Fluidigm Dynamic Array (South San Francisco, CA, USA) and a HumanHap317 BeadChip (Illumina) according to a standard protocol and using quality control measures.



3) Harvard University: Genotyping for externally validated SNPs was performed using the Illumina HumanHap610 chip following standard protocol, as previously described (123). Quality control measures were similar to those used in the MD Anderson populations: only SNPs and samples with a genotyping call rate >95% and SNPs with MAF >0.01 were included in the analysis.

### 2.3 Statistical analyses

Demographic and clinical variables by vital status were selected compared using the  $\chi^2$  and Fisher's exact test. The multivariable Cox proportional hazards regression models, with corresponding hazard ratios (HRs) and 95% confidence intervals (CIs), were used to estimate the effect of single SNPs on overall survival (the time between diagnosis and death or last follow-up) and progression (time from start of treatment to progression or last follow-up) based on the best fitting model. Kaplan-Meier survival curves and corresponding log-rank tests were used to estimate the effect of each SNP on time to death. Patients who had smoked fewer than 100 cigarettes over their lifetime were defined as never-smokers; ever-smokers were defined as patients who had smoked more or equal to 100 cigarettes over their lifetime, including former smokers (those who had quit smoking more than 1 year before diagnosis), and current smokers and recent quitters (those who had quit smoking within a year before diagnosis). Meta-analysis of the different populations was performed to obtain summary HRs and 95% CIs. Heterogeneity was tested using chi-square-based Q-statistics. A fixed-effect model was used when heterogeneity was absent (P for heterogeneity >0.05). The cumulative effect of the top 2 validated SNPs within each population was determined by counting the number of unfavorable genotypes (UFGs) each patient carried and using

patients without any UFGs within that population as a reference group. All the statistical analyses above were performing using STATA software (Stata Corporation, College Station, TX). Survival tree analysis was performed to identify higher-order gene-gene interactions affecting progression and/or survival using the STREE program (<http://masal.med.yale.edu/stree/>). STREE uses a log-rank statistic method to select the optimal split and subsequent split of the data set, each terminal node represented a group of patients who had the same genotype combination and risk profile. Multiple hypothesis testing was performed using R package with a q value (181), adjustment for multiple comparisons was based on a false discovery rate (FDR) of 5%. Bootstrap re-sampling method (by generating sample with duplicates for 1000 times) was used to internal validate the associations remained significant after multiple comparison ( $q < 0.05$ ). In case multiple SNPs were highly linked ( $R^2 > 0.8$ ), only one was kept for multiple SNPs analysis. Polyphen-2 (<http://genetics.bwh.harvard.edu/pph2/index.shtml>) and SIFT (<http://sift.bii.a-star.edu.sg/>) were used in silico to predict the influence of the validated missense SNP on protein function (182, 183).

#### **2.4 Luciferase reporter assay**

Selected miRNA binding site SNPs, *FAS*:rs2234978 and *SPI*:rs17695156, were evaluated *in vitro* using the dual-luciferase reporter assay. Due to the characteristics of the genomic sequence of the 3'UTR of *FZD4*, this region was unable to be cloned.

Luciferase reporter constructs for wildtype and variant allele containing binding site regions were generated. Briefly, a part of 3'UTR of each gene was amplified by PCR from genomic DNA and then cloning restriction sites were generated by nested PCR. The *FAS*

SNP (rs2234978) is located in exon 7, which serves as 3'UTR of a nonsense-mediated mRNA decay transcript (NCBI dbSNP database; [www.ncbi.nlm.nih.gov/projects/SNP](http://www.ncbi.nlm.nih.gov/projects/SNP)). Therefore, the entire exon sequence was generated by oligo hybridization. The PCR product or DNA fragment was digested with XbaI and FseI restriction enzymes (New England Biolabs, Ipswich, MA) and ligated into similarly digested pGL3 vector attached to 3' end of luciferase reporting gene. The variant allele containing vectors were generated by site-directed mutagenesis. All the constructs were sequenced to ensure the correct sequence. Primers and oligos used in reporter construct cloning are available upon request. Two lung cancer cell lines NCI-H460 (large cell carcinoma) and NCI-H2444 (adenocarcinoma) were cultured in RPMI-1640 medium (Mediatech, Manassas, VA) supplemented with 10% fetal bovine serum (Invitrogen, Carlsbad, CA) in 48-well tissue culture plates. Cells were transfected with 0.5 mg of each reporter construct, 5 pmol of negative control (scrambled sequence), or predicted targeting miRNAs (Sigma-Aldrich, St. Louis, MO) and 8 ng of pGL4 (Ambion, Austin, TX) Renilla luciferase reporter using Lipofectamine 2000 (Invitrogen). After 36 hours of incubation, cell lysates were harvested and measured for luciferase activity using the Dual-Luciferase Reporter Assay System (Promega, Madison, WI) and a FLUOstar Optima microplate reader (BMG Labtech, Cary, NC). Each assay was repeated independently at least two times with four replicates. The firefly luciferase activity was normalized to the Renilla luciferase activity to derive the relative luciferase activity.

## *Chapter 3: Results and Discussion*

### 3.1 miRNA-related genetic variations and clinical outcomes in NSCLC patients

#### 3.1.1 miRNA-related genetic variations and survival and recurrence in early stage NSCLC patients

##### 3.1.1.1 Patients characteristics

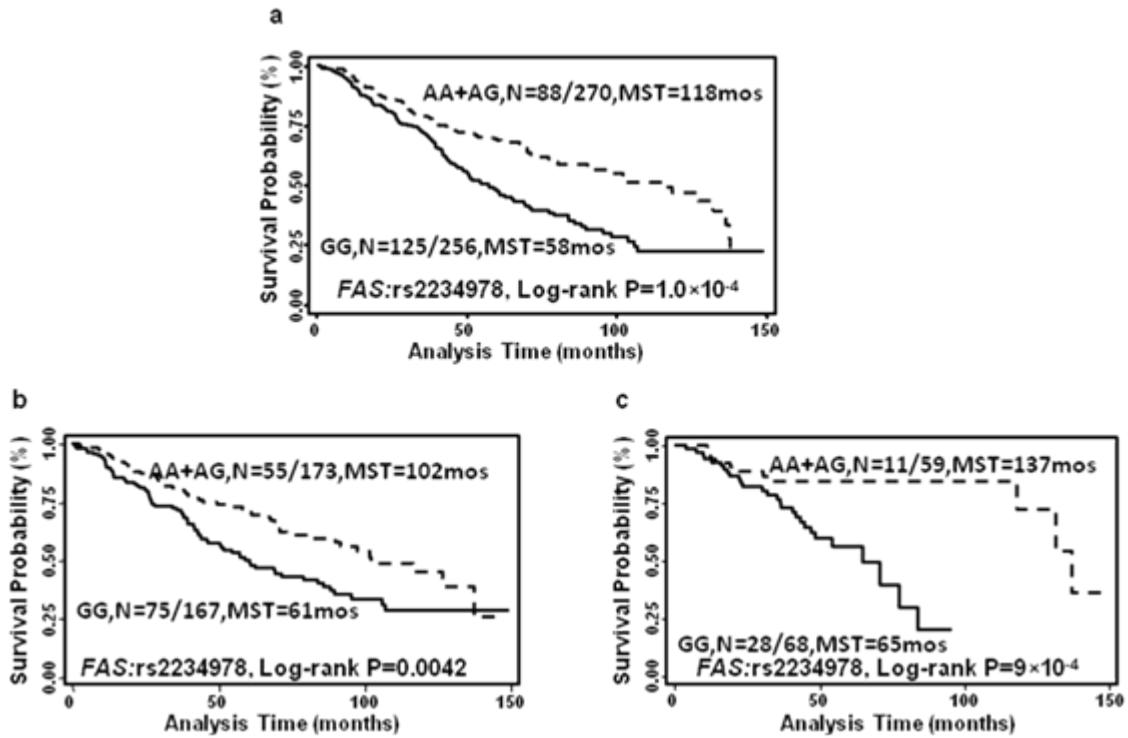
This study included 535 early stage (I and II) NSCLC patients with an overall median survival time of 90.2 months and median follow-up time of 62.1 months. Characteristics of the study population are shown in Table 5. Mean ages for surgery-only and surgery plus chemotherapy treated patients were 65.8 years and 62.9 years, respectively. At the time of analysis, 322 (60%) of the patients were alive and 360 (67%) did not have a progression of their disease. Nearly equal numbers of male and female participants were included (49% and 51% respectively) with a majority of the study population consisting of Caucasian patients (88%). The clinical stage distribution is stage IA (46%), stage IB (35%), stage IIA (5%) and stage IIB (14%). A majority of the NSCLC cases were adenocarcinomas (59%) with 28% squamous cell carcinoma and the remaining 13% unclassified or other NSCLC. Of the 535 participants, 340 patients received surgery-only, 127 patients were treated with surgery plus neoadjuvant and/or adjuvant chemotherapy, and the remainder (68 patients) treated with only radiation therapy with/without surgery.

**Table 5: Host characteristics of early stage NSCLCs**

Variables	All early stage	Surgery only	Surgery & chemo
	No. of patients (%)	No. of patients (%)	No. of patients (%)
<b>Total Patients</b>	535	340	127
<b>Median survival time(mos)</b>	90.2	102.0	118.3
<b>Median follow-up time(mos)</b>	62.1	71.6	50.7
<b>Age, mean(sd)</b>	65.7(10.1)	65.8(9.9)	62.9(10.2)
<b>Gender</b>			
<i>Male</i>	262(49)	166(49)	68(54)
<i>Female</i>	273(51)	174(51)	59(46)
<b>Ethnicity</b>			
<i>Caucasian</i>	469(88)	305(90)	109(86)
<i>African-American</i>	42(8)	25(7)	10(8)
<i>Others</i>	24(4)	10(3)	8(6)
<b>Pack year, mean(sd)</b>	44.9(36.6)	45.1(37.6)	39.4(35.2)
<b>Histology</b>			
<i>Adenocarcinoma</i>	315(59)	213(63)	74(58)
<i>Squamous cell carcinoma</i>	149(28)	87(26)	34(27)
<i>Unclassified or other</i>	71(13)	40(12)	19(15)
<b>Clinical stage</b>			
<i>Stage IA</i>	245(46)	181(53)	23(18)
<i>Stage IB</i>	188(35)	113(33)	55(43)
<i>Stage IIA</i>	26(5)	10(3)	14(11)
<i>Stage IIB</i>	76(14)	36(11)	35(28)
<b>Treatment</b>			
<i>Surgery only</i>	340(64)	340(100)	N/A
<i>Surgery &amp; other treatment</i>	142(27)	N/A	127(100)
<i>Treatment without surgery</i>	53(10)	N/A	N/A
<b>Surgery result</b>			
<i>Complete</i>	470(98)	340(100)	123(97)
<i>Residual</i>	12(2)	N/A	4(3)
<b>Vital status</b>			
<i>Alive</i>	213(40)	210(62)	88(69)
<i>Dead</i>	322(60)	130(38)	39(31)
<b>Progression</b>			
<i>No</i>	360(67)	233(69)	85(67)
<i>Yes</i>	175(33)	107(31)	42(33)

### 3.1.1.2 Associations between individual SNPs and NSCLC clinical outcomes

Among all the variants analyzed, 11 processing and 23 binding site SNPs were significantly associated with altered risk of dying. The most significant association with survival for early stage NSCLC was *FAS*:rs2234978 (HR:0.59, 95% CI:0.44-0.77,  $P=1.67 \times 10^{-4}$ ,  $q=0.018$ ), which remained significant after multiple comparison corrections, and resulted in a significant increase in median survival time (MST) from 59 to 118 months (log rank  $P=1.0 \times 10^{-4}$ ; Figure 4a).

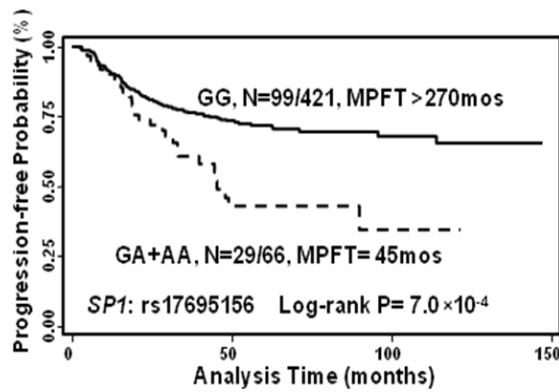


**Figure 4 Kaplan-Meier estimates of FAS:rs2234978 on overall survival:**

(a) total population; (b) surgery-only patients; (c) surgery plus chemotherapy patients. MST: median survival time in months. N=A/B, A: number of patients with event, B: total number of patients.



Five SNPs in processing genes and 23 SNPs in binding sites were significantly associated with time to progression. The most significant association, which remained significant after correcting for multiple comparisons, was *SP1*:rs17695156 (HR:2.22, 95% CI:1.44-3.41,  $P=3.00 \times 10^{-4}$ ,  $q=0.034$ ). Patients with at least one variant allele had more than 224 months decreased median progression-free time compared to patients who had common homozygous genotype (45.3 months vs >270 months, log rank  $P=7.0 \times 10^{-4}$ , Figure 5).



**Figure 5 Kaplan-Meier estimates on effect of *SP1*:rs17695156 on time to progression among early stage patients.**

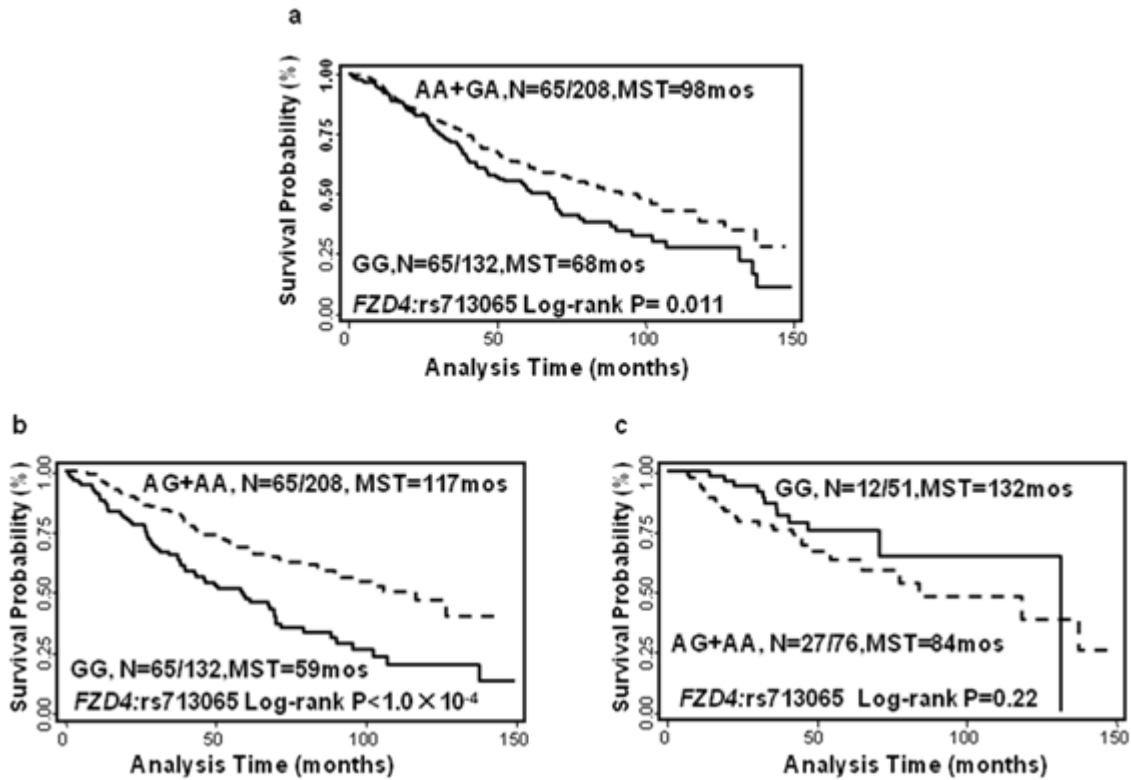
MPFT: median progression-free time in months.  $N=A/B$ , A: number of patients with event, B: number of all patients.

### 3.1.1.1 Effects of treatments on association of clinical outcomes

Different treatment regimens may function through different mechanisms to affect clinical outcomes. We performed subgroup analysis focusing on two groups of patients with relatively homogeneous treatment regimens: surgery-only and surgery plus chemotherapy.

#### *Effect on overall survival*

Eighteen SNPs were significantly associated with overall survival in surgery-only patients. *FZD4*:rs713065 (HR:0.46, 95% CI:0.32-0.65,  $P=2\times 10^{-5}$ ,  $q=0.002$ ), located in the 3' UTR of *FZD4*, remained significant after adjustment for multiple comparisons. Patients with at least one variant allele have significantly decreased risk of death and increased MST from 59 to 117 months compared those patients with the common genotype (log rank  $P=1.05\times 10^{-5}$ ; Figure 6a). Notably, in agreement with the overall population, for patients who received surgery plus chemotherapy, *FAS*:rs2234978 (HR:0.19, 95% CI:0.07-0.46,  $P=1.84\times 10^{-5}$ ) displayed the most significant association with survival. Patients with at least one variant allele had 81% lower risk of death (HR:0.19, 95% CI:0.07-0.46) with their MST increased by 2-fold, compared to patients who carry the homozygous common genotype (65 months vs. 137 months, log rank  $P=1.05\times 10^{-4}$ , Figure 4c). The association of this SNP with survival was borderline significant after correction for multiple comparisons in surgery-only patients (HR:0.59, 95% CI:0.42-0.84,  $P=0.004$ ,  $q=0.069$ ), with increased median survival time (61 months vs. 102 months, log rank  $P=4.02\times 10^{-3}$ , Figure 4b).



**Figure 6 Kaplan-Meier estimates of effect of FZD4:rs713065 on overall survival**

(a) total population; (b) surgery-only patients. (c) surgery plus chemotherapy patients. MST: median survival time in months. N=A/B, A: number of patients with event, B: total number of patients.

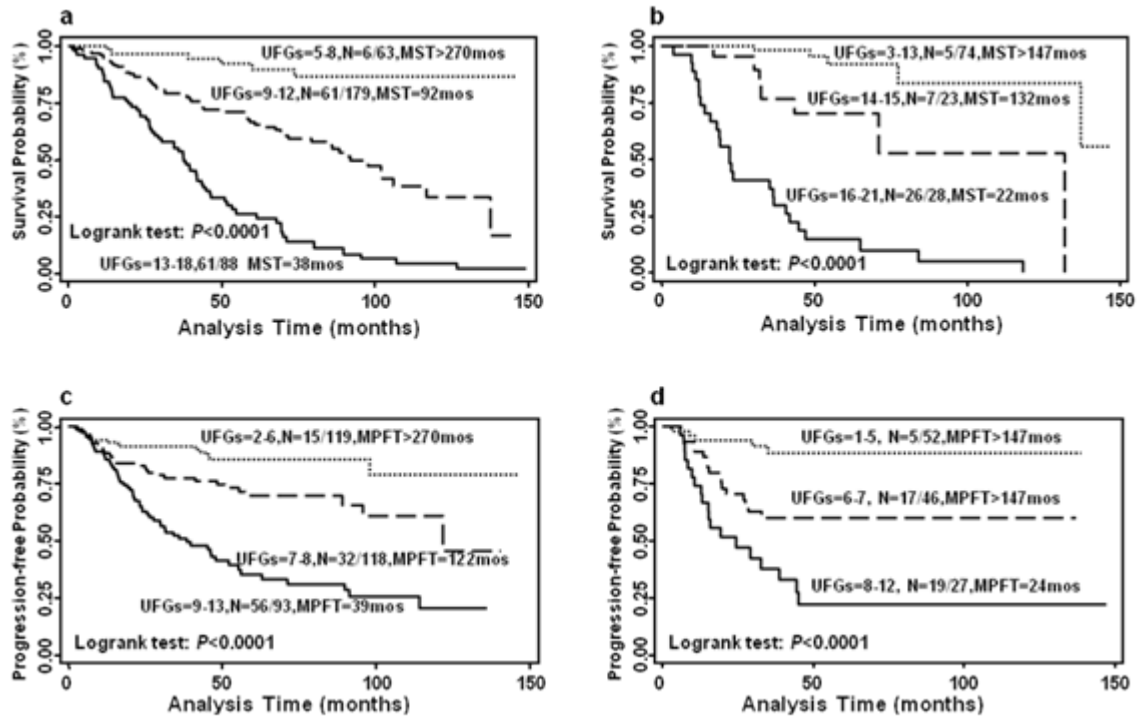
By comparing the findings between two subgroups, we identified two distinct clusters of SNPs that showed statistically significant association with survival in only one specific treatment subgroup. Seventeen SNPs (13 binding site and 4 processing SNPs) were found to have significant effects on risk of dying in surgery-only patients but no significant associations with survival in patients receiving surgery plus chemotherapy. In contrast, 28 SNPs (18 binding site and 10 processing SNPs) were found to be significantly associated with risk of death only in patients receiving surgery plus chemotherapy, but not in surgery-only patients. Intriguingly, within each cluster of SNPs, we identified SNPs with differential directions of their effect. A group of 29 SNPs has the same trend in both treatment subgroups (either protective or adverse), while 15 SNPs conferred opposite effects between two subgroups (Table 6). For example, *FZD4*:rs713065 was associated with significantly decreased risk of dying and prolonged survival time in surgery-only patients; however, in the surgery plus chemotherapy subgroup, this SNP was associated with increased risk of dying and a shortened median survival time (Figure 6b and 6c). Significant dose-dependent effects on risk of dying were identified in the two treatment subgroups with patients carrying increased number of UFGs showing a significant trend toward poorer survival and shortened median survival time ( $P$  for trend  $< 1 \times 10^{-4}$ , Figure 7a and 7b).

**Table 6: Effect of selected SNPs on survival in early stage NSCLC patients**

Gene	SNP	Curative Intent		Surgery-only		Surgery & chemotherapy	
		HR (95%CI)**	P	HR (95%CI)**	P	HR (95%CI)**	P
FAS	rs2234978	0.59(0.44-0.77)	<b>2×10<sup>-4</sup>*</b>	0.59(0.42-0.84)	<b>4×10<sup>-3</sup></b>	0.19(0.07-0.46)	<b>3×10<sup>-4</sup>*</b>
FZD4	rs713065	0.68(0.51-0.90)	<b>0.01</b>	0.46(0.32-0.65)	<b>2×10<sup>-5</sup>*</b>	1.50(0.71-3.19)	0.29
PON1	rs854552	1.94(1.29-2.92)	<b>2×10<sup>-3</sup></b>	2.29(1.36-3.88)	<b>2×10<sup>-3</sup></b>	2.00(0.87-4.60)	0.10
WNT2B	rs3790611	1.62(0.97-2.71)	0.06	2.60(1.39-4.86)	<b>3×10<sup>-3</sup></b>	0.45(0.09-2.11)	0.31
DDX20	rs197412	1.66(1.18-2.32)	<b>3×10<sup>-3</sup></b>	1.87(1.24-2.82)	<b>3×10<sup>-3</sup></b>	0.97(0.32-2.98)	0.96
ATP5A1	rs12954944	1.52(1.07-2.16)	<b>0.02</b>	1.74(1.14-2.65)	<b>0.01</b>	0.60(0.20-1.87)	0.38
DGCR8	rs11089328	1.38(0.96-1.98)	0.08	1.79(1.14-2.79)	<b>0.01</b>	0.90(0.31-2.62)	0.84
SMC1L2	rs3747238	1.55(1.11-2.16)	<b>0.01</b>	1.77(1.17-2.68)	<b>0.01</b>	0.96(0.37-2.47)	0.93
RAN	rs872396	1.50(1.08-2.09)	<b>0.02</b>	1.79(1.17-2.75)	<b>0.01</b>	1.28(0.57-2.85)	0.55
CDK4	rs1048691	0.43(0.21-0.89)	<b>0.02</b>	0.31(0.11-0.86)	<b>0.02</b>	1.33(0.38-4.63)	0.65
DROSHA	rs7719666	0.87(0.72-1.06)	0.16	0.74(0.57-0.96)	<b>0.02</b>	1.09(0.68-1.74)	0.73
NEIL2	rs1043180	1.50(1.07-2.10)	<b>0.02</b>	1.63(1.07-2.47)	<b>0.02</b>	1.39(0.63-3.06)	0.42
RPS6KA3	rs12010722	0.83(0.67-1.03)	0.08	0.73(0.55-0.97)	<b>0.03</b>	0.91(0.54-1.53)	0.72
SST	rs4988514	0.75(0.51-1.10)	0.15	0.60(0.37-0.98)	<b>0.04</b>	1.16(0.48-2.80)	0.75
ADH5	rs7669660	0.78(0.56-1.08)	0.13	0.63(0.41-0.97)	<b>0.04</b>	1.00(0.43-2.34)	0.99
RPA1	rs1131636	0.94(0.77-1.15)	0.57	0.75(0.58-0.98)	<b>0.04</b>	0.77(0.47-1.27)	0.30
GSTM3	rs15864	0.82(0.65-1.02)	0.08	0.75(0.56-0.99)	<b>0.04</b>	0.96(0.50-1.86)	0.91
TLR2	rs7695605	1.28(0.97-1.70)	0.09	1.44(1.01-2.05)	<b>0.05</b>	1.58(0.80-3.13)	0.19
SULT1C1	rs1047312	1.15(0.98-1.35)	0.08	1.08(0.88-1.32)	0.45	2.01(1.30-3.10)	<b>2×10<sup>-3</sup></b>
DROSHA	rs669702	1.26(0.91-1.75)	0.16	1.08(0.71-1.66)	0.72	3.49(1.59-7.64)	<b>2×10<sup>-3</sup></b>
GPR30	rs1133043	1.13(0.75-1.69)	0.56	0.67(0.37-1.21)	0.19	3.57(1.55-8.21)	<b>3×10<sup>-3</sup></b>
FANCD2	rs3172417	0.90(0.62-1.32)	0.59	0.71(0.42-1.19)	0.19	3.32(1.41-7.81)	<b>0.01</b>
NDUFA6	rs7245	1.08(0.78-1.50)	0.64	0.83(0.52-1.31)	0.42	2.76(1.29-5.88)	<b>0.01</b>
CDC7	rs12125947	1.22(0.99-1.49)	0.06	1.15(0.88-1.52)	0.3	1.90(1.14-3.17)	<b>0.01</b>
PDGFC	rs1425486	1.43(1.06-1.93)	<b>0.02</b>	1.18(0.81-1.71)	0.39	3.13(1.39-7.06)	<b>0.01</b>
FOXO1A	rs9532558	1.71(0.79-3.68)	0.17	1.44(0.52-4.04)	0.48	9.03(1.85-44.1)	<b>0.01</b>
BIRC4	rs17330637	1.57(1.04-2.38)	<b>0.03</b>	1.15(0.64-2.08)	0.64	2.79(1.25-6.25)	<b>0.01</b>
SMC1L2	rs3747240	1.29(0.94-1.77)	0.12	1.08(0.69-1.69)	0.73	2.59(1.30-5.19)	<b>0.01</b>
TNFRSF10D	rs7957	1.24(0.92-1.65)	0.15	1.01(0.69-1.49)	0.94	2.55(1.23-5.29)	<b>0.01</b>
RPS6KB2	rs10274	0.83(0.68-1.02)	0.08	0.85(0.66-1.10)	0.21	0.51(0.30-0.87)	<b>0.01</b>
FZD3	rs352222	0.82(0.68-1.00)	<b>0.05</b>	0.86(0.67-1.09)	0.21	0.46(0.25-0.82)	<b>0.01</b>
DROSHA	rs10035440	1.01(0.76-1.34)	0.95	1.27(0.89-1.82)	0.19	0.37(0.16-0.88)	<b>0.02</b>
IGF2BP1	rs6504593	1.30(0.94-1.80)	0.11	1.24(0.83-1.87)	0.3	2.90(1.16-7.28)	<b>0.02</b>
DROSHA	rs673019	1.27(0.92-1.77)	0.15	1.25(0.81-1.95)	0.31	2.40(1.13-5.08)	<b>0.02</b>
SP1	rs17695156	1.52(1.03-2.24)	<b>0.03</b>	1.20(0.68-2.13)	0.53	2.44(1.13-5.29)	<b>0.02</b>
RAN	rs10848238	1.40(1.06-1.86)	<b>0.02</b>	1.19(0.82-1.72)	0.36	2.15(1.11-4.17)	<b>0.02</b>
PMS2L3	rs1167829	0.86(0.68-1.10)	0.24	0.97(0.72-1.31)	0.84	0.44(0.22-0.86)	<b>0.02</b>
DROSHA	rs639174	0.96(0.56-1.66)	0.89	0.80(0.39-1.65)	0.55	2.86(1.09-7.49)	<b>0.03</b>
DROSHA	rs2302905	1.04(0.78-1.39)	0.81	0.91(0.62-1.33)	0.62	2.25(1.11-4.58)	<b>0.03</b>
MDM4	rs10900596	1.31(1.05-1.63)	<b>0.02</b>	1.29(0.97-1.72)	0.07	1.66(1.04-2.66)	<b>0.03</b>
ICAM1	rs281437	1.44(1.09-1.89)	<b>0.01</b>	1.38(0.96-1.98)	0.08	2.20(1.10-4.39)	<b>0.03</b>
DICER1	rs1187642	1.35(0.91-2.01)	0.14	1.00(0.58-1.73)	0.99	2.61(1.09-6.27)	<b>0.03</b>
DGCR8	rs2073778	0.78(0.54-1.11)	0.16	0.89(0.57-1.39)	0.61	0.30(0.10-0.88)	<b>0.03</b>
DGCR8	rs720012	0.76(0.53-1.09)	0.13	0.89(0.56-1.41)	0.62	0.30(0.10-0.90)	<b>0.03</b>
RAN	rs11061209	1.50(1.14-1.98)	<b>4×10<sup>-3</sup></b>	1.44(1.00-2.05)	0.05	2.13(1.04-4.36)	<b>0.04</b>
DNMT3B	rs6058896	1.39(0.93-2.07)	0.11	1.59(0.88-2.89)	0.13	2.37(1.00-5.59)	<b>0.05</b>

\* Remain significant after multiple comparisons using FDR of 5%

\*\*Adjusted by age, gender, ethnicity, stage, pack year and treatment regimens.



**Figure 7 Kaplan-Meier estimates of overall survival and time to progression in early stage NSCLC patients grouped by the number of unfavorable genotypes (UFG)**

(a) estimates for survival in surgery-only patients; (b) estimates for survival in patients receiving surgery plus chemotherapy; (c) estimates for time to progression in surgery-only patients; (d) estimates for time to progression in patients receiving surgery plus chemotherapy; MST: median survival time in months. MPFT: median progression-free time in months.  $N=A/B$ , A: number of patients with unfavorable event, B: total number of patients in subgroup.

### *Effect on progression*

In the subgroup analysis of treatment regimen for progression risk, *SPI:rs17695156*, which is the top SNP associated with progression in the combined population, was the most significant association with progression in surgery plus chemotherapy group with a borderline significant q value (HR:3.36, 95% CI:1.62-6.69,  $P=1.10\times 10^{-3}$ ,  $q=0.089$ ) (Table 7). One processing SNP, *DROSHA:rs6886834* was significantly associated with more than 6 times increased risk for progression in surgery-only patients (HR:6.38, 95%CI:2.49-16.31,  $P=1.10\times 10^{-4}$ ,  $q=0.011$ ) (Table 7). Patients who carried at least one variant allele of this SNP had significant reduction in progression-free time compared with patients with common homozygous genotype (23 months vs >270 months, log rank  $P=5.0\times 10^{-4}$ ; Figure 8). This association remained significant after correction for multiple comparisons (Table 7).

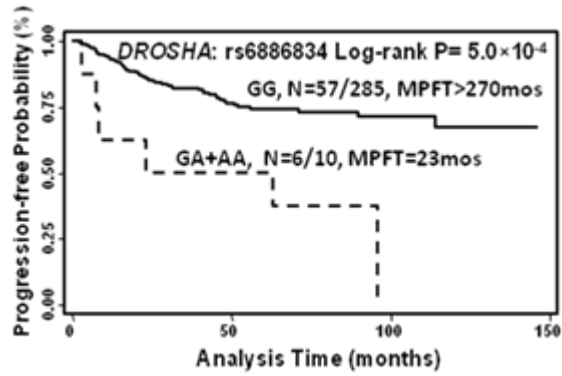
**Table 7: Effect of selected SNPs on progression in early stage NSCLC patients**

Gene	SNP	Model	Curative Intent		Surgery-only		Surgery & chemotherapy	
			HR (95%CI)**	P	HR (95%CI)**	P	HR (95%CI)**	P
DROSHA	rs6886834	REC	2.28(1.20-4.32)	<b>0.012</b>	6.38(2.49-16.31)	<b>1.1×10<sup>-4*</sup></b>		
MBD1	rs11663629	DOM	1.67(1.12-2.50)	<b>0.012</b>	2.33(1.35-4.02)	<b>0.002</b>	1.33(0.60-2.96)	0.487
ATP6V1C1	rs2453994	DOM	1.44(1.01-2.05)	<b>0.041</b>	2.07(1.24-3.46)	<b>0.006</b>	1.10(0.53-2.28)	0.793
RAN	rs11061209	REC	1.38(0.84-2.28)	0.201	2.40(1.25-4.61)	<b>0.008</b>	0.87(0.29-2.62)	0.804
DGCR8	rs1633445	ADD	1.28(0.96-1.71)	0.088	1.66(1.14-2.42)	<b>0.009</b>		
BAG3	rs8946	DOM	1.40(0.96-2.04)	0.084	2.20(1.21-4.00)	<b>0.010</b>	1.19(0.57-2.45)	0.644
DROSHA	rs502267	REC			3.11(1.29-7.51)	<b>0.012</b>		
CDC7	rs12125947	REC	1.91(1.27-2.88)	<b>0.002</b>	2.07(1.17-3.67)	<b>0.012</b>	2.10(0.92-4.79)	0.079
FEN1	rs4246215	DOM	0.75(0.52-1.08)	0.118	0.51(0.29-0.88)	<b>0.017</b>	1.21(0.61-2.40)	0.579
IGF2AS	rs10770125	DOM	0.77(0.53-1.11)	0.161	0.54(0.32-0.90)	<b>0.020</b>	1.17(0.56-2.43)	0.673
DGCR8	rs3757	ADD	1.28(0.95-1.71)	0.099	1.59(1.08-2.34)	<b>0.020</b>		
HSPB8	rs1133026	DOM	1.53(1.07-2.18)	<b>0.020</b>	1.81(1.09-3.00)	<b>0.022</b>	1.67(0.83-3.37)	0.148
WNT2B	rs3790611	REC	1.87(0.99-3.52)	0.052	2.79(1.15-6.73)	<b>0.023</b>		
ATP6V1C1	rs2248718	DOM	1.29(0.88-1.88)	0.195	1.81(1.09-3.02)	<b>0.023</b>	0.67(0.28-1.59)	0.362
DROSHA	rs7712155	REC	1.63(0.75-3.57)	0.218	2.94(1.16-7.45)	<b>0.023</b>		
DROSHA	rs12186785	DOM	1.77(1.15-2.71)	<b>0.009</b>	1.94(1.09-3.45)	<b>0.024</b>	1.33(0.54-3.25)	0.531
GSTM3	rs15864	REC	0.32(0.11-0.87)	<b>0.025</b>	0.19(0.05-0.81)	<b>0.025</b>	0.53(0.06-4.26)	0.547
DROSHA	rs10035440	DOM	0.71(0.48-1.06)	0.091	0.50(0.27-0.92)	<b>0.025</b>	1.33(0.66-2.68)	0.425
RING1	rs107822	DOM	0.82(0.57-1.17)	0.269	0.54(0.31-0.93)	<b>0.027</b>	1.41(0.72-2.74)	0.313
DGCR8	rs270014	ADD	1.25(0.93-1.67)	0.136	1.54(1.04-2.27)	<b>0.031</b>		
RRM1	rs1042927	DOM	1.27(0.78-2.06)	0.339	2.06(1.06-3.98)	<b>0.032</b>	0.53(0.20-1.38)	0.191
SMC1L2	rs3747238	DOM	0.78(0.53-1.14)	0.195	0.57(0.34-0.96)	<b>0.036</b>	1.03(0.49-2.16)	0.931
DROSHA	rs2287584	REC	0.91(0.44-1.89)	0.801	2.40(1.05-5.52)	<b>0.039</b>	0.44(0.10-2.01)	0.290
FEN1	rs174546	DOM	0.77(0.54-1.11)	0.159	0.58(0.34-0.98)	<b>0.042</b>	1.30(0.66-2.54)	0.445
DDX20	rs197412	REC	1.31(0.83-2.08)	0.249	1.85(1.01-3.38)	<b>0.045</b>	1.20(0.45-3.18)	0.716
FGF2	rs1048201	DOM	1.22(0.82-1.81)	0.325	1.72(1.01-2.92)	<b>0.047</b>	0.74(0.37-1.50)	0.408
IL1R1	rs3917328	DOM	0.63(0.32-1.26)	0.193	0.35(0.12-0.99)	<b>0.048</b>	1.48(0.44-5.01)	0.528
BIRC6	rs2710625	ADD	0.85(0.66-1.09)	0.207	0.69(0.48-0.99)	<b>0.049</b>	1.03(0.66-1.62)	0.895
SP1	rs17695156	DOM	2.22(1.44-3.41)	<b>3x10<sup>-4*</sup></b>	1.56(0.77-3.17)	0.213	3.36(1.62-6.96)	<b>0.001</b>
CASP7	rs1127687	DOM	1.61(1.12-2.29)	<b>0.009</b>	1.40(0.83-2.37)	0.212	3.15(1.58-6.30)	<b>0.001</b>
DROSHA	rs669702	DOM	1.41(0.94-2.13)	0.099	1.39(0.78-2.48)	0.261	2.93(1.39-6.20)	<b>0.005</b>
CASP7	rs10787498	ADD	0.85(0.66-1.11)	0.233	0.97(0.67-1.40)	0.871	0.50(0.30-0.86)	<b>0.012</b>
ANGPTL1	rs10913632	DOM	1.77(1.14-2.73)	<b>0.011</b>	1.39(0.70-2.76)	0.349	2.61(1.20-5.67)	<b>0.015</b>
DDX20	rs197377	DOM	1.36(0.95-1.96)	0.096	1.13(0.65-1.95)	0.663	2.21(1.16-4.18)	<b>0.015</b>
ARNTL	rs17452383	DOM	1.09(0.73-1.62)	0.682	0.77(0.41-1.43)	0.405	2.32(1.15-4.68)	<b>0.019</b>
DICER1	rs3742330	DOM	0.53(0.29-0.95)	<b>0.033</b>	0.77(0.36-1.63)	0.494	0.18(0.04-0.76)	<b>0.020</b>
DICER1	rs8006416	DOM	0.53(0.29-0.96)	<b>0.037</b>	0.80(0.38-1.70)	0.559	0.18(0.04-0.76)	<b>0.020</b>
MDM4	rs10900596	REC	1.70(0.99-2.91)	0.054	0.83(0.31-2.20)	0.706	2.49(1.14-5.44)	<b>0.022</b>
ICAM1	rs281437	DOM	1.56(1.09-2.23)	<b>0.015</b>	1.30(0.78-2.16)	0.306	2.20(1.11-4.36)	<b>0.023</b>
RPS6KB2	rs10274	REC	0.54(0.31-0.94)	<b>0.030</b>	0.74(0.37-1.48)	0.398	0.29(0.10-0.86)	<b>0.025</b>
RRM2B	rs5005121	DOM	1.10(0.62-1.94)	0.746	0.67(0.30-1.48)	0.320	2.79(1.13-6.88)	<b>0.026</b>
MTHFR	rs10779765	REC	1.20(0.70-2.03)	0.507	1.00(0.46-2.14)	0.992	2.62(1.05-6.53)	<b>0.038</b>
NDUFA6	rs7245	REC	1.19(0.79-1.80)	0.413	0.81(0.42-1.56)	0.525	2.12(1.04-4.35)	<b>0.040</b>
TNFRSF10D	rs7957	DOM	1.45(1.01-2.07)	<b>0.044</b>	1.59(0.96-2.63)	0.073	2.09(1.03-4.23)	<b>0.040</b>

\* Remain significant after multiple comparisons using FDR of 5%

\*\*Adjusted by age, gender, ethnicity, stage, pack year and treatment regimens.

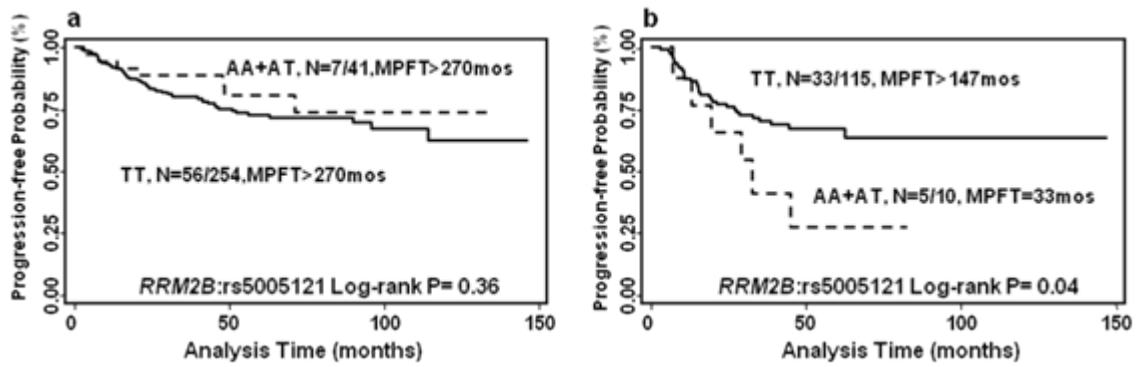




**Figure 8 Kaplan-Meier estimates on effect of DROSHA:rs6886834 on early stage progression among surgery-only patients**

MPFT: median progression-free time in months. N=A/B, A: number of patients with event, B: total number of patients in subgroup.

When comparing results of subgroup analysis, 28 SNPs (14 in binding sites and 14 in processing genes) and 16 SNPs (12 in binding sites and 4 in processing genes) were exclusively associated with altered risk for progression in surgery-only patients or surgery plus chemotherapy patients, respectively. Of these, 19 SNPs showed the same direction of the effects in both subgroups while 17 SNPs were found to have opposite effects in both subgroups (Table 7). For example, in patients received surgery plus chemotherapy, *RRM2B*:rs5005121 was associated with significantly increased risk for developing progressive disease and a shortened progression-free time, but in surgery-only patients, this SNP showed a protective effect against progression with an increased time to progression (Figure 6). We also observed a significant cumulative effect for each group of SNPs on modulating risk of progression - with increased number of UFGs, there is a gradual trend of increased risk for progression and corresponding shortened progression-free time in surgery-only and surgery plus chemotherapy subgroups ( $P$  for trend  $< 1 \times 10^{-4}$ , Figures 9c and 9d).

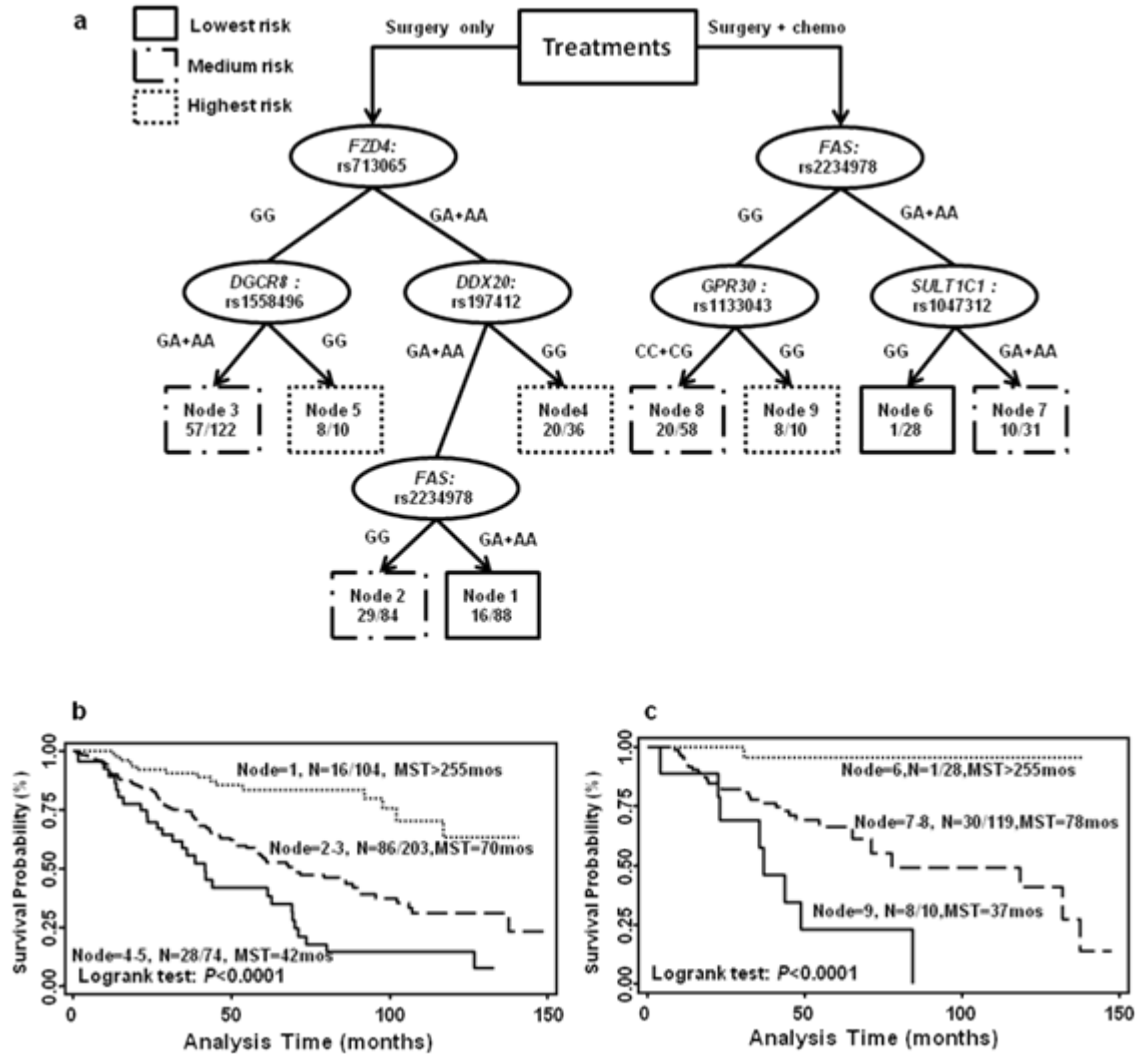


**Figure 9 Kaplan-Meier estimates for the effect of RRM2B:rs5005121 genotypes on NSCLC progression in two treatment subgroups based on the dominant model:**

(a) surgery-only patients; (b) surgery plus chemotherapy patients. MPFT: median progression-free time in months. N=A/B, A: number of patients with event, B: number of patients in subgroup.

### 3.1.1.3 Survival tree analysis of SNPs associated with NSCLC survival

Figure 10 shows the survival-tree structure identifying potential higher-order gene-gene interactions among miRNA-related genes in modulating overall survival. SNPs that displayed at least borderline significant association with survival after multiple comparison adjustment ( $q < 0.1$ ) were included in the analysis. The terminal nodes from the analyses were able to classify patients into three risk groups with significantly different survival probabilities. The MSTs based on these groupings varied from greater than 86 months for the low risk group to 41.7 months in the high risk group (log rank  $P < 0.0001$ ) in surgery-only patients, and from more than 118 months to 36.8 months for low and high risk groups respectively (log rank  $P < 0.0001$ ) in patients receiving surgery plus chemotherapy. Moreover, the initial split in the tree structure for each subgroup, *FZD4*:rs713065 and *FAS*:2234978, were also the two SNPs that remained significant after multiple comparison correction in the two treatment subgroup analyses. Because of the limited number of SNPs that were at least borderline significant after adjustment for multiple comparisons in the progression analysis, survival tree analysis was not performed for this endpoint.



**Figure 10 Potential gene-gene interactions among SNPs identified in the survival analysis in early stage NSCLC patients**

(a) Survival tree analysis showing higher-order gene-gene interactions; (b) Kaplan-Meier curves of survival time for surgery-only patients in three risk groups identified by the survival tree analysis; (c) Kaplan-Meier curves of survival time for surgery plus chemotherapy patients among the three risk groups. MST: median survival time in months. N=A/B, A: number of patients with event, B: total number of patients.

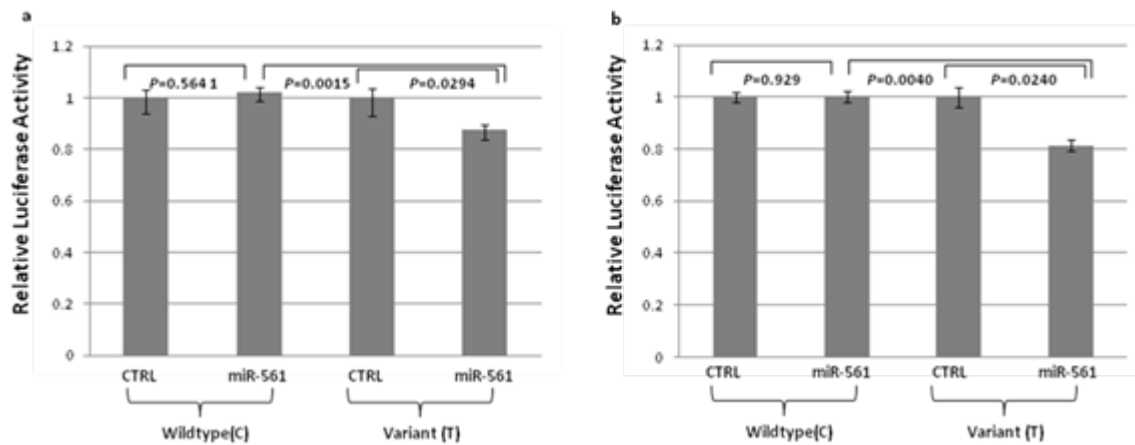
#### 3.1.1.4 Internal validation using bootstrap re-sampling method

In order to further validate our results and to exclude potentially false positive associations, a bootstrap re-sampling method was also used to examine the associations that remained significant after correcting for multiple comparisons. For each SNP, we used bootstrap re-sampling. All the SNPs that were significant after multiple comparisons at an FDR of 5% remained significant in the bootstrap analysis for at least 450 out of 500 re-samplings. Among them, two SNPs (*FZD4*:rs713065 in analysis of overall survival in surgery only subgroup; *SPI*:rs17695156 in progression analysis of all early stage patients) reached  $P < 0.05$  for each of the 500 iterations. *FAS*:rs2234978 reached  $P < 0.01$  for 500 iterations in analysis of overall survival in all early stage patients, this finding was consistent in the surgery plus chemotherapy subgroup as well with over 65% of the re-sampled datasets remaining significant at  $P < 0.01$ . Bootstrap re-sampling analysis was also performed for unfavorable genotype and survival tree analyses. The results were significant in all the subgroups analysis for entire 500 re-samplings at  $P < 0.05$ .

#### 3.1.1.5 The effect of selected miRNA binding site variants on miRNA-regulation

Since several SNPs located within predicted miRNA binding sites were significantly associated with clinical outcomes after correction for multiple comparisons, we then performed luciferase reporter assays to determine whether these predicted binding site variants truly result in altered miRNA regulation. *FAS*:rs2234978, which was consistently associated with a beneficial effect on prognosis, was predicted to create a new miRNA binding site in *FAS* for miR-561. Thus, we expected to observe a decrease in luciferase activity for variant allele-containing construct in the presence of miR-561. In both lung cancer cell lines, a significant decrease of luciferase signal was observed when miRNA-561

was transfected with variant allele-containing reporter (T) (NCI-H460: $P=0.029$ , Figure 8a; NCI-H2444: $P=0.025$ , Figure 11b), but not observed with wildtype allele (C) construct ( $P>0.5$ ). Comparison of the luciferase activities between co-transfections of miRNA-561 with the variant allele-containing and the wildtype-containing constructs also showed significant difference (NCI-H460: $P=0.002$ , Figure 11a; NCI-H2444: $P=0.004$ , Figure 11b). *SPI:rs17695156* was predicted to disrupt a conserved miRNA site; however, in our *in vitro* assays the miRNA-induced suppression of luciferase activity was observed in both variant and wildtype allele-containing constructs co-transfected with miRNA-545. There was no significant difference in reporter activities between the two alleles and the extent of signal decrease varied between cell lines (data not shown).



**Figure 11 Effect of the FAS variant allele on miR-561 targeting and luciferase reporter expression:**

(a) Relative luciferase reporter activity of the wildtype and variant *FAS* allele in the presence of control (Ctrl) or miR-561 in lung cancer cell line NCI-H460; (b) Relative luciferase reporter activity of the wildtype and variant *FAS* allele in the presence of control (Ctrl) or miR-561 in lung cancer cell line NCI-H2444.



### 3.1.2 miRNA-related genetic variations and survival in late stage NSCLC patients

#### 3.1.2.1 Patients characteristics

598 stage III and stage IV patients were identified in our study. Around half of all the patients had a adenocarcinoma histology, and majority patients are ever smokers (81%). Around 56% of all patients have been treated with any form of chemotherapy. Median survival time for late stage patients are 11.8 months. Between vital groups (dead vs. alive), at the time of this analysis, there are significant differences in the distribution of gender ( $P = 0.002$ ), clinical stage ( $P = 0.004$ ), and performance status ( $P = 0.001$ ), all of which have been included in the adjustment for multi-variant analysis. Histology, ethnicity, smoking status, and age were not significant different between patients who have died and still alive. (Table 8)

**Table 8: Host characteristics of late stage NSCLCs**

Variables	Dead, No (%)	Alive, No (%)	P-value
<b>Total Patients</b>	456	142	
<b>Age, mean(sd)</b>	59.6(10.0)	59.7(10.6)	0.884
<b>Gender</b>			
<i>Male</i>	262(57)	61(43)	
<i>Female</i>	194(43)	81(57)	<b>0.002</b>
<b>Ethnicity</b>			
<i>Caucasian</i>	358(79)	112(79)	
<i>African-American</i>	72(16)	23(16)	
<i>Others</i>	26(6)	7(5)	0.940
<b>Pack year, mean(sd)</b>	37(31)	37(30)	0.925
<b>Smoking status</b>			
<i>Never</i>	84(18)	28(20)	
<i>Former</i>	184(40)	59(42)	
<i>Current &amp; RQ</i>	188(41)	55(39)	0.860
<b>Histology</b>			
<i>Adenocarcinoma</i>	231(51)	73(51)	
<i>Squamous cell ca</i>	88(19)	37(26)	
<i>Unclassified/other</i>	137(29)	32(23)	0.246
<b>Clinical stage</b>			
<i>Stage IIIA</i>	57(13)	25(18)	
<i>Stage IIIB(dry)</i>	100(22)	42(30)	
<i>Stage IIB(wet)</i>	25(5)	14(10)	
<i>Stage IV</i>	274(60)	61(43)	<b>0.004</b>
<b>Performance status</b>			
<i>0</i>	96(21)	47(33)	
<i>1</i>	254(56)	80(56)	
<i>2-4</i>	66(14)	8(6)	
<i>missing</i>	40(9)	7(5)	<b>0.001</b>

### 3.1.1.6 Associations between individual SNPs and late stage patients survival

All 240 SNPs have been analyzed with survival in the 598 late stage NSCLC patients, 9 processing and 17 binding site SNPs were significantly associated with survival. The top SNP was rs15561 (HR=1.70, 95%CI=1.22-2.36), in predicted miRNA binding site of gene *NAT1*. The most significant processing gene SNP is rs7735863 (HR=1.38, 96%CI=1.1-1.74), an intronic SNP in *DROSHA*. However, none SNP remained significant after correcting multiple comparisons.

When comparing the SNPs remained significant after multiple comparison correction identified in early stage patients, rs713065 was significant associated with decreased risk for death in the late stage patients (HR=0.78, 95%CI=0.64-0.94). However, this association has not passed the multiple comparison criteria.

### 3.1.1.7 Associations between individual SNPs and survival in late stage patients treated with chemotherapy

Because the majority of patients have been treated with chemotherapy, we then performed a subgroup analysis focusing on patients treated with chemotherapy. A total of 24 SNP (20 binding site, 4 processing gene) were significantly associated with survival in this subgroup, among them five binding site SNPs remained significant after correcting for multiple comparisons (Table 9).

The most significant SNP was rs4796033 (HR=1.41, 96%CI=1.13-1.75), which is in *RAD51L3*. Patients had at least one variant of this SNP had around 8 month shortened MST (log-rank P=  $3.2 \times 10^{-4}$ , Figure 12a).

Interestingly, rs15561 (HR=1.98, 96%CI=1.32-2.94), which is the top SNPs in the overall late stage patients, remained significant after correcting for multiple comparisons. This SNP was also significantly associated with an altered survival time – patients carrying homogeneous variant genotype have six month shorter MST compared to those who have common allele (log-rank P=  $6.4 \times 10^{-3}$ , Figure 12b). Another *NATI* SNP rs4986993 remained significant after multiple comparisons, which is in highly linked with rs15561 ( $R^2=1.00$ ).

Rs10278782 (HR=1.58, 96%CI=1.21-1.06), in gene *CAV2*, was also significant after multiple comparison corrections. Also, compared to patients carrying at least one variant allele, those patients with two variants allele have around 4 months shortened survival time (Log-rank P= $6.6 \times 10^{-3}$ , Figure 12c). Rs17749202 (HR=2.03, 96%CI=1.29-3.17), binding sites SNP in *WNT11*, showed a significant risk effect after multiple comparison corrections, patients had a GG genotype had 5 month shortened median survival time compare to those

carrying at least one A allele, and the difference was significant (Log-rank  $P=0.016$ , Figure 12c)

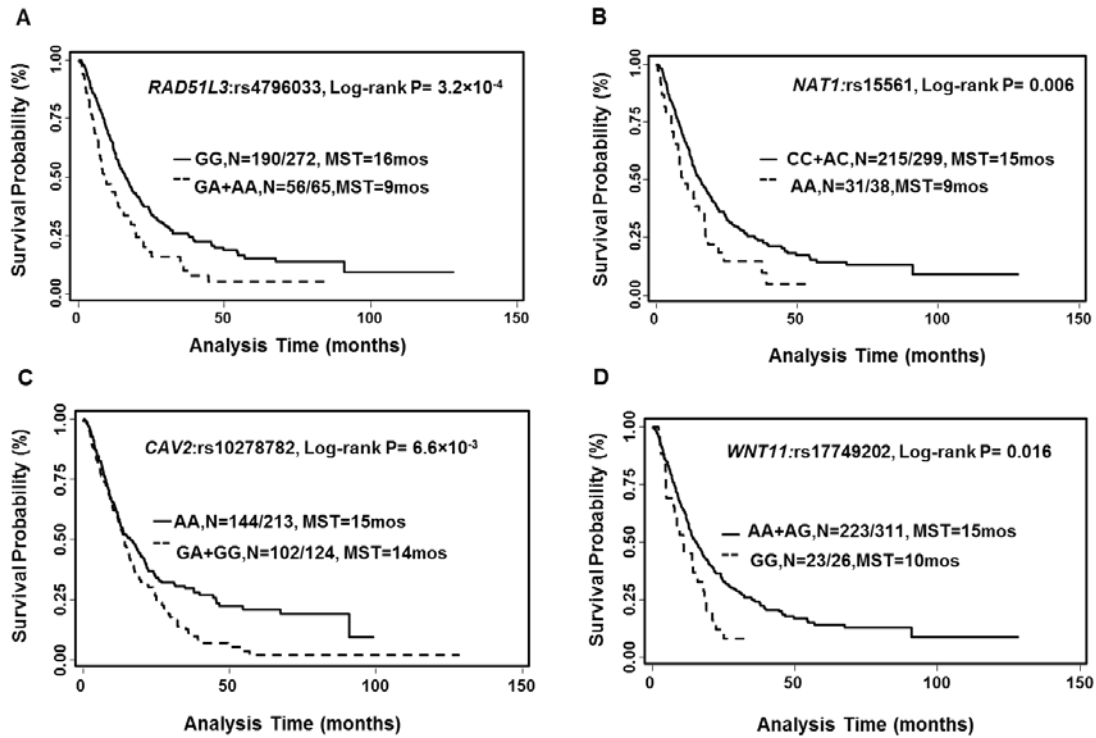
In the UFG analysis, significant cumulative effect was also observed for the four (removed rs4986993) SNPs, with increased in the number of UFG patients carries, a progressively increased risk was observed ( $P$  for trend  $=2.36 \times 10^{-7}$ ). Patients had at least two UFGs had nearly four-fold increased risk for death ( $P=1.6 \times 10^{-11}$ ) with a 9.2 month median survival time, compared to 21.6 month of patients without any UFG (Log-rank  $P=6.1 \times 10^{-8}$ ).

**Table 9: Selected SNPs with survival in late stage NSCLC patients**

Gene	SNP	Model	HR(95%CI)	P-value	Q-value
<i>RAD51L3</i>	rs4796033	DOM	1.93(1.42-2.63)	$2.86 \times 10^{-5}$ *	0.003
<i>NAT1</i>	rs15561	REC	1.98(1.32-2.94)	$8.36 \times 10^{-4}$ *	0.025
<i>NAT1</i>	rs4986993	REC	1.98(1.32-2.94)	$8.36 \times 10^{-4}$ *	0.025
<i>CAV2</i>	rs10278782	DOM	1.58(1.21-2.06)	$7.05 \times 10^{-4}$ *	0.025
<i>WNT11</i>	rs17749202	REC	2.03(1.29-3.17)	$2.01 \times 10^{-3}$ *	0.047

\* Remain significant after multiple comparisons using FDR of 5%

\*\* Adjusted by age, gender, clinical stage, and performance status.



**Figure 12 Kaplan-Meier estimates for the effect of selected SNPs on NSCLC survival in patients treated chemotherapy:**

(A) RAD51L3:rs4796033; (B) NAT1:rs15561; (C) CAV2:rs10278782; (D) WNT11:rs17749202. MST: median survival time in months. N=A/B, A: number of patients with event, B: number of patients in subgroup.

### 3.1.1.8 Internal validation using bootstrap re-sampling method

We then also adopted a bootstrap re-sampling method was to further examine the associations that remained significant after correcting for multiple comparison at an FDR of 5%. For these five each SNP and also UFG analysis, we did bootstrap re-sampling for 500 times with duplicates. All the SNPs that were significant after multiple comparisons remained significant in the bootstrap analysis for at least 465 out of 500 re-samplings at a P-value less than 0.05.

### 3.1.2 Discussion

In this study, we identified genetic variants in miRNA processing genes and binding sites for cancer-related genes that modulated overall survival and progression in early stage NSCLC patients. Because majority of late stage patients have metastatic disease at the time of diagnosis, thus analysis was only performed on survival for late stage patients. Panels of treatment subgroup-specific predictive markers were identified and the significance of top associations in our study was confirmed by controlling for false discoveries through multiple comparison corrections and internal validation. *FAS*: rs2234978 was identified as a potential prognostic factor in our results and functional data provides evidence that this SNP alters miRNA regulation of *FAS*. These results suggested that identified genetic variants in miRNA processing genes and miRNA binding sites may serve as potential prognostic markers for patients' clinical outcomes and predictive markers of response to treatment for future investigation and clinical applications.



Several studies have identified associations between polymorphisms in miRNA-binding sites and human disease, including cancer (149, 184, 185). In the current study, we identified miRNA-binding site SNPs that significantly modulated risk of either death or progression. Specifically, *FAS*:rs2234978 was observed to be significantly associated with decreased risk of death. We found a reduction in risk of dying in surgery only patients who carried the variant allele of this SNP. This protective effect was even stronger in patient treated with surgery plus chemotherapy. Moreover, its appearance at the top of the tree structure in the survival tree analysis also confirmed this locus as an important marker responsible for the largest proportion of the variation in predicting patient's overall survival. These consistent associations highlighted the potential importance of this SNP in modulating NSCLC risk of dying and its potential prognostic role. *FAS* (member 6 of TNF receptor superfamily) is a cell-surface receptor of the tumor necrosis family which plays an important role in the regulation of apoptosis signaling. Interestingly, rs2234978 is a synonymous SNP located in the seventh exon of *FAS* on chromosome 10q23. Typically, miRNA binding sites are located in 3'UTRs, which for *FAS* is located in exon 9 of the full length transcript. However, alternative splicing of *FAS* results in several transcribed isoforms that are involved in nonsense-mediated mRNA decay (NMD), including the transcript where exon 7 serves as the 3'UTR. NMD plays important roles in limiting the synthesis of truncated or mutant proteins which can negatively regulate the apoptosis mediated by the full length protein as well as global gene expression (NCBI dbSNP database). The nucleotide change from this polymorphism is predicted to create a new miRNA binding site for miR-561, resulting in decreased expression of the *FAS* alternative transcripts. We validated this function *in vitro* by luciferase assay in two lung cancer cell lines. Since the NMD transcripts may negatively

regulate normal *FAS* expression, this would ultimately result in increased level of *FAS* in tissues that express the targeting miRNA. It has also been reported that cisplatin treatment can increase *FAS*-mediated apoptosis (186). It is possible that in patients who carry the variant allele, higher expression of *FAS* in the presence of cisplatin treatment could increase tumor cell death resulting in better overall survival independent of treatment regimen, and this locus might even be synergistic with chemotherapy, thus conferred an more extreme protective effect. However, further studies will be needed to confirm whether this SNP has any influence on *FAS* protein level *in vivo* and whether it affects apoptotic activity in normal and tumor cells during treatment. *FZD4*:rs713065 is the only SNP associated with significant decreased risk of dying after adjustment for multiple comparisons in surgery-only patients. This SNP was also the top split in the survival tree analysis of survival for this subgroup, which suggests its importance in predicting survival for NSCLC patients. *FZD4* (frizzled homolog 4) is a member of the frizzled gene family of trans-membrane receptors, which help to transduce WNT signals and activate downstream WNT-pathway components. WNT is a major pathway involved in normal and cancer stem cell development (187). This *FZD4* binding site SNP may down-regulate *FZD4* expression by creating a miRNA binding site, thereby inhibiting transduction of the WNT signal. This effect could lead to enhanced survival in these patients due to decreased WNT signaling. Moreover, this SNP showed an opposite effect, although not significant, on survival in subgroup treated with surgery plus chemotherapy, thus indicating the interaction of these variant with chemotherapy on regulation of Wnt signaling. However, *in vitro* assessment of the effect of this variant on miRNA binding was not possible due to difficult sequencing characteristics of this region. Other appropriate functional assays will be needed to explore the function of this SNP.

We further evaluated in a late stage population of all the associations remained significant in early stage patients. However, none of them were significant at  $FDR < 0.05$  level, indicating a stronger effect of these identified SNPs in early stage NSCLCs. One *NAT1* binding site SNP, showed significant effect on survival in all late stage patients and remained significant after multiple comparison corrections in chemotherapy subgroup. NAT1 encoded an enzyme catalyzing an acetyl group from acetyl-CoA transfer. This enzyme assists drug metabolism as well as other xenobiotic and folate catabolism (188, 189). Thus, it is possible that the real causal SNP would influence chemotherapy compound metabolism through modulating NAT1 function, and influencing chemotherapy response and eventually have an impact on patients' survival, especially in late stage patients, who are standardly treated with chemotherapy.

In the progression analysis, since majority of late stage patients had metastatic disease at the time of diagnosis, analysis was only performed in early stage patients. *SP1*:rs17695156 showed significant association after multiple comparison corrections with increased risk for disease progression in the overall study population and surgery plus chemotherapy patients. SP1 is a transcription factor, which can regulate the expression of many genes, thus having a general, regulatory role within the cell. This SNP is predicted to disrupt a conserved miRNA site; however, in our *in vitro* experiments, we did not observe any significant difference between the two alleles in miRNA-induced repression of reporter activities. Nevertheless, it is possible that the SNP might show differential effect of miRNA targeting *in vivo*, where the expression of various cellular components are at physiological levels. Alternatively, this

3'UTR SNP might affect *SP1* expression independent of its putative role as a miRNA target site (e.g. affecting RNA stability or post-transcriptional regulation). Any alterations in *SP1*, because of its key regulatory role, would potentially have an effect on mechanisms of disease progression. A variant in *DROSHA*, a key biosynthesis pathway component, was associated with significantly increased risk for progression. This result suggests that this SNP, or variant tagged by it, may alter normal function of *DROSHA*, thus influencing overall miRNA processing and affecting different downstream biological processes.

The different patterns of associations with clinical outcomes among the two treatment subgroups observed in our study suggest that treatment context is important and the interactions between SNPs and response to treatment may play a major role. In this study, we identified panels of SNPs exclusively associated with clinical outcomes in either of the two subgroups with relatively homogeneous treatment regimens. These panels consist of markers from major pathways related to cancer and provide potential treatment-specific predictive markers for future investigation. Furthermore, we identified SNPs with different trends of associations for patients treated with different treatment regimens. This indicates that miRNA-related regulation of these genes function differently in the context of the cellular response to chemotherapeutic agents. This suggests that a subgroup of patients may benefit from chemotherapy, while other patients would not receive this same benefit based on their genetic background, which may take into consideration in the selection of treatment regimens.

Cancer is a multi-factorial disease, meaning that a single gene may not have great influence on disease risk or clinical outcomes. For the miRNA-related pathway, each miRNA can regulate a group of genes and miRNA processing genes work together to produce mature miRNA. Moreover, if these binding site SNPs are present in several important genes within a cancer-related pathway, we would predict that the cumulative effect would be much greater than the effect of an individual SNP in isolation. Therefore, it is reasonable to analyze the potential cumulative effect of these SNPs related to miRNA function on clinical outcomes. In our study, we observed a strong polygenetic effect and multiple potential gene-gene interactions for these miRNA-related SNPs. Similar analyses have been shown in several previous studies to have sufficient power in analyzing patients for cancer risk and clinical outcomes (190, 191). Interestingly, the most significant SNPs in the main effect analysis were also those predicted to be involved in potential gene-gene analyses, demonstrating that these top miRNA-related SNPs identified in our study are not only important genetic factors individually but they also function in a network to influence patient's overall prognosis.

Overall, the current study provides evidence that genetic variants in the miRNA processing pathway and miRNA binding sites influence clinical outcomes for early stage NSCLC patients. Specifically, we identified the potential prognostic role of FAS in predicting overall survival in these patients and supported this observation with *in vitro* functional genomic analyses. These findings can help to identify patients who will receive the most benefit from specific curative therapy regimens, thereby aiming to maximize treatment efficacy. These results provide a basis for future personalized medicine whereby

those early stage NSCLC patients with high probability for favorable outcomes can be identified and appropriately treated.

## 3.2 Genetic variations in inflammation related genes and survival in late stage NSCLC patients

### 3.2.1 Patient characteristics

A total of 502 patients were included in discovery phase from the MD Anderson population, all of which were non-Hispanic Caucasians, were included in the discovery phase (Table 10). At the time of analysis, 68% of the patients had died. The median follow-up time for the discovery-phase patients was 30.5 months, with a median survival time (MST) of 16.5 months. The differences in mean age at diagnosis and sex distribution between patients who had died and those who were alive at the time of analysis were not significant (Table 10). Half of the patients in the discovery population received radiation therapy. At the time of analysis, more patients who were current or former smokers than patients who had never smoked were alive, and more patients with stage III disease than patients with stage IV disease were alive ( $P < 0.05$ ; Table 10). The internal validation population included 335 patients, with a MFT of 89.6 months and a MST of 16.8 months (Table 10). Of these patients, 56% had received radiation therapy. At the time of analysis, more female patients than male patients were alive, and more patients with stage III disease than patients with stage IV disease were alive ( $P < 0.05$ ).

In the Harvard population (the external validation set), 371 patients were included, with a MFT of 60 months and a MST of 12.2 months. The patients who were alive at the time of analysis were slightly younger at diagnosis than were those who had died (Table 10). One hundred fifty-three of the 371 patients (41%) had received radiation therapy. The

distributions of sex, smoking status, and clinical stage were similar between patients who had died and those who were alive at the time of analysis.

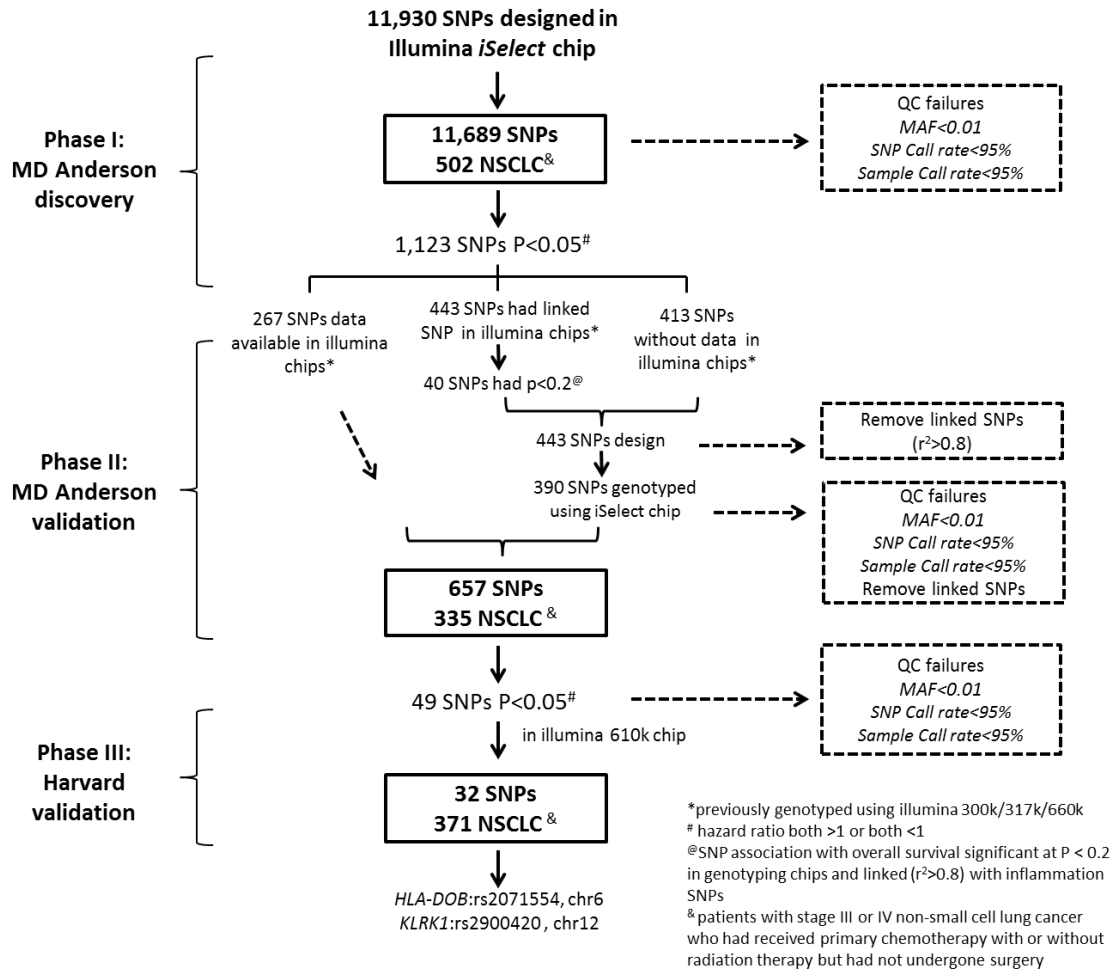


**Table 10: Characteristics of the study populations at the time of analysis**

Variables	MD Anderson Discovery			MD Anderson Validation			Harvard Validation		
	Dead(%)	Alive(%)	P	Dead(%)	Alive(%)	P	Dead(%)	Alive(%)	P
<b>MST (months)</b>	16.5			16.8			12.2		
<b>MFT (months)</b>	30.5			89.6			60.0		
<b>Age, mean(sd)</b>	60.7(11.2)	62.4(10.5)	0.099	59.3(10.4)	57.5(9.0)	0.374	63.58(10.55)	60.45(10.76)	0.053
<b>Sex</b>									
Male	166(49)	80(49)		196(64)	12(41)		171(54)	22(42)	
Female	174(51)	82(51)	0.907	110(36)	17(59)	<b>0.016</b>	147(46)	31(58)	0.098
<b>Smoking status</b>									
Never	129(38)	41(25)		4(1)	0(0)		25(8)	8(15)	
Former	117(34)	74(46)		145(47)	13(45)		154(48)	23(43)	
Current & RQ	94(28)	47(29)	<b>0.012</b>	157(51)	16(55)	0.782	139(44)	22(42)	0.227
<b>Clinical stage</b>									
Stage III	99(29)	72(44)		142(46)	20(69)		118(37)	22(42)	
Stage IV	241(71)	90(56)	<b>0.001</b>	164(54)	9(31)	<b>0.020</b>	200(63)	31(58)	0.540
<b>Radiotherapy</b>									
Yes	160(47)	92(57)		165(54)	22(76)		128(72)	25(47)	
No	180(53)	70(43)	<b>0.041</b>	141(46)	7(24)	<b>0.023</b>	90(28)	28(53)	0.344
<b>Total</b>	340	162		306	29		318	53	

### 3.2.2 Effects of inflammation-related SNPs on overall survival

A detailed workflow of our genotyping procedures is presented in Figure 13. A total of 11,930 SNPs from 904 genes were genotyped (see Figure 1), of which 11,689 passed quality control measures and were included in the MD Anderson discovery phase analysis. We observed that 1,123 SNPs had significant associations with overall survival in this group ( $P < 0.05$ ). Among these, 267 SNPs were found in previously published GWAS chips, for which data were ready for analysis. Of the remainder, 443 SNPs were associated with linked SNPs ( $r^2 > 0.8$ ) in previously published GWAS chips; we genotyped the 40 for which the association was significant ( $P < 0.2$ ). We also genotyped 413 SNPs that were not found in, or were not associated with linked SNPs in previously published GWAS chips. After genotyping (or using existing genotype data) the 657 of these SNPs that passed quality control measures in the internal validation population, we confirmed the association with overall survival for 49 SNPs (HRs both  $>1$  or  $<1$ ,  $P < 0.05$ ). We then performed a fast-track external validation of 32 of the 49 internally validated SNPs (those that had existing data available in previously published GWAS chips in the Harvard population. Seventeen SNPs were found to have consistent effects on overall survival in all 3 populations (Table 11).



**Figure 13 Study design and workflow.**

SNP indicates single nucleotide polymorphism; QC, quality control; MAF, minor allele frequency; GWAS, genome-wide association study

**Table 11: Inflammation-related single nucleotide polymorphisms (SNPs) that were found to affect overall survival in patients with late-stage non-small cell lung cancer**

SNP	Gene	Model	MD Anderson				Harvard		Combined		Phet
			Discovery		Validation		HR (95% CI)*	P	HR (95% CI)**	P	
			HR (95% CI)*	P	HR (95% CI)*	P					
rs2071554	HLA-DOB	DOM	1.46 (1.02-2.09)	<b>0.040</b>	1.51 (1.02-2.25)	<b>0.041</b>	1.52 (1.01-2.29)	<b>0.045</b>	1.49 (1.19-1.87)	<b>4.32×10<sup>-4</sup></b>	0.99
rs2900420	KLRK1	DOM	0.76 (0.60-0.96)	<b>0.021</b>	0.77 (0.61-0.99)	<b>0.038</b>	0.80 (0.63-1.02)	<b>0.069</b>	0.78 (0.68-0.89)	<b>3.51×10<sup>-4</sup></b>	0.94
rs12141256	FAF1	DOM	0.75 (0.57-0.97)	<b>0.031</b>	0.71 (0.52-0.97)	<b>0.033</b>	0.87 (0.66-1.13)	0.295	0.78 (0.66-0.91)	<b>2.27×10<sup>-3</sup></b>	0.60
rs1986649	FOXO1A	DOM	0.76 (0.60-0.96)	<b>0.020</b>	0.75 (0.59-0.95)	<b>0.018</b>	0.88 (0.69-1.13)	0.322	0.79 (0.69-0.91)	<b>9.43×10<sup>-4</sup></b>	0.58
rs7972757	KLRK1	DOM	0.73 (0.55-0.98)	<b>0.035</b>	0.67 (0.49-0.92)	<b>0.012</b>	0.87 (0.66-1.15)	0.331	0.76 (0.64-0.90)	<b>1.42×10<sup>-3</sup></b>	0.45
rs17446614	FOXO1A	DOM	0.72 (0.56-0.93)	<b>0.011</b>	0.69 (0.53-0.90)	<b>0.006</b>	0.89 (0.68-1.16)	0.386	0.76 (0.65-0.88)	<b>3.34×10<sup>-4</sup></b>	0.36
rs216136	CSF1R	ADD	1.21 (1.03-1.42)	<b>0.023</b>	1.17 (1.00-1.37)	<b>0.046</b>	1.07 (0.91-1.25)	0.410	1.15 (1.05-1.25)	<b>3.46×10<sup>-3</sup></b>	0.53
rs2189521	IL21R	REC	1.41 (1.03-1.94)	<b>0.032</b>	1.43 (1.08-1.89)	<b>0.014</b>	1.13 (0.85-1.50)	0.415	1.31 (1.10-1.55)	<b>1.85×10<sup>-3</sup></b>	0.44
rs1509	CAPN10	ADD	0.83 (0.69-0.99)	<b>0.038</b>	0.83 (0.68-1.00)	<b>0.048</b>	0.93 (0.78-1.11)	0.433	0.86 (0.78-0.96)	<b>5.53×10<sup>-3</sup></b>	0.57
rs10964912	IFNA14	REC	1.49 (1.01-2.19)	<b>0.044</b>	2.00 (1.26-3.17)	<b>0.003</b>	1.16 (0.78-1.72)	0.462	1.47 (1.16-1.86)	<b>1.38×10<sup>-3</sup></b>	0.21
rs971768	IL17RA	DOM	1.47 (1.09-1.98)	<b>0.012</b>	1.46 (1.00-2.12)	<b>0.047</b>	1.16 (0.78-1.74)	0.465	1.38 (1.13-1.69)	<b>1.71×10<sup>-3</sup></b>	0.63
rs10000856	IRF2	ADD	1.26 (1.07-1.50)	<b>0.007</b>	1.22 (1.03-1.44)	<b>0.020</b>	1.06 (0.90-1.25)	0.506	1.18 (1.07-1.29)	<b>1.07×10<sup>-3</sup></b>	0.29
rs2133092	TLN2	DOM	1.30 (1.04-1.63)	<b>0.023</b>	1.30 (1.03-1.64)	<b>0.027</b>	1.08 (0.84-1.38)	0.543	1.23 (1.07-1.41)	<b>2.88×10<sup>-3</sup></b>	0.47
rs11903566	PRKCE	DOM	1.60 (1.15-2.24)	<b>0.006</b>	1.45 (1.04-2.03)	<b>0.029</b>	1.11 (0.74-1.67)	0.625	1.41 (1.15-1.73)	<b>1.05×10<sup>-3</sup></b>	0.38
rs908742	PRKCZ	DOM	1.28 (1.03-1.60)	<b>0.024</b>	1.33 (1.06-1.67)	<b>0.015</b>	1.03 (0.82-1.29)	0.794	1.21 (1.06-1.37)	<b>4.44×10<sup>-3</sup></b>	0.23
rs3749166	CAPN10	REC	1.41 (1.04-1.92)	<b>0.029</b>	1.42 (1.02-1.99)	<b>0.038</b>	1.00 (0.71-1.41)	0.992	1.27 (1.06-1.54)	<b>0.012</b>	0.25

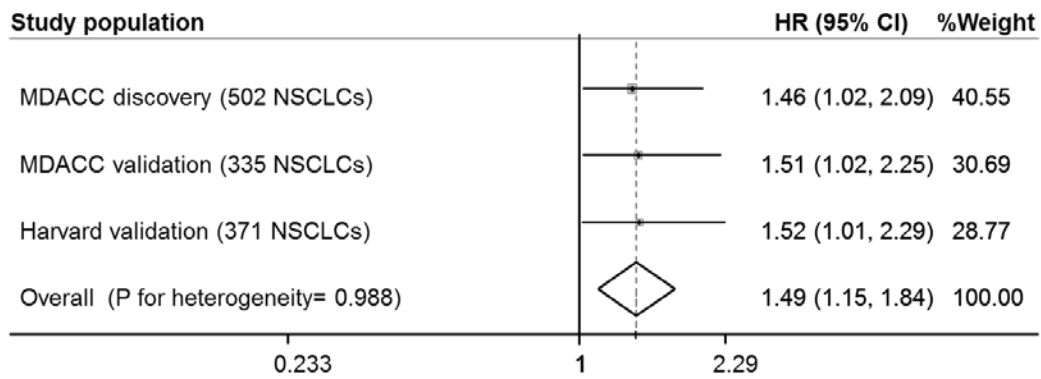
\* Adjusted for age, sex, smoking status, clinical stage, and treatment regimen.

\*\* Combined (meta-analysis) is based on the fixed-effects model.

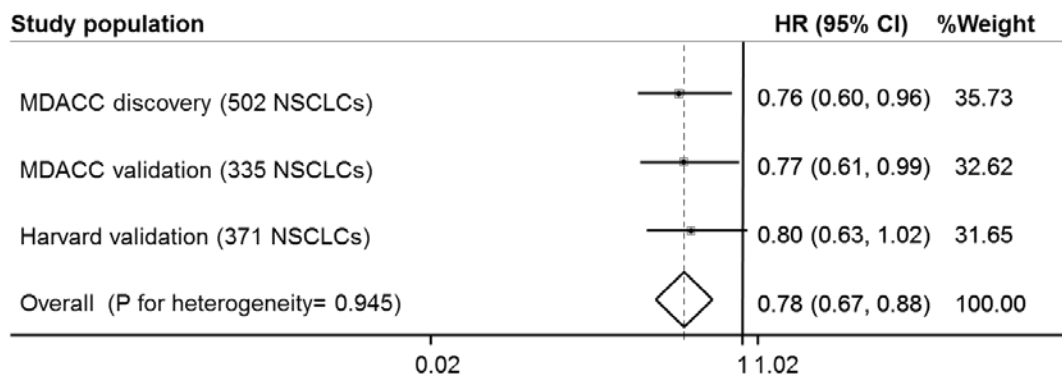
Abbreviations: Chr indicates chromosome; HR, hazard ratio; CI, confidence interval; P-het, P for heterogeneity test; DOM, dominant model; REC, recessive model; and ADD, additive model. Boldface indicates P < 0.1.

Rs2071554, a missense variation in the first exon of *HLA-DOB* (major histocompatibility complex class II, DO beta), was associated with poorer survival in all 3 populations, with similar HRs (Figure 14a). In the discovery population (HR = 1.46, 95% CI = 1.02- 2.09, P = 0.040), patients carrying at least 1 variant allele (AG or AA) had a significantly survival disparity of six months from 17 months to 11 months, compared with those who were homozygous for the common allele (GG), whose median overall survival time was 17 months (P for log rank test = 0.009; Figure 15a).

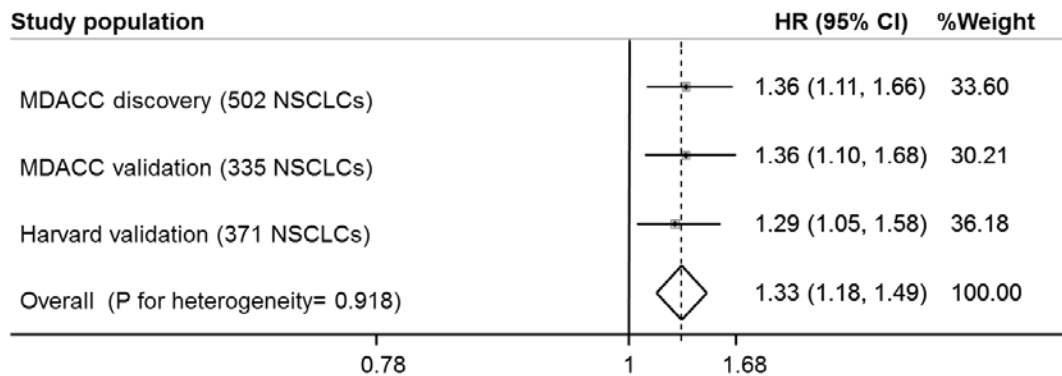
**A. HLA-DOB:rs2071554**



**B. KLRK1:rs2900420**

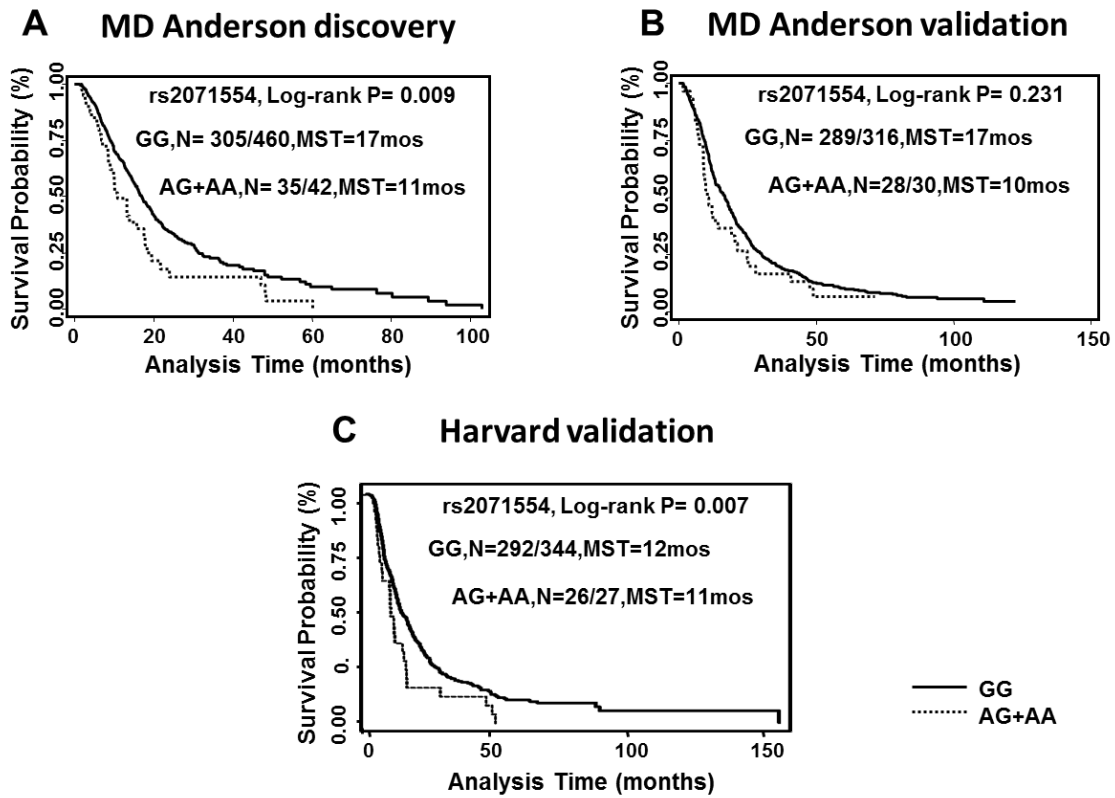


**C. UFG analysis**



**Figure 14 Forest plot for meta-analysis of the association of single nucleotide polymorphisms**

(A) *HLA-DOB:rs2071554* and (B) *KLRK1:rs2900420*, as well as (C) unfavorable genotypes (UFGs), with overall survival in discovery and internal validation populations



**Figure 15 Kaplan-Meier estimates of HLA-DOB:rs2071554 genotypes and risk of death in late-stage patients treated with chemotherapy**

(A) MD Anderson discovery; (B) MD Anderson validation; (C) Harvard validation. N=A/B, A: number of patients dead, B: number of all patients. MST: median survival time

In the internal validation population, rs2071554 was also associated with shortened overall survival (HR = 1.51, 95% CI = 1.02- 2.25, P = 0.041), and a non-significant, but appreciable seven month shortened survival time (Figure 15b). A similar effect was observed in the external validation population. The variant allele was associated with shortened overall survival (HR = 1.52, 95% CI = 1.01- 2.29, P = 0.045); patients carrying at least 1 copy of the variant allele had a shorter median overall survival time than patients who were homozygous for the common allele (P for log-rank test = 0.007; Figure 15c).

Meta-analysis of the association of rs2071554 with overall survival under the fixed effects model showed a P value of  $4.3 \times 10^{-4}$  (HR = 1.49, 95% CI = 1.19-1.87, P for heterogeneity = 0.988 Figure 14a). Rs2071554 is a missense variation that results in an arginine to glutamine substitution in the first exon of *HLA-DOB* a gene involved in phagocytosis and antigen presentation. To determine the potential consequences of this variant, we used Polyphen2 and SIFT to *in silico* evaluate the influence of rs2071554 on protein structure and function. In Polyphen2 analysis of this missense SNP, the amino acid change had a Polyphen2 score of 0.923 (sensitivity: 0.80; specificity: 0.94), suggesting that it may damage protein function; SIFT confirmed this SNP to be deleterious (SIFT score = 0.02).

*KLRK1*:rs2900420, which is located in the 3' flanking region of the *KLRK1* (killer cell lectin-like receptor subfamily K, member 1) gene, a component of the natural killer cell signaling pathway, was associated with prolonged overall survival in the discovery population (HR = 0.76, 95% CI = 0.60-0.96, P = 0.021) and in the internal validation population (HR = 0.77, 95% CI = 0.61-0.99, P = 0.038; Figure 14b). Significant overall survival time advantages were observed for patients who carried at least 1 variant allele



compared with patients who were homozygous for the common allele (discovery phase: GG, 15 months; AG and AA, 20 months; P for log-rank test = 0.011; internal validation phase: GG, 15 months; AG and AA, 18 months; P for log-rank test = 0.087). In the Harvard population, the association of rs2900420 with overall survival reached borderline significance (HR = 0.80, 95% CI = 0.63-1.02, P = 0.069), and in the meta-analysis, the effect was highly significant at  $3.5 \times 10^{-4}$  (HR = 0.78, 95% CI = 0.68-0.89, P for heterogeneity = 0.945).

Because most of the patients had died at the time of analysis, one year and three year survival were evaluated for these two validated SNPs in the MD Anderson population; similar results were found for the 2 SNPs at both durations (data not shown).

### 3.2.3 Stratified analyses

We next performed stratified analyses for rs2071554 and rs2900420 by smoking status. Similar effects on overall survival were observed in ever-smokers compared with the overall population group for each phase for both rs2071554 (discovery: HR = 1.68, 95% CI = 1.05-2.71, P = 0.092; internal validation: HR = 1.52, 95% CI = 1.02-2.26, P = 0.040; external validation: HR = 1.47, 95% CI = 0.96-2.25, P = 0.074) and rs2900420 (discovery: HR = 0.68, 95% CI = 0.50-0.93, P = 0.014; internal validation: HR = 0.79, 95% CI = 0.61-1.00, P = 0.52; external validation: HR = 0.81, 95% CI = 0.63-1.03, P = 0.086). Because of the limited number of never-smokers, stratified analysis was not performed for this group. When populations were stratified by stage at diagnosis, the two SNPs showed the same

effects on overall survival in stage III and stage IV patients as those observed in the overall population for each population (data not shown).

Because the majority of the patients received platinum-based chemotherapy regimens, we further did a subgroup analysis of the two SNPs in patients treated with platinum-based chemotherapy, and yield similar effect as in overall populations (data not shown).

#### 3.2.4 Cumulative effects of the top two SNPs

In the cumulative effects analysis, UFGs were defined as GA or AA for rs2071554 and GG for rs2900420. Using patients without any UFGs as a reference group within each population, we observed a significant “gene-dosage” effect of these SNPs on overall survival: the more UFGs a patient carried, the greater the deleterious effects on overall survival (discovery: HR = 1.36, 95% CI = 1.11-1.66, P-trend= 0.003; internal validation: HR = 1.36, 95% CI = 1.10-1.68, P-trend = 0.005; external validation: HR = 1.29, 95% CI = 1.05-1.58, P-trend = 0.015; Figure 14c).

### 3.2.5 Discussion

NSCLC patients with advanced stage disease are treated with primary chemotherapy as standard of care (192). Evidence has demonstrated that inflammation plays a role not only in lung cancer development, but also clinical outcomes such as response to chemotherapy (2, 193, 194). Thus, it follows that change in patients' inflammatory responses due to germline genetic polymorphisms might lead to variations in prognosis. In this analysis, we systematically evaluated the effect of SNPs from major inflammation-related genes on overall survival of advanced NSCLC patients who received primary chemotherapy without resection of their tumor. In our 3-phase pathway-based association study, we found 2 potential prognostic biomarkers for late-stage NSCLC: a *HLA-DOB* SNP was associated with poor survival in all 3 populations, and a *KLRK1* SNP was associated with prolonged overall survival in the MD Anderson populations (the association reached borderline significance in the Harvard population).

*HLA-DOB* is the beta subunit of the *HLA-DO* (*DO*) class II paralogs. It functions as negative regulator of major histocompatibility complex class II molecules by inhibiting *HLA-DM* (*DM*) molecules in a pH-dependent manner. The *DO:DM* ratio dictates major histocompatibility complex class II restricted-antigen presentation efficiency. Evidence has shown that dysregulation of the antigen presentation pathway related to the inflammatory response is involved in cancer development (195). Moreover, major histocompatibility complex class II molecules are key immune response molecules, which have been reported to have a positive relationship with prognosis in various cancers (196, 197). In our study, we found that the missense SNP *HLA-DOB:rs2071554* may damage protein structure and function, and we found that it had a robust adverse effect on survival across all 3 populations.

Hazard ratios indicated that patients with at least 1 variant allele of this SNP had nearly a 50% increase in risk of death compared with patients carrying no copies of the allele, and Kaplan-Meier survival analysis showed correspondingly decreased median overall survival times for carriers of the SNP. Our results suggest a potential prognostic role of this gene in lung cancer patients, making it worthy of future deep sequencing and functional analysis in vitro.

*KLRK1* (member 1 of the killer cell lectin-like receptor subfamily K) encodes for a transmembrane protein that interacts with various ligands to activate natural killer and T cells, leading to lysis of tumor cells. This gene has been shown to be involved in chemoresistance (198). Studies have reported that lung adenocarcinoma cells were able to escape from the innate immune response of natural killer cells by expressing heterogeneous ligands for KLRK1 (199). This gene has been identified as a promising target for immunotherapy for cancer (200, 201). *KLRK1*:rs2900420 is located 3 kilobases 3' to the *KLRK1* gene. In our study, it was associated with prolonged overall survival in the MD Anderson populations and its association with prolonged overall survival was nearly significant in the Harvard population. It is very likely that with a larger sample size the results would have reached statistical significance in the external validation population. Further exploration of the potential underlying biological mechanism(s) of this association would increase our understanding of this relationship.

This is the largest study to date to investigate the effects of inflammation-related genetic variations on clinical outcome. The major strength of this study was the 3-phase screening and validation approach using 2 independent patient populations, which were drawn from the largest lung cancer clinical outcome studies in the United States. Because the study populations were both well defined, with extensive clinical data collection, we were able to

identify a large sample of patients with relatively homogeneous treatment regimens to identify the most favorable replication population. This is of key importance when identifying biomarkers predictive of clinical outcome in pathway-based association studies.

Furthermore, instead of limiting our study to top SNPs, an approach usually adopted in pathway-based association studies to reduce cost and labor, we extensively genotyped almost all significant SNPs during our internal validation. This strict validation approach substantially improved the power of our study to detect candidate loci for subsequent analysis. In addition, we developed a comprehensive panel of inflammation-related genetic variations, which covered major cellular processes involved in inflammation responses and regulations. With this extensive coverage, our results provided a broad overview of the role of genetic variation in these essential genes within the overall inflammation network in modulating patients' clinical outcomes.

In conclusion, we identified and validated 2 potential genetic markers within the inflammation pathway that may affect clinical outcome in patients with late-stage NSCLC treated with chemotherapy. Given the important role of inflammation throughout the cancer continuum, these genetic markers may be good potential prognostic markers to help in tailoring treatment regimens in the clinic.

### **3.3 Genetic variations in inflammation pathway and survival in NSCLC patients in never smokers**

#### **3.3.1 Patient characteristics**

In the MD Anderson study, we identified 411 never-smokers with NSCLC (Table 12). Sixty-seven percent of them were women, and adenocarcinoma was the most common histology (77%). The mean age at diagnosis was 61.5 years. The median survival time (MST) was 23.2 months, and the median follow-up time (MFT) was 54.2 months. Most of the cases (77%) were diagnosed at a late stage (stage III/ IV). Fifty-three percent of the patients received chemotherapy only, 33% underwent surgery, and 24% received radiation-therapy. At the time of the current study, 276 (67%) of the patients had died. In the Mayo Clinic study, 311 never-smokers with NSCLC were identified and included as the validation population (Table 12). The mean age was 61.7 years, with the majority being female (73%). Sixty-one percent of the patients had late-stage disease at diagnosis. Fifty-nine percent of the patients received chemotherapy, 53% underwent surgery, and 25% received radiation-therapy. At the time of this study, 59% of the patients had died. Because of the greater proportions of patients with early-stage disease and who had undergone surgery in the Mayo Clinic population than in the MD Anderson population, the MST (44.6 months) and MFT (73.6 months) were longer in the former population than in the latter.

**Table 12: Characteristics of the never-smokers with lung cancer**

Characteristic	MD Anderson	Mayo Clinic
	No. of patients (%)	No. of patients (%)
MST, months	23.2	44.6
MFT, months	54.2	73.6
Mean age, years (SD)	61.5(13.0)	61.7(13.1)
Sex		
Male	135(33)	84(27)
Female	276(67)	227(73)
Stage		
I	93(23)	105(34)
II	15(4)	15(5)
III	91(22)	90(29)
IV	212(52)	101(32)
Histology		
Adenocarcinoma	316(77)	213(68)
Squamous cell carcinoma	28(7)	14(5)
Non-small cell carcinoma	41(10)	18(6)
Bronchoalveolar carcinoma	20(5)	11(4)
Other	6(1)	55(18)
Treatment		
Surgery	135(33)	165(53)
Radiation therapy	100(24)	77(25)
Chemotherapy	218(53)	182(59)
Concurrent chemoradiation	38(9)	36(12)
Vital status		
Dead	276(67)	182(59)
Alive	135(33)	129(41)
Total	411	311

SD=standard deviation.

### 3.3.2 Main effect of individual SNP on survival in the discovery, replication, and combined analysis

In the discovery phase, after carrying out quality control measures, 11,689 SNPs were included in our analysis. Of these SNPs, 1,538 were significantly associated with overall survival ( $p < 0.05$ ), with 14 of these variants being significant at the  $p < 10^{-4}$  level.

We selected 37 top SNPs for validation in the Mayo Clinic population. Eighteen SNPs had a consistent direction of the effect (HR same direction) in both populations (table 13). Of these 18, three SNPs ((interleukin 17 receptor A [*IL17RA*]:rs879576, bone morphogenetic protein 8A [*BMP8A*]:rs698141, and spleen tyrosine kinase [*SYK*]:rs290229) in the Mayo population were significant ( $p < 0.05$ ) with an additional two (*CD74*:rs1056400 and *CD38*:rs10805347) reaching borderline significance ( $p < 0.1$ )

The most significant SNP was rs879576, a synonymous variant in the last exon of the proinflammatory cytokine *IL17RA*. Rs879576 was associated with a significantly decreased risk of death in the discovery phase (hazard ratio [HR], 0.57; 95% confidence interval [CI], 0.41-0.78;  $p = 5.49 \times 10^{-4}$ ), validation phase (HR, 0.65; 95% CI, 0.44-0.94;  $p = 0.023$ ) and combined population (HR, 0.60; 95% CI, 0.47-0.77;  $p = 4.13 \times 10^{-5}$ ) (Table 13). This decreased risk of dying resulted in enhanced survival duration. Compared to patients with variant genotypes, a prolonged MST was observed in patients with the common homozygous genotype in both discovery (31 vs. 20 months,  $p = 0.066$ , log-rank test) and validation (46 vs. 34 months,  $p = 0.069$ , log-rank test) phases. (Figures 16a and 16b)

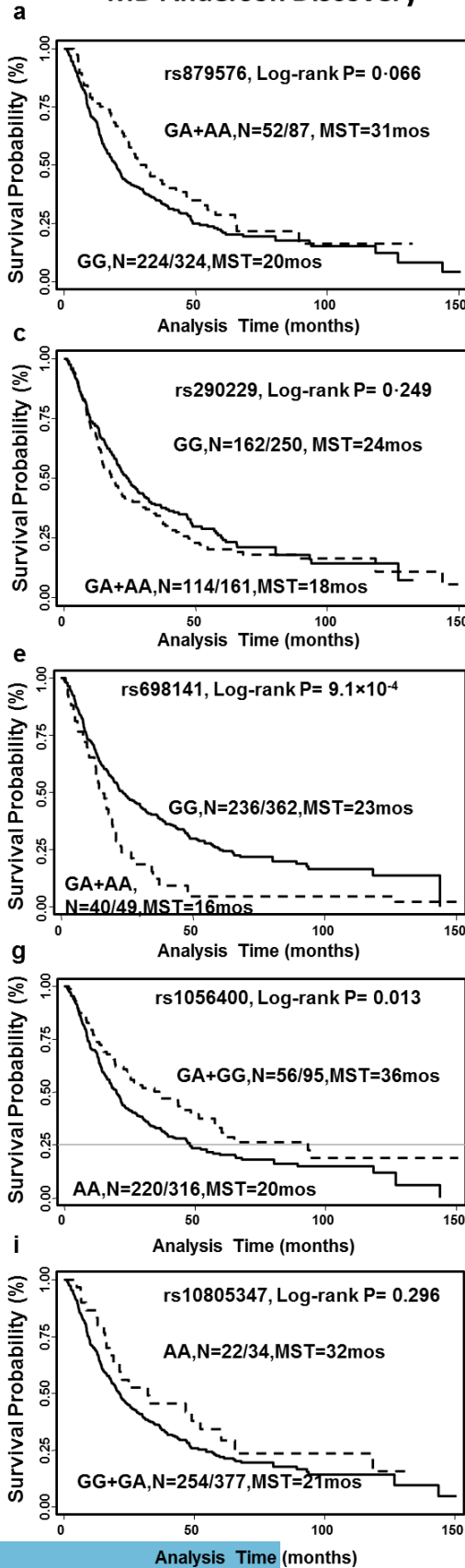


**Table 13: SNPs with the same trend in both the MD Anderson and Mayo Clinic populations**

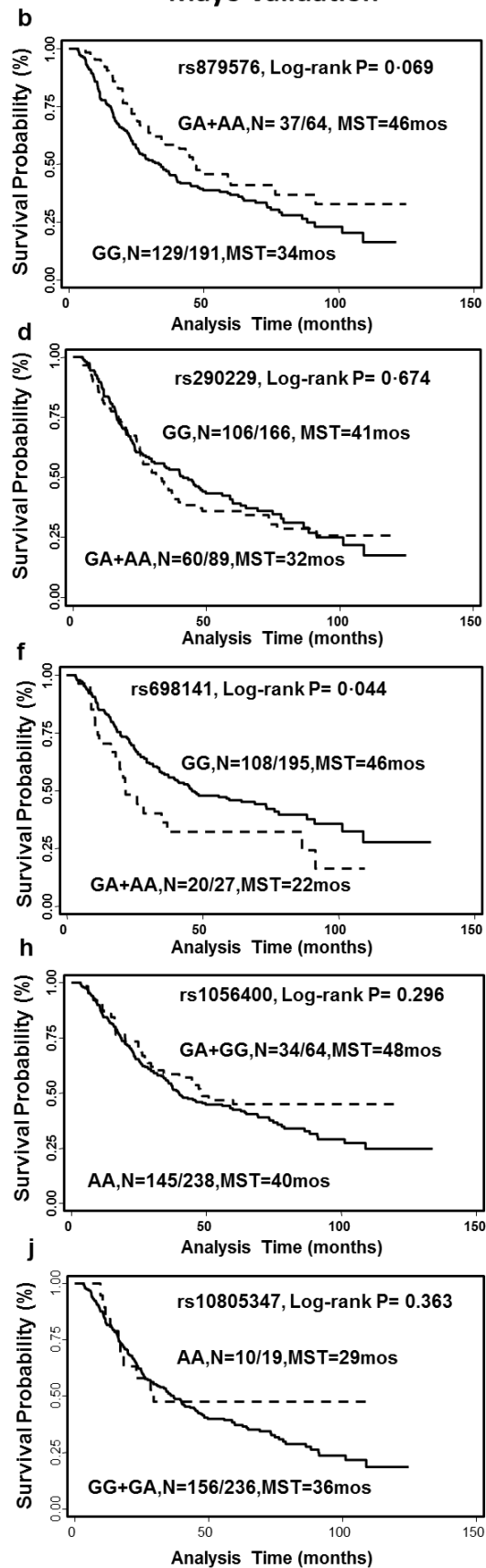
Position	Gene	SNP	Model	MD Anderson(discovery)		Mayo Clinic (validation)		Combined analysis**		
				HR (95% CI)*	p	HR (95% CI)*	p	HR (95% CI)**	p	p-het
Chr22:15969246	<i>IL17RA</i>	rs879576	DOM	0.57 (0.41-0.78)	<b>5.49 × 10<sup>-4</sup></b>	0.65 (0.44-0.94)	<b>0.023</b>	0.60 (0.47-0.77)	<b>4.13 × 10<sup>-5</sup></b>	0.610
Chr9:92674234	<i>SYK</i>	rs290229	DOM	1.58 (1.23-2.03)	<b>3.03 × 10<sup>-4</sup></b>	1.43 (1.01-2.02)	<b>0.046</b>	1.53 (1.25-1.87)	<b>4.15 × 10<sup>-5</sup></b>	0.635
Chr1:39738348	<i>BMP8A</i>	rs698141	DOM	1.89 (1.33-2.68)	<b>4.04 × 10<sup>-4</sup></b>	1.73 (1.03-2.91)	<b>0.038</b>	1.84 (1.37-2.46)	<b>4.29 × 10<sup>-5</sup></b>	0.789
Chr9:21397604	<i>IFNA8</i>	rs4978115	REC	1.79 (1.35-2.36)	<b>4.39 × 10<sup>-5</sup></b>	1.24 (0.86-1.79)	0.240	1.56 (1.25-1.95)	<b>7.43 × 10<sup>-5</sup></b>	0.123
Chr5:149800000	<i>CD74</i>	rs1056400	DOM	0.58 (0.43-0.78)	<b>3.54 × 10<sup>-4</sup></b>	0.71 (0.48-1.04)	<b>0.080</b>	0.62 (0.49-0.79)	<b>1.00 × 10<sup>-4</sup></b>	0.403
Chr9:21403703	<i>IFNA8</i>	rs13296822	REC	1.83 (1.36-2.47)	<b>7.76 × 10<sup>-5</sup></b>	1.23 (0.75-2.04)	0.412	1.65 (1.28-2.13)	<b>1.36 × 10<sup>-4</sup></b>	0.188
Chr4:15449937	<i>CD38</i>	rs10805347	REC	0.45 (0.28-0.72)	<b>7.93 × 10<sup>-4</sup></b>	0.55 (0.28-1.07)	<b>0.080</b>	0.48 (0.33-0.70)	<b>1.75 × 10<sup>-4</sup></b>	0.616
Chr20:36392996	<i>BPI</i>	rs5743539	DOM	2.77 (1.61-4.77)	<b>2.45 × 10<sup>-4</sup></b>	1.59 (0.87-2.92)	0.134	2.16 (1.44-3.25)	<b>1.90 × 10<sup>-4</sup></b>	0.184
Chr13:101700000	<i>FGF14</i>	rs1336726	ADD	0.71 (0.58-0.86)	<b>5.34 × 10<sup>-4</sup></b>	0.80 (0.60-1.07)	0.133	0.74 (0.63-0.87)	<b>2.10 × 10<sup>-4</sup></b>	0.474
Chr16:86446704	<i>SLC7A5</i>	rs4240803	DOM	0.66 (0.52-0.84)	<b>8.39 × 10<sup>-4</sup></b>	0.82 (0.60-1.12)	0.219	0.72 (0.59-0.87)	<b>6.78 × 10<sup>-4</sup></b>	0.290
Chr9:92684769	<i>SYK</i>	rs1755938	DOM	1.63 (1.25-2.14)	<b>3.68 × 10<sup>-4</sup></b>	1.17 (0.80-1.73)	0.417	1.47 (1.17-1.83)	<b>7.17 × 10<sup>-4</sup></b>	0.168
Chr7:2744970	<i>GNA12</i>	rs11971014	DOM	0.59 (0.44-0.80)	<b>4.77 × 10<sup>-4</sup></b>	0.87 (0.58-1.30)	0.488	0.67 (0.53-0.86)	<b>1.18 × 10<sup>-3</sup></b>	0.141
Chr21:33599261	<i>IL10RB</i>	rs2834178	DOM	0.66 (0.51-0.84)	<b>6.96 × 10<sup>-4</sup></b>	0.86 (0.63-1.18)	0.363	0.73 (0.60-0.88)	<b>1.19 × 10<sup>-3</sup></b>	0.178
Chr6:152500000	<i>ESR1</i>	rs9341066	DOM	1.96 (1.37-2.79)	<b>2.01 × 10<sup>-4</sup></b>	1.13 (0.74-1.75)	0.568	1.57 (1.20-2.07)	<b>1.20 × 10<sup>-3</sup></b>	0.056
Chr9:92665566	<i>SYK</i>	rs1572104	DOM	0.63 (0.49-0.81)	<b>3.12 × 10<sup>-4</sup></b>	0.92 (0.65-1.31)	0.660	0.72 (0.59-0.88)	<b>1.43 × 10<sup>-3</sup></b>	0.082
Chr12:6766579	<i>LAG3</i>	rs11064386	DOM	1.68 (1.26-2.23)	<b>3.74 × 10<sup>-4</sup></b>	1.10 (0.76-1.57)	0.618	1.42 (1.14-1.78)	<b>1.92 × 10<sup>-3</sup></b>	0.070
Chr5:172100000	<i>DUSP1</i>	rs4868204	DOM	0.69 (0.53-0.91)	<b>9.40 × 10<sup>-3</sup></b>	0.84 (0.58-1.23)	0.373	0.74 (0.59-0.93)	<b>8.67 × 10<sup>-3</sup></b>	0.421
Chr7:41713523	<i>INHBA</i>	rs12532252	REC	1.34 (1.00-1.79)	<b>0.050</b>	1.17 (0.79-1.72)	0.442	1.27 (1.01-1.61)	<b>0.042</b>	0.580

\*Adjusted according to age, sex, clinical stage, and treatment regimen.\*\*Combined (Meta) analysis based on fixed effects model. Boldface: p<0.1. p-het=P for heterogeneity test; DOM=dominant model; REC=recessive model; ADD=additive model.

### MD Anderson Discovery



### Mayo Validation



**Figure 16 Kaplan-Meier estimates of the effect of selected SNPs on survival probability in never-smokers with lung cancer**

(A) *IL17RA*:rs879576 in the MD Anderson population (discovery phase). (B) *IL17RA*:rs879576 in the Mayo Clinic population (validation phase). (C) *SYK*:rs290229 in the MD Anderson population (discovery phase). (D) *SYK*:rs290229 in the Mayo Clinic population (validation phase). (E) *BMP8A*:rs698141 in the MD Anderson population (discovery phase). (F) *BMP8A*:rs698141 in the Mayo Clinic (validation phase). (G) *CD74*:rs1056400 in the MD Anderson population (discovery phase). (H) *CD74*:rs1056400 in the Mayo Clinic population (validation phase). (I) *CD38*:rs10805347 in the MD Anderson population (discovery phase). (J) *CD38*:rs10805347 in the Mayo Clinic population (validation phase); MST: median survival time in months.  $N=A/B$ , A: number of patients with event, B: total number of patients.

Rs290229 is an intronic SNP in *SYK*, a gene that encodes for a non-receptor type Tyr protein kinase. This SNP was associated with a significantly increased risk of death in both the MD Anderson (HR, 1.58; 95% CI, 1.23-2.03;  $p=3.03 \times 10^{-4}$ ), Mayo Clinic (HR, 1.43; 95% CI, 1.01-2.02;  $p=0.046$ ), and combined (HR, 1.53; 95% CI, 1.25-1.87;  $p=4.15 \times 10^{-5}$ ) populations. Although not significant, both study populations had the same trend of decreased MST (Figures 16c and 16d).

Rs698141 is located in intron of *BMP8A*, a gene involved in cytokine signaling transduction. Patients who had at least one variant allele had a nearly two-fold increase in risk of death in both the MD Anderson (HR, 1.89; 95% CI, 1.33-2.68;  $p=4.04 \times 10^{-4}$ ) and Mayo Clinic (HR, 1.73; 95% CI, 1.03-2.91;  $p=0.038$ ) and combined (HR, 1.84; 95% CI, 1.37-2.46;  $p=4.29 \times 10^{-5}$ ) populations (Table 13). The MST was 23 months in patients with the common homozygous genotype and 16 months in patients with the heterozygous or homozygous variant genotypes in the MD Anderson population ( $p=9.1 \times 10^{-4}$ , log-rank test) (figure 1e). We also observed a similar longer MST (24 months) in the Mayo population ( $p=0.044$ , log-rank test) (Figure 16f).

*CD74*:rs1056400 (3'-untranslated region) and *CD38*:rs10805347 (intronic) were significantly associated with an increased risk of death in the MD Anderson population but were borderline significant in the Mayo Clinic population (Table 13). Although not statistically significant, the trend of differing survival times by genotype was observed (Figures 16g and 16h).

### 3.3.3 Main effects of individual SNPs on survival stratified by histology and stage

Because the majority of never-smokers with lung cancer have adenocarcinoma, we performed a subgroup analysis of survival in patients with adenocarcinoma. The results were similar to those of the overall analysis of all study patients (Table 14). We further performed a stratified analysis of the five top SNPs according to disease stage. Specifically, we combined the MD Anderson and Mayo Clinic patients and stratified them according to early-stage (I and II) and late-stage (III and IV) lung cancer. The results showed that all five SNPs were significantly associated with survival in the late-stage patients, an association that was comparable with or even stronger than that in the overall population. Because of the limited sample size and number of deaths in the early-stage patients, this association was not as robust. However, the same trend of effect for all five SNPs was observed in the early-stage patients (data not shown).

**Table 14: Effect of selected SNPs on survival in adenocarcinoma patients**

Gene	SNP	Model	MD Anderson (discovery)		Mayo Clinic (validation)		Combined analysis**		
			HR (95% CI)*	p	HR (95% CI)*	p	HR (95% CI)**	p	p-het
<i>CD74</i>	rs1056400	DOM	0.60 (0.42-0.85)	<b>3.95 × 10<sup>-3</sup></b>	0.57 (0.36-0.91)	<b>0.017</b>	0.59 (0.44-0.78)	<b>1.86 × 10<sup>-4</sup></b>	0.902
<i>CD38</i>	rs10805347	REC	0.27 (0.14-0.52)	<b>9.77 × 10<sup>-5</sup></b>	0.64 (0.30-1.33)	0.228	0.40 (0.24-0.65)	<b>2.01 × 10<sup>-4</sup></b>	0.093
<i>BMP8A</i>	rs698141	DOM	2.04 (1.38-3.04)	<b>4.09 × 10<sup>-4</sup></b>	1.52 (0.80-2.87)	0.199	1.88 (1.34-2.64)	<b>2.34 × 10<sup>-4</sup></b>	0.438
<i>SYK</i>	rs290229	DOM	1.58 (1.18-2.11)	<b>2.20 × 10<sup>-3</sup></b>	1.43 (0.97-2.12)	<b>0.073</b>	1.52 (1.21-1.93)	<b>4.22 × 10<sup>-4</sup></b>	0.696
<i>IL17RA</i>	rs879576	DOM	0.57 (0.40-0.83)	<b>3.31 × 10<sup>-3</sup></b>	0.65 (0.42-1.00)	<b>0.050</b>	0.61 (0.46-0.80)	<b>4.57 × 10<sup>-4</sup></b>	0.660

\*Adjusted according to age, sex, clinical stage, and treatment regimen. \*\* Combined (Meta) analysis based on fixed effects model. Boldface: p<0.1. p-het=P for heterogeneity test; DOM=dominant model; REC=recessive model.

### 3.3.4 Main effects of individual SNPs on survival in ever-smokers

We next analyzed overall survival in the 996 ever-smokers at MD Anderson to assess the effects of the five SNPs described above on survival according to smoking status. The ever-smokers were slightly older than the never-smokers (mean age, 64.8 years *vs* 61.5 years) and had a smaller proportion of women (42% *vs* 67%) and adenocarcinoma cases (52% *vs* 77%). The treatment regimens in the two groups were similar. None of the SNPs validated in the never-smokers were significantly associated with survival in the ever-smokers (Table 15). We further stratified the ever-smoker patients into former and current smokers and did not observe any significant associations within these subgroups (data not shown).

**Table 15: Effect of selected SNPs on survival according to smoking status in the MD Anderson population**

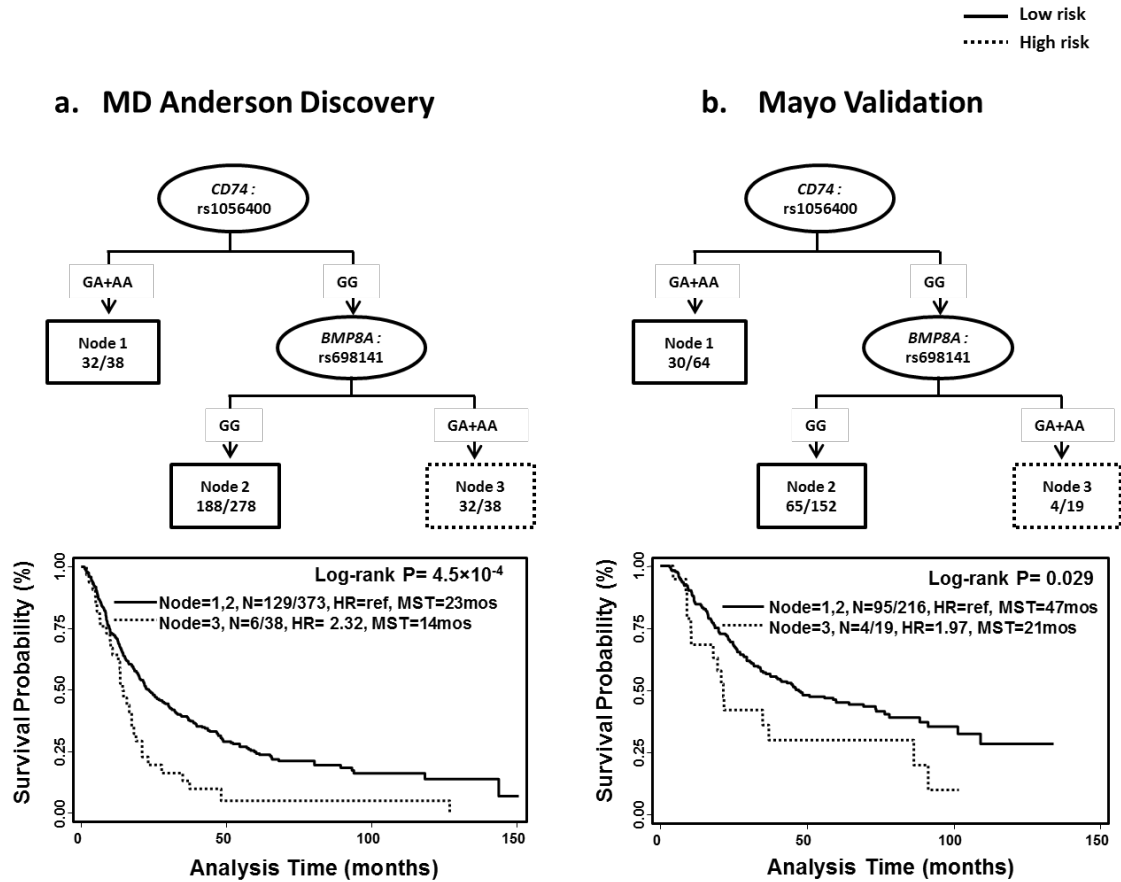
Gene	SNP	Model	Never-smokers		Ever-smokers	
			HR (95% CI)*	p	HR (95% CI)*	p
<i>SYK</i>	rs290229	DOM	1.58 (1.23-2.03)	<b>3.03 × 10<sup>-4</sup></b>	1.12 (0.94-1.33)	0.214
<i>CD74</i>	rs1056400	DOM	0.58 (0.43-0.78)	<b>3.54 × 10<sup>-4</sup></b>	0.93 (0.75-1.15)	0.487
<i>BMP8A</i>	rs698141	DOM	1.89 (1.33-2.68)	<b>4.04 × 10<sup>-4</sup></b>	1.06 (0.81-1.40)	0.655
<i>IL17RA</i>	rs879576	DOM	0.57 (0.41-0.78)	<b>5.49 × 10<sup>-4</sup></b>	0.89 (0.72-1.10)	0.266
<i>CD38</i>	rs10805347	REC	0.45 (0.28-0.72)	<b>7.93 × 10<sup>-4</sup></b>	0.91 (0.67-1.24)	0.557

\*Adjusted according to age, sex, clinical stage, and treatment regimen. Boldface: p<0.1.  
DOM=dominant model; REC=recessive model.



### 3.3.5 Survival tree analysis

Survival tree analysis was used to identify higher order gene-gene interactions among these five SNPs in modulating risk of death. Using the MD Anderson never-smoker population as a training set, we identified two SNPs (*CD74*:rs1056400 and *BMP8A*:rs698141) potentially having gene-gene interactions. Patients with the rs1056400\_GG/rs698141\_GA+AA genotype (node 3) had a 2.32-fold greater risk of death (HR, 2.32; 95% CI, 1.58-3.41;  $p=1.72 \times 10^{-5}$ ) and significantly shorter MST (14 months vs 23 months;  $p=4.5 \times 10^{-4}$ , log-rank test) than did patients with the rs1056400\_GG/rs698141\_GG or rs1056400\_GA+AA genotype (nodes 1 and 2). This tree model was validated in the Mayo Clinic population: patients with the rs1056400\_GG/rs698141\_GA+AA genotype (node 3) had a nearly twofold greater risk of death (HR, 1.97; 95% CI, 1.11-3.50;  $p=0.02$ ) and a strikingly shorter MST (by 26 months) than did patients with the rs1056400\_GG/rs698141\_GG or rs1056400\_GA+AA genotype (nodes 1 and 2) ( $p=0.029$ , log-rank test) (Figure 17).



**Figure 17 Potential gene-gene interactions among SNPs validated in the survival tree analysis**

(A) Survival tree analysis results and Kaplan-Meier estimates in the MD Anderson population (discovery phase). (B) Survival tree analysis results and Kaplan-Meier estimates in the Mayo Clinic population (validation phase). MST: median survival time in months. N=A/B, A: number of patients with event, B: total number of patients.

### 3.3.6 Discussion

NSCLC in never-smokers is unique from that in ever-smokers due to distinct clinical, histological, and genetic characteristics. These attributes warrant specific investigation of never-smokers. Although we are in the era of the genome-wide association study (GWAS), the coverage of certain genetic region on commercial available GWAS chips is not sufficient for detailed genetic analysis; this limits the power of GWAS to identify all genetic determinants. Thus study design based on prior knowledge focusing on known cancer relations is indispensable. In this context, we conducted a two-stage, discovery-validation study to identify genetic predictors of overall survival in never-smokers with lung cancer using a pathway-based approach. By systematically evaluating SNPs in major inflammatory pathways, we found five SNPs in *CD74*, *CD38*, *SYK*, *BMP8A*, and *IL17RA* that were significantly associated with overall survival in these patients. Furthermore, we analyzed and validated a survival tree model in predicting survival that takes gene-gene interactions into consideration. In comparing the associations of SNPs with survival in ever- and never-smokers, we provided evidence of distinct roles for inflammatory genetic determinants of prognosis in never-smokers with lung cancer.

Two SNPs—*IL17RA*:rs879576 and *BMP8A*:rs698141—are related to cytokine signaling. *IL17RA* is an isoform of the interleukin (IL)-17 receptors. In the presence of IL-17 ligands, these receptors can activate various downstream signaling pathways to induce macrophage recruitment, angiogenesis, and inflammatory lung diseases.(202, 203) In our study, *IL17RA*:rs879576 was associated with a consistent protective effect against death and corresponding prolonged MSTs in both the MD Anderson and Mayo Clinic populations. This is a synonymous SNP located in the last exon of *IL17RA* that may influence the

structure and/or regulation of its host gene. *BMP8A* is a member of the transforming growth factor  $\beta$  superfamily (204). BMP proteins play important roles in cell differentiation, proliferation, survival, and apoptosis and are implicated in tumor cell migration, metastasis, and angiogenesis in various cancers (205-208). Rs698141 is located in the first intron of *BMP8A*, and not in any obvious functional elements. Therefore, it is most likely linked with other functional SNPs that result in *BMP8A* altered function. Authors have reported that tobacco smoking can lead to immunosuppression and downregulation of proinflammatory cytokines specifically in the lung tissues, suggesting important roles for cytokines in lung pathology.(209) Cytokine signaling pathway variants were predominant in our validated SNPs highlighted the potential roles of cytokines in determining prognosis for lung cancer in never-smokers.

*SYK* belongs to the Syk family of tyrosine kinases and plays an oncogenic role in different cancers.(210) In lung cancer cells, *SYK* is silenced owing to hypermethylation in its promoter region.(211) *SYK*:rs290229 was associated with an increased risk of death and reduced survival in our populations. This SNP is located in an intron; it is possible that this SNP tagged another causal variant that affects the function of *SYK*. Further deep sequencing would be warranted to identify the potential casual locus responsible for this finding.

Two other SNPs—*CD38*:rs10805347 and *CD74*:rs1056400—were borderline significant in our validation. *CD74* is a member of a class of polypeptides involved in antigen presentation that is a potential therapeutic target and prognostic factor for cancer (212-215) with involvement in lung adenocarcinoma. Our results suggested the potential prognostic role of *CD74*:rs1056400 regarding overall survival in lung cancer patients. *CD38* is a multifunctional single-chain type II transmembrane glycoprotein, related to the development

of viral infections, diabetes, and cancer.(216) Studies have shown a prognostic role for CD38 in leukemia patients.(217) We observed a consistent protective effect for *CD38:rs10805347* against death, which indicated a potential role for this gene in solid cancers in addition to leukemia.

In the current study, we aimed at identify specific prognostic markers for never smokers. Although incidence is increasing, lung cancer in never smokers represents only ~10% of all lung cancer cases. Thus, to identify a homogeneous never smoking patient cohort with adequate demographic/clinical variables is a challenge. In this study, we were able to identify relatively large and well-characterized study populations from two study sites with complete collection of clinical and epidemiological data that enabled us to recruit a sufficient study population. This provided an important resource contributing to the understanding of this disease which has emerged as a major public health problem tracking smoking and smoking cessation rate. Interestingly, none of the five SNPs were significantly associated with overall survival in ever-smokers, providing additional evidence of lung cancer in never smoker as a distinct disease and requires identifying specific prognostic markers.

Moreover, the multi-stage study design with two independent patient populations largely reduced the likelihood of false-positive results for the SNPs that were significant in both populations. Therefore, although a portion of the findings would not be judged significant due to multiple comparisons, the replication provides a mechanism to address these concerns, attenuating the need for strict multiple comparisons correction. Another significant finding in our study was the identification and validation of a survival tree, which has proven to be a powerful analytical tool regarding survival in cancer patients based on higher order gene-

gene interactions (37-39). The survival tree analysis stratified the Mayo Clinic patients into significantly different risk subgroups in a manner similar to that in the MD Anderson patients. Beyond the effect of a single SNP on survival, the survival tree takes into account the complicated interactions of genes which are yet not discovered and has high predictive power regarding patients' prognosis that may be clinically applicable.

In conclusion, this is the first large-scale study to examine the association of SNPs in 800 inflammation-related genes with survival in never-smokers with lung cancer. The identified individual SNPs and the survival tree may be applicable to future modeling of clinical outcome for prediction of survival following validation in other independent populations of never-smokers with lung cancer.

## *Chapter 4: Conclusions*

Overall, the current study provides evidence that genetic variants in the miRNA and inflammation related pathways could influence clinical outcomes for NSCLC patients.

We evaluated miRNA pathway SNPs for their potential prognostic role and have identified some significant findings. Specifically, we identified a FAS gene binding site SNP that may predict overall survival in these patients and supported this observation with *in vitro* functional genomic analyses. We also identified and validated potential genetic markers within the inflammation pathway that may affect clinical outcome in NSCLC patients, particularly in never smokers and late-stage patients. Moreover, we have identified and validated a survival tree which has proven to be a powerful analytical tool regarding survival in never smoker cancer patients based on higher order gene-gene interactions. Given the important role of miRNA and inflammation throughout the cancer continuum, these genetic markers may be good potential prognostic markers that can help tailor treatment regimens in the clinic.

With further functional analysis and validations, these findings can help increase the prediction accuracy for traditional prognostic factors in predicting patients' prognosis through identification of optimal treatment and follow-up care regimens.



## *Chapter 5: Strength and Limitations*

One of the greatest strengths of our studies is the relatively large sample size. Studying never smoking lung cancer patients, a population that accounts for only ~10% of all lung cancer cases, can be difficult because identifying a homogeneous never smoking patient cohort with adequate demographic/clinical variables is usually a challenge. In the current study, we were able to identify relatively large and well-characterized study populations from two study sites with complete collection of clinical and epidemiological data that enabled us to recruit a sufficient study population. Tracking smoking and smoking cessation rates in large populations of lung cancer patients like this one can play a major role in understanding the disease. The detailed clinical information collected for these study subjects has enabled us to further investigate association in specific subgroups and helped us identify several treatment specific markers that may help to evaluate potential risks and/or benefits of different treatment regimens for patient subgroups with specific genotypes. Further studies of these SNPs in an independent population would be valuable in confirming our results. Moreover, the comprehensive query of SNPs from genes involved in both miRNA regulating and inflammatory responses provide a broad overview and investigation of the role of these genetic variations in modulating patients' clinical outcomes.

False discovery is an inherited issue for large scale association studies. We are aware that there is a chance of false discovery in our results. To correct for this beyond controlling for false discovery with a statistical strategy (FDR), we adopted several other approaches to further validate and study our findings. Since inflammation related SNPs are tagging SNPs, commonly as intronic polymorphisms, we adopted a multi-stage study design with independent patient populations identified from three of the largest lung cancer studies in the

US (MD Anderson, Mayo Clinic, and Harvard University) to which largely reduced the likelihood of false-positive results for the SNPs that were significant in all populations. Therefore, although a portion of the findings would not be judged significant due to multiple comparisons, the replication provides a mechanism to address these concerns attenuating the need for strict multiple comparisons correction. Furthermore, instead of limiting our study to top SNPs, an approach usually adopted in pathway-based association studies to reduce cost and labor, we extensively genotyped almost all significant SNPs during our internal validation. This strict validation approach substantially improved the power of our study to detect candidate loci for subsequent analysis. For those potentially functional (miRNA binding sites, non-synonymous) SNPs, either *in vitro* (luciferase reporter assay) or *in silico* (SIFT/Polyphen), functional analysis was performed to evaluate influence of these variants to gene or protein function to help better understanding our results.

An additional limitation of our study is due to the location of our SNPs. Most of our validated SNPs are located in intron- or intra-genic regions so their function to host or nearby genes are not clear. As a result, further fine-mapping or deep sequencing would be needed to discover the causal allele. In addition, although we performed functional analysis for miRNA binding sites polymorphisms, assays have not been performed in protein level or *in vivo*. Following-up deeper analysis for their functions is warranted.

## *Chapter 6: Future Directions*

In the current studies, we only focused on two critical pathways. There are other pathways also important for the understanding of lung cancer clinical outcomes. Thus, we will continue to identify and analyze more interesting pathways to gain a better overview of genetic variations contributing to patients' clinical outcomes.

We will make effort to seek collaborations and identify other independent populations with adequate and comparable repository of clinical and epidemiological data, to provide additional statistical power, and further validate our results. At the meantime, with the continuous recruitment of cases and longer follow-up time in our study and other ongoing GWAS of lung cancer in the field, we will have sufficient power to conduct GWAS of clinical outcomes.

After validated our findings, functional characterization and phenotypic analysis will be the major focus for our future studies. Deep sequencing or fine-mapping will be used to identify real causal allele tagged by intronic SNPs found in our study. Then, functional analysis will be designed accordingly. Luciferase assays will be designed and performed where feasible. For those SNPs that have already undergone functional analysis, further and deeper biological characterization will be done to test for their functions at the gene and protein level. When sample available, phenotypic assay, such as gene expression or protein array, will be designed to further explore the prognostic values of the identified genes. Mice models can be developed in collaboration with other basic science laboratories if feasible.

With more solid and comprehensive results as well as deeper understanding of the influence of the genetic variations on NSCLC clinical outcomes, these identified markers

could be incorporated into a prognosis prediction model to increase prediction accuracy in both population and individual level.

## References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin.* 2012;62:10-29.
2. Gomperts BN, Spira A, Massion PP, Walser TC, Wistuba, II, Minna JD, Dubinett SM. Evolving concepts in lung carcinogenesis. *Semin Respir Crit Care Med.* 2011;32:32-43.
3. Young JL, Jr., Percy CL, Asire AJ, Berg JW, Cusano MM, Gloeckler LA, Horm JW, Lourie WI, Jr., Pollack ES, Shambaugh EM. Cancer incidence and mortality in the United States, 1973-77. *Natl Cancer Inst Monogr.* 1981:1-187.
4. Chaudhuri MR. Primary pulmonary cavitating carcinomas. *Thorax.* 1973;28:354-66.
5. Woodring JH, Stelling CB. Adenocarcinoma of the lung: a tumor with a changing pleomorphic character. *AJR Am J Roentgenol.* 1983;140:657-64.
6. Glazer HS, Kaiser LR, Anderson DJ, Molina PL, Emami B, Roper CL, Sagel SS. Indeterminate mediastinal invasion in bronchogenic carcinoma: CT evaluation. *Radiology.* 1989;173:37-42.
7. Ferlay j, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. GLOBOCAN 2008: cancer incidence and mortality worldwide: IARC Cancer Base. no 10 International Agency for Research on Cancer. Lyon2010.
8. Thun MJ, Henley SJ, Burns D, Jemal A, Shanks TG, Calle EE. Lung cancer death rates in lifelong nonsmokers. *J Natl Cancer Inst.* 2006;98:691-9.
9. Wang JB, Jiang Y, Wei WQ, Yang GH, Qiao YL, Boffetta P. Estimation of cancer incidence and mortality attributable to smoking in China. *Cancer Causes Control.* 2010;21:959-65.

10. Peto R, Lopez A, Boreham J, Thun M. Mortality from smoking in developed countries 1950–2000. 2nd ed. Oxford: Oxford University Press; 2006.
11. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*. 2000;321:323-9.
12. Subramanian J, Govindan R. Lung cancer in never smokers: a review. *J Clin Oncol*. 2007;25:561-70.
13. Yano T, Miura N, Takenaka T, Haro A, Okazaki H, Ohba T, Kouso H, Kometani T, Shoji F, Maehara Y. Never-smoking nonsmall cell lung cancer as a separate entity: clinicopathologic features and survival. *Cancer*. 2008;113:1012-8.
14. Bryant A, Cerfolio RJ. Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer. *Chest*. 2007;132:185-92.
15. Meguid RA, Hooker CM, Harris J, Xu L, Westra WH, Sherwood JT, Sussman M, Cattaneo SM, 2nd, Shin J, Cox S, Christensen J, Prints Y, Yuan N, Zhang J, Yang SC, Brock MV. Long-term survival outcomes by smoking status in surgical and nonsurgical patients with non-small cell lung cancer: comparing never smokers and current smokers. *Chest*. 138:500-9.
16. Vital signs: nonsmokers' exposure to secondhand smoke --- United States, 1999-2008. *MMWR Morb Mortal Wkly Rep*. 2010;59:1141-6.
17. National Institutes of Health NCI. Smoking and Tobacco Control Monograph 10: Health Effects of Exposure to Environmental Tobacco Smoke. 1999.



18. Ellis PM, Vandermeer R. Delays in the diagnosis of lung cancer. *J Thorac Dis.* 2011;3:183-8.
19. Hyer JD, Silvestri G. Diagnosis and staging of lung cancer. *Clin Chest Med.* 2000;21:95-106, viii-ix.
20. Muers MF, Robertson RJ. Diagnosis of lung cancer: FOB before CT or CT before FOB? *Thorax.* 2000;55:350-1.
21. Pastorino U. Lung cancer: diagnosis and surgery. *Eur J Cancer.* 2001;37 Suppl 7:S75-90.
22. Petty TL. The early diagnosis of lung cancer. *Dis Mon.* 2001;47:204-64.
23. Mountain CF. Revisions in the International System for Staging Lung Cancer. *Chest.* 1997;111:1710-7.
24. Benson MK. Age and the treatment of lung cancer. *Thorax.* 1997;52:203.
25. Krauss S, Perez C, Lowenbraun S, Sonoda T, Bartolucci A, Buchanan R. Combined modality treatment of localized small-cell lung carcinoma. A randomized prospective study of the Southeastern Cancer Study Group. *Cancer Clin Trials.* 1980;3:297-305.
26. Westeel V, Depierre A. Combined modality treatment of non-small-cell lung cancer. *Am J Respir Med.* 2003;2:477-90.
27. Burdett S, Stewart LA, Rydzewska L. A systematic review and meta-analysis of the literature: chemotherapy and surgery versus surgery alone in non-small cell lung cancer. *J Thorac Oncol.* 2006;1:611-21.

28. Martini N, Bains MS, Burt ME, Zakowski MF, McCormack P, Rusch VW, Ginsberg RJ. Incidence of local recurrence and second primary tumors in resected stage I lung cancer. *J Thorac Cardiovasc Surg.* 1995;109:120-9.
29. Albain KS, Crowley JJ, LeBlanc M, Livingston RB. Survival determinants in extensive-stage non-small-cell lung cancer: the Southwest Oncology Group experience. *J Clin Oncol.* 1991;9:1618-26.
30. Macchiarini P, Fontanini G, Hardin MJ, Chuanchieh H, Bigini D, Vignati S, Pingitore R, Angeletti CA. Blood vessel invasion by tumor cells predicts recurrence in completely resected T1 N0 M0 non-small-cell lung cancer. *J Thorac Cardiovasc Surg.* 1993;106:80-9.
31. Ichinose Y, Yano T, Asoh H, Yokoyama H, Yoshino I, Katsuda Y. Prognostic factors obtained by a pathologic examination in completely resected non-small-cell lung cancer. An analysis in each pathologic stage. *J Thorac Cardiovasc Surg.* 1995;110:601-5.
32. Fontanini G, Bigini D, Vignati S, Basolo F, Mussi A, Lucchi M, Chine S, Angeletti CA, Harris AL, Bevilacqua G. Microvessel count predicts metastatic disease and survival in non-small cell lung cancer. *J Pathol.* 1995;177:57-63.
33. Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, Postmus PE, Rusch V, Sobin L. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol.* 2007;2:706-14.
34. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest.* 2002;122:1037-57.

35. Ginsberg RJ, Hill LD, Eagan RT, Thomas P, Mountain CF, Deslauriers J, Fry WA, Butz RO, Goldberg M, Waters PF, et al. Modern thirty-day operative mortality for surgical resections in lung cancer. *J Thorac Cardiovasc Surg.* 1983;86:654-8.
36. Gilligan D, Nicolson M, Smith I, Groen H, Dalesio O, Goldstraw P, Hatton M, Hopwood P, Manegold C, Schramel F, Smit H, van Meerbeeck J, Nankivell M, Parmar M, Pugh C, Stephens R. Preoperative chemotherapy in patients with resectable non-small cell lung cancer: results of the MRC LU22/NVALT 2/EORTC 08012 multicentre randomised trial and update of systematic review. *Lancet.* 2007;369:1929-37.
37. Postoperative radiotherapy for non-small cell lung cancer. *Cochrane Database Syst Rev.* 2005:CD002142.
38. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. *CA Cancer J Clin.*60:277-300.
39. Gurubhagavatula S, Liu G, Park S, Zhou W, Su L, Wain JC, Lynch TJ, Neuberg DS, Christiani DC. XPD and XRCC1 genetic polymorphisms are prognostic factors in advanced non-small-cell lung cancer patients treated with platinum chemotherapy. *J Clin Oncol.* 2004;22:2594-601.
40. Tibaldi C, Giovannetti E, Vasile E, Mey V, Laan AC, Nannizzi S, Di Marsico R, Antonuzzo A, Orlandini C, Ricciardi S, Del Tacca M, Peters GJ, Falcone A, Danesi R. Correlation of CDA, ERCC1, and XPD polymorphisms with response and survival in gemcitabine/cisplatin-treated advanced non-small cell lung cancer patients. *Clin Cancer Res.* 2008;14:1797-803.

41. Rosell R, Cecere F, Santarpia M, Reguart N, Taron M. Predicting the outcome of chemotherapy for lung cancer. *Curr Opin Pharmacol.* 2006;6:323-31.
42. Miller JI, Jr., Phillips TW. Neodymium:YAG laser and brachytherapy in the management of inoperable bronchogenic carcinoma. *Ann Thorac Surg.* 1990;50:190-5; discussion 5-6.
43. Lester JF, Macbeth FR, Toy E, Coles B. Palliative radiotherapy regimens for non-small cell lung cancer. *Cochrane Database Syst Rev.* 2006:CD002143.
44. Mountain CF. New prognostic factors in lung cancer. Biologic prophets of cancer cell aggression. *Chest.* 1995;108:246-54.
45. Salgia R, Skarin AT. Molecular abnormalities in lung cancer. *J Clin Oncol.* 1998;16:1207-17.
46. Riely GJ, Marks J, Pao W. KRAS mutations in non-small cell lung cancer. *Proc Am Thorac Soc.* 2009;6:201-5.
47. van Zandwijk N, Mathy A, Boerrigter L, Ruijter H, Tielen I, de Jong D, Baas P, Burgers S, Nederlof P. EGFR and KRAS mutations as criteria for treatment with tyrosine kinase inhibitors: retro- and prospective observations in non-small-cell lung cancer. *Ann Oncol.* 2007;18:99-103.
48. Eberhard DA, Johnson BE, Amler LC, Goddard AD, Heldens SL, Herbst RS, Ince WL, Janne PA, Januario T, Johnson DH, Klein P, Miller VA, Ostland MA, Ramies DA, Sebisanovic D, Stinson JA, Zhang YR, Seshagiri S, Hillan KJ. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J Clin Oncol.* 2005;23:5900-9.

49. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*. 2004;304:1497-500.
50. Beroukhi R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Taberero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010;463:899-905.
51. Testa JR, Siegfried JM. Chromosome abnormalities in human non-small cell lung cancer. *Cancer Res*. 1992;52:2702s-6s.
52. Yoshino I, Osoegawa A, Yohena T, Kameyama T, Oki E, Oda S, Maehara Y. Loss of heterozygosity (LOH) in non-small cell lung cancer: difference between adenocarcinoma and squamous cell carcinoma. *Respir Med*. 2005;99:308-12.
53. Otterson G, Lin A, Kay F. Genetic etiology of lung cancer. *Oncology (Williston Park)*. 1992;6:97-104, 7; discussion 8, 11-2.
54. Burbee DG, Forgacs E, Zochbauer-Muller S, Shivakumar L, Fong K, Gao B, Randle D, Kondo M, Virmani A, Bader S, Sekido Y, Latif F, Milchgrub S, Toyooka S,

- Gazdar AF, Lerman MI, Zabarovsky E, White M, Minna JD. Epigenetic inactivation of RASSF1A in lung and breast cancers and malignant phenotype suppression. *J Natl Cancer Inst.* 2001;93:691-9.
55. Maruyama R, Sugio K, Yoshino I, Maehara Y, Gazdar AF. Hypermethylation of FHIT as a prognostic marker in nonsmall cell lung carcinoma. *Cancer.* 2004;100:1472-7.
56. Bremnes RM, Veve R, Gabrielson E, Hirsch FR, Baron A, Bemis L, Gemmill RM, Drabkin HA, Franklin WA. High-throughput tissue microarray analysis used to evaluate biology and prognostic significance of the E-cadherin pathway in non-small-cell lung cancer. *J Clin Oncol.* 2002;20:2417-28.
57. Slebos RJ, Kibbelaar RE, Dalesio O, Kooistra A, Stam J, Meijer CJ, Wagenaar SS, Vanderschueren RG, van Zandwijk N, Mooi WJ, et al. K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung. *N Engl J Med.* 1990;323:561-5.
58. Bunn PA, Jr., Doebele RC. Genetic testing for lung cancer: reflex versus clinical selection. *J Clin Oncol.* 2011;29:1943-5.
59. Paik PK, Arcila ME, Fara M, Sima CS, Miller VA, Kris MG, Ladanyi M, Riely GJ. Clinical Characteristics of Patients With Lung Adenocarcinomas Harboring BRAF Mutations. *J Clin Oncol.* 2011;29:2046-51.
60. Kern JA, Schwartz DA, Nordberg JE, Weiner DB, Greene MI, Torney L, Robinson RA. p185neu expression in human lung adenocarcinomas predicts shortened survival. *Cancer Res.* 1990;50:5184-7.

61. Laudanski J, Chyczewski L, Niklinska WE, Kretowska M, Furman M, Sawicki B, Niklinski J. Expression of bcl-2 protein in non-small cell lung cancer: correlation with clinicopathology and patient survival. *Neoplasma*. 1999;46:25-30.
62. Zheng Z, Chen T, Li X, Haura E, Sharma A, Bepler G. DNA synthesis and repair genes RRM1 and ERCC1 in lung cancer. *N Engl J Med*. 2007;356:800-8.
63. Olaussen KA, Dunant A, Fouret P, Brambilla E, Andre F, Haddad V, Taranchon E, Filipits M, Pirker R, Popper HH, Stahel R, Sabatier L, Pignon JP, Tursz T, Le Chevalier T, Soria JC. DNA repair by ERCC1 in non-small-cell lung cancer and cisplatin-based adjuvant chemotherapy. *N Engl J Med*. 2006;355:983-91.
64. Ceppi P, Volante M, Novello S, Rapa I, Danenberg KD, Danenberg PV, Cambieri A, Selvaggi G, Saviozzi S, Calogero R, Papotti M, Scagliotti GV. ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine. *Ann Oncol*. 2006;17:1818-25.
65. Reynolds C, Obasaju C, Schell MJ, Li X, Zheng Z, Boulware D, Caton JR, Demarco LC, O'Rourke MA, Shaw Wright G, Boehm KA, Asmar L, Bromund J, Peng G, Monberg MJ, Bepler G. Randomized phase III trial of gemcitabine-based chemotherapy with in situ RRM1 and ERCC1 protein levels for response prediction in non-small-cell lung cancer. *J Clin Oncol*. 2009;27:5808-15.
66. Minev B, Hipp J, Firat H, Schmidt JD, Langlade-Demoyen P, Zanetti M. Cytotoxic T cell immunity against telomerase reverse transcriptase in humans. *Proc Natl Acad Sci U S A*. 2000;97:4796-801.

67. Marrogi AJ, Travis WD, Welsh JA, Khan MA, Rahim H, Tazelaar H, Pairolero P, Trastek V, Jett J, Caporaso NE, Liotta LA, Harris CC. Nitric oxide synthase, cyclooxygenase 2, and vascular endothelial growth factor in the angiogenesis of non-small cell lung carcinoma. *Clin Cancer Res.* 2000;6:4739-44.
68. Khuri FR, Wu H, Lee JJ, Kemp BL, Lotan R, Lippman SM, Feng L, Hong WK, Xu XC. Cyclooxygenase-2 overexpression is a marker of poor prognosis in stage I non-small cell lung cancer. *Clin Cancer Res.* 2001;7:861-7.
69. Hida T, Yatabe Y, Achiwa H, Muramatsu H, Kozaki K, Nakamura S, Ogawa M, Mitsudomi T, Sugiura T, Takahashi T. Increased expression of cyclooxygenase 2 occurs frequently in human lung cancers, specifically in adenocarcinomas. *Cancer Res.* 1998;58:3761-4.
70. Wolff H, Saukkonen K, Anttila S, Karjalainen A, Vainio H, Ristimaki A. Expression of cyclooxygenase-2 in human lung carcinoma. *Cancer Res.* 1998;58:4997-5001.
71. Salden M, Splinter TA, Peters HA, Look MP, Timmermans M, van Meerbeeck JP, Foekens JA. The urokinase-type plasminogen activator system in resected non-small-cell lung cancer. Rotterdam Oncology Thoracic Study Group. *Ann Oncol.* 2000;11:327-32.
72. Malmstrom P, Bendahl PO, Boiesen P, Brunner N, Idvall I, Ferno M. S-phase fraction and urokinase plasminogen activator are better markers for distant recurrences than Nottingham Prognostic Index and histologic grade in a prospective study of premenopausal lymph node-negative breast cancer. *J Clin Oncol.* 2001;19:2010-9.



73. Nelson AR, Fingleton B, Rothenberg ML, Matrisian LM. Matrix metalloproteinases: biologic activity and clinical implications. *J Clin Oncol.* 2000;18:1135-49.
74. Sellers TA, Ooi WL, Elston RC, Chen VW, Bailey-Wilson JE, Rothschild H. Increased familial risk for non-lung cancer among relatives of lung cancer patients. *Am J Epidemiol.* 1987;126:237-46.
75. Ooi WL, Elston RC, Chen VW, Bailey-Wilson JE, Rothschild H. Increased familial risk for lung cancer. *J Natl Cancer Inst.* 1986;76:217-22.
76. Mayne ST, Buenconsejo J, Janerich DT. Familial cancer history and lung cancer risk in United States nonsmoking men and women. *Cancer Epidemiol Biomarkers Prev.* 1999;8:1065-9.
77. Broman K, Pohlabeln H, Jahn I, Ahrens W, Jockel KH. Aggregation of lung cancer in families: results from a population-based case-control study in Germany. *Am J Epidemiol.* 2000;152:497-505.
78. Lerman C, Shields AE. Genetic testing for cancer susceptibility: the promise and the pitfalls. *Nat Rev Cancer.* 2004;4:235-41.
79. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. *Oncogene.* 2004;23:6471-6.
80. Bailey-Wilson JE, Amos CI, Pinney SM, Petersen GM, de Andrade M, Wiest JS, Fain P, Schwartz AG, You M, Franklin W, Klein C, Gazdar A, Rothschild H, Mandal D, Coons T, Slusser J, Lee J, Gaba C, Kupert E, Perez A, Zhou X, Zeng D, Liu Q, Zhang Q, Seminara D, Minna J, Anderson MW. A major lung cancer susceptibility locus maps to chromosome 6q23-25. *Am J Hum Genet.* 2004;75:460-74.

81. You M, Wang D, Liu P, Vikis H, James M, Lu Y, Wang Y, Wang M, Chen Q, Jia D, Liu Y, Wen W, Yang P, Sun Z, Pinney SM, Zheng W, Shu XO, Long J, Gao YT, Xiang YB, Chow WH, Rothman N, Petersen GM, de Andrade M, Wu Y, Cunningham JM, Wiest JS, Fain PR, Schwartz AG, Girard L, Gazdar A, Gaba C, Rothschild H, Mandal D, Coons T, Lee J, Kupert E, Seminara D, Minna J, Bailey-Wilson JE, Amos CI, Anderson MW. Fine mapping of chromosome 6q23-25 region in familial lung cancer families reveals RGS17 as a likely candidate gene. *Clin Cancer Res.* 2009;15:2666-74.
82. Sierra DA, Gilbert DJ, Householder D, Grishin NV, Yu K, Ukidwe P, Barker SA, He W, Wensel TG, Otero G, Brown G, Copeland NG, Jenkins NA, Wilkie TM. Evolution of the regulators of G-protein signaling multigene family in mouse and human. *Genomics.* 2002;79:177-85.
83. Chung CC, Magalhaes WC, Gonzalez-Bosquet J, Chanock SJ. Genome-wide association studies in cancer--current and future directions. *Carcinogenesis.* 31:111-20.
84. Park JY, Park JM, Jang JS, Choi JE, Kim KM, Cha SI, Kim CH, Kang YM, Lee WK, Kam S, Park RW, Kim IS, Lee JT, Jung TH. Caspase 9 promoter polymorphisms and risk of primary lung cancer. *Hum Mol Genet.* 2006;15:1963-71.
85. Zhang X, Miao X, Sun T, Tan W, Qu S, Xiong P, Zhou Y, Lin D. Functional polymorphisms in cell death pathway genes FAS and FASL contribute to risk of lung cancer. *J Med Genet.* 2005;42:479-84.
86. Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. *Nat Rev Cancer.* 2004;4:850-60.

87. Dong LM, Potter JD, White E, Ulrich CM, Cardon LR, Peters U. Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA*. 2008;299:2423-36.
88. Sifri R, Gangadharappa S, Acheson LS. Identifying and testing for hereditary susceptibility to common cancers. *CA Cancer J Clin*. 2004;54:309-26.
89. Guan P, Huang D, Yin Z, Zhou B. Association of the hOGG1 Ser326Cys polymorphism with increased lung cancer susceptibility in Asians: a meta-analysis of 18 studies including 7592 cases and 8129 controls. *Asian Pac J Cancer Prev*. 2011;12:1067-72.
90. Sugimura H, Kohno T, Wakai K, Nagura K, Genka K, Igarashi H, Morris BJ, Baba S, Ohno Y, Gao C, Li Z, Wang J, Takezaki T, Tajima K, Varga T, Sawaguchi T, Lum JK, Martinson JJ, Tsugane S, Iwamasa T, Shinmura K, Yokota J. hOGG1 Ser326Cys polymorphism and lung cancer susceptibility. *Cancer Epidemiol Biomarkers Prev*. 1999;8:669-74.
91. Qian B, Zhang H, Zhang L, Zhou X, Yu H, Chen K. Association of genetic polymorphisms in DNA repair pathway genes with non-small cell lung cancer risk. *Lung Cancer*. 2011;73:138-46.
92. Zienolddiny S, Campa D, Lind H, Ryberg D, Skaug V, Stangeland L, Phillips DH, Canzian F, Haugen A. Polymorphisms of DNA repair genes and risk of non-small cell lung cancer. *Carcinogenesis*. 2006;27:560-7.
93. Schneider J, Classen V, Helmig S. XRCC1 polymorphism and lung cancer risk. *Expert Rev Mol Diagn*. 2008;8:761-80.

94. Hao B, Miao X, Li Y, Zhang X, Sun T, Liang G, Zhao Y, Zhou Y, Wang H, Chen X, Zhang L, Tan W, Wei Q, Lin D, He F. A novel T-77C polymorphism in DNA repair gene XRCC1 contributes to diminished promoter activity and increased risk of non-small cell lung cancer. *Oncogene*. 2006;25:3613-20.
95. Schneider J, Classen V, Bernges U, Philipp M. XRCC1 polymorphism and lung cancer risk in relation to tobacco smoking. *Int J Mol Med*. 2005;16:709-16.
96. Hu Z, Ma H, Lu D, Zhou J, Chen Y, Xu L, Zhu J, Huo X, Qian J, Wei Q, Shen H. A promoter polymorphism (-77T>C) of DNA repair gene XRCC1 is associated with risk of lung cancer in relation to tobacco smoking. *Pharmacogenet Genomics*. 2005;15:457-63.
97. Kiyohara C, Yoshimasu K. Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis. *Int J Med Sci*. 2007;4:59-71.
98. Liu J, Liao Q, Zhang Y, Sun S, Zhong C, Liu X. Cyclin D1 G870A polymorphism and lung cancer risk: a meta-analysis. *Tumour Biol*. 2012.
99. Wang W, Spitz MR, Yang H, Lu C, Stewart DJ, Wu X. Genetic variants in cell cycle control pathway confer susceptibility to lung cancer. *Clin Cancer Res*. 2007;13:5974-81.
100. Zhang X, Miao X, Guo Y, Tan W, Zhou Y, Sun T, Wang Y, Lin D. Genetic polymorphisms in cell cycle regulatory genes MDM2 and TP53 are associated with susceptibility to lung cancer. *Hum Mutat*. 2006;27:110-7.
101. Sasaki H, Okuda K, Shimizu S, Takada M, Kawahara M, Kitahara N, Okumura M, Matsumura A, Iuchi K, Kawaguchi T, Kubo A, Kawano O, Yukiue H, Yano M, Fujii

- Y. EGFR R497K polymorphism is a favorable prognostic factor for advanced lung cancer. *J Cancer Res Clin Oncol.* 2009;135:313-8.
102. Zhang W, Stabile LP, Keohavong P, Romkes M, Grandis JR, Traynor AM, Siegfried JM. Mutation and polymorphism in the EGFR-TK domain associated with lung cancer. *J Thorac Oncol.* 2006;1:635-47.
103. Dubey S, Stephenson P, Levy DE, Miller JA, Keller SM, Schiller JH, Johnson DH, Kolesar JM. EGFR dinucleotide repeat polymorphism as a prognostic indicator in non-small cell lung cancer. *J Thorac Oncol.* 2006;1:406-12.
104. Liao WY, Shih JY, Chang GC, Cheng YK, Yang JC, Chen YM, Yu CJ. Genetic Polymorphism of XRCC1 Arg399Gln Is Associated With Survival in Non-Small-Cell Lung Cancer Patients Treated With Gemcitabine/Platinum. *J Thorac Oncol.* 2012.
105. Krawczyk P, Wojas-Krawczyk K, Mlak R, Kucharczyk T, Biernacka B, Milanowski J. Predictive value of ERCC1 single-nucleotide polymorphism in patients receiving platinum-based chemotherapy for locally-advanced and advanced non-small cell lung cancer - a pilot study. *Folia Histochem Cytobiol.* 2012;50:80-6.
106. Wu W, Li H, Wang H, Zhao X, Gao Z, Qiao R, Zhang W, Qian J, Wang J, Chen H, Wei Q, Han B, Lu D. Effect of Polymorphisms in XPD on Clinical Outcomes of Platinum-Based Chemotherapy for Chinese Non-Small Cell Lung Cancer Patients. *PLoS One.* 2012;7:e33200.
107. Cui Z, Yin Z, Li X, Wu W, Guan P, Zhou B. Association between polymorphisms in XRCC1 gene and clinical outcomes of patients with lung cancer: a meta-analysis. *BMC Cancer.* 2012;12:71.

108. Kim MJ, Kang HG, Lee SY, Jeon HS, Lee WK, Park JY, Lee EB, Lee JH, Cha SI, Kim DS, Kim CH, Kam S, Jung TH. AKT1 polymorphisms and survival of early stage non-small cell lung cancer. *J Surg Oncol.* 2012;105:167-74.
109. Pu X, Hildebrandt MA, Lu C, Lin J, Stewart DJ, Ye Y, Gu J, Spitz MR, Wu X. PI3K/PTEN/AKT/mTOR pathway genetic variation predicts toxicity and distant progression in lung cancer patients receiving platinum-based chemotherapy. *Lung Cancer.* 71:82-8.
110. Giovannetti E, Zucali PA, Peters GJ, Cortesi F, D'Incecco A, Smit EF, Falcone A, Burgers JA, Santoro A, Danesi R, Giaccone G, Tibaldi C. Association of polymorphisms in AKT1 and EGFR with clinical outcome and toxicity in non-small cell lung cancer patients treated with gefitinib. *Mol Cancer Ther.* 2010;9:581-93.
111. Campayo M, Navarro A, Vinolas N, Tejero R, Munoz C, Diaz T, Marrades R, Cabanas ML, Gimferrer JM, Gascon P, Ramirez J, Monzo M. A dual role for KRT81: a miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma. *PLoS One.* 2011;6:e22509.
112. Rotunno M, Zhao Y, Bergen AW, Koshiol J, Burdette L, Rubagotti M, Linnoila RI, Marincola FM, Bertazzi PA, Pesatori AC, Caporaso NE, McShane LM, Wang E, Landi MT. Inherited polymorphisms in the RNA-mediated interference machinery affect microRNA expression and lung cancer survival. *Br J Cancer.* 103:1870-4.
113. Hu Z, Shu Y, Chen Y, Chen J, Dong J, Liu Y, Pan S, Xu L, Xu J, Wang Y, Dai J, Ma H, Jin G, Shen H. Genetic polymorphisms in the precursor MicroRNA flanking region and non-small cell lung cancer survival. *Am J Respir Crit Care Med.* 2011;183:641-8.

114. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299:1335-44.
115. Pahl R, Schafer H, Muller HH. Optimal multistage designs--a general framework for efficient genome-wide association studies. *Biostatistics*. 2009;10:297-309.
116. Spinola M, Leoni VP, Galvan A, Korsching E, Conti B, Pastorino U, Ravagnani F, Columbano A, Skaug V, Haugen A, Dragani TA. Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. *Cancer Lett*. 2007;251:311-6.
117. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcatova I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452:633-7.
118. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. Genome-wide association scan of tag

- SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008;40:616-22.
119. Liu P, Vikis HG, Wang D, Lu Y, Wang Y, Schwartz AG, Pinney SM, Yang P, de Andrade M, Petersen GM, Wiest JS, Fain PR, Gazdar A, Gaba C, Rothschild H, Mandal D, Coons T, Lee J, Kupert E, Seminara D, Minna J, Bailey-Wilson JE, Wu X, Spitz MR, Eisen T, Houlston RS, Amos CI, Anderson MW, You M. Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J Natl Cancer Inst.* 2008;100:1326-30.
120. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, Spitz MR, Eisen T, Amos CI, Houlston RS. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet.* 2008;40:1407-9.
121. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, McLaughlin J, Shepherd F, Montpetit A, Narod S, Krokan HE, Skorpen F, Elvestad MB, Vatten L, Njolstad I, Axelsson T, Chen C, Goodman G, Barnett M, Loomis MM, Lubinski J, Matyjasik J, Lener M, Oszutowska D, Field J, Liloglou T, Xinarianos G, Cassidy A, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, Gonzalez CA, Ramon Quiros J, Martinez C, Navarro C, Ardanaz E, Larranaga N, Kham KT, Key T, Bueno-de-Mesquita HB, Peeters PH, Trichopoulou A, Linseisen J, Boeing H, Hallmans G, Overvad K, Tjonneland A, Kumle M, Riboli E, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. Lung cancer susceptibility locus at 5p15.33. *Nat Genet.* 2008;40:1404-6.



122. Broderick P, Wang Y, Vijayakrishnan J, Matakidou A, Spitz MR, Eisen T, Amos CI, Houlston RS. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res.* 2009;69:6633-41.
123. Li Y, Sheu CC, Ye Y, de Andrade M, Wang L, Chang SC, Aubry MC, Aakre JA, Allen MS, Chen F, Cunningham JM, Deschamps C, Jiang R, Lin J, Marks RS, Pankratz VS, Su L, Sun Z, Tang H, Vasmatazis G, Harris CC, Spitz MR, Jen J, Wang R, Zhang ZF, Christiani DC, Wu X, Yang P. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol.* 2010;11:321-30.
124. Wu C, Xu B, Yuan P, Ott J, Guan Y, Liu Y, Liu Z, Shen Y, Yu D, Lin D. Genome-wide examination of genetic variants associated with response to platinum-based chemotherapy in patients with small-cell lung cancer. *Pharmacogenet Genomics.* 2010;20:389-95.
125. Yoon KA, Park JH, Han J, Park S, Lee GK, Han JY, Zo JI, Kim J, Lee JE, Takahashi A, Kubo M, Nakamura Y, Lee JS. A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. *Hum Mol Genet.* 19:4948-54.
126. Sato Y, Yamamoto N, Kunitoh H, Ohe Y, Minami H, Laird NM, Katori N, Saito Y, Ohnami S, Sakamoto H, Sawada JI, Saijo N, Yoshida T, Tamura T. Genome-Wide Association Study on Overall Survival of Advanced Non-small Cell Lung Cancer Patients Treated with Carboplatin and Paclitaxel. *J Thorac Oncol.*
127. Wu C, Xu B, Yuan P, Miao X, Liu Y, Guan Y, Yu D, Xu J, Zhang T, Shen H, Wu T, Lin D. Genome-wide interrogation identifies YAP1 variants associated with survival of small-cell lung cancer patients. *Cancer Res.* 2010;70:9721-9.

128. Wu X, Ye Y, Rosell R, Amos CI, Stewart DJ, Hildebrandt MA, Roth JA, Minna JD, Gu J, Lin J, Buch SC, Nukui T, Ramirez Serrano JL, Taron M, Cassidy A, Lu C, Chang JY, Lippman SM, Hong WK, Spitz MR, Romkes M, Yang P. Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *J Natl Cancer Inst.* 2011;103:817-25.
129. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, Li Z, Deng Q, Wang J, Wu W, Jin G, Jiang Y, Yu D, Zhou G, Chen H, Guan P, Chen Y, Shu Y, Xu L, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Ma H, Chen J, Chu M, Lu F, Zhang Z, Chen F, Wang X, Jin L, Lu J, Zhou B, Lu D, Wu T, Lin D, Shen H. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet.* 2011;43:792-6.
130. Tan XL, Moyer AM, Fridley BL, Schaid D, Niu N, Batzler A, Jenkins GD, Abo R, Li L, Cunningham JM, Sun Z, Yang P, Wang L. Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clin Cancer Res.* 2011.
131. Ahn MJ, Won HH, Lee J, Lee ST, Sun JM, Park YH, Ahn JS, Kwon OJ, Kim H, Shim YM, Kim J, Kim K, Kim YH, Park JY, Kim JW, Park K. The 18p11.22 locus is associated with never smoker non-small cell lung cancer susceptibility in Korean populations. *Hum Genet.* 2012;131:365-72.
132. Ricketts C, Zeegers MP, Lubinski J, Maher ER. Analysis of germline variants in CDH1, IGFBP3, MMP1, MMP3, STK15 and VEGF in familial and sporadic renal cell carcinoma. *PLoS One.* 2009;4:e6037.

133. Wu X, Gu J, Grossman HB, Amos CI, Etzel C, Huang M, Zhang Q, Millikan RE, Lerner S, Dinney CP, Spitz MR. Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes. *Am J Hum Genet.* 2006;78:464-79.
134. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Gronberg H. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med.* 2008;358:910-9.
135. Loza MJ, McCall CE, Li L, Isaacs WB, Xu J, Chang BL. Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS One.* 2007;2:e1035.
136. Bethke L, Murray A, Webb E, Schoemaker M, Muir K, McKinney P, Hepworth S, Dimitropoulou P, Lophatananon A, Feychting M, Lonn S, Ahlbom A, Malmer B, Henriksson R, Auvinen A, Kiuru A, Salminen T, Johansen C, Christensen HC, Kosteljanetz M, Swerdlow A, Houlston R. Comprehensive analysis of DNA repair gene variants and risk of meningioma. *J Natl Cancer Inst.* 2008;100:270-6.
137. Rudd MF, Sellick GS, Webb EL, Catovsky D, Houlston RS. Variants in the ATM-BRCA2-CHEK2 axis predispose to chronic lymphocytic leukemia. *Blood.* 2006;108:638-44.
138. Deng S, Calin GA, Croce CM, Coukos G, Zhang L. Mechanisms of microRNA deregulation in human cancer. *Cell Cycle.* 2008;7:2643-6.
139. Hammond SM. RNAi, microRNAs, and human disease. *Cancer Chemother Pharmacol.* 2006;58 Suppl 1:s63-8.
140. Croce CM. Oncogenes and cancer. *N Engl J Med.* 2008;358:502-11.

141. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM, Harris CC. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*. 2006;9:189-98.
142. Gregory RI, Shiekhattar R. MicroRNA biogenesis and cancer. *Cancer Res*. 2005;65:3509-12.
143. Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nat Rev Cancer*. 2006;6:857-66.
144. Yang H, Dinney CP, Ye Y, Zhu Y, Grossman HB, Wu X. Evaluation of genetic variants in microRNA-related genes and risk of bladder cancer. *Cancer Res*. 2008;68:2530-7.
145. Ye Y, Wang KK, Gu J, Yang H, Lin J, Ajani JA, Wu X. Genetic variations in microRNA-related genes are novel susceptibility loci for esophageal cancer risk. *Cancer Prev Res (Phila)*. 2008;1:460-9.
146. Horikawa Y, Wood CG, Yang H, Zhao H, Ye Y, Gu J, Lin J, Habuchi T, Wu X. Single nucleotide polymorphisms of microRNA machinery genes modify the risk of renal cell carcinoma. *Clin Cancer Res*. 2008;14:7956-62.
147. Winter J, Jung S, Keller S, Gregory RI, Diederichs S. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol*. 2009;11:228-34.
148. Saunders MA, Liang H, Li WH. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A*. 2007;104:3300-5.
149. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. *Trends Genet*. 2008;24:489-97.

150. Ye Y, Wang KK, Gu J, Yang H, Lin J, Ajani JA, Wu X. Genetic variations in microRNA-related genes are novel susceptibility loci for esophageal cancer risk. *Cancer Prev Res (Phila Pa)*. 2008;1:460-9.
151. Liang D, Meyer L, Chang DW, Lin J, Pu X, Ye Y, Gu J, Wu X, Lu K. Genetic variants in MicroRNA biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response. *Cancer Res*.70:9765-76.
152. Saetrom P, Biesinger J, Li SM, Smith D, Thomas LF, Majzoub K, Rivas GE, Alluin J, Rossi JJ, Krontiris TG, Weitzel J, Daly MB, Benson AB, Kirkwood JM, O'Dwyer PJ, Sutphen R, Stewart JA, Johnson D, Larson GP. A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Res*. 2009;69:7459-65.
153. Ruzzo A, Canestrari E, Galluccio N, Santini D, Vincenzi B, Tonini G, Magnani M, Graziano F. Role of KRAS let-7 LCS6 SNP in metastatic colorectal cancer patients. *Ann Oncol*. 2011;22:234-5.
154. Zhang W, Winder T, Ning Y, Pohl A, Yang D, Kahn M, Lurje G, Labonte MJ, Wilson PM, Gordon MA, Hu-Lieskovan S, Mauro DJ, Langer C, Rowinsky EK, Lenz HJ. A let-7 microRNA-binding site polymorphism in 3'-untranslated region of KRAS gene predicts response in wild-type KRAS patients with metastatic colorectal cancer treated with cetuximab monotherapy. *Ann Oncol*. 2011;22:104-9.
155. Graziano F, Canestrari E, Loupakis F, Ruzzo A, Galluccio N, Santini D, Rocchi M, Vincenzi B, Salvatore L, Cremolini C, Spoto C, Catalano V, D'Emidio S, Giordani P, Tonini G, Falcone A, Magnani M. Genetic modulation of the Let-7 microRNA binding to KRAS 3'-untranslated region and survival of metastatic colorectal cancer

- patients treated with salvage cetuximab-irinotecan. *Pharmacogenomics J.* 2010;10:458-64.
156. Christensen BC, Moyer BJ, Avissar M, Ouellet LG, Plaza SL, McClean MD, Marsit CJ, Kelsey KT. A let-7 microRNA-binding site polymorphism in the KRAS 3' UTR is associated with reduced survival in oral cancers. *Carcinogenesis.* 2009;30:1003-7.
157. Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, Muller RU, Straka E, Su L, Burki EA, Crowell RE, Patel R, Kulkarni T, Homer R, Zeltermann D, Kidd KK, Zhu Y, Christiani DC, Belinsky SA, Slack FJ, Weidhaas JB. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. *Cancer Res.* 2008;68:8535-40.
158. Takamizawa J, Konishi H, Yanagisawa K, Tomida S, Osada H, Endoh H, Harano T, Yatabe Y, Nagino M, Nimura Y, Mitsudomi T, Takahashi T. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.* 2004;64:3753-6.
159. Patnaik SK, Kannisto E, Knudsen S, Yendamuri S. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. *Cancer Res.* 2010;70:36-45.
160. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. *Cell.* 140:883-99.
161. Coussens LM, Werb Z. Inflammation and cancer. *Nature.* 2002;420:860-7.
162. Khatami M. Inflammation, aging, and cancer: tumoricidal versus tumorigenesis of immunity: a common denominator mapping chronic diseases. *Cell Biochem Biophys.* 2009;55:55-79.

163. Sorrentino C, Di Carlo E. Expression of IL-32 in human lung cancer is related to the histotype and metastatic phenotype. *Am J Respir Crit Care Med.* 2009;180:769-79.
164. Gocheva V, Wang HW, Gadea BB, Shree T, Hunter KE, Garfall AL, Berman T, Joyce JA. IL-4 induces cathepsin protease activity in tumor-associated macrophages to promote cancer growth and invasion. *Genes Dev.*24:241-55.
165. Zeng L, O'Connor C, Zhang J, Kaplan AM, Cohen DA. IL-10 promotes resistance to apoptosis and metastatic potential in lung tumor cell lines. *Cytokine.*49:294-302.
166. Proctor MJ, Morrison DS, Talwar D, Balmer SM, O'Reilly DS, Foulis AK, Horgan PG, McMillan DC. An inflammation-based prognostic score (mGPS) predicts cancer survival independent of tumour site: a Glasgow Inflammation Outcome Study. *Br J Cancer.*104:726-34.
167. Ishizuka M, Nagata H, Takagi K, Horie T, Kubota K. Inflammation-based prognostic score is a novel predictor of postoperative outcome in patients with colorectal cancer. *Ann Surg.* 2007;246:1047-51.
168. Hara M, Matsuzaki Y, Shimuzu T, Tomita M, Ayabe T, Enomoto Y, Onitsuka T. Preoperative serum C-reactive protein level in non-small cell lung cancer. *Anticancer Res.* 2007;27:3001-4.
169. O'Dowd C, McRae LA, McMillan DC, Kirk A, Milroy R. Elevated preoperative C-reactive protein predicts poor cancer specific survival in patients undergoing resection for non-small cell lung cancer. *J Thorac Oncol.*5:988-92.
170. Achiwa H, Yatabe Y, Hida T, Kuroishi T, Kozaki K, Nakamura S, Ogawa M, Sugiura T, Mitsudomi T, Takahashi T. Prognostic significance of elevated

- cyclooxygenase 2 expression in primary, resected lung adenocarcinomas. *Clin Cancer Res.* 1999;5:1001-5.
171. Mascaux C, Martin B, Paesmans M, Berghmans T, Dusart M, Haller A, Lothaire P, Meert AP, Lafitte JJ, Sculier JP. Has Cox-2 a prognostic role in non-small-cell lung cancer? A systematic review of the literature with meta-analysis of the survival results. *Br J Cancer.* 2006;95:139-45.
172. Minamiya Y, Miura M, Hinai Y, Saito H, Ito M, Imai K, Ono T, Motoyama S, Ogawa J. The CRP 1846T/T genotype is associated with a poor prognosis in patients with non-small cell lung cancer. *Tumour Biol.* 2010;31:673-9.
173. Pine SR, Mechanic LE, Ambis S, Bowman ED, Chanock SJ, Loffredo C, Shields PG, Harris CC. Lung cancer survival and functional polymorphisms in MBL2, an innate-immunity gene. *J Natl Cancer Inst.* 2007;99:1401-9.
174. Asomaning K, Miller DP, Liu G, Wain JC, Lynch TJ, Su L, Christiani DC. Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer.* 2008;61:13-20.
175. Yang P, Allen MS, Aubry MC, Wampfler JA, Marks RS, Edell ES, Thibodeau S, Adjei AA, Jett J, Deschamps C. Clinical features of 5,628 primary lung cancer patients: experience at Mayo Clinic from 1997 to 2003. *Chest.* 2005;128:452-62.
176. Li Y, Sheu CC, Ye Y, de Andrade M, Wang L, Chang SC, Aubry MC, Aakre JA, Allen MS, Chen F, Cunningham JM, Deschamps C, Jiang R, Lin J, Marks RS, Pankratz VS, Su L, Sun Z, Tang H, Vasmatazis G, Harris CC, Spitz MR, Jen J, Wang R, Zhang ZF, Christiani DC, Wu X, Yang P. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol.* 11:321-30.



177. Wu X, Spitz MR, Lee JJ, Lippman SM, Ye Y, Yang H, Khuri FR, Kim E, Gu J, Lotan R, Hong WK. Novel susceptibility loci for second primary tumors/recurrence in head and neck cancer patients: large-scale evaluation of genetic variants. *Cancer Prev Res (Phila Pa)*. 2009;2:617-24.
178. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet*. 2005;37:1217-23.
179. Spitz MR, Gorlov IP, Amos CI, Dong Q, Chen W, Etzel CJ, Gorlova OY, Chang DW, Pu X, Zhang D, Wang L, Cunningham JM, Yang P, Wu X. Variants in Inflammation Genes Are Implicated in Risk of Lung Cancer in Never Smokers Exposed to Second-hand Smoke. *Cancer Discov*. 2011;1:10.
180. Wu X, Ye Y, Rosell R, Amos CI, Stewart DJ, Hildebrandt MA, Roth JA, Minna JD, Gu J, Lin J, Buch SC, Nukui T, Ramirez Serrano JL, Taron M, Cassidy A, Lu C, Chang JY, Lippman SM, Hong WK, Spitz MR, Romkes M, Yang P. Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *J Natl Cancer Inst*. 103:817-25.
181. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003;100:9440-5.
182. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248-9.
183. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7:61-80.

184. Landi D, Gemignani F, Barale R, Landi S. A catalog of polymorphisms falling in microRNA-binding regions of cancer genes. *DNA Cell Biol.* 2008;27:35-43.
185. Mishra PJ, Humeniuk R, Longo-Sorbello GS, Banerjee D, Bertino JR. A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc Natl Acad Sci U S A.* 2007;104:13513-8.
186. Huang CR, Jin ZX, Dong L, Tong XP, Yue S, Kawanami T, Sawaki T, Sakai T, Miki M, Iwao H, Nakajima A, Masaki Y, Fukushima Y, Tanaka M, Fujita Y, Nakajima H, Okazaki T, Umehara H. Cisplatin augments FAS-mediated apoptosis through lipid rafts. *Anticancer Res.*30:2065-71.
187. Katoh M. WNT signaling in stem cell biology and regenerative medicine. *Curr Drug Targets.* 2008;9:565-70.
188. Hein DW. Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. *Mutat Res.* 2002;506-507:65-77.
189. Ward A, Summers MJ, Sim E. Purification of recombinant human N-acetyltransferase type 1 (NAT1) expressed in *E. coli* and characterization of its potential role in folate metabolism. *Biochem Pharmacol.* 1995;49:1759-67.
190. Kathiresan S, Melander O, Anevski D, Guiducci C, Burt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altshuler DM, Newton-Cheh C, Orholm Melander M. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med.* 2008;358:1240-9.
191. Hildebrandt MA, Gu J, Wu X. Pharmacogenomics of platinum-based chemotherapy in NSCLC. *Expert Opin Drug Metab Toxicol.* 2009;5:745-55.

192. Provencio M, de Las Penas R, Camps C, Artal A, Massuti B, Cobo M, Perez FJ, Sanchez A, Rosell R. Cisplatin plus vinorelbine as first-line treatment for advanced non-small-cell lung cancer: Is a hemogram on day 8 essential? *Lung Cancer*. 2010;68:415-9.
193. Gioulbasanis I, Patrikidou A, Kitikidou K, Papadimitriou K, Vlachostergios PJ, Tsatsanis C, Margioris AN, Papandreou CN, Mavroudis D, Georgoulas V. Baseline Plasma Levels of Interleukin-8 in Stage IV Non-Small-Cell Lung Cancer Patients: Relationship With Nutritional Status and Prognosis. *Nutr Cancer*. 2011.
194. Vlachostergios PJ, Gioulbasanis I, Kamposioras K, Georgoulas P, Baracos VE, Ghosh S, Maragouli E, Georgoulas V, Papandreou CN. Baseline insulin-like growth factor-I plasma levels, systemic inflammation, weight loss and clinical outcome in metastatic non-small cell lung cancer patients. *Oncology*. 2011;81:113-8.
195. Seliger B. Different regulation of MHC class I antigen processing components in human tumors. *J Immunotoxicol*. 2008;5:361-7.
196. Chamuleau ME, Ossenkoppele GJ, van de Loosdrecht AA. MHC class II molecules in tumour immunology: prognostic marker and target for immune modulation. *Immunobiology*. 2006;211:619-25.
197. Souwer Y, Chamuleau ME, van de Loosdrecht AA, Tolosa E, Jorritsma T, Muris JJ, Dinnissen-van Poppel MJ, Snel SN, van de Corput L, Ossenkoppele GJ, Meijer CJ, Neefjes JJ, Marieke van Ham S. Detection of aberrant transcription of major histocompatibility complex class II antigen presentation genes in chronic lymphocytic leukaemia identifies HLA-DOA mRNA as a prognostic factor for survival. *Br J Haematol*. 2009;145:334-43.

198. Buddingh EP, Schilham MW, Ruslan SE, Berghuis D, Szuhai K, Suurmond J, Taminiou AH, Gelderblom H, Egeler RM, Serra M, Hogendoorn PC, Lankester AC. Chemotherapy-resistant osteosarcoma is highly susceptible to IL-15-activated allogeneic and autologous NK cells. *Cancer Immunol Immunother.* 2011;60:575-86.
199. Le Maux Chansac B, Misse D, Richon C, Vergnon I, Kubin M, Soria JC, Moretta A, Chouaib S, Mami-Chouaib F. Potentiation of NK cell-mediated cytotoxicity in human lung adenocarcinoma: role of NKG2D-dependent pathway. *Int Immunol.* 2008;20:801-10.
200. Zhang T, Sentman CL. Cancer immunotherapy using a bispecific NK receptor fusion protein that engages both T cells and tumor cells. *Cancer Res.* 2011;71:2066-76.
201. Park MJ, Bae JH, Chung JS, Kim SH, Kang CD. Induction of NKG2D ligands and increased sensitivity of tumor cells to NK cell-mediated cytotoxicity by hematoporphyrin-based photodynamic therapy. *Immunol Invest.* 2011;40:367-82.
202. Kolls JK, Linden A. Interleukin-17 family members and inflammation. *Immunity.* 2004;21:467-76.
203. Zhou Y, Toh ML, Zrioual S, Miossec P. IL-17A versus IL-17F induced intracellular signal transduction pathways and modulation by IL-17RA and IL-17RC RNA interference in AGS gastric adenocarcinoma cells. *Cytokine.* 2007;38:157-64.
204. Katoh Y, Katoh M. Comparative integromics on BMP/GDF family. *Int J Mol Med.* 2006;17:951-5.
205. Ye L, Bokobza SM, Jiang WG. Bone morphogenetic proteins in development and progression of breast cancer and therapeutic potential (review). *Int J Mol Med.* 2009;24:591-7.

206. Singh A, Morris RJ. The Yin and Yang of bone morphogenetic proteins in cancer. *Cytokine Growth Factor Rev.* 21:299-313.
207. Corey E, Vessella RL. Bone morphogenetic proteins and prostate cancer: evolving complexities. *J Urol.* 2007;178:750-1.
208. Ye L, Lewis-Russell JM, Kyanaston HG, Jiang WG. Bone morphogenetic proteins and their receptor signaling in prostate cancer. *Histol Histopathol.* 2007;22:1129-47.
209. Chen H, Cowan MJ, Hasday JD, Vogel SN, Medvedev AE. Tobacco smoking inhibits expression of proinflammatory cytokines and activation of IL-1R-associated kinase, p38, and NF-kappaB in alveolar macrophages stimulated with TLR2 and TLR4 agonists. *J Immunol.* 2007;179:6097-106.
210. Medves S, Demoulin JB. Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies. *J Cell Mol Med.* 2011.
211. Ma L, Dong S, Zhang P, Xu N, Yan H, Liu H, Li Y, Zhou Q. The relationship between methylation of the Syk gene in the promoter region and the genesis of lung cancer. *Clin Lab.* 2010;56:407-16.
212. Nagata S, Jin YF, Yoshizato K, Tomoeda M, Song M, Iizuka N, Kitamura M, Takahashi H, Eguchi H, Ohigashi H, Ishikawa O, Tomita Y. CD74 is a novel prognostic factor for patients with pancreatic cancer receiving multimodal therapy. *Ann Surg Oncol.* 2009;16:2531-8.
213. Borghese F, Clanchy FI. CD74: an emerging opportunity as a therapeutic target in cancer and autoimmune disease. *Expert Opin Ther Targets.* 2011;15:237-51.

214. Stein R, Mattes MJ, Cardillo TM, Hansen HJ, Chang CH, Burton J, Govindan S, Goldenberg DM. CD74: a new candidate target for the immunotherapy of B-cell neoplasms. *Clin Cancer Res.* 2007;13:5556s-63s.
215. Burton JD, Ely S, Reddy PK, Stein R, Gold DV, Cardillo TM, Goldenberg DM. CD74 is expressed by multiple myeloma and is a promising target for therapy. *Clin Cancer Res.* 2004;10:6606-11.
216. Malavasi F, Deaglio S, Funaro A, Ferrero E, Horenstein AL, Ortolan E, Vaisitti T, Aydin S. Evolution and function of the ADP ribosyl cyclase/CD38 gene family in physiology and pathology. *Physiol Rev.* 2008;88:841-86.
217. Mougalian SS, O'Brien S. Adverse prognostic features in chronic lymphocytic leukemia. *Oncology (Williston Park).* 2011;25:692-6, 9.

## Appendix A: Other Peer-reviewed Publications during Ph.D. Study

1. Wu X, Scelo G, Purdue MP, Rothman N, Johansson M, Ye Y, Wang Z, Zelenika D, Moore LE, Wood CG, Prokhorov E, Gaborieau V, Jacobs KB, Chow WH, Toro JR, Zaridze D, Lin J, Lubinski J, Trubicka J, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Jinga V, Bencko V, Slamova A, Holcatova I, Navratilova M, Janout V, Boffetta P, Colt JS, Davis FG, Schwartz KL, Banks RE, Selby PJ, Harnden P, Berg CD, Hsing AW, Grubb RL, 3rd, Boeing H, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, Duell EJ, Quiros JR, Sanchez MJ, Navarro C, Ardanaz E, Dorronsoro M, Khaw KT, Allen NE, Bueno-de-Mesquita HB, Peeters PH, Trichopoulos D, Linseisen J, Ljungberg B, Overvad K, Tjonneland A, Romieu I, Riboli E, Stevens VL, Thun MJ, Diver WR, Gapstur SM, Pharoah PD, Easton DF, Albanes D, Virtamo J, Vatten L, Hveem K, Fletcher T, Koppova K, Cussenot O, Cancel-Tassin G, Benhamou S, Hildebrandt MA, **Pu X**, Foglio M, Lechner D, Hutchinson A, Yeager M, Fraumeni JF, Jr., Lathrop M, Skryabin KG, McKay JD, Gu J, Brennan P, Chanock SJ. A genome-wide association study identifies a novel susceptibility locus for renal cell carcinoma on 12p11.23. *Hum Mol Genet.* 2011.
2. Wilkinson AV, Bondy ML, Wu X, Wang J, Dong Q, D'Amelio AM, Jr., Prokhorov AV, **Pu X**, Yu RK, Etzel CJ, Shete S, Spitz MR. Cigarette Experimentation in Mexican Origin Youth: Psychosocial and Genetic Determinants. *Cancer Epidemiol Biomarkers Prev.* 2011.
3. Tan W, Hildebrandt MA, **Pu X**, Huang M, Lin J, Matin SF, Tamboli P, Wood CG, Wu X. Role of inflammatory related gene expression in clear cell renal cell carcinoma development and clinical outcomes. *J Urol.* 2011;186:2071-7.

4. Spitz MR, Gorlov IP, Amos CI, Dong Q, Chen W, Etzel CJ, Gorlova OY, Chang DW, **Pu X**, Zhang D, Wang L, Cunningham JM, Yang P, Wu X. Variants in Inflammation Genes Are Implicated in Risk of Lung Cancer in Never Smokers Exposed to Second-hand Smoke. *Cancer Discov.* 2011;1:10.
5. Purdue MP, Johansson M, Zelenika D, Toro JR, Scelo G, Moore LE, Prokhortchouk E, Wu X, Kiemeny LA, Gaborieau V, Jacobs KB, Chow WH, Zaridze D, Matveev V, Lubinski J, Trubicka J, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Bucur A, Bencko V, Foretova L, Janout V, Boffetta P, Colt JS, Davis FG, Schwartz KL, Banks RE, Selby PJ, Harnden P, Berg CD, Hsing AW, Grubb RL, 3rd, Boeing H, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, Duell EJ, Quiros JR, Sanchez MJ, Navarro C, Ardanaz E, Dorronsoro M, Khaw KT, Allen NE, Bueno-de-Mesquita HB, Peeters PH, Trichopoulos D, Linseisen J, Ljungberg B, Overvad K, Tjonneland A, Romieu I, Riboli E, Mukeria A, Shangina O, Stevens VL, Thun MJ, Diver WR, Gapstur SM, Pharoah PD, Easton DF, Albanes D, Weinstein SJ, Virtamo J, Vatten L, Hveem K, Njolstad I, Tell GS, Stoltenberg C, Kumar R, Koppova K, Cussenot O, Benhamou S, Oosterwijk E, Vermeulen SH, Aben KK, van der Marel SL, Ye Y, Wood CG, **Pu X**, Mazur AM, Boulygina ES, Chekanov NN, Foglio M, Lechner D, Gut I, Heath S, Blanche H, Hutchinson A, Thomas G, Wang Z, Yeager M, Fraumeni JF, Jr., Skryabin KG, McKay JD, Rothman N, Chanock SJ, Lathrop M, Brennan P. Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nat Genet.* 2011;43:60-5.
6. **Pu X**, Hildebrandt MA, Lu C, Lin J, Stewart DJ, Ye Y, Gu J, Spitz MR, Wu X. PI3K/PTEN/AKT/mTOR pathway genetic variation predicts toxicity and distant progression in lung cancer patients receiving platinum-based chemotherapy. *Lung Cancer.* 2011;71:82-8.



7. Garcia-Closas M, Ye Y, Rothman N, Figueroa JD, Malats N, Dinney CP, Chatterjee N, Prokunina-Olsson L, Wang Z, Lin J, Real FX, Jacobs KB, Baris D, Thun M, De Vivo I, Albanes D, Purdue MP, Kogevinas M, Kamat AM, Lerner SP, Grossman HB, Gu J, **Pu X**, Hutchinson A, Fu YP, Burdett L, Yeager M, Tang W, Tardon A, Serra C, Carrato A, Garcia-Closas R, Lloreta J, Johnson A, Schwenn M, Karagas MR, Schned A, Andriole G, Jr., Grubb R, 3rd, Black A, Jacobs EJ, Diver WR, Gapstur SM, Weinstein SJ, Virtamo J, Hunter DJ, Caporaso N, Landi MT, Fraumeni JF, Jr., Silverman DT, Chanock SJ, Wu X. A genome-wide association study of bladder cancer identifies a new susceptibility locus within SLC14A1, a urea transporter gene on chromosome 18q12.3. *Hum Mol Genet.* 2011;20:4282-9.
8. Liang D, Meyer L, Chang DW, Lin J, **Pu X**, Ye Y, Gu J, Wu X, Lu K. Genetic variants in MicroRNA biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response. *Cancer Res.* 2010;70:9765-76.
9. **Pu X**, Lippman SM, Yang H, Lee JJ, Wu X. Cyclooxygenase-2 gene polymorphisms reduce the risk of oral premalignant lesions. *Cancer.* 2009;115:1498-506.
10. Lin J, **Pu X**, Wang W, Matin S, Tannir NM, Wood CG, Wu X. Case-control analysis of nucleotide excision repair pathway and the risk of renal cell carcinoma. *Carcinogenesis.* 2008;29:2112-9.

## Appendix B: supplementary table 1 miRNA related SNPs selected

Function	Gene	SNP	chromosome	Major allele	Minor allele
binding	ACVR1B	rs2854464	12	A	G
binding	ADH5	rs7669660	4	A	G
binding	ADH6	rs12507078	4	G	A
binding	ALDH18A1	rs4037	10	G	A
binding	ANGPT4	rs1888087	20	C	A
binding	ANGPTL1	rs10913632	1	A	G
binding	ARNTL	rs17452383	11	A	G
binding	ATG4A	rs5973822	X	A	G
binding	ATG9A	rs2276635	2	A	G
binding	ATP5A1	rs12954944	18	A	G
binding	ATP5L	rs3194726	11	A	G
binding	ATP6V1C1	rs2248718	8	G	A
binding	ATP6V1C1	rs2453994	8	G	A
binding	ATP6V1C1	rs4734684	8	G	A
binding	BAG1	rs542912	9	C	G
binding	BAG3	rs8946	10	G	C
binding	BAG5	rs7154948	14	G	A
binding	BAX	rs4645900	19	G	A
binding	BCL2L11	rs724710	2	G	A
binding	BCL2L2	rs1884056	14	G	A
binding	BIRC4	rs17330637	X	A	C
binding	BIRC5	rs1042489	17	A	G
binding	BIRC5	rs2071214	17	A	G
binding	BIRC6	rs2710625	2	G	A
binding	BMF	rs10518679	15	A	G
binding	BNIP3L	rs1042992	8	G	A
binding	CA9	rs17259350	9	G	A
binding	CASP2	rs4647342	7	A	T
binding	CASP7	rs10787498	10	A	C
binding	CASP7	rs1127687	10	G	A
binding	CASP8	rs10931936	2	G	A
binding	CAV1	rs8713	7	A	C
binding	CAV1	rs9920	7	A	G
binding	CAV2	rs10278782	7	A	G
binding	CD34	rs7572	1	G	A
binding	CD4	rs1045261	12	A	G
binding	CD44	rs11821102	11	G	A
binding	CDC7	rs12125947	1	A	G
binding	CDK4	rs1048691	12	G	A

binding	CDKN1B	rs7330	12	A	C
binding	CDKN2A	rs3088440	9	G	A
binding	CDKN2C	rs12855	1	G	A
binding	COL18A1	rs7499	21	G	A
binding	COX4NB	rs8587	16	C	A
binding	DDB2	rs1050244	11	G	A
binding	DNMT3B	rs6058896	20	G	A
binding	E2F2	rs2075993	1	A	G
binding	E2F7	rs2279575	12	G	A
binding	E2F7	rs7958377	12	G	A
binding	EIF2C1	rs11263830	1	G	A
binding	EIF2C1	rs11584005	1	A	G
binding	EIF2C1	rs2057606	1	G	C
binding	EIF2C1	rs595055	1	A	G
binding	EIF2C1	rs617673	1	A	C
binding	EPHX2	rs1042032	8	A	G
binding	EPHX2	rs1042064	8	A	G
binding	ERN1	rs8078549	17	G	A
binding	EZH1	rs7214055	17	C	G
binding	FANCD2	rs3172417	3	G	A
binding	FAS	rs2234978	10	G	A
binding	FEN1	rs174546	11	G	A
binding	FEN1	rs4246215	11	C	A
binding	FGF2	rs1048201	4	G	A
binding	FGF2	rs1476215	4	T	A
binding	FGF2	rs6854081	4	A	C
binding	FGF5	rs3733336	4	A	G
binding	FGF5	rs4690150	4	C	G
binding	FGF5	rs6838203	4	T	A
binding	FGF9	rs546782	13	A	T
binding	FLJ35220	rs8065843	17	A	C
binding	FLJ38991	rs16849151	4	A	C
binding	FOXO1A	rs9532558	13	A	G
binding	FZD3	rs352222	8	C	A
binding	FZD4	rs713065	11	G	A
binding	GHITM	rs7576	10	A	C
binding	GHRHR	rs2741	7	A	C
binding	GPR30	rs1133043	7	C	G
binding	GPX3	rs4661	5	G	A
binding	GPX7	rs1047635	1	A	C
binding	GSTM3	rs15864	1	G	C
binding	GSTM5	rs17024661	1	A	G
binding	HSPB8	rs1133026	12	G	A

binding	ICAM1	rs281437	19	G	A
binding	IGF2AS	rs10770125	11	A	G
binding	IGF2BP1	rs11655950	17	G	A
binding	IGF2BP1	rs2969	17	G	A
binding	IGF2BP1	rs6504593	17	G	A
binding	IGFBP2	rs6413492	2	T	A
binding	IGFBP5	rs3276	2	G	A
binding	IL1R1	rs3917328	2	G	A
binding	IL1R1	rs3917329	2	C	A
binding	KRAS	rs10771184	12	T	A
binding	MBD1	rs11663629	18	A	C
binding	MDM4	rs10900596	1	G	A
binding	MDM4	rs4252745	1	C	G
binding	MLL	rs573971	11	G	A
binding	MTHFR	rs10779765	1	G	A
binding	MTR	rs2853523	1	C	A
binding	NAT1	rs15561	8	C	A
binding	NAT1	rs4986993	8	C	A
binding	NDUFA6	rs7245	22	A	G
binding	NEIL2	rs1043180	8	G	A
binding	NEIL2	rs7015453	8	G	A
binding	NFKBIB	rs3136642	19	A	G
binding	NODAL	rs7909303	10	A	C
binding	NOTCH1	rs3124591	9	A	G
binding	NQO1	rs11641233	16	G	A
binding	NQO1	rs9980	16	C	G
binding	NR1I2	rs3732360	3	A	G
binding	NR1I2	rs3814058	3	T	C
binding	OGG1	rs1052133	3	C	G
binding	PDGFC	rs1425486	4	G	A
binding	PGRMC2	rs4016	4	A	T
binding	PLK1	rs7588	16	G	A
binding	PMS2L3	rs1167829	7	G	A
binding	POLH	rs6941583	6	A	T
binding	POLH	rs9333555	6	A	G
binding	PON1	rs854552	7	A	G
binding	RAD51L3	rs4796033	17	G	A
binding	RET	rs2075912	10	G	A
binding	RICTOR	rs443039	5	A	C
binding	RING1	rs107822	6	G	A
binding	RING1	rs213210	6	A	G
binding	RPA1	rs1131636	17	A	G
binding	RPA1	rs5030740	17	A	G

binding	RPS6KA3	rs12010722	X	G	A
binding	RPS6KA3	rs7051161	X	T	A
binding	RPS6KB1	rs1051424	17	A	G
binding	RPS6KB2	rs10274	11	G	A
binding	RRM1	rs1042927	11	A	C
binding	RRM2B	rs16869269	8	A	G
binding	RRM2B	rs5005121	8	T	A
binding	RXRA	rs4842194	9	A	G
binding	SETD1A	rs11076	16	G	A
binding	SIRT3	rs12226697	11	G	A
binding	SMAD1	rs6537355	4	A	G
binding	SMAD3	rs12900401	15	G	A
binding	SMAD3	rs3743342	15	G	A
binding	SMAD7	rs16950113	18	A	G
binding	SMC1L2	rs3747238	22	G	A
binding	SMC1L2	rs3747240	22	A	G
binding	SMO	rs1061280	7	A	G
binding	SMO	rs1061285	7	C	A
binding	SNAI1	rs1047920	20	G	A
binding	SP1	rs17695156	12	G	A
binding	SPP1	rs1126772	4	A	G
binding	SST	rs4988514	3	A	G
binding	SSTR1	rs12889916	14	A	G
binding	SSTR2	rs7210080	17	A	G
binding	SUFU	rs11594179	10	G	A
binding	SULT1C1	rs1047312	2	G	A
binding	SULT4A1	rs138056	22	C	A
binding	TLR2	rs7695605	4	G	C
binding	TLR4	rs7869402	9	G	A
binding	TNFRSF10D	rs7957	8	A	G
binding	TNFRSF21	rs9473029	6	C	G
binding	TNFSF10	rs17600346	3	A	G
binding	TSC1	rs2073869	9	G	A
binding	TXN2	rs139999	22	C	A
binding	UGT2A3	rs17147016	4	T	A
binding	UGT3A2	rs10472999	5	G	A
binding	VDR	rs739837	12	A	C
binding	VEGF	rs3025039	6	G	A
binding	VEGF	rs3025040	6	G	A
binding	WNT11	rs17749202	11	A	G
binding	WNT2B	rs2273368	1	G	A
binding	WNT2B	rs3790611	1	A	G
binding	XRCC5	rs1051685	2	A	G

processing	DDX20	rs197377	1	G	A
processing	DDX20	rs197383	1	A	G
processing	DDX20	rs197412	1	A	G
processing	DDX20	rs563002	1	A	G
processing	DDX20	rs85276	1	A	G
processing	DGCR8	rs11089328	22	A	G
processing	DGCR8	rs1558496	22	A	G
processing	DGCR8	rs1633445	22	A	G
processing	DGCR8	rs1640299	22	A	C
processing	DGCR8	rs2073778	22	G	A
processing	DGCR8	rs2286928	22	G	A
processing	DGCR8	rs3757	22	G	A
processing	DGCR8	rs417309	22	G	A
processing	DGCR8	rs446059	22	G	A
processing	DGCR8	rs720012	22	G	A
processing	DGCR8	rs720014	22	A	G
processing	DGCR8	rs8139591	22	A	G
processing	DGCR8	rs9606248	22	A	G
processing	DICER1	rs10149095	14	A	G
processing	DICER1	rs1057035	14	A	G
processing	DICER1	rs11160231	14	A	C
processing	DICER1	rs11624081	14	G	A
processing	DICER1	rs1187642	14	G	A
processing	DICER1	rs1187652	14	A	G
processing	DICER1	rs12881840	14	G	A
processing	DICER1	rs17784006	14	A	C
processing	DICER1	rs3742330	14	A	G
processing	DICER1	rs4905275	14	G	A
processing	DICER1	rs8006416	14	G	A
processing	GEMIN4	rs1062923	17	A	G
processing	GEMIN4	rs2291778	17	C	A
processing	GEMIN4	rs2740349	17	A	G
processing	GEMIN4	rs2740351	17	A	G
processing	GEMIN4	rs3087833	17	G	A
processing	GEMIN4	rs3744741	17	G	A
processing	GEMIN4	rs7813	17	A	G
processing	RAN	rs10773831	12	G	C
processing	RAN	rs10848238	12	T	A
processing	RAN	rs11061209	12	G	A
processing	RAN	rs12318549	12	G	A
processing	RAN	rs872396	12	A	G
processing	RNASEN	rs10035440	5	A	G
processing	RNASEN	rs10719	5	G	A

processing	RNASN	rs10805564	5	A	G
processing	RNASN	rs11958935	5	A	G
processing	RNASN	rs12186785	5	A	G
processing	RNASN	rs13183642	5	C	A
processing	RNASN	rs16901165	5	G	A
processing	RNASN	rs17408716	5	A	G
processing	RNASN	rs17410035	5	C	A
processing	RNASN	rs2287584	5	A	G
processing	RNASN	rs2302905	5	G	A
processing	RNASN	rs3095825	5	A	G
processing	RNASN	rs3792830	5	A	G
processing	RNASN	rs3805500	5	A	G
processing	RNASN	rs3805502	5	A	T
processing	RNASN	rs3805525	5	A	G
processing	RNASN	rs4867329	5	C	A
processing	RNASN	rs502267	5	C	A
processing	RNASN	rs573010	5	C	A
processing	RNASN	rs639174	5	G	A
processing	RNASN	rs669702	5	G	A
processing	RNASN	rs673019	5	A	G
processing	RNASN	rs6884823	5	G	A
processing	RNASN	rs6886834	5	G	A
processing	RNASN	rs7712155	5	G	A
processing	RNASN	rs7719666	5	G	A
processing	RNASN	rs7735863	5	G	A
processing	XPO5	rs1106841	6	A	C
processing	XPO5	rs17287964	6	A	G
processing	XPO5	rs2227301	6	G	A
processing	XPO5	rs2257082	6	G	A

## VITA

Xia Pu, the daughter of Xiaowei Pu and Shuhui Xia, was born in Nanjing, Jiangsu, People's Republic of China in 1982. In August 2000, she entered China Pharmaceutical University in Nanjing, and received her Bachelor of Science degree in 2004. After three years as a graduate student in the department of analytic chemistry, she was granted Master of Science degree in June 2007. In the same year, she was enrolled in the Ph.D. program at The University of Texas Graduate School of Biomedical Science at Houston. Under the supervision of Dr. Xifeng Wu, She did her dissertation work in the Department of Epidemiology, The University of Texas MD Anderson Cancer Center.