SINGLE NUCLEOTIDE POLYMORPHISMS IN CANCER RELATED GENES LEAD TO INTER-INDIVIDUAL RESPONSE TO PROGNOSIS, DISEASE RISK AND ENVIRONMENTAL AGENT METABOLISM IN MULTIPLE MYELOMA AND LUNG CANCER

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

MAJDA HAZNADAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BRIAN G VAN NESS

August, 2010

ACKNOWLEDGMENTS

I first came to the Van Ness Lab in February of 2006 as a first year PhD student after my undergraduate education at St. Olaf College. Even though I had already lived in the United States for five years by that time, I still felt like a young girl from Sarajevo, Bosnia and Herzegovina, who had just come to a foreign country in search of a better education and better opportunities. I had much to learn and was eager to begin my journey. My five-year education and experiences in the Van Ness Lab are now nearing an end. Looking back, I am in awe at how much I I've learned, how many priceless skills I've gained and how many wonderful and supportive people I've met. Brian has been a tremendously supportive mentor who gave me unlimited support and freedom to develop and apply my intellectual abilities and creativity. Brian, thank you for the opportunities to become involved in so many collaborative projects—I've gained a great amount intellectually from all of those collaborations. I am extremely thankful and lucky for having such a remarkable mentor. I am enormously thankful for such an outstanding Thesis Committee—your guidance has been essential. I would like to thank Christine Ramos for having been a supportive lab mate and someone who taught me a great deal. Christine, thank you for your advice and countless rewarding discussions. I would also like to thank Dr. Tim Church and Dr. Steve Hecht for their support and mentorship while developing my lung cancer project. Dr. Vipin Kumar's group, thank you for your collaboration and support. There have been many others who have helped me along the way and I am thankful to all of them. Van Ness Lab members past and present, thank you for making my journey as a PhD student unforgettable and rewarding. I would also like to thank my family and friends for having been there, continuously supporting me. When I look back on the past five years, I have not only grown as a scientist, but also as a person. I am enormously appreciative of everyone who has helped me along my journey and supported me in every way possible.

ABSTRACT

Multiple myeloma is a chronic disease for which there is presently no cure. Because of the significant genetic heterogeneity in this disease and the fact that it is rare, it has been difficult to study genetic variations that contribute to disease risk and clinical outcomes. Nevertheless, there are apoptotic and oncogenic signaling pathways that constitute common themes in genetic deregulation leading to myeloma. Therefore, we biologically guided single nucleotide (SNP) association studies by pre-identifying important pathways. We managed this by designing a targeted SNP chip panel containing genes in functional groups crucial to various cancer processes, especially to myeloma, and applied it to SNP association studies represented in this thesis.

Lung cancer is one of the most common cancers in the world. It is a leading cause of cancer death in men and women in the United States. Cigarette smoking causes most lung cancers. Therefore, it is one of the rare diseases for which there is a known environmental exposure. The occurrence of lung cancer is thus attributed to a complex interplay of genetic factors and environmental exposure. Since many polymorphic genetic variations produce proteins with increased, decreased or a complete loss of enzymatic activity, they are relevant factors in the gene—environment interplay.

The first part of this thesis explores impact of interindividual variations resulting in variable bone disease, prognosis (progression-free survival) and disease risk in myeloma. In the bone disease association study, we demonstrated that there are genetic variants in genes important in the inflammatory response, Wnt signaling, and in growth factors previously linked to etiology of myeloma. The novelty of this study is in combining gene expression profile of DKK1 with a SNP profile that resulted in a better prediction of bone disease. We then investigated genetic variants in relation to progression-free survival and risk in myeloma. This study resulted in the identification of polymorphisms in genes involved in drug metabolism and detoxification, immunity, DNA repair and signaling cascades important to multiple myeloma (MM) risk and survival. This was done by using novel combinatorial search algorithms that can robustly identify markers that associate with the studied outcomes, and decrease false discovery rates.

The second part of this thesis explores impact of interindividual variations on the metabolism of tobacco-smoke carcinogens. Variations in genes involved in tobacco-smoke carcinogen metabolism can result in variable amounts of harmful DNA adducts that can ultimately lead to cancer. We first demonstrated that there is a variation in CYP1B1 gene, previously shown to result in decreased cellular protein levels. This polymorphism appears to function as a protection from lung cancer at low levels of exposure, whereas it loses its protectiveness at high exposure levels. This was important to unraveling gene—environment interactions, especially relevant in the initiation of lung cancer. We then demonstrated that there are SNP-SNP interactions that associate with lung cancer risk by applying a novel data mining combinatorial search algorithm. This approach will serve as a useful tool to more robustly study SNP associations with outcomes of interest, while minimizing the false discovery rate. Our findings will help aid in further understanding etiology of myeloma and lung cancer. The novel computational method we helped to develop and first applied will serve as an important tool in further identifying and validating genetic variability that leads to differential response to disease risk and outcomes.

TABLE OF CONTENTS Page

Page

Appendix

LIST OF FIGURES Page

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

APPENDIX I

APPENDIX II

APPENDIX III

LIST OF TABLES Page

CHAPTER 2

CHAPTER 3

CHAPTER 4

Page

CHAPTER 5

APPPENDIX I

APPENDIX II

APPENDIX III

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Multiple myeloma:

Multiple myeloma (MM) is a plasma cell malignancy characterized by accumulation of clonal plasma cells, primarily in the bone marrow. MM leads to one or more clinical complications such as bone destruction, anemia, hypercalcemia, and renal insufficiency. It is the second most frequent malignancy of the blood in the USA after non-Hodgkin lymphoma. MM causes about 1% of neoplastic diseases and 13% of hematological malignancies. It affects over 20,000 patients each year in the United States, with about 11,000 deaths annually (1).

The median age at diagnosis is about 62 years for men and 61 years for women (range 20-92); however, only 2% of patients are younger than 40 years. The incidence varies from 1 in 100,000 people in China, to about 4 in 100,000 people in most developed countries. The occurrence of the disease is more prevalent in men than women and is twice as high in black than in white American people (1, 2). The reason for the uneven race and sex distribution has not been uncovered, but may be due to genetic variability, environmental influences, or both. Median survival after conventional treatments is 3-4 years. High dose treatment followed by autologous stem-cell transplantation can extend median survival to 5-7 years (3). Novel drugs used for treating MM are constantly being assessed, used alone and in combination with already existing treatments, with a goal to further improve survival (4, 5).

Direct interaction between MM cells and bone marrow stromal cells, or between extracellular matrix proteins of MM cells, are mediated through cell surface receptors, i.e. integrins, cadherins, selectins, syndecansm and the immunoglobulin superfamily of cell adhesion molecules (6). These types of interactions increase growth, survival, migration, and drug resistance of MM cells, and modulate functions of bone marrow stromal cells, such as by enhancing cytokine secretion. Adhesion of MM cells to extracellular matrix proteins (collagen, fibronectin, laminin) triggers survival and drug resistance (7), as well as production and secretion of urokinase-type plasminogen activator (PLAUR), metalloproteinase-2 (MMP2), and metalloproteinase-9 (MMP9) (8). The cells in the bone marrow microenvironment, including MM cells, are regulated by autocrine and paracrine loops, and they produce and secrete cytokines and growth factors such as interleukin 6 (IL6), vascular endothelial growth factor (VEGF), insulin-like growth factor 1 (IGF1), members of the TNF-superfamily, transforming growth factor, beta 1 (TGFB1), chemokine (C-C motif) ligand 3 (CCL3), KIT ligand (KITLG or SCF), hepatocyte growth factor (HGF), and IL10 (9, 10).

In the cellular bone marrow, MM cells interact with hemopoietic and non-hemopoietic cells, which results in immune suppression and lytic bone lesions. Bone marrow stromal cells in turn send signals (through cell-cell contact or through secretion of soluble factors), which then affects MM cell growth, survival, migration and drug resistance. Bone disease is a very serious consequence of MM. Osteolytic bone lesions and osteoporosis are diagnosed by metastatic bone surveys, CT, fluorodeoxyglucose-PET, and MRI in more than 90% of MM patients. These complications are associated with bone pain and fractures, and are caused by increased osteoclast formation and activity and reduced numbers of osteoblasts (11). Understanding the mechanisms by which myeloma cells impair osteoblastogenesis will help in development of effective interventions for MM bone disease.

Patients with monoclonal gammopathy of undetermined significance (MGUS), a benign disease with an annual conversion rate of 1% to over myeloma, do not have typical osteolytic lesions seen in patients with MM (12). Bataille *et. al*. first reported that increased bone osteoblast activity coupled with increased bone resorption rate is an early event in conversion from MGUS to MM (13). They also found that patients who maintained high osteoblast activity did not develop osteolytic disease (13). The concept of MM cells actively suppressing osteoblastogenesis through production of soluble factors was first suggested 20 years ago by Evans et.al. (14). Additional studies since have supported this. It is now evident that osteoblast differentiation is inhibited by factors secreted by both MM cells (i.e. Wnt-signaling inhibitors DKK1, SFRP2, IL7, HGF) and by microenvironmental cells within myelomatous bone (i.e. IL3) (15-20).

Canonical Wnt ligands promote osteoblast differentiation by triggering phosphorylation of the GSK3/Axin complex and subsequent prevention of β –catenin degradation by the proteosome (21). Canonical Wnt signaling appears to be inhibited in the MM microenvironment mainly through production of DKK1 by myeloma cells; DKK1 inhibits Wnt signaling by binding to Wnt receptors LRP5/6 and Kermen (22). Groups have shown that circulating DKK1 levels are highly correlated with degree of bone loss in patients with MM (14, 22) and MGUS (23). MM patients with high levels of DKK1 in sera effectively suppress Wnt3a-induced β –catenin stabilization in osteoblast precursors and prevent their differentiation (24). DKK1 inhibition of Wnt signaling indirectly promotes osteoclastogenesis by increasing the RANKL/OPG ratio in preosteoblasts (25). These studies suggest DKK1 may be a potential target for interventions intended at targeting MM bone disease.

There are studies suggesting that increasing Wnt signaling to induce osteoblastogenesis is a potential therapeutic mechanism. Edwards et al. investigated the consequences of increasing Wnt signaling in the 5TGM1 myeloma mouse model by using lithium chloride to inhibit GSK-3 β (26). Increased Wnt signaling effectively prevented bone disease and reduced MM cell burden in the bone marrow. Another group showed that increased Wnt signaling reduces bone disease (27). It is therefore well-established that Wnt signaling effectively prevents MM bone disease and reduced MM tumor burden in bone, partially by simultaneously increasing osteoblastogenesis and reducing osteoclastogenesis

MM is an incurable disease and understanding its biology aids the main goal of myeloma therapy: designing novel treatment regimens that extend patient survival while decreasing occurrence of adverse effects. Initial treatment of MM has changed substantially as a result of drug development (28). Thalidomide, lenalidomide, and bortezomib have improved overall and duration of response, and progression-free and overall survival for patients with newly diagnosed, relapsed and refractory MM (29).

The ubiquitin-proteasome pathway plays a critical role in regulating growth and survival of MM cells. This provides strong rationale for proteasome targeting and a successful induction of bortezomib into clinical management of myeloma (30). It has been reported that the inhibition of the ubiquitin-proteasome pathway modulates osteoblast differentiation through upregulating expression of bone morphogenetic protein 2 (BMP2) (31) and by preventing proteolytic degradation of RUNX2 (32, 33), but also β-catenin (34). It has been suggested that proteosome inhibitor bortezomib promotes bone formation by inhibiting DKK1 expression in osteogenic cells (35). This drug also directly inhibits osteoclast differentiation, possibly by inhibiting NF-ƙB activity in osteoclast precursors (36).

It is well-established that risk and outcomes in MM rise as a consequence of complex interplay of genetic factors and tumor deregulation. Bone disease alone, a serious complication in MM patients, is a complex event which clearly shows interactions between variable individual genetic backgrounds with the bone microenvironment, various ligands and other factors produced by MM cells. Moreover, therapeutic outcomes are impacted by not only tumor genetics, but also, importantly, by the individual genomic variability. The first part of this thesis focuses on the genetic variations that contribute to inter-individual variability to disease risk and bone complications in MM.

Lung cancer:

Lung cancer is a disease of uncontrolled cell growth in tissues of the lung. This growth may lead to metastasis to adjacent tissues and infiltration beyond the lungs. The majority of primary lung cancers are derived from epithelial cells. Lung cancer is the most common cause of cancer-related death in men and women, and is responsible for 1.3 million deaths worldwide annually, as of 2004 (37).

The main types of lung cancer are small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). Depending on the type of lung cancer, the treatment varies: NSCLC is sometimes treated with surgery, while SCLC often responds better to chemotherapy and radiation (38). The most common cause of lung cancer is long-term exposure to tobacco smoke (39). The occurrence of lung cancer in nonsmokers, who account for as many as 15% of cases (40), is often attributed to a combination of genetic factors (41, 42) radon gas (43), asbestos (44) and air pollution (45-47) including secondhand smoke (48, 49). Undoubtedly, tobacco-smoke caused lung cancer is impacted by inter-individual genetic variability that in turn affects the metabolism of tobacco-smoke carcinogens, similar to genetic impact on the metabolism of therapeutic agents. In this case, however, variable metabolism results in variable levels of harmful DNA adducts, and ultimately to cancer-causing mutations, rather than to adverse effects, as is case from chemotherapeutic treatments.

The use of tobacco products continues to be an immense public health problem, and one might argue is the largest voluntary source of human exposure to carcinogens in the world. However, much progress has been made in the past 20 years, not only in understanding mechanisms of tobacco carcinogenesis, but also in tobacco control. However, there are still 1.3 billion smokers in the world and hundreds of millions of smokeless tobacco users (50). Cigarette smoking causes 30% of all cancer mortality in developed countries (51).

4-(methylnitrosamino)-1-(30pyridyl)-1-butanone (NNK) is arguably one of the most important carcinogens in tobacco products. There is a considerable amount of evidence indicating that this tobacco-specific amine is one cause of lung cancer in smokers. NNK requires metabolic activation to form DNA adducts that are critical for its carcinogenicity (52). Moreover, polycyclic aromatic hydrocarbons (PAHs) are also widely regarded as important causes of lung cancer. In addition to the activation pathways, there are competing detoxification pathways which lead to harmless excretion of NNK and PAHs. Multiple cytochrome P450 (CYP) enzymes and phase II enzumes are involved in the metabolic activation and detoxification of NNK, its major metabolite 4-(Methylnitrosamino)-1-(3-Pyridyl)-1-Butanol (NNAL) and PAHs (52-54).

Most of the chemical carcinogens in the environment are chemically inert in themselves and require metabolic activation by CYP enzymes to more reactive metabolites in order to exhibit carcinogenicity in experimental animals and humans (55, 56). Of the 17 families of human CYPs identified to date, the CYP1, 2, and 3 family membes play major roles in the metabolic activation of a variety of environmental carcinogens. It has been implicated that CYP1A1 and CYP1B1 activate most PAHs to epoxide intermediates, which are then converted to more reactive diol-epoxides with the aid of epoxide hydrolase (55, 57). Both CYP1A1 and CYP1B1 are expressed mainly in extrahepatic organs and thus make a major contribution to the incidence of cancers in those organs, when PAHs and other carcinogens are ingested into animal’s body (57). PAHs and polyhalogenated hydrocarbons such as 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) induce several xenobiotic-metabolizing enzymes, including CYP1A1 and CYP1B1, through the arylhydrocarbon receptor (AhR) and the increased synthesis of these enzymes, as well as genetic variations in CYP1A1 and CYP1B1 genes, may determine difference in susceptibility of individuals to carcinogenesis caused by PAHs and other carcinogens.

CYP1B1 metabolizes numerous PAHs, and many N-heterocyclic amines, arylamines, amino azo dyes, and several other carcinogens (58). Unlike CYP1A1, CYP1B1 often shows substantial constitutive levels, and its expression is high in vascular endothelial cells, breast, prostate, uterus, epithelial lining of the head and neck, lung, various types of tumors, adrenal cortex, and many other tissues. It is thought that the highest benzo[*a*]pyrene (BaP) hydroxylase activity in the first trimester, which occurs in human adrenal cortex, reflects constitutive CYP1B1 expression (59). It is not clear what are constitutive and what are inducible CYP1B1 protein levels among smokers, non-smokers and ex-smokers in human lung, however, large interindividual differences (as well as CYP1A1) have been reported (60).

As might be expected, in vitro studies have uncovered that the *Cyp1b1* (-/-) mouse exhibits increased protection against 7,12-dimethylbenzo[*a*]antracene (DMBA)-induced lymphomas (61), DMBA-induced marrow toxicity and pre-lekemia (62), and dibenzo[*a,l*]pyrene-induced tumors (63). Therefore, presence of CYP1B1 is essential in activating these environmental chemicals, and if absent, less toxicity or neoplasia is seen. CYP1B1 is emerging as an important player in development of many different types of cancers, including lung cancer. Studies considering genetic variations within this gene are beginning to emerge, and are further emphasizing its importance (64-67).

Both, environmental and genetic factors play a role in a development of most diseases, and therefore it is important to understand the interactions between these factors. In all discussions of environmental exposure (environment being defined as any external agent, including medications), the dose to which an individual is exposed is critical. Therefore, it is crucial to understand gene—environment interaction. The idea of genetic variants modifying risk for cancer upon exposure to varying levels of external agents first came from studies of metabolic genes in environmental carcinogenesis (68, 69). A portion of this thesis investigates gene—environment relationship in lung cancer by studying variations in the known environmental exposure pathways.

Human genome variation:

Since the draft and the improved version of the human genome sequence published in 2001 and 2004 by the International Human Genome Consortium, scientists have become aware of large genetic variations, also called polymorphisms (70-72). Polymorphisms are stable and heritable, thus distinguishable from somatic mutations. They include single nucleotide polymorphisms (SNPs), micro- and minisatellites, insertions, and deletions (indels). About 90% of human genome variation comes in the form of SNPs, single nucleotide changes that can alter amino acid coding (coding-nonsynonymous vs. coding-synonymous).

Since the unveiling of the human genome sequence, it has been possible to more closely investigate disease risk, drug metabolism and various outcomes. A notion that genetic variations impact diseases risk and clinical outcomes became an important principle to consider. The overarching hypothesis of this dissertation is that there are common genetic variations that alter disease risk, clinical outcomes and response to environmental agents, either drugs or various carcinogens.

With a goal of studying genetic variations in the context of disease risk and outcomes, we have engaged in an international program designated *Bank On A Cure* (BOAC). This program was established to bank DNA from multiple groups and institutional trials, as well as to develop a genotyping platform for examining genetic variation in association with disease risk and clinical outcomes. We developed a novel custom SNP panel based on the Affymetrix/Gene Chip Targeted Genotyping Platform, which contains 3,404 SNPs in 983 genes, variations important in many cancer-related processes, drug metabolism and in transcriptional regulatory regions (i.e. promoter) (73). Design and the first applications of BOAC panel are presented in Part 1 of the Appendix in this thesis.

The HapMap Project (74), established to identify genetic variations and map out the SNP pattern haplotype of the human genome, has made available a great deal of information to the SNP research community. SNP genotyping allows for further characterization of the human genome and a better understanding of the hereditary transfer of diseases. The study of SNPs is of great value for the medical and pharmaceutical communities as they can help predict disease association and how an individual may respond to a given drug. Human SNP genotyping has lots of implications in the context of medical significance of diseases, clinical diagnostics, improving current practice of medicine, and in drug discovery.

The candidate gene approach can be a powerful tool for the identification of genetic associations to complex disease traits. Candidate gene studies using targeted SNPs are faster and more cost-effective than alternative approaches that may require complete re-sequencing of candidate genes or larger studies that require whole-genome genotyping panels. Projects such as the National Cancer Institute's Cancer Genome Anatomy Project SNP500 Cancer Database allow for SNP chip designs that are utilized for studying cancer and other diseases (75). The goals of such projects are not only to provide a central resource of newly discovered variants, but also for validating existing variants.

Identifying genes involved in the development of cancer is crucial to fully understanding cancer biology, for developing novel therapeutics for cancer treatment and for providing methods for cancer prevention and early diagnosis.The use of polymorphic markers, specifically single nucleotide polymorphisms (SNPs), promises to provide a comprehensive tool for analysing the human genome and identifying those genes and genomic regions contributing to the cancer phenotype.

Rationale and specific aims:

Compared to other lymphoid malignancies, multiple myeloma is difficult to study and treat due to not only genetic heterogeneity, but also due to the fact that it is a rare disease. Absence of large clinical trials enrolling significant numbers of subjects and repetitions of such trials with similar treatments has made it difficult to validate findings from SNP association studies.

The underlyining principle of the research presented in this thesis is that there are variations in cancer-relevant pathways that lead to inter-individual variability to disease risk and outcomes. We have developed a targeted SNP panel that consists of genes in cancer-important functional categories, and 3,404 variations important to cancer initiation and progression pathways, mostly targeted to MM (73).

The second chapter of this thesis examines variations in pathways involved in the pathogenesis of myeloma bone disease, including Wnt pathway, as well as growth factors important to the tumor microenvironment and bone disease progression. The novelty comes from combining gene expression signature of DKK1, previously associated with bone disease prognosis, and a newly identified SNP profile. This approach resulted in a better prediction of bone disease in MM patients. These results not only demonstrate a novel approach in studying genetic markers in association with clinical outcomes, but also identify novel variations that can differentiate between subjects with poor outcomes from those with a better bone disease prognosis.

The third chapter of this thesis looks at inter-individual genetic variability and its impact on myeloma progression-free survival and risk. We applied novel data mining algorithms to efficiently and robustly identify SNP-SNP interactions among 3,404 polymorphisms on the custom panel, that associate with risk and prognosis. With limited sample sizes and lack of multiple clinical trials with similar treatments, the need for robust computational methods is essential. We successfully applied novel computational algorithms and uncovered additional genetic variants important to myeloma risk and progression-free survival, as well as validated some of our previous findings.

The fourth chapter investigates inter-individual genetic variability in the context of metabolism of tobacco-smoke carcinogens and the ultimate impact on lung cancer risk. We identified a polymorphism in CYP1B1 gene, which has an interaction with levels of a known exposure agent, tobacco smoke carcinogen NNAL, metabolite of NNK. This polymorphism has been functionally shown to decrease protein cellular levels, thus thought to alter protein's enzymatic activity and ultimately metabolism of harmful carcinogens. In this chapter, we show a gene—environment interaction and significance of this interplay on disease risk.

The fifth chapter in this thesis investigates all 3,404 SNPs in association with lung cancer risk. We applied a novel data mining combinatorial search algorithm to identify significant SNP-SNP interactions, which distinguish cases from the controls. The identified variations are in genes involved in drug metabolism, and demonstrated by other groups to associate with lung cancer risk and metabolism of environmental carcinogens.

These findings implicate that genetic variations have a crucial role in disease initiation and progression, and they impact disease risk, as well as alter drug and environmental agents metabolism.

Our hypothesis is that there are genetic variants that affect inter-individual risk and response, and the metabolism of environmental agents, thus altering disease risk. Furthermore, our research relied on developing novel computational algorithms to robustly identify variants and their interactions.

Contribution to projects:

In chapter two, I conducted the genotyping of all 3,404 SNPs in all subjects presented in the study. I helped with review and editing of the manuscript.

In chapter three, I conducted the genotyping of all 3,404 SNPs in most of the subjects represented in the study, mostly for the samples in the mouthwash BOAC patient-provided samples. I guided the data mining approach designed by our collaborators, and wrote the manuscript. I provided biological interpretations and insights to our data mining collaborators, and helped them direct their analytical design. I conducted data analysis upon receiving the results produced by the algorithms. This is the first application of these novel computational algorithms on myeloma genotyping data.

In chapter four, I processed all the samples, from DNA extraction to genotyping using BOAC SNP panel and Sequenom. I guided the statistical analysis by having weekly meetings with Dr. Tim Church and Ms. Mindy Geisser, who analyzed the data. I was primarily responsible for coordinating statistical analysis by Dr. Church's group, and biochemical analysis by Dr. Hecht's group. I biologically interpreted the results and guided the way data was being handled and processed. I co-first authored the manuscript with Dr. Tim Church.

In chapter five, I processed and genotyped all the samples. The same as in chapter three, I helped guide the process of designing novel combinatorial search algorithms. This is the first application of this novel data mining method developed by our collaborators on lung cancer on genotyping data.

In each of the projects presented in chaptes 1-5, I was the primary contact point for the laboratory, and maintained contact with collaborating lab personnel as well as their PIs.

Appendix contains three additional parts to which I contributed during the course of my research. In Appendix I, I helped genotype some of the samples represented in the study, and was marginally involved in writing the manuscript by contributing with some data analysis and editing. I provided the analysis of linkage disequilibrium variations, which served as an internal control (all variations in the same LD block are expected to associate together). In Appendix II, I contributed to genotyping of some clinical trials represented in the study, as well as with editing of the manuscript. This study was a large collaborative effort. In Appendix III, I helped guide the design of novel data mining algorithms represented here. I regularly met with our collaborators and provided them with biological interpretation. I conducted the data analysis following the production of the results by the algorithm.

References:

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. **Cancer statistics**. 2009. *CA Cancer J Clin*. 2009; 59: 225-49.

2. Cohen HJ, Crawford J, Rao MK, Pieper CF, Currie MS. **Racial differences in the prevalence of monoclonal gammopathy in a community-based sample of the elderly**. *Am J Med*. 1998; 104: 439-44.

3. Munshi NC. **Plasma cell disorders: an historical perspective.** *Hematology Am Soc Hematol Educ Program*. 2008: 297.

4. Kumar SK, Rajkumar SV, Dispenzieri A, Lacy MQ, Hayman SR, Buadi FK, Zeldenrust SR, Dingli D, Russell SJ, Lust JA, Greipp PR, Kyle RA, Gertz MA. **Improved survival in multiple myeloma and the impact of novel therapies**. *Blood*. 2008; 111: 2516-20.

5. Brenner H, Gondos A, Pulte D. **Recent major improvement in long-term survival of younger patients with multiple myeloma**. *Blood*. 2008; 111: 2521-6.

6. Michigami T, *et al*. **Cell-cell contact between marrow stromal cells andmyeloma cells via VCAM-1 and alpha(4)beta(1)-integrin enhances production ofosteoclast-stimulating activity**. *Blood*. 2000; 96: 1953–1960.

7. Hazlehurst LA, Damiano JS, Buyuksal I, Pledger WJ, Dalton WS. Adhesion **to fibronectin via beta1 integrins regulates p27kip1 levels and contributes to cell adhesion mediated drug resistance (CAM-DR)**. *Oncogene*. 2000; 19: 4319-27.

8. Vacca A, Ria R, Presta M, Ribatti D, Iurlaro M, Merchionne F, Tanghetti E, Dammacco F. **alpha(v)beta(3) integrin engagement modulates cell adhesion, proliferation, and protease secretion in human lymphoid tumor cells**. *Exp Hematol*. 2001; 29: 993-1003.

9. Anderson KC, Lust JA. **Role of cytokines in multiple myeloma**. *Semin Hematol*. 1999; 36(1 Suppl 3): 14-20.

10. Podar K, Hideshima T, Chauhan D, Anderson KC. **Targeting signalling pathways for the treatment of multiple myeloma**. *Expert Opin Ther Targets*. 2005; 9: 359-81.

11. Giuliani N, Rizzoli V, Roodman GD. **Multiple myeloma bone disease: Pathophysiology of osteoblast inhibition.** *Blood*. 2006; 108: 3992-6.

12. Kyle RA, Therneau TM, Rajkumar SV, Offord JR, Larson DR, Plevak MF, Melton LJ 3rd. **A long-term study of prognosis in monoclonal gammopathy of undetermined significance**. *N Engl J Med*. 2002; 346: 564-9.

13. Bataille R, Chappard D, Marcelli C, Dessauw P, Baldet P, Sany J, Alexandre C. **Recruitment of new osteoblasts and osteoclasts is the earliest critical event in the pathogenesis of human multiple myeloma**. *J Clin Invest*. 1991; 88: 62-6.

14. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD Jr. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483-94.

15. Oshima T, Abe M, Asano J, Hara T, Kitazoe K, Sekimoto E, Tanaka Y, Shibata H, Hashimoto T, Ozaki S, Kido S, Inoue D, Matsumoto T. **Myeloma cells suppress bone formation by secreting a soluble Wnt inhibitor, sFRP-2**. *Blood*. 2005; 106: 3160-5.

16. Giuliani N, Colla S, Morandi F, Lazzaretti M, Sala R, Bonomini S, Grano M, Colucci S, Svaldi M, Rizzoli V. **Myeloma cells block RUNX2/CBFA1 activity in human bone marrow osteoblast progenitors and inhibit osteoblast formation and differentiation**. *Blood*. 2005; 106: 2472-83.

17. Standal T, Abildgaard N, Fagerli UM, Stordal B, Hjertner O, Borset M, Sundan A. **HGF inhibits BMP-induced osteoblastogenesis: possible implications for the bone disease of multiple myeloma**. *Blood*. 2007; 109: 3024-30.

18. Lee JW, Chung HY, Ehrlich LA, Jelinek DF, Callander NS, Roodman GD, Choi SJ. **IL-3 expression by myeloma cells increases both osteoclast formation and growth of myeloma cells**. *Blood*. 2004; 103: 2308-15.

19. Ehrlich LA, Chung HY, Ghobrial I, Choi SJ, Morandi F, Colla S, Rizzoli V, Roodman GD, Giuliani N. **IL-3 is a potential inhibitor of osteoblast differentiation in multiple myeloma**. *Blood*. 2005; 106: 1407-14.

20. Clevers H. **Wnt/beta-catenin signaling in development and disease**. *Cell*. 2006; 127: 469-80.

21. Pinzone JJ, Hall BM, Thudi NK, Vonau M, Qiang YW, Rosol TJ, Shaughnessy JD Jr. **The role of Dickkopf-1 in bone development, homeostasis, and disease.** *Blood*. 2009; 113: 517-25.

22. Kaiser M, Mieth M, Liebisch P, Oberländer R, Rademacher J, Jakob C, Kleeberg L, Fleissner C, Braendle E, Peters M, Stover D, Sezer O, Heider U. **Serum concentrations of DKK-1 correlate with the extent of bone disease in patients with multiple myeloma**. *Eur J Haematol*. 2008; 80: 490-4.

23. Drake M, Ng A, Kumar S, et al. **Increases in serum levels of dickkopf 1 are associated with alterations in skeletal microstructure in monoclonal gammopathy of undetermined significance** [abstract]. In: T*he 31st Annual Meeting of the American Society for Bone and Mineral Research*, September 2009, Denver USA.

24. Qiang YW, Barlogie B, Rudikoff S, Shaughnessy JD Jr. **Dkk1-induced inhibition of Wnt signaling in osteoblast differentiation is an underlying mechanism of bone loss in multiple myeloma**. *Bone*. 2008; 42: 669-80

25. Qiang YW, Chen Y, Stephens O, Brown N, Chen B, Epstein J, Barlogie B, Shaughnessy JD Jr. **Myeloma-derived Dickkopf-1 disrupts Wnt-regulated osteoprotegerin and RANKL production by osteoblasts: a potential mechanism underlying osteolytic bone lesions in multiple myeloma.** *Blood*. 2008; 112: 196-207.

26. Edwards CM, Edwards JR, Lwin ST, Esparza J, Oyajobi BO, McCluskey B, Munoz S, Grubbs B, Mundy GR. **Increasing Wnt signaling in the bone marrow microenvironment inhibits the development of myeloma bone disease and reduces tumor burden in bone in vivo**. *Blood*. 2008; 111: 2833-42.

27. Qiang YW, Shaughnessy JD Jr, Yaccoby S. **Wnt3a signaling within bone inhibits multiple myeloma bone disease and tumor growth**. *Blood*. 2008; 112: 374-82.

28. Raab MS, Breitkreutz I, Anderson KC. **Targeted treatments to improve stem cell outcome: old and new drugs**. *Bone Marrow Transplant*. 2007; 40: 1129-37.

29. Richardson PG, Hideshima T, Mitsiades C, Anderson KC. **The emerging role of novel therapies for the treatment of relapsed myeloma**. *J Natl Compr Canc Netw*. 2007; 5: 149-62.

30. Richardson PG, Barlogie B, Berenson J, Singhal S, Jagannath S, Irwin D, Rajkumar SV, Srkalovic G, Alsina M, Alexanian R, Siegel D, Orlowski RZ, Kuter D, Limentani SA, Lee S, Hideshima T, Esseltine DL, Kauffman M, Adams J, Schenkein DP, Anderson KC. **A phase 2 study of bortezomib in relapsed, refractory myeloma**. *N Engl J Med*. 2003; 348: 2609-17.

31. Garrett IR, Chen D, Gutierrez G, Zhao M, Escobedo A, Rossini G, Harris SE, Gallwitz W, Kim KB, Hu S, Crews CM, Mundy GR. **Selective inhibitors of the osteoblast proteasome stimulate bone formation in vivo and in vitro**. *J Clin Invest*. 2003; 111: 1771-82.

32. Bellido T, Ali AA, Plotkin LI, Fu Q, Gubrij I, Roberson PK, Weinstein RS, O'Brien CA, Manolagas SC, Jilka RL. **Proteasomal degradation of Runx2 shortens parathyroid hormone-induced anti-apoptotic signaling in osteoblasts. A putative explanation for why intermittent administration is needed for bone anabolism**. *J Biol Chem*. 2003; 278: 50259-72.

33. Giuliani N, Morandi F, Tagliaferri S, Lazzaretti M, Bonomini S, Crugnola M, Mancini C, Martella E, Ferrari L, Tabilio A, Rizzoli V. **The proteasome inhibitor bortezomib affects osteoblast differentiation in vitro and in vivo in multiple myeloma patients**. *Blood*. 2007; 110: 334-8.

34. Qiang YW, Hu B, Chen Y, Zhong Y, Shi B, Barlogie B, Shaughnessy JD Jr. **Bortezomib induces osteoblast differentiation via Wnt-independent activation of beta-catenin/TCF signaling**. *Blood*. 2009; 113: 4319-30.

35. Oyajobi BO, Garrett IR, Gupta A, Flores A, Esparza J, Muñoz S, Zhao M, Mundy GR. **Stimulation of new bone formation by the proteasome inhibitor, bortezomib: implications for myeloma bone disease**. *Br J Haematol*. 2007; 139: 434-8.

36. Zavrski I, Krebbel H, Wildemann B, Heider U, Kaiser M, Possinger K, Sezer O. **Proteasome inhibitors abrogate osteoclast differentiation and osteoclast function.** *Biochem Biophys Res Commun.* 2005; 333: 200-5.

37. WHO (February 2006). "Cancer". World Health Organization. http://www.who.int/mediacentre/factsheets/fs297/en/

38. Vaporciyan AA, Nesbitt JC, Lee JS et al. **Cancer Medicine**. *B C Decker*. 2000; pp. 1227–1292.

39. Merck Manual Professional Edition, Online edition. **Lung Carcinoma: Tumors of the Lung**. http://www.merck.com/mmpe/sec05/ch062/ch062b.html#sec05-ch062-ch062b-1405.

40. Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee IM, Marugame T, Palmer JR, Riboli E, Sobue T, Avila-Tang E, Wilkens LR, Samet JM. **Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies**. *PLoS Med*. 2008; 5: e185.

41. Gorlova OY, Weng SF, Zhang Y, Amos CI, Spitz MR. **Aggregation of cancer among relatives of never-smoking lung cancer patients**. *Int J Cancer*. 2007; 121: 111-8.

42. Hackshaw AK, Law MR, Wald NJ. **The accumulated evidence on lung cancer and environmental tobacco smoke**. *BMJ*. 1997; 315: 980-8.

43. Catelinois O, Rogel A, Laurier D, Billon S, Hemon D, Verger P, Tirmarche M. **Lung cancer attributable to indoor radon exposure in france: impact of the risk models and uncertainty analysis**. *Environ Health Perspect*. 2006; 114: 1361-6.

44. O'Reilly KM, Mclaughlin AM, Beckett WS, Sime PJ. **Asbestos-related lung disease**. *Am Fam Physician.* 2007; 75: 683-8.

45. Kabir Z, Bennett K, Clancy L. **Lung cancer and urban air-pollution in Dublin: a temporal association?** *Ir Med J.* 2007; 100: 367-9.

46. Coyle YM, Minahjuddin AT, Hynan LS, Minna JD. **An ecological study of the association of metal air pollutants with lung cancer incidence in Texas**. *J Thorac Oncol*. 2006; 1: 654-61.

47. Chiu HF, Cheng MH, Tsai SS, Wu TN, Kuo HW, Yang CY. **Outdoor air pollution and female lung cancer in Taiwan**. *Inhal Toxicol*. 2006; 18: 1025-31.

48. Carmona, RH. **Secondhand smoke exposure causes disease and premature death in children and adults who do not smoke.** *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General*. U.S. Department of Health and Human Services. 2006. http://www.surgeongeneral.gov/library/secondhandsmoke.

49. WHO International Agency for Research on Cancer. **Tobacco Smoke and Involuntary Smoking**. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans 83*. 2002. http://monographs.iarc.fr/ENG/Monographs/vol83/volume83.pdf.

50. World Health Organization. **The World Health Report 2003: Shaping the Future**. World Health Organization Geneva, Switzerland. 2003; pp91-94.

51. International Union Against Cancer. 2007; www.deathsfromsmoking.net

52. Hecht SS. **Biochemistry, biology, and carcinogenicity of tobacco-specific *N*-nitrosamines**. *Chem. Res. Toxicol*. 1998; 11, 559-603.

53. Jalas JR, Hecht SS, Murphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem Res Toxicol* 2005; 18: 95-110.

54. Cooper CS, P. L. Grover, and P. Sims. The **metabolism and activation of benzo[a]pyrene**. *Prog Drug Metab.* 1983; 7: 295-396.

55. Conney AH. **Induction of microsomal enzymes by foreign chemicals and carcinogenesis by polycyclic aromatic hydrocarbons: G. H. A. Clowes Memorial Lecture**. *Cancer Res*. 1982; 42: 4875-917.

56. Guengerich FP, Shimada T. **Oxidation of toxic and carcinogenic chemicals by human cytochrome P-450 enzymes**. *Chem Res Toxicol*. 1991; 4: 391-407.

57. Shimada T, Hayes CL, Yamazaki H, Amin S, Hecht SS, Guengerich FP, Sutter TR. **Activation of chemically diverse procarcinogens by human cytochrome P-450 1B1**. *Cancer Res*. 1996; 56: 2979-84.

58. Guengerich FP. **Metabolism of chemical carcinogens**. *Carcinogenesis*. 2000; 21: 345-51.

59. Rane A, Sjöqvist F, Orrenius S. **Cytochrome P-450 in human fetal liver microsomes**. *Chem Biol Interact*. 1971; 3 :305.

60. Kim JH, Sherman ME, Curriero FC, Guengerich FP, Strickland PT, Sutter TR. **Expression of cytochromes P450 1A1 and 1B1 in human lung from smokers, non-smokers, and ex-smokers**. *Toxicol Appl Pharmacol*. 2004; 199: 210-9.

61. Buters JT, Sakai S, Richter T, Pineau T, Alexander DL, Savas U, Doehmer J, Ward JM, Jefcoate CR, Gonzalez FJ. **Cytochrome P450 CYP1B1 determines susceptibility to 7, 12-dimethylbenz[a]anthracene-induced lymphomas**. *Proc Natl Acad Sci U.S.A*. 1999; 96: 1977-82.

62. Page TJ, O'Brien S, Holston K, MacWilliams PS, Jefcoate CR, Czuprynski CJ. **7,12-Dimethylbenz[a]anthracene-induced bone marrow toxicity is p53-dependent**. *Toxicol Sci*. 2003; 74: 85-92.

63. Buters JT, Mahadevan B, Quintanilla-Martinez L, Gonzalez FJ, Greim H, Baird WM, Luch A. **Cytochrome P450 1B1 determines susceptibility to dibenzo[a,l]pyrene-induced tumor formation**. *Chem Res Toxicol*. 2002; 15: 1127-35.

64. Han JF, He XY, Herrington JS, White LA, Zhang JF, Hong JY. **Metabolism of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) by human CYP1B1 genetic variants**. *Drug Metab Dispos*. 2008; 36: 745-52.

65. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol*. 2005; 67: 435-43.

66. Aklillu E, Øvrebø S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res*. 2005; 65: 5105-11.

67. Mammen JS, Pittman GS, Li Y, Abou-Zahr F, Bejjani BA, Bell DA, Strickland PT, Sutter TR. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis*. 2003; 24: 1247-55.

68. Vineis P, Bartsch H, Caporaso N, Harrington AM, Kadlubar FF, Landi MT, Malaveille C, Shields PG, Skipper P, Talaska G, et al. **Genetically based N-acetyltransferase metabolic polymorphism and low-level environmental exposure to carcinogens**. *Nature*. 1994; 369: 154-6.

69. Garte S, Zocchetti C, Taioli E. Garte S, Zocchetti C, Taioli E. **Gene--environment interactions in the application of biomarkers of cancer susceptibility in epidemiology**. *IARC Sci Publ*. 1997; 142: 251-64.

70. International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome**. *Nature*. 2001; 409: 860-921.

71. Venter JC, *et.al*. **The sequence of the human genome**. *Science*. 2001; 291: 1304-51.

72. International Human Genome Sequencing Consortium. **Finishing the euchromatic sequence of the human genome**. *Nature*. 2004; 431: 931-45.

73. Van Ness B, Ramos C, Haznadar M, *et.al*. **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med*. 2008; 6:26.

74. The International HapMap Consortium. **The International HapMap Project**. *Nature*. 2003; 426, 789-796. http:/www.hapmap.org

75. Packer, B.R., et al. **SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes**. *Nucleic Acids Res*. 2004; 32 Database issue: D528-32. http://snp500cancer.nci.nih.gov/

CHAPTER 2

GENETIC POLYMORPHISMS OF EPHX1, GSK3β, TNFSF8 AND MYELOMA CELL DKK-1 EXPRESSION LINKED TO BONE DISEASE IN MYELOMA

BGM Durie[1], B Van Ness[2], C Ramos[2], O Stephens[3], M Haznadar[2], A Hoering[4], J Haessler[4], MS Katz[5], GR Mundy[6], RA Kyle[7], GJ Morgan[8], J Crowley[4], B Barlogie[3] and J Shaughnessy Jr[3]

[1]Hematology/Oncology, Cedars-Sinai Outpatient Cancer Center at the Samuel Oschin Comprehensive Cancer Institute & Aptium Oncology, Los Angeles, CA, USA; [2]Institute of Human Genetics, University of Minnesota, MN, USA; [3]Myeloma Institute for Research and Therapy, University of Arkansas, Little Rock, AR, USA; [4]Southwest Oncology Group Statistical Centre, Cancer Research And Biostatistics, Seattle, WA, USA; [5]Information Technology, International Myeloma Foundation, North Hollywood, CA, USA; [6]Cancer Biology, Vanderbilt Center for Bone Biology, Nashville, TN, USA; [7]Laboratory Medicine/Pathology, Mayo Clinic College of Medicine, Rochester, MN, USA and [8]Section of Haemato-Oncology, Royal Marsden Hospital, London, Surrey, UK

This investigation was supported in part by an unrestricted grant from the International Myeloma Foundation (Bank on a Cure project), as well as by the following PHS Cooperative Agreement grant numbers awarded by the National Cancer Institute, DHHS: CA32102 and CA38926 (SWOG); and CA21115 (ECOG); plus CA 97513 (JDS and BB).

Reprinted with permission (license #2436120352436) from the Nature Publishing Group, Leukemia 2009, 23(10): 1913-9.

I conducted the genotyping of all 3,404 SNPs in all subjects presented in the study. I helped with review and editing of the manuscript.

Bone disease in myeloma occurs as a result of complex interactions between myeloma cells and the bone marrow microenvironment. A custom-built DNA single nucleotide polymorphism (SNP) chip containing 3404 SNPs was used to test genomic DNA from myeloma patients classified by the extent of bone disease. Correlations identified with a Total Therapy 2 (TT2) (Arkansas) data set were validated with Eastern Cooperative Oncology Group (ECOG) and Southwest Oncology Group (SWOG) data sets, where the primary phenotype was progression-free survival. Univariate correlates with severity of bone disease included: EPHX1, IGF1R, IL-4 and Gsk3b. SNP signatures were linked to the number of bone lesions, log2 DKK-1 myeloma cell expression levels and patient survival. Using stepwise multivariate regression analysis, the following SNPs: EPHX1 (P=0.0026); log2 DKK-1 expression (P=0.0046); serum lactic dehydrogenase (LDH) (P=0.0074); Gsk3b (P=0.02) and TNFSF8 (P=0.04) were linked to bone disease. This assessment of genetic polymorphisms identifies SNPs with both potential biological relevance and utility in prognostic models of myeloma bone disease.

Introduction:

Multiple myeloma is a tumor of plasma cells that depends on the bone marrow microenvironment for growth and survival (1, 2). Bone disease in myeloma occurs as a result of the complex interactions between myeloma cells and the bone marrow osteoclasts, osteoblasts plus other accessory cells and microenvironmental components (2). Myeloma bone disease is characterized by a unique combination of enhanced osteoclast numbers and function plus reduced osteoblast differentiation and function (1–12). The important elements in osteoclast activation are myeloma cellderived MIP-1a, which activates osteoclast CCRX-5 plus microenvironmental-derived RANK-ligand (RANK-L), which activates osteoclast RANK and competes with stromal-derived osteoprotegrin (OPG) (10–12). Recent studies have emphasized the central role of the Wnt (Wingless-type MMTV integration site family (mammalian homologue))-signaling inhibitor

DKKK-1 in the pathogenesis of the osteolytic bone lesions in myeloma (6). DKK-1 inhibits both osteoblast differentiation and function and increases osteoclast activity. Attention is focused both on the mechanisms responsible for the upregulation of DKK-1 synthesis in plasma cells and the interactions with the microenvironment (7–10). Expression of DKK-1 is regulated by a combination of intrinsic genomic factors and interactions with the bone marrow microenvironment (8). To assess the predilection to bone disease, it was elected to study the effect of single nucleotide DNA polymorphisms (SNP) in a well-characterized population of myeloma patients for whom DKK-1 expression and gene expression profile (GEP) gene signature data were also available (13). We focused on several pathways involved in the pathogenesis of myeloma bone disease, including the Wnt pathway, in particular GSK3, as well as insulin growth factor, interleukin 4, bradykinin receptors and b3 adrenergic receptors. Peripheral blood DNA from 282 patients enrolled in the UARK 2003–33 'Total Therapy 2' (TT2) protocol was studied using the previously reported Affymetrix 3k BOAC custom DNA chip to assess the presence or absence of relevant genetic polymorphisms (14–16). Here, we present evidence that several SNPs significantly correlate with both the clinical extent of the bone disease, as well as DKK-1 expression.

Materials and Methods:

Patients

These analyses included 282 patients with previously untreated myeloma enrolled in the TT2 trial between October 1998 and February 2004. The Arkansas group has developed the so-called Total Therapy programs during the past 20 years by using all the available drugs throughout all the up-front treatment steps. Total Therapy 2 included the new drug thalidomide in both induction and maintenance. Details of patient characteristics plus treatment and clinical outcomes have been reported (14). All participants had provided written informed consent in keeping with institutional and National Cancer Institute (NIH, Bethesda, MD, USA) guidelines. All details of the protocol had been approved by institutional guidelines and the United States Food and Drug Administration, and were monitored by a data safety and monitoring board as required for Phase III trials. The multiple myeloma baseline evaluation included serum and urine protein electrophoresis, quantitative immunoglobulin measurements, total 24-h urine protein excretion, serum b2-microglobulin (Sb2M), C-reactive protein, and lactic dehydrogenase (LDH) plus bone marrow aspirate and biopsy evaluations.

Bone studies

Imaging included baseline magnetic resonance imaging (MRI) and complete skeletal survey radiological examination (myeloma bone survey (MBS)) in a prospective manner (14). The MRI included the axial skeleton and pelvis plus any additional areas requiring diagnostic evaluation for pain or other medical issues. MRI studies were carried out with a series of sequences to permit identification of focal or diffuse bone marrow involvement, including spin echo (T2-wt), short T, inversion recovery (STIR) and gadolinium-enhanced spin echo sequences with and without fat suppression. Myeloma bone survey encompassed the long bones and were carried out with digital radiographs incorporating two views of the chest; views of ribs, lateral skull, vertebral column; anteroposterior views of the pelvis, shoulders; and the extremities including hands and feet. Focal lesions on both MRI and myeloma bone survey were identified as areas with an axial diameter of at least 0.5 cm. The MRIs were reviewed independently by four individuals who recorded the size, number and location of all focal lesions compatible with myeloma. Full details have been previously published (14).

Classification of bone disease

X-ray was the primary classification system for bone disease. The exception was 12 patients with extensive focal MRI disease, but no focal changes on X-ray. On the basis of detailed previous analyses (14), this 4% subset was added to the ‘extensive bone disease’ category to give 183/282 (65%) within this extensive bone disease group. The remaining 99 patients (35%) all had negative X-rays and no extensive focal disease on MRI. Using X-ray results only, validation of the TT2 findings was conducted comparing results in separate Eastern Cooperative Oncology Group (ECOG) and Southwest Oncology Group (SWOG) data sets (15, 16). For these analyses, patients with completely negative X-rays were compared with those having 43 focal lesions on X-ray.

Genotyping

Peripheral blood was collected in heparinized green top tubes and centrifuged to recover mononuclear cell pellets. DNA was extracted from the mononuclear cell pellets and genotyped using the Affymetrix (Santa Clara, CA, USA) Genchip scanner 3000 Targeted Genotyping System (GCS 3000 TG System) using molecular inversion probes to simultaneously identify the 3404 pre-selected SNPs in 983 genes (15, 16). All genotyping experiments were carried out in strict adherence to the manufacturer’s protocol.

Custom SNP Chip design and content

A directed, custom SNP chip design was developed with specific criteria from public and commercial databases. Full details are described elsewhere (15, 16). In essence, a custom SNP chip was developed, focusing on functionally relevant polymorphisms known to have a role in normal and abnormal cellular functions related to inflammation, immunity and drug responses.

Statistical analysis

Overview

Several methods were used to assess possible correlations between SNPs and the presence or absence of bone disease.

Univariate correlations of individual SNPs were assessed. This was first carried out for the TT2 data set and then for validation with the Eastern Cooperative Oncology Group and Southwest Oncology Group data sets.

Recursive partitioning was used to identify the best combinations of SNPs correlated with bone disease.

The validity of correlations with individual SNPs and combinations of SNPs was assessed using multivariate logistic regression analyses that incorporated known standard prognostic factors, gene expression profile results (risk groups: TT2 only) and Dkk-1expression results (TT2 only).

Correlations between individual SNPs as combinations of SNPs and patient outcomes were assessed including progression-free (PFS) and overall survivals (OS).

The Eastern Cooperative Oncology Group and Southwest Oncology Group data sets were evaluated with respect to SNP signatures identified in the TT2 data.

Statistical analysis details

We used Fisher's exact test as a univariate screening tool to determine the association of SNPs with bone disease. The top 50 rank-ordered SNPs were selected and a recursive-partitioning algorithm was carried out to determine the combination of SNPs that best distinguished the bone disease subgroups. In recursive partitioning, each genotype was evaluated on its ability to make a correct prediction, creating a decision node (17). Recursive partitioning allowed for interactions of SNPs and also included SNPs further down the rank-ordered list. Univariate association between clinical parameters was assessed using continuous and categorical variables (18). The non-parametric Kruskal–Wallace test was used for continuous variables and the w2-test was used for categorical variables (19). Multivariate logistic regression was used to test for associations of

SNPs and clinical parameters with bone disease (20). Survival curves were constructed according to Kaplan and Meier (21).

Results:

Classification of bone disease

The 282 patients were divided into 99 patients (35%) with no bone disease (X-rays negative) and 183 patients (65%) with definite/extensive bone disease (X-rays positive and/or extensive focal lesions on MRI (12 patients)). This separation best identified the two sub-populations in detailed analyses of the imaging results for the TT2 data set (14).

Univariate correlations between bone disease and SNPs in TT2 data set: Fisher's exact test was used as a univariatescreening tool to determine the association of SNPs with bone disease. Results are shown in Table 1, which displays the top SNPs most highly correlated with bone disease. The top-ranked SNP, EPHX1 (P=0.0003), is rs3766934 (GG), which is an expoxide hydrolase SNP. Several SNPs linked to bone biology were among the top-ranked SNPs, including IGFIR (P=0.003: #6), IL-4 (P=0.009: #16) and Gsk3b (P=0.015: #23).

Recursive partitioning: The top 50 SNPs with the lowest P-values were selected for recursive partitioning analysis. The results of recursive partitioning analysis are shown in Figure 1.

The 4 SNPs providing the best correlation were: rs3766934, EPHX1, RANK #1; rs3783408, Gsk3b, RANK #23; rs1052637, DDX18, RANK #26; and rs3181366, TNFSF8, RANK #29 in the univariate correlations (Table 1). The 4 SNP combination was then used as a search engine to identify further correlations. The results are shown in Figures 2a and b, respectively. There were excellent correlations with both numbers of individual focal bone lesions (P values=0.001) and the directly measured DKK-1 expression levels for individual patients (P=0.05).

Stepwise multivariate regression analyses: Several logistic regression models were used to further assess the correlations with the identified SNPs. Results are displayed in Table 2. Again, the previously identified SNPs prove to be statistically significantly associated with the bone disease status. The individual SNPs (EPHX1, Gsk3b and TNSF8), DKK-1 and lactic dehydrogenase (serum level) are predictive in the displayed multivariate analysis.

Correlations with progression-free and overall survival: Figure 3 shows the correlations between SNP pattern and outcomes. The cross correlations between known and predicted survivals are highly significant.

Cross-validation in additional clinical data sets with bone disease defined by X-ray only. These statistical analyses used 207 patients with zero or more than three X-ray focal lesions from the original TT2 data set plus 62 patients from Southwest Oncology Group (S9321) and 69 patients from Eastern Cooperative Oncology Group (E1A00 and E9486). Collectively, there were 163 patients with no X-ray evidence of bone disease and 175 patients with more then three focal lesions evident on X-ray. A majority of the SNPs from the TT2 only analyses were again highly ranked in combined analyses. For example, EPHX1 (previously ranked #1, now #9); IGFIR (previously ranked #6, now #13) and IL-4 (previously ranked #16, now #19) again showed significant correlations. Conversely, the SNPs for BDKRB1, ADRB3 and DDX18 were not highly ranked. Stepwise multivariate logistic regression analysis was then repeated incorporating top SNPs identified by cross-validation. The results are displayed in Table 3. This further cross-validation assessment shows that the EPHX1 SNP is still the top SNP in both the univariate and multivariate regressions. The TREX1 SNP, previously ranked number 11, acquires greater significance in these univariate and multivariate regressions. Other significant correlations were with DDK-1, lactic dehydrogenase, the 17-gene expression profile high risk, plus again the SNPs for Gsk3b and TNFSF8.

Discussion:

In this study, several SNPs are correlated with the likelihood of bone disease. The top SNP is EPHX1 (rs3766934: GG genotype versus GT/TT), an epoxide hydrolase. Although EPHX1 has been evaluated in multiple studies of genetic polymorphisms of biotransformation enzymes related to cancer, the functional significance of this specific GG genotype is currently unclear (22). Nonetheless, it is known that epoxide hydrolase is involved in both the inflammatory response linked to the bioactivation of leukotoxins (23) and the activation of the dioxin response element by benzo[a] pyrene compounds (24). Further studies are necessary to investigate the potential significance of this EPHX1 SNP genotype in laboratory, clinical and epidemiological studies. The Gsk3b SNP (Table 1 and Figure 1) was the second SNP selected as part of the recursive partitioning decision tree. This SNP is especially interesting as binding of GSK3bi with axin and APC forms a critical complex involved in Wnt-activated release or stabilization of b-catenin (25–29). This pathway is central to osteoblast function (30). Increased Wnt signaling through Wnt 3A results in an increase in the bone mineral density and a decrease in the osteoclast/osteoblast ratio (31–34). Gsk3b is the target of upregulation by thalidomide and is central to reactive oxygen species-mediated thalidomide-induced apoptosis (28). Other identified SNPs linked to bone related pathways (see Table 4) included the following: insulin-like growth factor 1 receptor (ranked number 6: Table 1) (35–39); bradykinin receptor B1 (ranked number 10: Table 1) (40); adrenergic receptor B3 (ranked number 14: Table 1) (41, 42); and interleukin-4 (ranked number 16: Table 1) (43). Several SNPs linked to drug and/or toxin metabolism and/or DNA metabolism and repair were noted and are summarized in Table 4. As dioxins have been linked to the etiology of myeloma (44), it is noteworthy that EPHX1 (45–47) is important in dioxin and polycyclic aromatic hydrocarbon metabolism. In addition, the DPYD SNP (rs1399291) ranked number 28 (Table 1) is involved with pyrimidine metabolism, and has, in addition, been identified in a separate recent large scale screening (48). Testing with the 3400

SNP custom chip has, thus, revealed several SNPs that are significantly correlated with the likelihood of bone disease in patients with myeloma. Larger studies are currently underway, for example, in collaboration with the National Cancer Institute (NCI) epidemiology branch, to further explore the relationships with identified SNPs (48).

References:

1. Hideshima T, Mitsiades C, Tonon G, Richardson PG, Anderson KC. **Understanding multiple myeloma pathogenesis in the bone marrow to identify new therapeutic targets**. *Nat Rev Cancer.* 2007; 7: 585–598.
2. Giuliani N, Rizzoli V, Roodman GD. **Multiple myeloma bone disease: pathophysiology of osteoblast inhibition**. *Blood*. 2006; 108: 3992–3996.
3. Durie BGM, Salmon SE, Mundy GR. **Relation of osteoclast activating factor production to extent of bone disease in multiple myeloma**. *Br J Hematol*. 1981; 47: 21–30.
4. Harada S, Rodan G. **Control of osteoblast function and regulation of bone mass**. *Nature*. 2003; 423: 349–355.
5. Westendorf JJ, Kahler RA, Schroeder TM. **Wnt signaling in osteoblasts and bone diseases.** *Gene.* 2004; 341: 19–39.
6. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B et al. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483–2494.
7. Yaccoby S, Ling W, Zhan F, Walker R, Barlogie B, Shaughnessy Jr JD. **Antibody-based inhibition of DKK1 suppresses tumor-induced bone resorption and multiple myeloma growth in vivo**. *Blood.* 2007; 109: 2106–2111.
8. Colla S, Zhan F, Xiong W, Wu X, Xu H, Stephens O et al. **The oxidative stress response regulates DKK1 expression through the JNK signaling cascade in multiple myeloma plasma cells**. *Blood*. 2007; 109: 4470–4477.
9. Qian J, Xie J, Hong S, Yang J, Zhang L, Han X et al. **Dickkopf-1 (DKK-1) is a widely expressed and potent tumor-associated antigen in multiple myeloma**. *Blood*. 2007; 110: 1587–1594.
10. Choi SJ, Oba Y, Gazitt Y, Alsina M, Cruz J, Anderson J et al. **Antisense inhibition of macrophage inflammatory protein 1-alpha blocks bone destruction in a model of myeloma bone disease**. *J Clin Invest.* 2001; 108: 1833–1841.
11. Lentzsch S, Gries M, Janz M, Bargou R, Dorken B, Mapara MY. **Macrophage inflammatory protein 1-alpha (MIP-1 alpha) triggers migration and signaling cascades mediating survival and proliferation in multiple myeloma (MM) cells**. *Blood*. 2003; 101: 3568–3573.
12. Vallet S, Raje N, Ishitsuka K, Hideshima T, Podar K, Chhetri S et al. **MLN3897, a novel CCR1 inhibitor, impairs osteoclastogenesis and inhibits the interaction of multiple myeloma cells and osteoclasts**. *Blood.* 2007; 110: 3744–3752.
13. Zhan F, Huang Y, Colla S, Stewart JP, Hanamura I, Gupta S et al. **The molecular classification of multiple myeloma**. *Blood.* 2006; 108: 2020–2028.
14. Walker R, Barlogie B, Haessler J, Tricot G, Anaissie E, Shaughnessy Jr JD et al. **Magnetic resonance imaging in multiple myeloma: diagnostic and clinical implications**. *J Clin Oncol*. 2007; 25: 1121–1128.
15. Johnson DC, Corthals S, Ramos C, Hoering A, Cocks K, Dickens NJ et al. **Genetic associations with thalidomide mediated venous thrombotic events in myeloma identified using targeted genotyping**. *Blood.* 2008; 112: 4924–4934.

16. Van Ness B, Ramos C, Haznadar M, Hoering A, Haessler J, Crowley J et al. **Genomic variation in myeloma: design, content, and initial application of the bank on a cure SNP panel to analysis of survival**. *BMC Med.* 2008; 6: 26 [pages not specified].
17. Terry MT, Elizabeth J, Atkinson R. **An introduction to recursive partitioning using the rpart routines.** 1997. *Technical report 61*, Mayo Clinic. 2Available at:

http://mayoresearch.mayo.edu/mayo/research/biostat/techreports.cfm##R package available at: http://cran.r-project.org/src/contrib/Descriptions/rpart.html.

18. Agresti A. **An introduction to categorical data analysis**. *Wiley*: NJ, USA, 1996.
19. Breiman L. **Random forests**. *Machine Learning*. 2001; 45: 5–32.
20. Efron B, Tibshirani R. **An introduction to the Bootstrap**. *Chapman & Hall/CRC*: FL, USA, 1994.
21. Kaplan EL, Meier P. **Nonparametric estimation from incomplete observations**. *J Am Stat Assoc*. 1958; 53: 457–481.
22. Hirschhorn JN, Lohmueller K, Byrne E, Hurschhorn K. **A comprehensive review of genetic association studies**. *Genet Med*. 2002; 4: 45–61.
23. Moghaddam MF, Grant DF, Cheek JM, Green JF, Williamson KC, Hammock BD. **Bioactivation of leukotoxins to their toxic diols by epoxide hydrolase**. *Nature Med*. 1997; 3: 562–566.
24. Burchiel SW, Thompson TA, Lauer FT, Oprea TI. **Activation of dioxin response element (DRE)-associated genes by benzo (A) pyrene3,6-quinone and benzo (A) pyrene1,6-quinone in MCF-10A human mammary epithelial cells**. *Toxicol Appl Pharmacol* 2007; 221: 203–214.
25. Shi C-S, Huang N-N, Harrison K, Han S-B, Kehrl JH. **The mitogenactivated protein kinase kinase kinase kinase GCKR positively regulates canonical and noncanonical Wnt signaling in B lymphocytes**. *Mol Cell Biol* 2006; 26: 6511–6521.
26. Robinson JA, Chatterjee-Kishore M, Yaworsky PJ, Cullen DM, ZhaoW LC et al. **Wnt/b signaling is a normal physiological response to mechanical loading in bone**. *J Biol Chem* 2006; 281: 31720–31728.
27. Shi C-S, Tuscano JM, Witte ON, Kehrl JH. **GCKR links the Bcr-Abl oncogene and Ras to the stress-activated protein kinase pathway**. *Blood* 1999; 93: 1338–1345.
28. Knobloch J, Reimann K, Klotz L-O, Ruther U. **Thalidomide resistance is based upon the capacity of the glutathione-dependent antioxidant defense**. *Mol Pharmaceutics* 2008; 5: 1138–1144.
29. Edwards CM, Edwards JR, Lwin ST, Mundy GR. **Target Wnt signaling in myeloma in vivo; differential effects on tumor burden and myeloma bone disease**. *Blood* 2008; 111: 2833–2842.
30. Caspi M, Zilberberg A, Eldar-Finkelman H, Rosin-Arbesfeld R. **Nuclear GSK-3b inhibits the canonical Wnt signalling pathway in a b-catenin phosphorylation-independent manner**. *Oncogene* 2008; 27: 3546–3555.
31. Staal FJT, Luis TC, Tiemessen MM. **WNT signalling in the immune system: WNT is spreading its wings**. *Immunology* 2008; 8: 581–593.
32. Qiang Y-W, Chen Y, Stephens O, Brown N, Chen B, Epstein J et al. **Myeloma-derived Dixkkopf-1 disrupts Wnt-regulated osteoprotegerin and RANKL production by**

osteoblasts: a potential mechanism underlying osteolytic bone lesions in multiple myeloma. *Blood* 2008; 112: 196–207.

33. Qiang Y-W, Shaughnessy JD, Yaccoby S. **Wnt3a signaling within bone inhibits multiple myeloma bone disease and tumor growth**. *Blood* 2008; 112: 374–382.
34. Edward CM. **Wnt signaling: bone's defense against myeloma**. *Blood* 2008; 112: 216–218.
35. Ferlin M, Noraz N, Hertogh C, Brochier J, Taylor N, Klein B. **Insulin-like growth factor induces the survival and proliferation of myeloma cells through an interleukin-6-independent transduction pathway.** *Br J Hematology* 2000; 111: 626–634.
36. Ge N-L, Rudikoff S. **Insulin-like growth factor 1 is a dual effector of multiple myeloma cell growth**. *Blood* 2000; 96: 2856–2861.
37. Mitsiades CS, Mitsiades N, Poulaki V, Schlossman R, Akivama M, Chauhan D et al. **Activation of NF-kB and upregulation of intracellular anti-apoptotic proteins via the IGF-1/Akt signaling in human multiple myeloma calls: therapeutic implications**. *Oncogene* 2002; 21: 5673–5683.
38. Podar K, Tai Y-T, Cole CE, Hideshima T, Sattler M, Hamblin A et al. **Essential role of caveolae in interleukin-6 and insulin-like growth factor I-triggered Akt-1-mediated survival of multiple myeloma cells**. *J Biol Chem* 2003; 278: 5794–5801.
39. DiGirolamo DJ, Mukherjee A, Fulzele K, Gan Y, Cao X, Frank SJ et al. **Mode of growth hormone action in osteoblasts**. *J Biol Chem* 2007; 282: 31666–31674.
40. Brechter AB. **Kinins-important regulators in inflammation induced bone resorption.** *Umea Univ Odontological Dissertation*. Department of Oral Cell Biology, Umea University: Umea, Sweden, 2006 (ISBN 91-7264-195-9).
41. Fujisawa T, Ikegami H, Kawaguchi Y, Ogihata T. **Meta-analysis of the association of Trp Arg polymorphism of b3-adrenergic receptor gene with body mass index**. *J Clin Endocrinol Metab* 1998; 83: 2441–2444.
42. Wang CY, Nguyen ND, Morrison NA, Eisman JA, Center JR, Nguyen TV. **b3-adrenergic receptor gene, body mass index, bone mineral density and fracture risk in elderly men and women: the dubbo osteoporosis epidemiology study (DOES)**. *BMC Med Genet* 2006; 7: 57.
43. Riancho JA, Zarrabeitia MT, Olmos JM, Amado JA, Gonzalez MJ. **Effects of interleukin-4 on human osteoblast-like cells**. *Bone Miner* 1993; 21: 53–61.
44. Durie BGM, Urnovitz HB, Murphy WH. **RT-PCR amplicons in the plasma of multiple myeloma patients, clinical relevance and molecular pathology.** *Acta Oncologica* 2000; 39: 789–796.
45. Hassett C, Robinson KB, Beck NB, Omiecinski CJ. **The human microsomal expoxide hydrolase gene (EPHX1): complete nucleotide sequence and structural characterization**. *Genomics* 1994; 23: 433–442.
46. Fretland AJ, Omiecinski CJ. **Epoxide hydrolases: biochemistry and molecular biology**. *Chem Biol Interact* 2000; 129: 41–59.
47. Omiecinski CJ, Hassett C, Hosagrahara V. **Epoxide hydrolasepolymorphism and role in toxicology**. *Toxicol Lett* 2000; 112–113: 365–370.

48. Berndt SI, Johnson D, Crowley J, Durie BGM, Hoover R, Katz M et al. **Large scale evaluation of genetic variation and the risk of multiple myeloma**. *Blood* 2008; 112, Abstract # 1679.

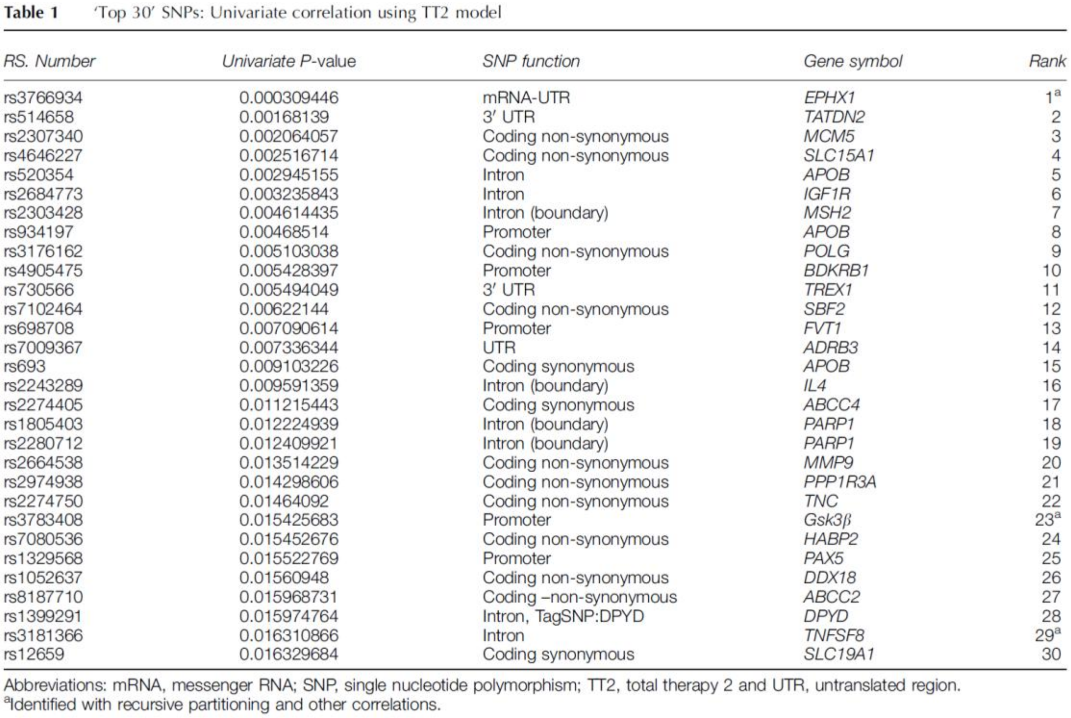


Table 1 'Top 30' SNPs: Univariate correlation using TT2 model

RS. Number	*Univariate P-value*	*SNP function*	*Gene symbol*	*Rank*
rs3766934	0.000309446	mRNA-UTR	*EPHX1*	1[a]
rs514658	0.00168139	3′ UTR	*TATDN2*	2
rs2307340	0.002064057	Coding non-synonymous	*MCM5*	3
rs4646227	0.002516714	Coding non-synonymous	*SLC15A1*	4
rs520354	0.002945155	Intron	*APOB*	5
rs2684773	0.003235843	Intron	*IGF1R*	6
rs2303428	0.004614435	Intron (boundary)	*MSH2*	7
rs934197	0.00468514	Promoter	*APOB*	8
rs3176162	0.005103038	Coding non-synonymous	*POLG*	9
rs4905475	0.005428397	Promoter	*BDKRB1*	10
rs730566	0.005494049	3′ UTR	*TREX1*	11
rs7102464	0.00622144	Coding non-synonymous	*SBF2*	12
rs698708	0.007090614	Promoter	*FVT1*	13
rs7009367	0.007336344	UTR	*ADRB3*	14
rs693	0.009103226	Coding synonymous	*APOB*	15
rs2243289	0.009591359	Intron (boundary)	*IL4*	16
rs2274405	0.011215443	Coding synonymous	*ABCC4*	17
rs1805403	0.012224939	Intron (boundary)	*PARP1*	18
rs2280712	0.012409921	Intron (boundary)	*PARP1*	19
rs2664538	0.013514229	Coding non-synonymous	*MMP9*	20
rs2974938	0.014298606	Coding non-synonymous	*PPP1R3A*	21
rs2274750	0.01464092	Coding non-synonymous	*TNC*	22
rs3783408	0.015425683	Promoter	*Gsk3β*	23[a]
rs7080536	0.015452676	Coding non-synonymous	*HABP2*	24
rs1329568	0.015522769	Promoter	*PAX5*	25
rs1052637	0.01560948	Coding non-synonymous	*DDX18*	26
rs8187710	0.015968731	Coding –non-synonymous	*ABCC2*	27
rs1399291	0.015974764	Intron, TagSNP:DPYD	*DPYD*	28
rs3181366	0.016310866	Intron	*TNFSF8*	29[a]
rs12659	0.016329684	Coding synonymous	*SLC19A1*	30

Abbreviations: mRNA, messenger RNA; SNP, single nucleotide polymorphism; TT2, total therapy 2 and UTR, untranslated region.
[a] Identified with recursive partitioning and other correlations.

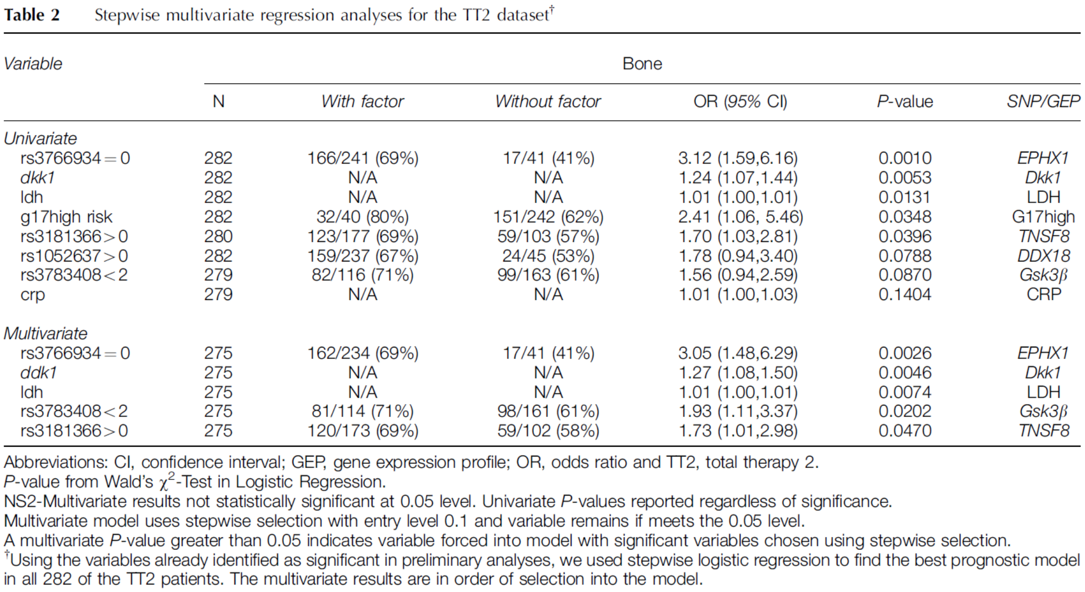


Table 2 Stepwise multivariate regression analyses for the TT2 dataset[†]

Variable	Bone					
	N	*With factor*	*Without factor*	OR (*95% CI*)	*P*-value	*SNP/GEP*
Univariate						
rs3766934 = 0	282	166/241 (69%)	17/41 (41%)	3.12 (1.59,6.16)	0.0010	*EPHX1*
dkk1	282	N/A	N/A	1.24 (1.07,1.44)	0.0053	*Dkk1*
ldh	282	N/A	N/A	1.01 (1.00,1.01)	0.0131	LDH
g17high risk	282	32/40 (80%)	151/242 (62%)	2.41 (1.06, 5.46)	0.0348	G17high
rs3181366 > 0	280	123/177 (69%)	59/103 (57%)	1.70 (1.03,2.81)	0.0396	*TNSF8*
rs1052637 > 0	282	159/237 (67%)	24/45 (53%)	1.78 (0.94,3.40)	0.0788	*DDX18*
rs3783408 < 2	279	82/116 (71%)	99/163 (61%)	1.56 (0.94,2.59)	0.0870	*Gsk3β*
crp	279	N/A	N/A	1.01 (1.00,1.03)	0.1404	CRP
Multivariate						
rs3766934 = 0	275	162/234 (69%)	17/41 (41%)	3.05 (1.48,6.29)	0.0026	*EPHX1*
ddk1	275	N/A	N/A	1.27 (1.08,1.50)	0.0046	*Dkk1*
ldh	275	N/A	N/A	1.01 (1.00,1.01)	0.0074	LDH
rs3783408 < 2	275	81/114 (71%)	98/161 (61%)	1.93 (1.11,3.37)	0.0202	*Gsk3β*
rs3181366 > 0	275	120/173 (69%)	59/102 (58%)	1.73 (1.01,2.98)	0.0470	*TNSF8*

Abbreviations: CI, confidence interval; GEP, gene expression profile; OR, odds ratio and TT2, total therapy 2.
P-value from Wald's χ^2-Test in Logistic Regression.
NS2-Multivariate results not statistically significant at 0.05 level. Univariate *P*-values reported regardless of significance.
Multivariate model uses stepwise selection with entry level 0.1 and variable remains if meets the 0.05 level.
A multivariate *P*-value greater than 0.05 indicates variable forced into model with significant variables chosen using stepwise selection.
[†] Using the variables already identified as significant in preliminary analyses, we used stepwise logistic regression to find the best prognostic model in all 282 of the TT2 patients. The multivariate results are in order of selection into the model.

Table 3 Stepwise multivariate regression analyses incorporating SNPs identified with bone disease classified by X-rays only (0 versus >3 lesions)[a]

Variable	*Bone*					
	N	*With factor*	*Without factor*	OR (*95%* CI)	*P*-value	*SNP/GEP*
Univariate						
rs3766934 = 0	282	166/241 (69%)	17/41 (41%)	312 (1.59,6.16)	0.0010	*EPHX1*
rs730566 < 2	282	172/254 (68%)	11/28 (39%)	3.24 (1.45,7.23)	0.0041	*TREX1*
dkk1	282	N/A	N/A	1.24 (1.07,1.44)	0.0053	*Dkk1*
ldh	282	N/A	N/A	1.01 (1.00,1.01)	0.0131	LDH
g17high	282	32/40 (80%)	151/242 (62%)	2.41 (1.06,5.46)	0.0348	G17high
rs3181366 > 0	280	123/177 (69%)	59/103 (57%)	1.70 (1.03,2.81)	0.0396	*TNSF8*
rs7120118 = 0	281	21/25 (84%)	161/256 (63%)	3.10 (1.03,9.30)	0.0437	*NRIH3*
rs1052637 > 0	282	159/237 (67%)	24/45 (53%)	1.78 (0.94,3.40)	0.0788	*DDX18*
rs3783408 < 2	279	82/116 (71%)	99/163 (61%)	1.56 (0.94,2.59)	0.0870	*Gsk3β*
crp	279	N/A	N/A	1.01 (1.00,1.03)	0.1404	CRP
Multivariate						
rs3766934 = 0	274	161/233 (69%)	17/41 (41%)	2.90 (1.36,6.17)	0.0057	*EPHX1*
rs730566 < 2	274	168/247 (68%)	10/27 (37%)	3.40 (1.38,8.37)	0.0077	*TREX1*
ldh	274	N/A	N/A	1.01 (1.00,1.01)	0.0105	LDH
rs3783408 < 2	274	80/113 (71%)	98/161 (61%)	2.09 (1.17,3.70)	0.0121	*GSK3β*
ddk1	274	N/A	N/A	1.23 (1.04,11.08)	0.0178	*Dkk1*
rs3181366 > 0	274	119/172 (69%)	59/102 (58%)	1.78 (1.02,3.10)	0.0423	*TNSF8*
rs7120118 = 0	274	21/25 (84%)	157/249 (63%)	3.39 (1.04,11.08)	0.0430	*NRIH3*

Abbreviations: CI, confidence interval; GEP, gene expression profile; OR, odds ratio and SNP, single nucleotide polymorphism.

P-value from Wald's χ^2-Test in Logistic Regression.

NS2-Multivariate results not statistically significant at 0.05 level. Univariate *P*-values reported regardless of significance.

Multivariate model uses stepwise selection with entry level 0.1 and variable remains if meets the 0.05 level.

A multivariate *P*-value greater than 0.05 indicates variable forced into model with significant variables chosen using stepwise selection.

[a] Logistic regression on all 282 Total Therapy 2 (TT2) patients. The variables considered in Table 2 together with top two SNPs from the X-ray only analysis are considered.

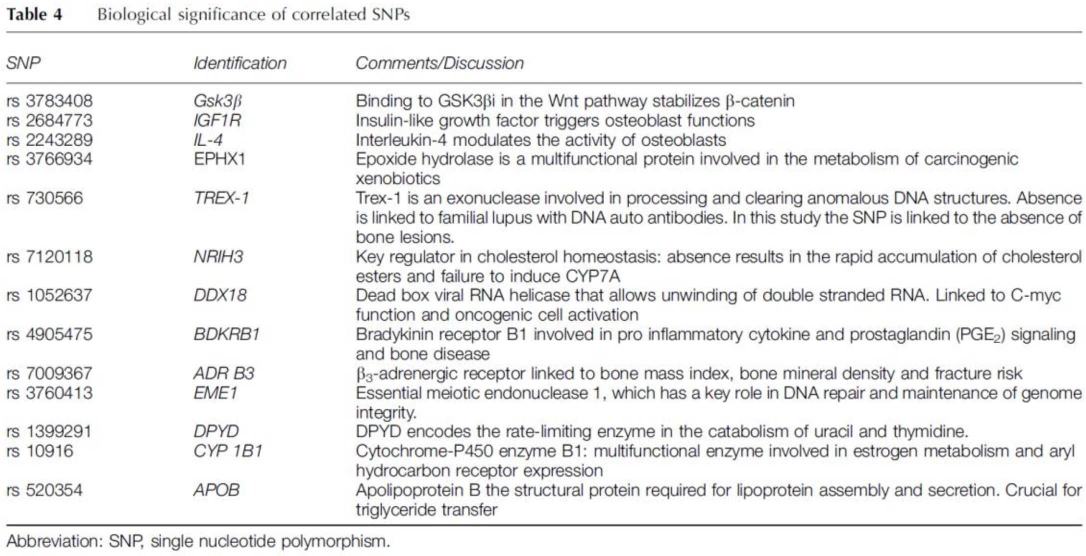
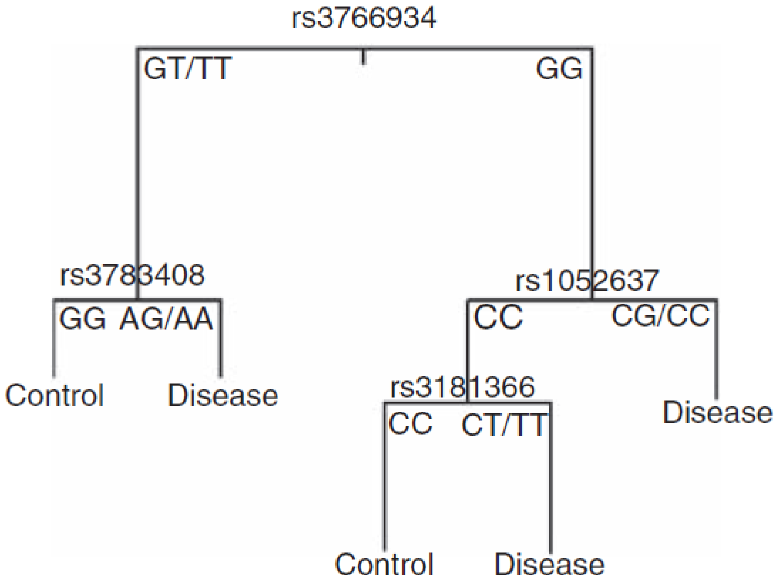


Table 4 Biological significance of correlated SNPs

SNP	*Identification*	*Comments/Discussion*
rs 3783408	*Gsk3β*	Binding to GSK3βi in the Wnt pathway stabilizes β-catenin
rs 2684773	*IGF1R*	Insulin-like growth factor triggers osteoblast functions
rs 2243289	*IL-4*	Interleukin-4 modulates the activity of osteoblasts
rs 3766934	EPHX1	Epoxide hydrolase is a multifunctional protein involved in the metabolism of carcinogenic xenobiotics
rs 730566	*TREX-1*	Trex-1 is an exonuclease involved in processing and clearing anomalous DNA structures. Absence is linked to familial lupus with DNA auto antibodies. In this study the SNP is linked to the absence of bone lesions.
rs 7120118	*NRIH3*	Key regulator in cholesterol homeostasis: absence results in the rapid accumulation of cholesterol esters and failure to induce CYP7A
rs 1052637	*DDX18*	Dead box viral RNA helicase that allows unwinding of double stranded RNA. Linked to C-myc function and oncogenic cell activation
rs 4905475	*BDKRB1*	Bradykinin receptor B1 involved in pro inflammatory cytokine and prostaglandin (PGE_2) signaling and bone disease
rs 7009367	*ADR B3*	β_3-adrenergic receptor linked to bone mass index, bone mineral density and fracture risk
rs 3760413	*EME1*	Essential meiotic endonuclease 1, which has a key role in DNA repair and maintenance of genome integrity.
rs 1399291	*DPYD*	DPYD encodes the rate-limiting enzyme in the catabolism of uracil and thymidine.
rs 10916	*CYP 1B1*	Cytochrome-P450 enzyme B1: multifunctional enzyme involved in estrogen metabolism and aryl hydrocarbon receptor expression
rs 520354	*APOB*	Apolipoprotein B the structural protein required for lipoprotein assembly and secretion. Crucial for triglyceride transfer

Abbreviation: SNP, single nucleotide polymorphism.

Figure 1. Recursive partitioning using ‘Top SNPs’ with Total Therapy 2 (TT2) model. Recursive partitioning branching tree displaying the four single nucleotide polymorphisms (SNPs) used in the model: rs3766934 (EPHX1); rs3783408 (Gsk3b); rs1052637 (DDX18); and rs3181366 (TNSF8). The SNP genotypes are identified: EPHX1(GT/TT versus GG); Gsk3b (GG versus AG/AA); DDX18 (CC versus CG/CC); and TNFSF8 (CC versus CT/TT). The appended table shows the univariate P-values for each SNP and SNP function.



RS Number	Rank	Univariate p-value	SNP Function	Gene Symbol
rs3766934	1	0.000309446	mrna-utr	*EPHX1*
rs3783408	23	0.015425683	promoter	*Gsk3β*
rs1052637	26	0.01560948	coding-nonsynonymous	*DDX18*
rs3181366	29	0.016310866	Intron	*TNFSF8*

Note: Control=Limited : Disease=Extensive

Figure 2. Baseline focal bone lesions and baseline log2 DKK-1 by predicted disease using the recursive-partitioning model. (a) The number of focal bone lesions (per patient) is plotted for patients with limited bone disease and extensive bone disease predicted by the four single nucleotide polymorphism (SNP) model illustrated in Figure 1. The mean values are identified. The P-value for the difference is P=0.001. (b) The directly measured log2 DKK-1 expression values are plotted for patients with limited bone disease and extensive bone disease predicted by the four SNP model illustrated in Figure 1. The Pvalue for the difference is P=0.05.

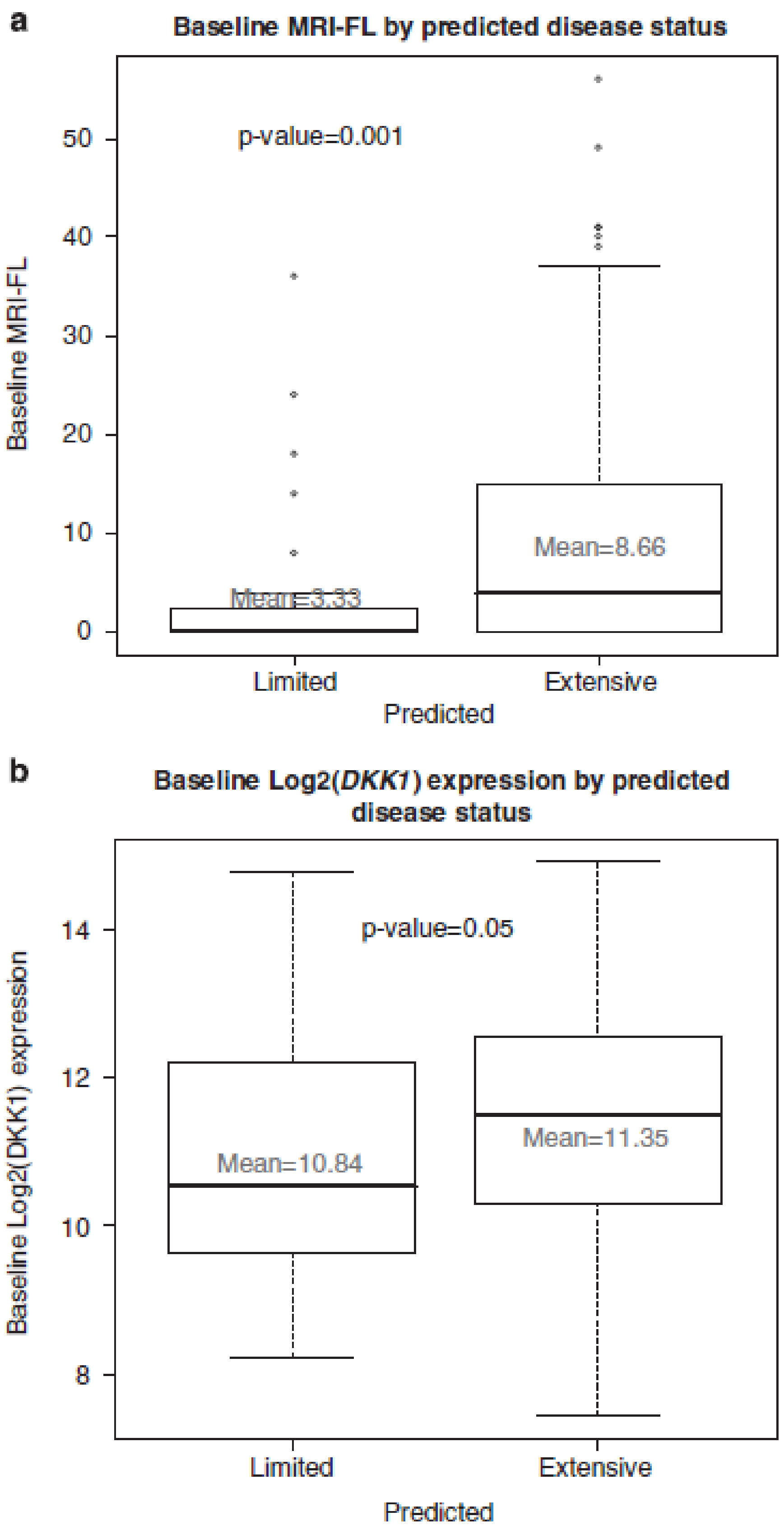
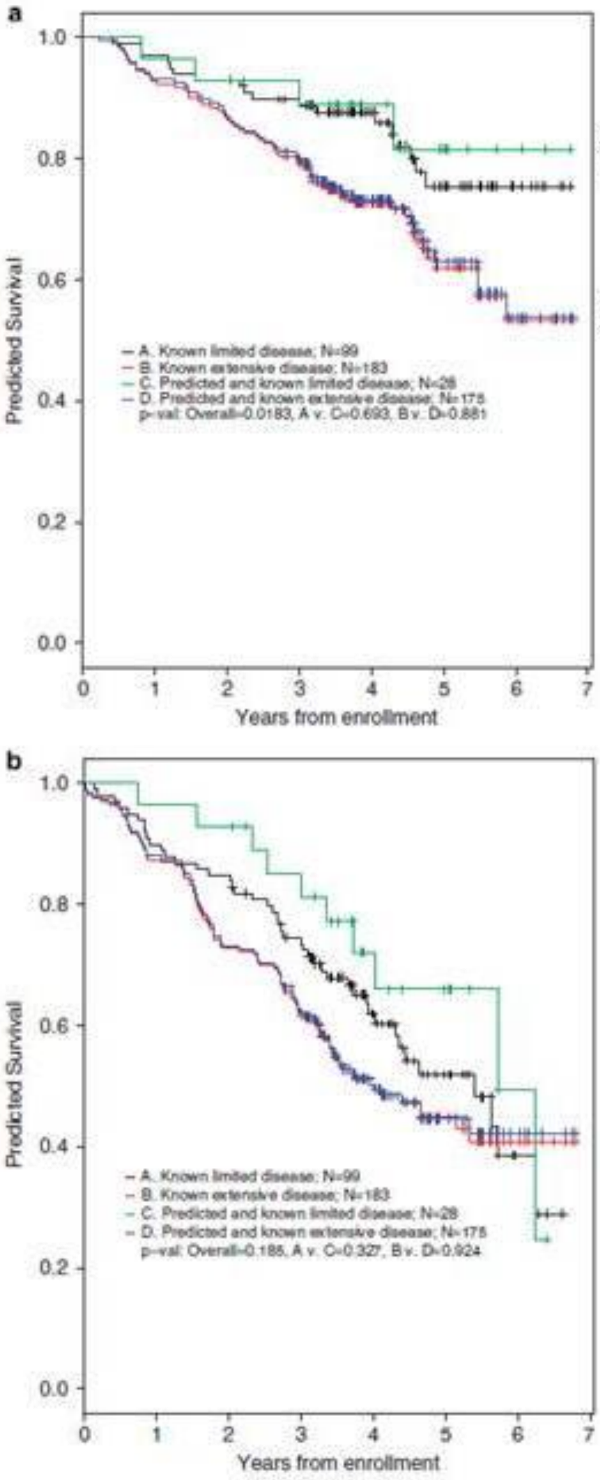

a
Baseline MRI-FL by predicted disease status
p-value=0.001
Baseline MRI-FL
50
40
30
20
10
0
Mean=3.33
Mean=8.66
Limited
Extensive
Predicted
b
Baseline Log2(*DKK1*) expression by predicted disease status
p-value=0.05
Baseline Log2(DKK1) expression
14
12
10
8
Mean=10.84
Mean=11.35
Limited
Extensive
Predicted

Figure 3. Overall survival (OS) and Event-free survival (EFS) for both actual and predicted bone disease (Total Therapy 2 (TT2) model). (a) OS is shown for patients with known limited and extensive bone disease and compared with the survival for patients predicted by the four single nucleotide polymorphism (SNP) model to have limited and extensive disease. The listed P-values indicate that OS is statistically inferior for patients with both actual and predicted extensive versus limited bone disease (P=0.0183). The actual versus predicted outcomes are not different (P-values 0.693 and 0.881, respectively). (b) EFS is shown for patients with known limited and extensive bone disease and compared with EFS for patients predicted to have limited and extensive bone disease based on the four SNP model (Figure 1). The P-values indicate that EFS is not different for limited versus extensive disease, but this is true for both the actual and predicted patient populations (P-values: overall 0.185; and 0.327 and 0.924 for comparisons).



predicted limited bone disease

known limited bone disease

predicted extensive bone disease
known extensive bone disease

predicted extensive bone disease
known extensive bone disease

known limited bone disease
predicted limited bone disease

CHAPTER 3

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISM INTERACTIONS ASSOCIATED WITH SURVIVAL AND RISK IN MULTIPLE MYELOMA USING NOVEL DATA MINING METHODS

Majda Haznadar[1], Gang Fang[2], Wen Wang[2], Vanja Paunic[2], Patrick Day[1], Michael Steinbach[2], Vipin Kumar[2], Brian Durie[3], Bart Barlogie[4], Brian Van Ness[1]

Affiliations:

[1]Institute of Human Genetics, Computer Science and Engineering[2], University of Minnesota, Minneapolis, Minnesota, [3]Cedars-Sinai Medical Center, Los Angeles, CA, [4]University of Arkansas Medical Center, Little Rock, AR, USA

I conducted the genotyping of all 3,404 SNPs in most of the subjects represented in the study, mostly for the samples in the mouthwash BOAC patient-provided samples. I guided the data mining approach designed by our collaborators, and wrote the manuscript. I provided biological interpretations and insights to our data mining collaborators, and helped them direct their analytical design. I conducted data analysis upon receiving the results produced by the algorithms. This is the first application of these novel computational algorithms on myeloma genotyping data.

Backround: Multiple myeloma (MM) is a disease that results in the accumulation of malignant plasma cells in the bone marrow and is characterized by marked genomic heterogeneity. To elucidate genetic complexity of MM by studying interplay of genomic variations, we have developed novel combinatorial search data mining methods, allowing us to search for high order variation and gene interactions.

Methods: We conducted an association study of 3,404 single nucleotide polymorphisms (SNPs) as a whole and constrained the number of SNPs by defined pathways and gene sets, thus biologically guiding and increasing the power of the analysis. Two primary phenotypes investigated are progression-free survival and disease risk.

Results: We identified individual and pairs of interacting SNPs in genes involved in drug metabolism and detoxification, immunity, DNA repair and signaling cascades important to MM risk and survival.

Conclusion: Due to a complex genetic interplay that leads to cancer, as well as a large amount of SNP genotyping data that can be generated, novel methods that successfully identify SNP-SNP interactions and associations with outcomes are needed. We applied a novel computational approach that efficiently identifies high order SNP interactions, while minimizing the false discovery rate.

Introduction:

Multiple Myeloma is a universally fatal disease that results in the accumulation of malignant plasma cells in the bone marrow (1). This disease is responsible for 2% of all cancer deaths and 15% of all hematologic malignancies, with approximately 13,000 deaths per year in the USA (1). In order for MM plasma cells to grow, there is a complex interplay among various growth and other factors in the tumor microenvironment.

Single nucleotide polymorphisms (SNPs) account for over 90% of genetic variation in the human genome and have emerged as important genetic factors shown to impact disease risk, outcome, drug metabolism and other clinical outcomes (2-4). Due to a large amount of data generated by SNP genotyping, researchers have struggled in analyzing and interpreting the results obtained in association studies with outcomes of interest.

Many complex diseases such as cancer are believed to be driven by a complex interplay of multiple genetic variations (5). There is a clear challenge in searching for combinations of SNPs that are associated with disease risk and outcomes from a large number of variations that result in an exponentially increasing number of combinations. This is a particular problem in the small sample numbers in which false discovery rates exceed the power in the study. Therefore, efficient, robust and scalable computational techniques are needed (5).

In order to identify the impact of genetic variation on MM, we engaged in an international program named Bank On A Cure (BOAC), whose mission is to bank DNA from collaborative groups and institutional trials and examine the associations of genetic variations with risk and outcomes (6). This effort was directed at developing a platform for examining polymorphisms in genes relevant in cancer initiation and progression pathways, predominantly targeted to MM, with an expectation that genetic variations would have a large impact on disease initiation, progression and response by altering pathways involved in these processes. The design and the first applications of the BOAC SNP chip containing 3,404 SNPs have been previously published (7).

MM is a complex disease with a wide heterogeneity among the tumors, and various serious complications i.e. bone disease, hypercalcemia, renal failure etc (1). We expected that the observed phenotypes are a result of a multifaceted interaction of variations in genes involved in crucial processes i.e. apoptosis, DNA repair, inflammation, etc. Thus, we developed novel computational methods to observe interplay of defined functional pathways in risk and survival. A manuscript describing the computational details is being published separately (8). In this manuscript we describe the application that has revealed SNP-SNP interactions associated with both outcome and risk.

Materials and Methods:

Patient and control samples

DNA samples were prepared from 143 MM patients enrolled in phase III clinical trials: Eastern Oncology Cooperative Group (ECOG) E9486 (n=52) and Southwest Oncology Group (SWOG) S9321 (n=91); all patients signed informed consent (9, 10). Patients enrolled in E9486 and S9321 had similar treatment regimens using vincristine, adrimycin, dexamethasone and melphlan (9, 10). Patients in the analysis were selected based on progression-free survival (PFS) of less than 1 year (n=70) or greater than 3 years (n=73), short-term and long-term survivors respectively, to give us a binary association analysis.

DNA samples were also prepared from 396 buccal cell mouthwash samples of MM patients and their spousal controls, all having signed informed consent (BOAC Epi dataset). There are 143 patients matched with spousal controls. The rest, 104 MM patients, and 6 controls are non-matched samples. These samples are a part of an ongoing collection in the BOAC program.

Genotyping

All samples were genotyped in the laboratory of one of the authors (Van Ness) using the Affymetrix/Gene Chip Targeted Genotyping Platform (7). Genotyping was performed using the Affymetrix® Gene-Chip® Scanner 3000 Targeted Genotyping System (GCS 3000 TG System), which utilizes molecular inversion probes (11) to simultaneously identify many SNPs. There was a total of 3,404 SNPs genotyped, in 983 genes.

Statistical/Data mining Methods

Fang et al. has recently proposed novel data mining algorithms for searching high-order SNP combinations that are associated with a disease phenotype (8). The algorithms were shown to be computationally more efficient and statistically more powerful than other approaches. The computational details will be reported elsewhere and will be available as a technical report (8). Briefly, methods for searching disease-associated SNP combinations are designed to address the following two key challenges:

The first challenge is to handle the computational complexity of the combinatorial search. Specifically, given thousands of SNPs, there are exponentially increasing number of SNP combinations in the search space, i.e. millions of size-2 combinations, billions of size-3 combinations, etc. To address the computational challenge, the proposed algorithm leverages the recent development on discriminative pattern mining (DPM) (12, 13), a data mining technique that searches for combinations of items (genotypes in our case) that occur with disproportionate frequency in one class versus another (i.e. short term survivors vs. long term survivors or in cases vs. controls). The better computational efficiency of DPM, over existing brute-force approaches such as multifactor dimensionality reduction MDR (14), comes from the systematic pruning of combinatorial search space based on the anti-monotonicity (14) of the objective function (i.e. chi-square statistic). Finding all frequent item-sets in a database is difficult since it involves searching all possible item-sets (item combinations). Since a search for all possible combinations in a given power-set increases exponentially, there is a need to minimize the sets for an efficient search. Efficient search is possible using the anti-monotonicity property (13), which guarantees that for a frequent item-set, all its subsets are frequent and thus for an infrequent item-set, all its supersets must be infrequent. What this means in the case of searching for SNP combinations is that an item-set (in this case a pair of SNPs occurring together) is frequent because the SNPs are frequent and therefore it is possible to efficiently search for those combinations in a given dataset.

Frequency is user-defined and this measure also has an anti-monotonicity property. Mathematical calculations for the aforementioned frequency can be found in the methods report published separately (8).

The second challenge is to improve the statistical power of high-order disease-associated SNP-combination mining. Specifically, due to a relatively low sample size, and a very large number of hypotheses tested (i.e. billions of size-3 SNP-combinations are tested given thousands of SNPs), discovered SNP combinations are not statistically significant after correcting for multiple hypothesis testing (i.e. via permutation test (13)). To address this statistical challenge, Fang et al. (8) further proposed to use known biological gene sets (i.e. pathways) as constraints in the search of SNP combinations. Specifically, the algorithm only allows the SNPs associated with at most one or two biological gene sets (15) to combine with each other. A computational capability of the method allows SNP combinations to be searched within one pathway or gene sets, or between two pathways/gene sets. Such a constraint-based formulation can effectively reduce false positive SNP combinations as well as further improve the computational efficiency, as shown in (8).

Among the discovered statistically significant SNP-combinations, it is specifically interesting to identify those combinations composed of SNPs with weak individual association with the phenotype, but strong association as a combination. For this purpose, Fang et al. (8) proposed to use a measure calculated as a difference between the chi-square statistic of the SNP-combination and the best chi-square statistic from all the subsets of the SNP combinations, i.e. chi-square jump. These SNP combinations may indicate phenotype-associated interactions between the genes that the SNPs affect. We refer to this as *p-value jump* associations.

Results:

There are four separate analyses conducted for each of the two datasets represented in this study, S9321/E9486 investigating progression-free survival, and BOAC Epi mouthwash datasets for myeloma disease risk analysis: 1) univariate on all 3404 BOAC panel SNPs, 2) pair-wise on all 3404 SNPs, 3) *p-value jump* on all 3404 SNPs and 4) a pathway-constraint analysis. All analyses were computed using a chi-square test. Significance was determined by a p-value of less than 0.05 and a false discovery rate (FDR) of less than 25%. The survival analyses were conducted by comparing long term (n=73) to short term survival (n=70) as outcomes. The BOAC Epi mouthwash dataset was analyzed by associating variations to a disease status (disease vs. no disease), thus investigating disease risk as an outcome. We compared cases (n=143) to their spousal controls (n=143), and in a separate analysis the same controls (n=143) to the non-spousal cases (n=104).

Univariate analysis on all 3,404 genotyped SNPs in the survival dataset resulted in one significant SNP, rs2108622 in cytochrome P450, family 4, subfamily F, polypeptide 2, CYP4F2 (p-value=0.00012) (Table 1). Direction of the effect was a higher frequency of the SNP in short term survivors, thus exhibiting a detrimental effect (minor allele frequencies of SNPs reported in tables). The survival pair-wise analysis did not result in any significant results that survived a permutation test.

There are two significant pairs of SNPs in the *p-value jump* analysis on the survival dataset: rs225278 in protein tyrosine phosphatase, receptor type, B, PTPRB and rs1043424 in PTEN induced putative kinase 1, PINK1 (p-value=$10^{\wedge -19}$) and rs1520663 in dihydropyrimidine dehydrogenase, DPYD and rs2020911 in mutS homolog 6 (E. coli), MSH6 (p-value=$10^{\wedge -19}$) (Table 2).

The survival dataset pathway-constraint analysis amounted to four significant pairs of SNPs: rs1549760 in cyclin-dependent kinase 5, CDK5 and rs2075685 in X-ray repair complementing defective repair in Chinese hamster cells 4, XRCC4 (p-value= 6*10$^{\wedge-6}$); rs3448 in glutathione peroxidase 1, GPX1 and rs673197 in glutathione S-transferase alpha 4, GSTA4 (p-value= 4*10$^{\wedge-6}$); rs25406 in proliferating cell nuclear antigen, PCNA and rs610308 in protein phosphatase 1, regulatory (inhibitor) subunit 15A, PPP1R15A (p-value=1.9*10$^{\wedge-5}$); rs762623 in cyclin-dependent kinase inhibitor 1A (p21, Cip1), CDKN1A and rs2108622 in CYP4F2 (p-value=1.7*10$^{\wedge-5}$). All SNP combinations seem detrimental, as they are more prevalent in short-term vs. long-term survivors (Table 3).

BOAC Epi mouthwash dataset univariate analysis comparing patients vs. age matched healthy controls resulted in two significant associations, rs1157745 in paraoxonase 1, PON1 and rs2410558 (intergenic), with p-values of 6.3*10$^{\wedge-7}$ and 6.6*10$^{\wedge-8}$ respectively (Table 4). Both SNPs seem to be protective, as they are more prevalent in the controls. Pair-wise analysis on BOAC Epi analysis amounted to a pair of SNPs distinguishing cases from the controls: rs25648 in vascular endothelial growth factor, VEGF and rs1157745 in PON1, p-value of 4.8*10$^{\wedge-8}$ (Table 5). Significantly, this pair came out in both of the subsets analyzed, matched-cases vs. matched-controls and non-matched cases vs. non-matched controls. A combination of these two SNPs seems protective, with a higher prevalence in the control subset.

P-value jump analysis on BOAC Epi dataset resulted in two pairs of significant SNPs: rs1913263 in microsomal glutathione S-transferase 1, MGST1 and rs1801406 in breast cancer 2, early onset, BRCA2 (p-value=0.00013) and rs1881420 in anaplastic lymphoma receptor tyrosine kinase, ALK and rs2234962 in BCL2-associated athanogene 3, BAG3 (p-value=0.00001). The first pair of SNPs appears detrimental, while the second pair appears protective, according to their prevalence in cases vs. controls (Table 6).

The pathway-constraint analysis also resulted in two significant pairs: rs2305948 in kinase insert domain receptor (a type III receptor tyrosine kinase), KDR and rs1157745 in PON1 (p-value=0.000035) and rs2839686 in chemokine (C-X-C motif) ligand 12, CXCL12 and rs1157745 in PON1 (p-value=0.00001). The first pair appears detrimental, whereas the second appears protective (Table 7). All the significant results survived a permutation analysis, and were corrected for multiple hypothesis testing by FDR, and are thus unlikely to have been associated with outcomes by chance occurrence only. Discriminative pattern mining from dense and high dimensional datasets is computationally challenging because existing techniques cannot effectively prune frequent non-discriminative patterns. By reusing the repeated computation in the permutation tests (8), FDR is reduced and the variations that rise above the random noise are not false positives. With the substantially improved efficiency, the novel algorithms are able to complete a thorough search of pair-wise SNP combinations, and identify those with significant p-values and reduced FDR (8).

Discussion:

Since SNP association studies have become a useful and a common tool in studying genetic associations with outcomes in diseases of interest, researchers have been encountering a problem in a lack of robust computational methods that predict risk markers and eliminate the false positives, especially when sample sizes are small. The novelty in this study lies in the ability to look for high order associations, not only single alleles but also combinations of alleles resulting in gene-gene interaction models with relatively small sample sizes. We were able to build new data mining combinatorial search algorithms for finding SNP-SNP associations by analyzing the entire genotyped data set, and by constraining the dataset to smaller subsets that are constrained by pathway/gene sets, thus increasing the significance of the associations. The *p-value jump* method further predicted SNP-SNP associations, resulting in more significant associations involving a pair vs. each SNP individually.

We have previously published a study involving the survival dataset analyzed in this manuscript. Using novel methods presented in this study, we were able to validate some of our results obtained by univariately ranking SNPs by Fisher's exact test (7). Validations on individual cohorts require a repetition of clinical trials with similar treatments which is highly unlikely, as new treatments are rapidly emerging. Nevertheless, as we demonstrated here, building novel and more robust data mining and computational models will help eliminate false positives and validate previously published results. Notably in our analyses, we had linkage disequilibrium (LD) variants associating with outcomes with similar p-values, thus serving as an internal analysis control. All LD SNPs are expected to be associated in an analysis if one of the variants shows an association. Those SNPs we were able to validate using novel data mining approaches presented in this manuscript are as follows: rs1043424 in PINK1, rs2198622 in CYP4F2 and rs2069456 in CDK5 (rs1549760 in CDK5 significantly associated with survival in this study is in LD with rs2069456 previously associated with survival (7)). These genes are involved in cell cycle and apoptosis, drug metabolism, inflammation and immunity. PINK1 and CDK5 were univariately ranked in our previous publication (7), but shared increased significance when interacted with other SNPs. This is not surprising because heterogeneous diseases such as cancer are initiated and progress as a result of complex variation and gene interplay.

The subjects in the survival clinical trial presented here were treated with a combination of drugs. It is important to consider differences in treatment protocols, as that may influce the association analysis and uncover variants distinguishing differential treatments. However, when multi-drug treatments are utilized, it is difficult to discern whether the associating polymorphic markers are involved in the metabolism of those drugs, and whether they impact the outcome. Relevant to the results in this study, reduced expression of MSH6, which along with a SNP in DPYD associated with short term survival, has been previously linked to a resistance to prednisone, a drug used in this clinical trial (16). Variants in DPYD have been linked to 5-fluorouracil chemotherapeutic toxicity in cancer patients, and could be potentially linked to other drug-related toxicities (17-19). A study reported an association between the variants in double-strand break DNA repair genes XRCC4 and XRCC5 and multiple myeloma (20). Another study demonstrated an association between an increased expression of XRCC4 and multiple myeloma

(21). An association with a variant from our analysis, rs2075685 in XRCC4 and myeloma has not been previously revealed. GSTA4 enzymes are involved in cellular defense against toxic, carcinogenic, and pharmacologically active electrophilic compounds, and have consistently appeared in our analyses associated with survival, in this and our previous study (7).

Alexandrakis et al. have correlated pretreatment PCNA elevated expression with bone marrow microvessel density and plasma cell infiltration (22). A group identified polymorphisms in CDKN1A, among a couple of other genes, associated with severe mucositis in patients treated with a high dose of melphalan, one of the drugs utilized in the survival clinical trial presented here (23). A variant in CDKN1A in our analysis also associated with poor prognosis and shorter survival. Thus, plausible biological associations have been found in genes we have identified in this novel approach.

In the BOAC Epi mouthwash dataset analysis, we investigated genetic variant associations with disease risk, using the same novel data mining approaches as in the survival analysis. Despite a very small dataset, a variant in PON1, a gene encoding paraoxonase 1 enzyme and involved in metabolism and response to exogenous chemicals, shows strong associations. It is now widely believed that interindividual differences in susceptibility to malignancy may be mediated in part through variability in the xenobiotic enzyme system. One group found a significant increase in incidences of a PON1 variant in myeloma cases compared with controls (24). There were a few genes in interaction with PON1 variant rs1157745 that associated with disease risk (data not published); we present only those results that were replicated in two subsets as described in Methods and Results. Vascular endothelial growth factor (VEGF) is involved in angiogenesis and associated with many different types of cancers (25-27), as well as with multiple myeloma. Down-regulation of VEGF signaling has been shown to inhibit myeloma cell growth (28). The FA/BRCA pathway has been implicated in melphalan resistance in multiple myeloma cells (29); we found an interaction between a SNP in BRCA2 and a SNP in MGST1 in the *p-value jump* analysis on BOAC Epi dataset, that together associate with risk. ALK, anaplastic lymphoma kinase, has been shown to become induced upon treatment with arsenic trioxide in multiple myeloma cell lines and may serve an important role in apoptosis (30), but has also been shown to mediate lymphomagenesis and was suggested as a potential therapeutic target (31). Proteasome inhibitors are becoming effective drugs for the treatment of relapsed multiple myeloma and possibly some solid tumors. Bcl-2-associated athanogene 3 (BAG3) is a survival protein that has been shown to be stimulated during cell response to stressful conditions. Caspase-dependent cleavage of BAG3 was shown to become induced by proteasome inhibitors and suggested to facilitate apoptosis in sensitive thyroid cancer cells (32). Co-expression of vascular endothelial growth factor receptor-2 (KDR) and CD133 was shown to characterize endothelial progenitor cells in myeloma, which are considered a pathogenic biomarker and a potential treatment target in myeloma (33). CXCL12 was also shown to be involved in the progression of myeloma. A study suggests a correlation between CXCL12 chemokine and bone marrow angiogenesis in patients with multiple myeloma (34). Thus, our computational approaches show plausible biomarkers associated with MM. Notably, the approaches we have taken are with small sample numbers.

Conclusions

We have successfully designed a novel method for investigating complex interplay among genetic variations and disease prognosis and risk. There is a large need for robust computational methods that are able to elucidate complex genetic interactions while decreasing a potential for false positive errors with limited datasets. As Figure 1 illustrates, results obtained from analyzing small sample sizes rarely reach significance (compare red line to black lines). In this case, *p-value jump* analysis resulted in two significant pairs of SNPs with high Jump, low p-value and a low FDR, as reported in Table 6 (one of the pairs was captured twice in the reverse order, SNP1+SNP2, then SNP2+SNP1, which is why three pairs may be observed in Figure 1).

Jump is a measure that shows a discriminative power of a pair of SNPs versus each SNP individually. Multiple hypothesis correction methods such as Bonferroni may be too conservative. We therefore used FDR as a false positive minimization method (Figure 1B). We were able to identify meaningful associations that rise above the random noise and survive permutation tests. When in lack of independent validation cohorts, permutation tests may be used as a tool for validation. This is especially significant when studying rare disease such as myeloma.

We have successfully identified variations that associate with disease prognosis (progression free survival) and multiple myeloma risk, which may help elucidate initiation, progression and therapeutic value and complications, using novel combinatiorial search data mining methods.

References:

1. Kyle RA, Rajkumar SV: **Plasma cell disorders**. In *Cecil textbook of medicine* 22nd edition. Edited by: Goldman L, Ausiello DA. Philadelphia: W.B. Saunders; 2004:1184-1195.
2. Venter JC, *et al*: **The sequence of the human genome**. *Science*. 2001, 291:1304-1351.
3. Lander ES, from the International Human Genome Sequencing Consortium *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
4. International Human Genome Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
5. Marchini J, Donnelly P, Cardon LR. **Genome-wide strategies for detecting multiple loci that influence complex diseases**. *Nat Genet* 2005; 37(4):413-7.
6. http://myeloma.org/ArticlePage.action?articleId=2099
7. Van Ness B, Ramos CR, Haznadar M, Hoering A, Haessler J, Crowley J, Jacobus S, Oken M, Rajkumar V, Greipp P, Barlogie B, Durie B, Katz M, Atluri G, Fang G, Gupta R, Steinbach M, Kumar V, Mushlin R, Johnson, D, Morgan G: **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med* 2008, **8**:6:26.
8. Fang G, Haznadar M, Wang W, Steinbach M, Van Ness B, Kumar V: **A Computationally Efficient and Statistically Powerful Framework for Searching High-order Epistasis with Systematic Pruning and Gene-set Constraints**. Technical Report 013, Department of Computer Science, University of Minnesota, 2010. (submitted)
9. Oken MM, Leong T, Lenhard RE, Greipp PR, Kay NE, Van Ness B, KeimowitzRM, Kyle RA: **The addition of interferon or high dosecyclophosphamide to standard chemotherapy in the treatmentof patients with multiple myeloma: Phase III EasternCooperative Oncology Group Clinical Trial EST 9486.** *Cancer* 1999, **86(6):**957-968.
10. Barlogie B, Kyle RA, Anderson KC, Greipp PR, Lazarus HM, HurdDD, McCoy J, Moore DF Jr, Sakhil SR, Lanier KS, Chapman RA,Cromer JN, Salmon SE, Durie B, Crowley JC: **Standard chemotherapycompared with high-dose chemoradiotherapy formultiple myeloma: final results of phase III US IntergroupTrial S9321.** *J Clin Oncol* 2006, **24(6):**929-936.
11. Hardenbol P, Baner J, Jain M *et al*. **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nat Biotechnol* 2003; 21(6):673-678.

12. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, Kumar V. **Subspace differential coexpression analysis: problem definition and a general approach**. *Pac Symp Biocomput*. 2010, 145-56.

13. Fang G, Pandey G, Gupta M, Steinbach M, Kumar V. **Mining Low-support Discriminative Patterns from Dense and High-dimensional Data**. Tech report; Department of Computer Science, University of Minnesota. 2009: 011. (Minor revision, IEEE Transactions on Data and Knowledge Engineering, 2010).

14. Hahn LW, Ritchie MD, Moore JH. **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions**. *Bioinformatics*. 2003; 12;19(3):376-82.

15. The Molecular Signature Database http://www.broadinstitute.org/gsea/msigdb/index.jsp

16. Yang JJ, Bhojwani D, Yang W, Cai X, Stocco G, Crews K, Wang J, Morrison D, Devidas M, Hunger SP, Willman CL, Raetz EA, Pui CH, Evans WE, Relling MV, Carroll WL. **Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia**. *Blood*. 2008; 112:4178-83.

17. Gross E, Busse B, Riemenschneider M, Neubauer S, Seck K, Klein HG, Kiechle M, Lordick F, Meindl A. **Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients**. *PLoS One*. 2008; 3: e4003.

18. Kleibl Z, Fidlerova J, Kleiblova P, Kormunda S, Bilek M, Bouskova K, Sevcik J, Novotny J. **Influence of dihydropyrimidine dehydrogenase gene (DPYD) coding sequence variants on the development of fluoropyrimidine-related toxicity in patients with high-grade toxicity and patients with excellent tolerance of fluoropyrimidine-based chemotherapy**. *Neoplasma*. 2009; 56:303-16.

19. Ofverholm A, Arkblad E, Skrtic S, Albertsson P, Shubbar E, Enerbäck C. **Two cases of 5-fluorouracil toxicity linked with gene variants in the DPYD gene**. *Clin Biochem*. 2010; 43:331-4.

20. Hayden PJ, Tewari P, Morris DW, Staines A, Crowley D, Nieters A, Becker N, de Sanjosé S, Foretova L, Maynadié M, Cocco PL, Boffetta P, Brennan P, Chanock SJ, Browne PV, Lawler M. **Variation in DNA repair genes XRCC3, XRCC4, XRCC5 and susceptibility to myeloma**. *Hum Mol Genet*. 2007; 15;16: 3117-27.

21. Roddam PL, Allan JM, Dring AM, Worrillow LJ, Davies FE, Morgan GJ. **Non-homologous end-joining gene profiling reveals distinct expression patterns associated with lymphoma and multiple myeloma**. *Br J Haematol*. 2010; 149:258-62.

22. Alexandrakis MG, Passam FH, Pappa CA, Dambaki C, Sfakiotaki G, Alegakis AK, Kyriakou DS, Stathopoulos E. **Expression of proliferating cell nuclear antigen (PCNA) in multiple myeloma: its relationship to bone marrow microvessel density and other factors of disease activity**. *Int J Immunopathol Pharmacol*. 2004; 17: 49-56.

23. Dumontet C, Landi S, Reiman T, Perry T, Plesa A, Bellini I, Barale R, Pilarski LM, Troncy J, Tavtigian S, Gemignani F. **Genetic polymorphisms associated with outcome in multiple myeloma patients receiving high-dose melphalan**. *Bone Marrow Transplant*. 2009. [Epub ahead of print]

24. Lincz LF, Kerridge I, Scorgie FE, Bailey M, Enno A, Spencer A**. Xenobiotic gene polymorphisms and susceptibility to multiple myeloma**. *Haematologica*. 2004; 89:628-9.

25. Liang J, Wang H, Xiao H, Li N, Cheng C, Zhao Y, Ma Y, Gao J, Bai R, Zheng H. **Relationship and prognostic significance of SPARC and VEGF protein expression in colon cancer**. J *Exp Clin Cancer Res*. 2010; 29(1):71. [Epub ahead of print]

26. Zhou Y, Li G, Wu J, Zhang Z, Wu Z, Fan P, Hao T, Zhang X, Li M, Zhang F, Li Q, Lu B, Qiao L. **Clinicopathological significance of E-cadherin, VEGF, and MMPs in gastric cancer**. *Tumour Biol*. 2010 [Epub ahead of print]

27. Donnem T, Al-Shibli K, Andersen S, Al-Saad S, Busund LT, Bremnes RM. **Combination of low vascular endothelial growth factor A (VEGF-A)/VEGF receptor 2 expression and high lymphocyte infiltration is a strong and independent favorable prognostic factor in patients with nonsmall cell lung cancer**. *Cancer*. 2010 [Epub ahead of print]

28. Zhang L, Hu Y, Sun CY, Li J, Guo T, Huang J, Chu ZB. **Lentiviral shRNA silencing of BDNF inhibits in vivo multiple myeloma growth and angiogenesis via down-regulated stroma-derived VEGF expression in the bone marrow milieu**. *Cancer Sci*. 2010; 101:1117-24.

29. Chen Q, Van der Sluis PC, Boulware D, Hazlehurst LA, Dalton WS. **The FA/BRCA pathway is involved in melphalan-induced DNA interstrand cross-link repair and accounts for melphalan resistance in multiple myeloma cells**. *Blood.* 2005; 106:698-705.

30. Wang M, Liu S, Liu P. **Gene expression profile of multiple myeloma cell line treated by arsenic trioxide**. *J Huazhong Univ Sci Technolog Med Sci*. 2007; 27:646-9.

31. Chiarle R, Simmons WJ, Cai H, Dhall G, Zamo A, Raz R, Karras JG, Levy DE, Inghirami G. **Stat3 is required for ALK-mediated lymphomagenesis and provides a possible therapeutic target**. *Nat Med*. 2005 Jun;11(6):623-9. Epub 2005 May 15.

32. Du ZX, Meng X, Zhang HY, Guan Y, Wang HQ. **Caspase-dependent cleavage of BAG3 in proteasome inhibitors-induced apoptosis in thyroid cancer cells**. *Biochem Biophys Res Commun*. 2008; 369:894-8.

33. Zhang H, Vakil V, Braunstein M, Smith EL, Maroney J, Chen L, Dai K, Berenson JR, Hussain MM, Klueppelberg U, Norin AJ, Akman HO, Ozçelik T, Batuman OA. **Circulating endothelial progenitor cells in multiple myeloma: implications and significance**. *Blood*. 2005; 105:3286-94.

34. Martin SK, Diamond P, Williams SA, To LB, Peet DJ, Fujii N, Gronthos S, Harris AL, Zannettino AC. **Hypoxia-inducible factor-2 is a novel regulator of aberrant CXCL12 expression in multiple myeloma plasma cells**. *Haematologica*. 2010; 95:776-84.

Table 1. Top SNPs in the progression-free survival analysis univariately ranked

SNP ID	Gene ID	Function	OR	p-value	MAF
rs2108622	CYP4F2	coding-nonsyn	3.6	0.00012	23.3%

OR=odds ratio

MAF=minor allele frequency

Table 2. Top SNPs in the progression-free survival analysis computed by *p-value jump*

SNP ID 1	SNP ID 2	Gene ID 1	Gene ID 2	Function 1	Function 2	p-value of jump	MAF 1	MAF 2
rs2252784	rs1043424	PTPRB	PINK1	coding-nonsyn	coding-nonsyn	$10^{\wedge-19}$	23.3%	34.2%
rs1520663	rs2020911	DPYD	MSH6	intron	intron	$10^{\wedge-19}$	39.8%	39.2%

OR=odds ratio

MAF=minor allele frequency

Table 3. Top SNPs in the progression-free survival analysis computed by pathway-constraints

SNP ID 1	SNP ID 2	Gene ID 1	Gene ID 2	Function 1	Function 2	OR	p-value	MAF 1	MAF 2
rs1549760	rs2075685	CDK5	XRCC4	nearGene-5	nearGene-5	6.3	$6*10^{\wedge-6}$	21.2%	44.2%
rs3448	rs673197	GPX1	GSTA4	nearGene-5	intron	11.3	$4*10^{\wedge-6}$	30.0%	44.4%
rs25406	rs610308	PCNA	PPP1R15A	intron	coding-nonsyn	9.9	$1.9*10^{\wedge-5}$	40.8%	31.7%
rs762623	rs2108622	CDKN1A	CYP4F2	nearGene-5	coding-nonsyn	4.4	$1.7*10^{\wedge-5}$	13.6%	23.3%

OR=odds ratio

MAF=minor allele frequency

Table 4. Top SNPs in the disease risk analysis univariately ranked

SNP ID	Gene ID	Function	OR	p-value	MAF
rs1157745	PON1	intron	3.6	$6.3*10^{\wedge-7}$	36.4%
rs2410558		intergenic	2.4	$6.6*10^{\wedge-8}$	30.8%

OR=odds ratio

MAF=minor allele frequency

Table 5. Top SNPs in the disease risk analysis ranked by the pair-wise method

SNP ID 1	SNP ID 2	Gene ID 1	Gene ID 2	Function 1	Function 2	OR	p-value	MAF 1	MAF 2
rs25648	rs1157745	VEGF	PON1	coding-syn	intron	8.5	4.8*10$^{\wedge -8}$	16.4%	36.4%

OR=odds ratio

MAF=minor allele frequency

Table 6. Top SNPs in the disease risk analysis computed by *p-value jump*

SNP ID 1	SNP ID 2	Gene ID 1	Gene ID 2	Function 1	Function 2	OR	p-value of jump	MAF 1	MAF 2
rs1913263	rs1801406	MGST1	BRCA2	nearGene-5	coding-syn	3	0.00013	11.1%	29.4%
rs1881420	rs2234962	ALK	BAG3	coding-nonsyn	coding-nonsyn	3.2	0.0003	25.4%	20.8%

OR=odds ratio

MAF=minor allele frequency

Table 7. Top SNPs in the disease risk analysis computed by pathway-constraints

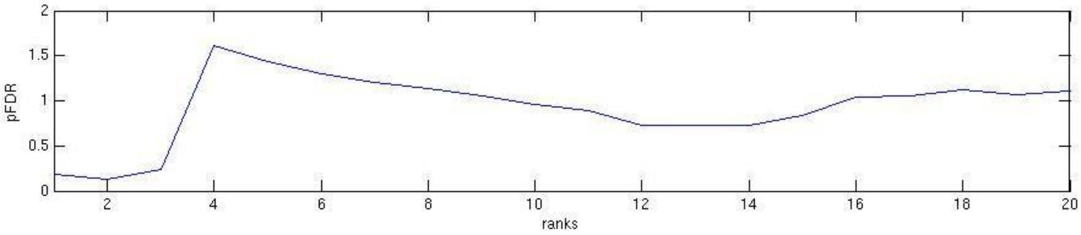
SNP ID 1	SNP ID 2	Gene ID 1	Gene ID 2	Function 1	Function 2	OR	p-value	MAF 1	MAF 2
rs2305948	rs1157745	KDR	PON1	coding-nonsyn	intron	2.7	0.000035	6.7%	36.4%
rs2839686	rs1157745	CXCL12	PON1	nearGene-5	intron	4.1	0.00001	24.6%	36.4%

OR=odds ratio

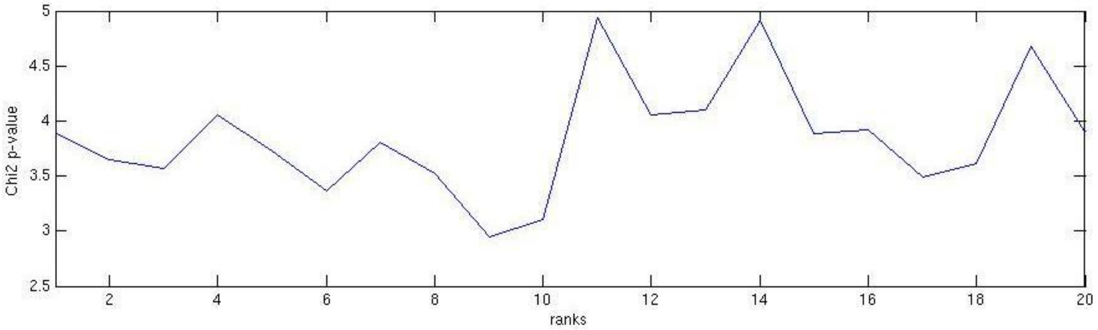
MAF=minor allele frequency

Figure 1. *P-value jump* graphical illustration on BOAC Epi dataset. A) Red line represents pairs of SNPs ranked by their significance based on true class labels; black lines represent 100 permutations of randomly assigned class labels. Y-axis represents Jump generated using chi-square test, a measure that shows a discriminative power of a pair of SNPs vs. each SNP individually; formula: pair_chi2_jump = pair_chi2 - max(individual chi2) / max(individual chi2); X-axis, ranks, represents pairs of SNPs: two highest ranked SNPs (ranks 1 and 2) represent the two statistically significant SNPs shown in Table 6. One of the pairs was captured twice in a reverse order, which is why 3 significant pairs may be observed in the figure. B) False discovery rate (FDR), where a dip in the line to the far left represents low FDR in the two pairs of statistically significant SNPs; formula: pFDR (X) = (number of (random_jumps >= X) / N) / number of (real_jumps >= X); where N is number of random datasets (datasets with shuffled class labels, in this case N=100). pFDR on the Y-axis was computed using DiffSup, measure that represents a difference of the SNP pair's frequency in cases vs. controls. DiffSup ranges from 0 to 1; 0 corresponds to a SNP pair equally present in both classes, while 1 corresponds to a SNP pair present in one class only. Higher DiffSup corresponds to a more significant result. C) p-values of each pair (Y-axis= -log10 p-value) computed using chi-square test. Low p-value in addition to a low pFDR and a high Jump represent significant pairs of SNPs that distinguish cases from controls. One may observe more significant p-values along the graph to the right; those SNP pairs are, however, statistically not significant due to a high pFDR and a low Jump. Detailed computational methods can be found in (8).

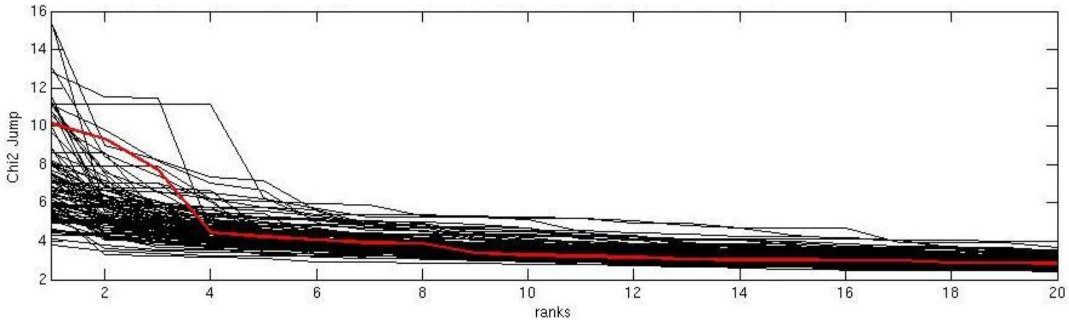
A)


pFDR
0
0.5
1
1.5
2
2
4
6
8
10
12
14
16
18
20
ranks

B)


Chi2 p-value
2.5
3
3.5
4
4.5
5
2
4
6
8
10
12
14
16
18
20
ranks

C)


Chi2 Jump
2
4
6
8
10
12
14
16
2
4
6
8
10
12
14
16
18
20
ranks

CHAPTER 4

INTERACTION OF CYP1B1, CIGARETTE-SMOKE CARCINOGEN METABOLISM, AND LUNG CANCER RISK

Timothy R. Church[1*], Majda Haznadar[2*], Mindy S. Geisser[1], Kristin E. Anderson[3], Neil E. Caporaso[4], Chap Le[5], Salwan B. Abdullah[2], Stephen S. Hecht[6], Martin M. Oken[7], Brian Van Ness[2]

Affiliations:

Divisions of [1]Environmental Health Sciences, [3]Epidemiology and Community Health, and [5]Biostatistics, University of Minnesota School of Public Health, [2]Institute of Human Genetics, [6]Masonic Cancer Center, University of Minnesota, and [7]Division of Hematology and Oncology, Department of Medicine, University of Minnesota Medical School, Minneapolis, Minnesota, USA; and [4]Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

*First co-authors

Disclosure of Potential Conflicts of Interest: The authors declare they have no competing financial interests in this study.

Grant Support: U.S. National Cancer Institute, NIH, Department of Health and Human Services (contract number N01-CN-25513); NIH (grant DA-13333).

I processed all the samples, from DNA extraction to genotyping using BOAC SNP panel and Sequenom. I guided the statistical analysis by having weekly meetings with Dr. Tim Church and Ms. Mindy Geisser, who analyzed the data. I was primarily responsible for coordinating statistical analysis by Dr. Church's group, and biochemical analysis by Dr. Hecht's group. I biologically interpreted the results and guided the way data was being handled and processed. I co-first authored the manuscript with Dr. Tim Church.

A previously published case-control study nested in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial found a significant relationship of serum levels of total NNAL (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol and its glucuronides) to prospective lung cancer risk. The present paper examines this relationship in the context of single-nucleotide polymorphisms (SNPs) in genes important in the metabolism of tobacco smoke carcinogens. DNA was extracted from the subjects' lymphocytes and analyzed for SNPs in 11 locations on four genes related to tobacco carcinogen metabolism. Logistic regressions on case-control status were used to estimate main effects of SNPs and biomarkers and their interactions adjusting for potential confounders. Of the 11 SNPs, only one, in CYP1B1, significantly interacted with total NNAL affecting risk for lung cancer. At low NNAL levels, the variant appeared protective. However, for those with the minor variant, the risk for lung cancer increased with increasing NNAL five times as rapidly compared to those without it, so that at high NNAL levels, this SNP's protection disappears. Analyzing only adenocarcinomas, the effect of the variant was even stronger, with the risk of cancer increasing six times as fast. A common polymorphism of CYP1B1 may play a role in the risk of NNK, a powerful lung carcinogen, in the development of lung cancer in smokers.

Introduction:

Among the multiple carcinogens in cigarette smoke, tobacco-specific nitrosamines such as 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and polycyclic aromatic hydrocarbons (PAH) are widely regarded as important causes of lung cancer, which kills an average of 3000 people per day in the world [1-3]. NNK and PAH require metabolic activation to exert their carcinogenic effects through the formation of DNA adducts which can cause mutations in critical growth control genes, leading ultimately to genomic instability and lung cancer [4]. There are competing detoxification reactions which lead to harmless excretion of NNK and PAH metabolites. Multiple cytochrome P450 enzymes and Phase II enzymes are involved in the metabolic activation and detoxification of NNK and PAH [5-7]. Single nucleotide polymorphisms (SNPs) in these enzymes could affect the balance of metabolic activation and detoxification in a given smoker, thus altering lung cancer risk upon exposure to NNK and PAH in cigarette smoke.

Previously, we reported the first investigation of the relationship between lung cancer and biomarkers of NNK and PAH exposure, using a nested case control design embedded in the National Cancer Institute-sponsored Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial [8]. We found that serum levels of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol and its glucuronides (total NNAL), an established biomarker of NNK exposure [4], were significantly related to lung cancer risk in smokers. In the same study, we also examined the relationship to lung cancer risk of r-1,t-2,3,c-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene (PheT), a metabolite of the PAH phenanthrene [9,10], but found no significant effect.

In the study reported here, we have examined the joint effects of SNPs in several enzymes involved in carcinogen metabolism and the biomarkers total NNAL and PheT as risk factors for lung cancer in the PLCO study. We report an unexpectedly strong effect of a CYP1B1 polymorphism interacting with total NNAL to affect lung cancer risk.

Materials and Methods:

Parent study

The PLCO is an NCI-funded multi-center, randomized, prospective trial of screening for cancers of the prostate, lung, colorectum and ovaries that began in 1993 and is projected to end in 2011 [11]. The screening in the trial includes 77,468 men and women, of whom approximately 25,000 are current or former smokers. In addition to annually screening participants and carefully abstracting cancers from medical records, the PLCO has prospectively collected extensive information from study participants, including smoking history, family history of cancer, and demographic information collected at randomization; and it maintains a bio-repository of blood samples drawn over six annual screening visits starting in 1993. The PLCO trial made available its prospectively collected blood samples from the first screening visit and its extensive baseline and clinical data, thus providing for the direct calculation of lung cancer risks in the groups with different baseline levels of biomarkers. In addition, in the PLCO screening cohort nearly all cases of lung cancer had been screened at least once and so the variability in diagnostic lead times and the potential confounding that variability can produce in unscreened or partially screened cohorts was substantially reduced. At the time our study was initiated, over 800 lung cancer cases had been diagnosed in the screening arm of the PLCO. We randomly selected cases and controls from subjects who reported currently smoking at least 10 cigarettes per day on the baseline questionnaire filled out at the time of randomization.

The PLCO was approved by the institutional review boards of each participating institution, and all subjects signed consents permitting the research represented here.

Case-control Study

We used a nested case-control approach wherein the source cohort consisted of PLCO participants who at randomization filled out a baseline questionnaire indicating they were free of cancer and currently smoking at least 10 cigarettes per day, and who contributed adequate blood samples to the biorepository. Cases were those smokers subsequently diagnosed with lung cancer and controls were those smokers with no diagnosis of lung cancer before the cut-off date (August 17, 2007). From this cohort, we randomly selected 100 incident lung cancer cases and 100 controls and obtained their demographic and other baseline data from the PLCO database, as well as serum samples adequate to measure total cotinine, total NNAL, and PheT. The intent was to determine whether or not biomarker levels in lung cancer subjects differ from those in non-lung cancer subjects. We hypothesized that higher levels of tobacco carcinogens and their specific metabolites among long-term current smokers predispose them to higher risks of developing lung cancer. We did not match on any characteristics, choosing to control for age, sex, family history, and smoking exposure by post-adjustment; this avoided over-matching and allowed us to examine the risks associated with these factors in comparison to those for the biomarkers [12]. Since all subjects were current smokers at baseline, no adjustment for time since quitting was necessary.

Choice of SNPs

The original protocol called for four non-synonymous SNPs to be analyzed (Table 1, asterisked SNPs). These SNPs were selected because they a) could be related to the metabolism of tobacco smoke carcinogens [10], b) had reported allele frequencies that afforded at least 80% power to detect an odds ratio (OR) of 1.5, and c) were on a custom BOAC SNP chip panel for the Affymetrix/Gene Chip Targeted Genotyping Platform [13] developed at the University of Minnesota by two of the authors [13]. After the study was approved by the PLCO, these four SNPs were augmented with seven additional common non-synonymous SNPs considered potentially important in carcinogen metabolism, based on the literature [10] but not appearing on the BOAC chip (Table 1). We genotyped coding-nonsynonymous SNPs in other genes involved in tobacco smoke carcinogen metabolism (CYP1A1, CYP2A13, CYP2A6). However, genotyping of those SNPs resulted in distributions which could not be analyzed (i.e., all homozygous major calls), and therefore are not included.

Tissue Samples from the PLCO Biorepository

For the first screening visit in the PLCO study, participants were asked to provide blood samples adequate for 10 ml of serum, 4 ml of plasma, 2 ml of red blood cells, and 2 ml of “buffy coat” (lymphocytes). These samples were stored at the central biorepository facility in Frederick, Maryland.

Laboratory Methods

Serum Biomarkers

The methods for assaying total NNAL and PheT in blood samples have been previously published [14]. Total cotinine (free cotinine plus cotinine N-glucuronide) concentration in serum was quantified by gas chromatography-mass spectrometry. The method was similar to that used previously to analyze urinary cotinine [15].

DNA Extraction

For some of the subjects, DNA was already isolated and provided by the PLCO. For the remainder, lymphocytes were requested and DNA extracted using Qiagen FlexiGene DNA kit (250) from buffy coat preparations provided by the PLCO.

Genetic Analysis

A directed, custom SNP chip design was developed at the University of Minnesota, and contains functionally relevant polymorphisms playing a role in normal and abnormal cellular functions, inflammation and immunity, and drug responses. The design, quality controls, and platform have been described [13]. While the full SNP panel consists of 3,404 SNPs in ~1,000 genes, in this study we focused initially only on 4 SNPs related to carcinogen metabolism. Further analysis of the total SNP pool will be reported elsewhere. Genotyping was performed using the Affymetrix® Gene-Chip® Scanner 3000 Targeted Genotyping System (GCS 3000 TG System), which utilizes molecular inversion probes [16] to simultaneously identify many SNPs.

In order to add coverage of relevant metabolic genes and SNPs, we selected an additional 7 SNPs from genes involved in NNK and phenanthrene metabolism (SNPs with frequencies in the controls too low to allow analysis are excluded). The genotyping was performed at the Genotyping Facility, part of the BioMedical Genomics Center, at the University of Minnesota, using the Sequenom platform. Among all assays, 14% of the samples were repeated, with an average repeatability of 99.8% concordance in SNP calls.

HPLC of H3-NNK Incubated Lymphoblastoid Cell Lines

Lymphoblastoid cell lines obtained from the Coriell Cell Repositories were established by Epstein-Barr Virus transformation of peripheral blood mononuclear cells using phytohemagluttinin as a mitogen. Six cell lines were selected based on their genotypes of CYP1B1N453S in order to investigate the variation's impact on the metabolism of NNK and NNAL, if any. Two cell lines of each genotype (for a total of six cell lines) were analyzed. Due to the missing information on the kinetics and the involvement of CYP1B1 in NNK/NNAL metabolism, two different NNK concentrations (low=0.092 μM and high=100.1 μM) and two different incubation times (2 hrs and 6 hrs) were selected. Cells were incubated in a sodium bicarbonate assay buffer (pH 7.4). The NNK metabolic activity was determined by radioflow HPLC. The HPLC column used was a Phenomenex Gemini C18 column (5 μm, 250 x 4.60 mm) eluted with a gradient from 100% A [20 mM sodium phosphate (pH 7) containing 1mM sodium EDTA] to 70% A over 30 min, and then to 50% A in 10 min; B was 100% acetonitrile. The eluant flow rate was 0.5 ml/min, scintillant [Monoflow, National Scientific, Rockwood, TN]. The [5-^{3}H] NNK was purchased from Moravek, Brea, CA (specific activity of 21.7 Ci/mmol). The standard metabolites were detected by UV absorbance at 254 nm. Cell metabolism was stopped with 300 μL of 100% acetonitrile, and samples were reconstituted with deionized water before injection onto the HPLC column.

Statistical Methods

The sample size for the original case-control study of biomarkers was determined to provide 80% power with 5% type I error rate to detect an OR of 1.5 for 1 standard-deviation difference in serum biomarker level. Similar power and type I error for that OR are obtained for a SNP that occurs in at least 45% of the population; for lower frequencies, the power is smaller.

Standard descriptive analyses were conducted on all continuous and discrete variables. The logistic regression on case-control status from the previous analysis - originally based on a hypothesized causal diagram and adjusting the biomarker effects for potential confounders [17] and for associated covariates to improve power - was modified to include the SNP data for this analysis. The original regression included sex, age at randomization, family history of lung cancer, cotinine, total NNAL, PheT, and years of cigarette smoking. Untransformed biomarker measurements were used in the regression, as the distributions were reasonably symmetric. We augmented this regression by including each SNP and its interaction with total NNAL (separate regressions with each SNP and its interaction with PheT were also performed, but are not presented here as none of them were significant). Parameter estimates and corresponding odds ratios with 95% Wald confidence limits were estimated and intervals excluding 1 were considered statistically significant. Joint significance of both main effect and interaction parameters were tested using the likelihood ratio test. Graphic display of interaction effects were computed as smoothed averages of case/control status indicators across values of log total NNAL by genotype subgroups to compare visually non-parametrically estimated dose-response curves. Further exploratory regressions were done using the same dataset on histological subtypes of the lung cancers. All computations were done in SAS v. 9.1 (SAS Institute, Inc., Cary, NC, USA) for Windows XP OS (Microsoft, Inc., Redmond, WA, USA).

Results:

Table 2 gives the distributions by case/control status and tests of significance for categorical variables, including age, sex, race/ethnicity, education, marital status, occupation, family history of lung cancer, the usual number of cigarettes smoked per day and the frequency of the genotypes for each of the eleven SNPs and Table 3 gives descriptive statistics by case/control status and tests of significance for continuous variables, including duration of smoking and measured serum levels of total cotinine and total NNAL and PheT. Only age ($p = 0.0039$), years of smoking ($p < 0.0001$), and serum level of total NNAL ($p = 0.0084$) were statistically significantly associated with lung cancer risk. Although not to a statistically significant degree, total cotinine and PheT differed in the direction expected between cases and controls (Table 3). None of the selected SNPs showed significant, independent association with case/control status.

When we estimated the association of CYP1N543S and lung cancer risk, and the SNP's interaction with total NNAL, while adjusting for PheT, total cotinine and potential confounders, the interaction effect was statistically significant in the logistic regression (OR = 1.020, 95% confidence interval = 1.002, 1.038) (Table 4a); this indicated that the SNP modifies the effect of NNAL on risk of lung cancer. At lower levels of NNAL the SNP appeared to be protective, although there was no significant difference between the risk of those with the SNP and those without it at an average NNAL exposure. This is because the protective effect diminishes as NNAL levels approach the mean and disappears altogether by the time the NNAL level reaches its average. The test of the joint effect of the main and interaction effects for CYP1B1N453S was marginally significant (χ_2^2 = 5.888, p = 0.0527).

In prior analyses without the SNP data [8], we estimated that each standard deviation (40 fmol/ml) increase in total NNAL is associated with an approximate 57% increase in lung cancer risk (95% CI: 8%, 128%). The regression that includes the CYP1B1N453S interaction indicates that subjects with at least one minor variant allele exhibit a different effect of NNAL on risk from that among those with both major alleles. For those with the minor allele, each standard deviation increase in NNAL increases lung cancer risk by 170%, more than three times the originally estimated effect. For those without a minor allele, the estimated increase in risk associated with a standard deviation increase in NNAL is 22%, about an eighth the increase in the minor allele group. Because the frequency of the minor allele is about 1/3, the overall rate of increase averages to the 57% shown in the previous paper [8].

The graph in Figure 1 estimates the trend for the unadjusted relationship between log total NNAL concentration and case/control status for subjects with at least on minor allele at CYP1B1N453S (black points and line) and for those without (red points and line). These trends are estimated non-parametrically by a weighted moving average between the cases, plotted on the vertical axis at y = 1, and the controls, plotted at y = 0. This graph is consistent with the logistic regression results and clearly shows a lower risk at the lower levels of NNAL and approximately the same risk at average and higher levels. Notably, CYP1N453S minor allele shows a cluster of cases at high NNAL levels (black dots). The SNP shows protection at low NNAL levels, whereas the protection decreases with increasing NNAL levels. We cannot determine with statistical significance whether the protection becomes completely lost at high NNAL levels.

To further investigate the effect of the CYP1B1N453S, the cancers were grouped into adenocarcinomas (N = 59) vs. non-adenocarcinomas (N = 41) and analyzed separately. The main effect of CYP1B1N453S and its interaction effect with total NNAL (Table 4b) were jointly statistically significant (χ_2^2 = 9.068, p = 0.0107) for the adenocarcinoma group and both of higher magnitude than the effects including all lung cancers (OR = 0.415, 95% confidence interval = 0.157, 1.098; and OR = 1.031, 95% confidence interval = 1.006, 1.056, respectively). The main and interaction effects were not statistically significant for non-adenocarcinomas as a group.

In the logistic regressions for the other 10 SNP (not shown) neither the main nor the interaction effects were significant.

In order to understand the potential effect of the SNP on total NNAL levels themselves, we calculated means, standard deviations, and ranges for total NNAL levels by case/control and CYP1B1N453S status (Table 5). Note that for homozygous major alleles, the total serum NNAL level for cases is only 2.5 fmol/ml higher than for controls, but those cases carrying at least one minor allele of CYP1B1N453S had a total serum NNAL level 42.5 fmol higher than the controls. This result is highly statistically significant, and consistent with the findings of the logistic regression. This reflects the greater impact of NNAL on lung cancer risk among those with the minor allele of CYP1B1. Notably, when this analysis is narrowed to the adencarcinoma cases, the impact of carrying a minor allele becomes even stronger, and the total serum NNAL level is 52.4 fmol/ml higher for cases than for controls.

CYP1B1 has never been shown to have an enzymatic activity in the metabolism of NNAL and until now its involvement in this pathway had never been tested. Coriell lymphoblastoid cell lines have CYP1B1 activity (http://hapmap.ncbi.nlm.nih.gov/) and the naturally occurring variants of CYP1B1N453S in selected cell lines provided a testable in vitro model for potential differences in NNK metabolism. Six Coriell lines, two of each of the three different genotypes in CYP1B1N453S, were tested for differences in the metabolism of NNK. While metabolism of NNK to NNAL was observed in all six lines, no significant differences were detected in the conversion, and no other metabolites were observed (information available on request from the corresponding authors). This suggests the interaction of CYP1B1N453S variants with levels of NNAL is not through a direct involvement of this CYP activity on NNK/NNAL metabolism.

Discussion:

Most lung cancers derive from cigarette tobacco smoke, which accounts for as much as 90% of all lung cancer cases in the US [18,19]. NNK is a powerful lung carcinogen associated with tobacco smoke, and total serum NNAL is a biomarker of its exposure that has been shown to be significantly associated with lung cancer risk [8]. In the present study, we found that the CYP1B1N453S has an interaction effect on the relation between total NNAL and lung cancer risk in addition to a main effect on risk. It is notable that the SNP would not have been identified had we looked first for a main effect alone. The main effect is statistically significant only in the presence of the interaction term. Further strengthening the result is the fact that the effect increases when the analysis is limited to adenocarcinomas, the histological subtype of lung cancer caused by NNK in laboratory animals, and the most common type of lung cancer in the U.S. This would not likely have been observed if the initial observation was a chance occurrence.

The involvement of CYP1B1 has not been previously implicated in NNK or NNAL metabolism. We considered the possibility that CYP1B1 has a direct involvement in the metabolism of NNK, but found no evidence for this. However, CYP1B1 could have an influence on the NNAL pathway by affecting transcription of CYP1A1, whose role in the metabolic activation of total NNAL has been previously described, even though its catalytic efficiency is not very great [5,20].

Transcription of both P450 family members CYP1A1 and CYP1B1 is induced upon activation of the aryl hydrocarbon receptor (AhR) pathway [21]. AhR is a cytosolic transcription factor that is normally inactive and bound to several co-chaperones. Following exposure to endogenous and exogenous chemicals, AhR acts as a ligand-activated receptor and transcription factor, activating the transcription of xenobiotic-metabolizing enzymes such as CYP1A1 and CYP1B1 as well as other genes [22-24]. There could be a signaling loop mechanism in which CYP1B1 can also act as a ligand and mediate the AhR signaling pathway, either in an activating or a suppressing fashion.

If the SNP implicated in this study, CYP1B1N453S, has a functional significance on the protein levels of CYP1B1 such that it downregulates or abrogates them, then this would be expected to enhance AhR activation. Significantly, one study did show that inhibition of CYP1B1 is linked to enhanced AhR activation [25]. Consequently, enhanced AhR activation leads to an enhanced transcription of CYP1A1. In fact, a recent study showed that CYP1B1N453S has a functional impact on the protein such that the protein displays lower intracellular levels and is degraded more rapidly than all other CYP1B1 variants tested in the study [26]. It is not clear what structural alterations are responsible for the increased rate of CYP1B1 degradation caused by the codon change Asn453SSer. This residue is located in the large meander region between the K- and L- helix and probably highly accessible to proteases. This so-called meander region is situated in the COOH terminal half of CYP1B1, important in the heme-binding and proper folding of the molecule. Moreover, it is interesting to note that the regions in which the putative non-synonymous SNPs reside in CYP1B1 are not highly conserved in mammals with the exception of the SNP at codon 453 [27]. Two different groups reported a 2-fold reduction in the cellular level of the protein containing this polymorphism, and a significantly reduced enzyme half-life [27,28]. It is therefore well established that this variation has a functional consequence on the protein cellular levels, its folding and stability. Due to CYP1B1's involvement in the metabolism of carcinogens, and the SNPs residence in a conserved region of the gene, it is not surprising that this SNP is emerging as an important player in carcinogenesis. Therefore, CYP1B1N453S connection to CYP1A1 and its consequent indirect involvement in the NNK/NNAL metabolism is a possible explanation for our findings (Figure 2). AhR mediated induction of CYP1 enzymes can lead to many cancer-related processes including genotoxicity, mutation and tumor initiation [29].

This indirect impact of CYP1B1 on NNK metabolism through CYP1A1 could involve other pathways that we are not aware of due to the complexity of tobacco smoke carcinogenesis. A relationship between CYP1 inducibility and cancer has been previously shown [30]. A group of researchers demonstrated an association between CYP1 inducibility and bronchogenic carcinoma [31]. Furthermore, in the context of hepatoma cells or in vitro studies, CYP1A1 is a primary determinant of the metabolism of benzo[a]pyrene, a PAH likely involved in tobacco-induced lung cancer [32]. Thus, CYP1A1's link to lung cancer has been proposed in many previous studies, although the possible relationship of our observations to CYP1A1 inducibility remains speculative.

As presented in this paper, we found an even stronger effect of CYP1B1N453S in a smaller adenocarcinoma group. A study by Chang et al. [33] found expression of AhR and CYP1B1 to be associated regardless of smoking status and AhR overexpression to up-regulate the expression of CYP1B1 in the early stage of lung adenocarcinoma. This finding may strengthen the results of our study. Therefore, the effect of the CYP1B1N453S we observed might be predicted—lower levels of CYB1B1 protein results in increased activation of AhR, which in turn increases CYP1A1 activity (Figure 2). Based on our analysis of HapMap variants we do not believe CYP1B1 to be directly involved in the metabolism of NNAL, although further functional studies on CYP1B1's involvement in NNAL and NNK metabolism are needed. Phenanthrene and other PAHs are substrates for CYP1B1 and CYP1A1 [32,34]. We did not observe an association between PheT levels and lung cancer, nor was there any interaction with CYP1B1 polymorphisms. This somewhat unexpected result may be due to the relatively small size of our study, and to the fact that phenanthrene, in contrast to NNK, is not tobacco-specific. Thus, substantial amounts of serum PheT are due to phenanthrene exposure from diet or general environment.

The study is limited by its small size, which required a focus on just a few SNPs, rather than on a broad array of polymorphisms. We chose a subset of the eleven most likely candidates for study, and found evidence that one of those SNPs may segregate the population by the risk conferred by NNAL exposure as well as by the underlying risk itself. The evidence of a strong interaction between total serum NNAL and the CYP1B1N453S SNP from this study was unexpected and, as yet, is not fully explained. If confirmed by appropriate additional molecular and epidemiologic studies, this outcome constitutes an important step in understanding how exposure to cigarette smoke leads to inter-individual variation in risk of lung cancer.

References:

1. International Agency for Research on Cancer. **Tobacco Smoke and Involuntary Smoking**. Lyon, FR: *IARC*; 2004. p 33-1187.

2. International Agency for Research on Cancer. **Smokeless Tobacco and Tobacco-Specific Nitrosamines**. Lyon, FR: *IARC*; 2007. p 548-553.

3. Straif K, Baan R, Grosse Y, Secretan B, El Ghissassi F, Cogliano V. **Carcinogenicity of polycyclic aromatic hydrocarbons**. *Lancet Oncol* 2005;6(12):931-932.

4. Hecht SS. **Tobacco carcinogens, their biomarkers and tobacco-induced cancer**. *Nat Rev Cancer* 2003;3(10):733-744.

5. Jalas JR, Hecht SS, Murphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem Res Toxicol* 2005;18(2):95-110.

6. Hecht SS. B**iochemistry, biology, and carcinogenicity of tobacco-specific N-nitrosamines**. *Chem Res Toxicol* 1998;11(6):559-603.

7. Cooper CS, P. L. Grover, and P. Sims. **The metabolism and activation of benzo[a]pyrene**. *ProgDrug Metab* 1983;7:295-396.

8. Church TR, Anderson KE, Caporaso NE et al. **A prospectively measured serum biomarker for a tobacco-specific carcinogen and lung cancer in smokers**. *Cancer Epidemiol Biomarkers Prev* 2009;18(1):260-266.

9. Hecht SS, Chen M, Yagi H, Jerina DM, Carmella SG. **r-1,t-2,3,c-4-Tetrahydroxy-1,2,3,4-tetrahydrophenanthrene in human urine: a potential biomarker for assessing polycyclic aromatic hydrocarbon metabolic activation**. *Cancer Epidemiol Biomarkers Prev* 2003;12(12):1501-1508.

10. Hecht SS, Carmella SG, Yoder A et al. **Comparison of polymorphisms in genes involved in polycyclic aromatic hydrocarbon metabolism with urinary phenanthrene metabolite ratios in smokers**. *Cancer Epidemiol Biomarkers Prev* 2006;15(10):1805-1811.

11. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. **The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status**. *Control Clin Trials* 2000;21(6 Suppl S):251S-272S.

12. Rothman KJ, Greenland S, editors. **Modern Epidemiology**. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998. xiii, 737 p.

13. Van Ness B, Ramos C, Haznadar M et al**. Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med* 2008;6:26.

14. Carmella SG, Yoder A, Hecht SS. C**ombined analysis of r-1,t-2,3,c-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol in smokers' plasm**a. *Cancer Epidemiol Biomarkers Prev* 2006;15(8):1490-1494.

15. Hecht SS. **Tobacco smoke carcinogens and lung cancer**. *J Natl Cancer Inst* 1999;91(14):1194-1210.

16. Hardenbol P, Baner J, Jain M et al. **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nat Biotechnol* 2003;21(6):673-678.

17. Greenland S, Pearl J, Robins JM. **Causal diagrams for epidemiologic research**. *Epidemiology* 1999;10(1):37-48.

18. Peto RLA, Boreham J et al. **Mortality from smoking in developed countries 1950-2000: Indirect estimates from National Vital Statistics**: Oxford University Press; 2006.

19. Biesalski HK, Bueno de Mesquita B, Chesson A et al. **European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel**. *CA Cancer J Clin* 1998;48(3):167-176; discussion 164-166.

20. Hecht SS. **Recent studies on mechanisms of bioactivation and detoxification of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), a tobacco-specific lung carcinogen**. *Crit Rev Toxicol* 1996;26(2):163-181.

21. Lin P, Hu SW, Chang TH. **Correlation between gene expression of aryl hydrocarbon receptor (AhR), hydrocarbon receptor nuclear translocator (Arnt), cytochromes P4501A1 (CYP1A1) and 1B1 (CYP1B1), and inducibility of CYP1A1 and CYP1B1 in human lymphocytes**. *Toxicol Sci* 2003;71(1):20-26.

22. Hoffman EC, Reyes H, Chu FF et al. **Cloning of a factor required for activity of the Ah (dioxin) receptor**. *Science* 1991;252(5008):954-958.

23. Jiang BH, Rue E, Wang GL, Roe R, Semenza GL. **Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1**. *J Biol Chem* 1996;271(30):17771-17778.

24. Whitlock JP, Jr**. Induction of cytochrome P4501A1**. *Annu Rev Pharmacol Toxicol* 1999;39:103-125.

25. Alexander DL, Zhang L, Foroozesh M, Alworth WL, Jefcoate CR. **Metabolism-based polycyclic aromatic acetylene inhibition of CYP1B1 in 10T1/2 cells potentiates aryl hydrocarbon receptor activity**. *Toxicol Appl Pharmacol* 1999;161(2):123-139.

26. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol* 2005;67(2):435-443.

27. Mammen JS, Pittman GS, Li Y et al. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis* 2003;24(7):1247-1255.

28. Aklillu E, Ovrebo S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res* 2005;65(12):5105-5111.

29. Nebert DW, Roe AL, Dieter MZ, Solis WA, Yang Y, Dalton TP. **Role of the aromatic hydrocarbon receptor and [Ah] gene battery in the oxidative stress response, cell cycle control, and apoptosis**. *Biochem Pharmacol* 2000;59(1):65-85.

30. Nebert DW, Benedict, W. F., and Kouri, R. E. *Chemical Carcinogenesis*. In: Ts'o POP, and DiPaolo, J. A, editor. New York: Marcel Dekker, Inc.; 1974. p 271-288.

31. Kellermann G, Shaw CR, Luyten-Kellerman M. **Aryl hydrocarbon hydroxylase inducibility and bronchogenic carcinoma**. *N Engl J Med* 1973;289(18):934-937.

32. Nebert DW, Dalton TP, Okey AB, Gonzalez FJ. **Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer**. *J Biol Chem* 2004;279(23):23847-23850.

33. Chang JT, Chang H, Chen PH, Lin SL, Lin P. **Requirement of aryl hydrocarbon receptor overexpression for CYP1B1 up-regulation and cell growth in human lung adenocarcinomas**. *Clin Cancer Res* 2007;13(1):38-45.

34. Shimada T, Hayes CL, Yamazaki H et al. **Activation of chemically diverse procarcinogens by human cytochrome P-450 1B1**. *Cancer Res* 1996;56(13):2979-2984.

35. Meyer BK, Pray-Grant MG, Vanden Heuvel JP, Perdew GH. **Hepatitis B virus X-associated protein 2 is a subunit of the unliganded aryl hydrocarbon receptor core complex and exhibits transcriptional enhancer activity**. *Mol Cell Biol* 1998;18(2):978-988.

36. Kazlauskas A, Poellinger L, Pongratz I. **Evidence that the co-chaperone p23 regulates ligand responsiveness of the dioxin (Aryl hydrocarbon) receptor**. *J Biol Chem* 1999;274(19):13519-13524.

37. Hord NG, Perdew GH. **Physicochemical and immunocytochemical analysis of the aryl hydrocarbon receptor nuclear translocator: characterization of two monoclonal antibodies to the aryl hydrocarbon receptor nuclear translocator**. *Mol Pharmacol* 1994;46(4):618-626.

38. Pollenz RS, Sattler CA, Poland A. **The aryl hydrocarbon receptor and aryl hydrocarbon receptor nuclear translocator protein show distinct subcellular**

localizations in Hepa 1c1c7 cells by immunofluorescence microscopy. *Mol Pharmacol* 1994;45(3):428-438.

39. Hankinson O. **The aryl hydrocarbon receptor complex**. *Annu Rev Pharmacol Toxicol* 1995;35:307-340.

40. Probst MR, Reisz-Porszasz S, Agbunag RV, Ong MS, Hankinson O. **Role of the aryl hydrocarbon receptor nuclear translocator protein in aryl hydrocarbon (dioxin) receptor action**. *Mol Pharmacol* 1993;44(3):511-518.

41. Denison MS, Elferink CF, Phelan D. **The Ah receptor signal transduction pathway**. In: *Denison* MS, W.G. H, editors. Toxicant-Receptor Interactions in the Modulation of Signal Transduction and Gene Expression. Philadelphia: Taylor & Francis; 1998. p 3-33.

42. Denison MS, Fisher JM, Whitlock JP, Jr. **The DNA recognition site for the dioxin-Ah receptor complex. Nucleotide sequence and functional analysis**. *J Biol Chem* 1988;263(33):17221-17224.

Table 1. A list of analyzed single nucleotide polymorphisms with detailed descriptions.

SNP ID	dbSNP allele	HUGO gene name	Function	Amino acid change	Position
rs1056836[a]	C/G	CYP1B1	coding-nonsyn	Leu432Val	g.10122C>G
rs1051740[a]	C/T	EPHX1	coding-nonsyn	Tyr113His	g.26837T>C
rs947894[a]	A/G	GSTP1	coding-nonsyn	Ile105Val	g.6624A>G
rs1799811[a]	C/T	GSTP1	coding-nonsyn	Ala114Val	g.7514C>T
rs10012	C/G	CYP1B1	coding-nonsyn	Arg48Gly	g.5935C>G
rs1056827	G/T	CYP1B1	coding-nonsyn	Ala119Ser	g.6148G>T
rs1800440	A/G	CYP1B1	coding-nonsyn	Asn453Ser	g.10186A>G
rs2234922	A/G	EPHX1	coding-nonsyn	His139Arg	g.33610A>G
rs744177	C/T	NA	intergenic	NA	g.33523236C>T
rs1105879	G/T	UGT1A6	coding-nonsyn	Arg184Ser	g.108813A>C
rs6759892	G/T	UGT1A6	coding-nonsyn	Ser7Ala	g.108280T>G

[a]Four SNPs identified in the original protocol.

Table 2. Comparing the age, sex, race/ethnicity, education, marital status, occupational status, family history of lung cancer, cigarettes per day and SNP calls of cases and controls.

VARIABLES	Controls N=100	Cases N=100	*P*	OR[a] (95% CI)[b]
Age at Randomization			0.0039	
≤59 years	51 (51%)	28 (28%)		reference
60-64 years	26 (26%)	27 (27%)		1.891 (0.931,3.843)
65-69 years	18 (18%)	36 (36%)		3.642 (1.756,7.557)
≥70years	5 (5%)	9 (9%)		3.278 (1.001,10.738)
Sex			0.2913	
Women	36 (36%)	29 (29%)		reference
Men	64 (64%)	71 (71%)		1.377 (0.760,2.495)
Race/Ethnicity			0.0958	
White, non-Hispanic	93 (93%)	84 (84%)		reference
Black, non-Hispanic	4 (4%)	13 (13%)		3.598 (1.129,11.465)
Other	3 (3%)	3 (3%)		1.107 (0.218,5.635)
Education			0.5518	
Less than 12 years	10 (10%)	15 (15%)		1.345 (0.515,3.510)
12 yrs or completed high school	26 (26%)	29 (29%)		reference
Post-high-school training other than college	16 (16%)	10 (10%)		0.560 (0.216,1.450)
Some college	20 (20%)	24 (24%)		1.076 (0.486,2.383)
College graduate	19 (19%)	13 (13%)		0.613 (0.254,1.482)
Post-graduate training	9 (9%)	9 (9%)		0.897 (0.309,2.600)
Marital Status			0.8507	
Married or living as married	75 (75%)	69 (69%)		reference
Widowed	10 (10%)	10 (10%)		1.087 (0.427,2.770)
Divorced	11 (11%)	16 (16%)		1.581 (0.686,3.642)
Separated	2 (2%)	2 (2%)		1.087 (0.149,7.928)
Never married	2 (2%)	3 (3%)		1.630 (0.264,10.051)
Occupation			0.7811	
Homemaker	6 (6%)	5 (5%)		reference
Working	44 (44%)	36 (36%)		0.982 (0.277,3.482)
Unemployed	3 (3%)	2 (2%)		0.800 (0.093,6.848)
Retired	39 (39%)	49 (49%)		1.508 (0.428,5.311)
Disabled	4 (4%)	5 (5%)		1.500 (0.255,8.817)
Other/not answered	4 (4%)	3 (3%)		0.900 (0.133,6.080)
Family History of Lung Cancer			0.1456	
No	94 (94%)	88 (88%)		reference
Yes	6 (6%)	12 (12%)		2.136 (0.769,5.937)
Cigarettes per Day			0.0576	
11-20	48 (48%)	49 (49%)		reference
21-30	37 (37%)	23 (23%)		0.609 (0.316,1.173)
31-40	13 (13%)	21 (21%)		1.582 (0.712,3.515)
41 +	2 (2%)	7 (7%)		3.429 (0.678,17.344)

Table 2. continued

VARIABLES	Controls N=100	Cases N=100	*P*	OR[a] (95% CI)[b]
RS1056836			0.3600	
CC	36 (36%)	26 (26%)		reference
CG	42 (42%)	43 (43%)		1.418 (0.733,2.742)
GG	21 (21%)	30 (30%)		1.978 (0.933,4.196)
No sample	1 (1%)	1 (1%)		1.385 (0.083,23.168)
RS1051740			0.7286	
TT	45 (45%)	53 (53%)		reference
CT	45 (45%)	38 (38 %)		0.717 (0.399,1.289)
CC	9 (9%)	8 (8%)		0.755(0.269,2.118)
No sample	1 (1%)	1 (1%)		0.849(0.052,13.964)
RS947894			0.5424	
AA	30 (30%)	36 (36 %)		reference
AG	45 (45%)	33 (33%)		0.611 (0.316,1.183)
GG	9 (9%)	12 (12 %)		1.111 (0.413,2.993)
No call	15 (15%)	18 (18%)		1.000 (0.432,2.315)
No sample	1 (1%)	1 (1%)		0.833 (0.050,13.895)
RS1799811			0.9774	
CC	87 (87%)	86 (86%)		reference
CT	12 (12%)	13 (13%)		1.096 (0.474,2.537)
No sample	1 (1%)	1 (1%)		1.012 (0.062,16.434)
RS10012			0.2343	
CC	51 (51%)	52 (52%)		reference
CG	32 (32%)	37 (37%)		1.134 (0.616,2.089)
GG	15 (15%)	6 (6%)		0.392 (0.141,1.091)
No call	1 (1%)	2 (2%)		1.962(0.173,22.311)
No sample	1 (1%)	3 (3%)		2.942 (0.296,29.227)
RS1056827			0.0826	
GG	49 (49%)	51 (51%)		reference
GT	35 (35%)	39 (39%)		1.071 (0.587,1.954)
TT	15 (15%)	5 (5%)		0.320(0.108,0.948)
No call	0 (0%)	2 (2%)		
No sample	1 (1%)	3 (3%)		2.882 (0.290,28.660)
RS1800440			0.4877	
AA	65 (65%)	70 (70%)		reference
GA	26 (26%)	23 (23%)		0.821 (0.427,1.581)
GG	5 (5%)	1 (1%)		0.186 (0.021,1.632)
No call	2 (2%)	3 (3%)		1.393 (0.226,8.603)
No sample	2 (1%)	3 (3%)		1.393(0.226,8.603)
RS2234922			0.1036	
AA	62 (62%)	47 (47%)		reference
GA	32 (32%)	45 (45%)		1.855 (1.027,3.349)
GG	5 (5%)	4 (4%)		1.055 (0.269,4.146)
No sample	1 (1%)	4 (4%)		5.277 (0.571,48.771)

Table 2. continued

VARIABLES	Controls N=100	Cases N=100	*P*	OR[a] (95% CI)[b]
RS744177			0.4530	
GG	25 (25%)	32 (32%)		reference
AG	46 (46%)	48 (48%)		0.815 (0.421, 1.579)
AA	28 (28%)	19 (19%)		0.530 (0.242, 1.160)
No call	0 (0%)	0 (0%)		
No sample	1 (1%)	1 (1%)		0.781 (0.047,13.117)
RS6759892			0.6673	
TT	34 (34%)	31 (31%)		reference
GT	49 (49%)	45 (45%)		1.007 (0.535, 1.897)
GG	16 (16%)	23 (23%)		1.577 (0.707, 3.518)
No call	0 (0%)	0 (0%)		
No sample	1 (0%)	1 (1%)		1.097 (0.066,18.294)
RS1105879			0.9689	
TT	36 (36%)	39 (39%)		reference
GT	38 (38%)	34 (34%)		0.826 (0.432, 1.578)
GG	14 (14%)	13 (13%)		0.857 (0.355, 2.067)
No call	11(11%)	13 (13%)		1.091 (0.434, 2.743)
No sample	1 (1%)	1 (1%)		0.923 (0.056,15.311)

[a] OR = odds ratio

[b] (95% CI) = 95% confidence interval, given as (lower limit, upper limit)

Table 3. Comparing cases and controls on serum levels of cotinine, NNAL, PheT, and years of smoking.

VARIABLES	Controls N=100	Cases N=100	Difference (controls-cases)	*P*
	Mean ± SD[a]	Mean ± SD	Mean ± SE[b]	
Years of Cigarette Smoking	41.6 ± 7.2	45.4 ± 6.5	-3.9 ± 1.0	0.0001
Cotinine (ng/ml)	217 ± 111	227 ± 93	-10 ± 15	0.4681
Total NNAL[c] (fmol/ml)	77.4 ± 39.3	92.4 ± 40.7	-15.0 ± 5.7	0.0084
PheT[d] (fmol/ml)	76.3 ± 66.8	92.5 ± 107.6	-16.1 ± 12.7	0.2039

[a] SD = standard deviation

[b] SE = standard error

[c] NNAL = 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol; total includes its glucuronides.

[d] PheT = *r*-1,*t*-2,3,*c*-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene

Table 4. Results of multiple logistic regression of lung cancer risk on sex, age at randomization, family history of lung cancer, years of cigarette smoking, cotinine, PheT[a], total NNAL[b], CYP1B1N453S and its interaction with total NNAL[b].

a. ALL LUNG CANCERS

VARIABLES		95% Confidence Limits		*P*
	OR[c]	Lower	Upper	
Sex (men vs. women)	1.384	0.675	2.837	0.3749
Age at randomization	1.086	1.005	1.174	0.0371
Family history of lung cancer	2.215	0.736	6.669	0.1574
Years of cigarette smoking	1.048	0.987	1.112	0.1263
Cotinine	0.998	0.995	1.002	0.3239
PheT[a]	1.002	0.998	1.006	0.2540
Total NNAL[b]	1.005	0.994	1.016	0.3784
CYP1B1N453S (not AA vs. AA)	0.693	0.345	1.394	0.3037
NNAL[b] × CYP1B1N453S interaction	1.020	1.002	1.038	0.0299

b. ADENOCARCINOMA

VARIABLES		95% Confidence Limits		*P*
	OR[c]	Lower	Upper	
Sex (men vs. women)	1.037	0.450	2.393	0.9318
Age at randomization	1.093	1.001	1.194	0.0471
Family history of lung cancer	2.962	0.896	9.789	0.0750
Years of cigarette smoking	1.051	0.982	1.125	0.1503
Cotinine	0.997	0.993	1.001	0.1552
PheT[a]	1.002	0.997	1.008	0.3608
Total NNAL[b]	1.006	0.994	1.018	0.3368
CYP1B1N453S	0.415	0.157	1.098	0.0765
NNAL[b] × CYP1B1N453S interaction	1.031	1.006	1.056	0.0144

[a] PheT = *r*-1,*t*-2,3,*c*-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene

[b] NNAL = 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol; total includes its glucuronides.

[c] OR = odds ratio

Table 5. Total serum NNAL values (fmol/ml) for cases and controls by CYP1B1N453S genotype.

A) All lung cancers

rs1800440 genotype	Status	N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
homozygous major	Case	70	83.7	4.1	75.7	91.8
	Control	65	81.2	5.0	71.2	91.2
not homozygous major	Case	30	112.8	8.8	94.9	130.8
	Control	35	70.3	6.2	57.7	82.9

B) Adenocarcinoma only

rs1800440 genotype	Status	N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
homozygous major	Case	43	84.0	5.0	73.9	94.0
	Control	65	81.2	5.0	71.2	91.2
not homozygous major	Case	16	122.7	9.9	101.6	143.9
	Control	35	70.3	6.2	57.7	82.9

Figure 1. Relation of total serum NNAL concentration and case-control status by CYP1B1 genotype. The dots plot the individual cases and controls by log (total NNAL) value and each line is the kernel-smoothed relation between the logarithm of the total serum NNAL value on the horizontal axis and the case-control variable, assigning case = 1 and control = 0; the red dots and lines are for the CYP1B1N453S homozygous major genotype and the black dots and lines are for the complement. For the subjects with a CYP1N453S minor allele (black dots) the fraction of cases compared to controls increases noticeably at high NNAL levels compared to lower levels and yields a upward sloping line; for those without this SNP (red dots), the fraction of cases compared to controls is about the same regardless of NNAL level. Thus within the study population, this SNP drives the increase in risk for lung cancer with each standard deviation increase in NNAL.

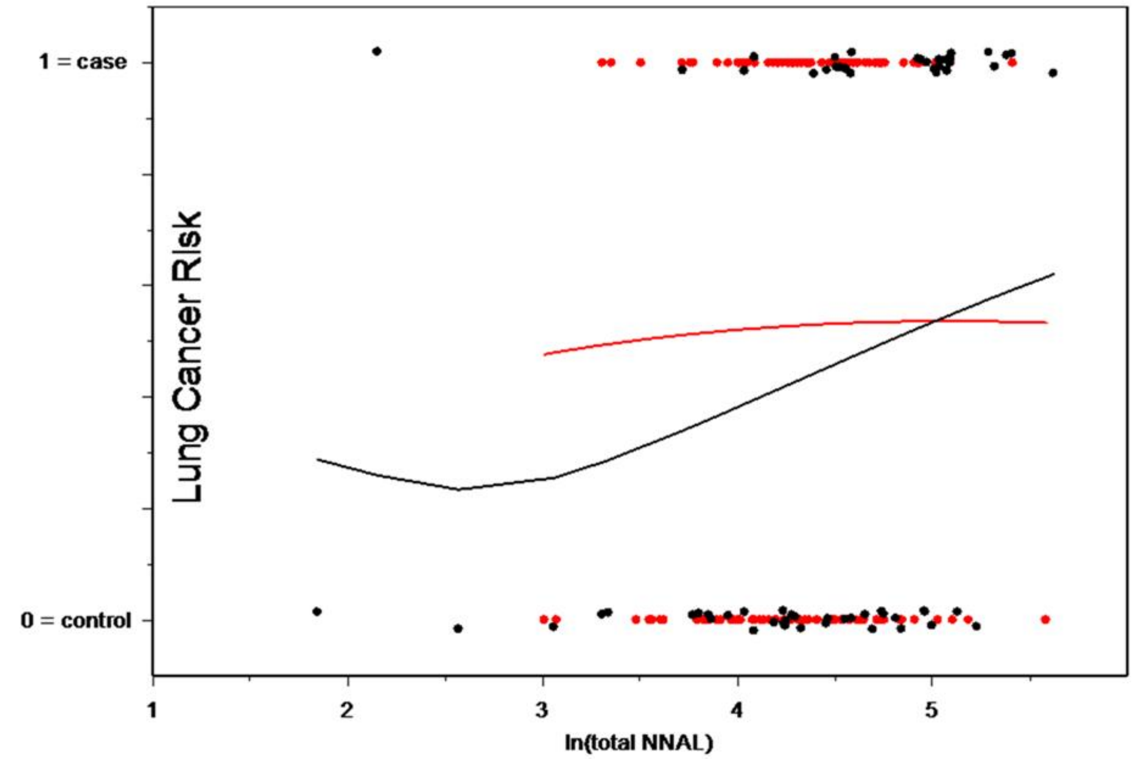
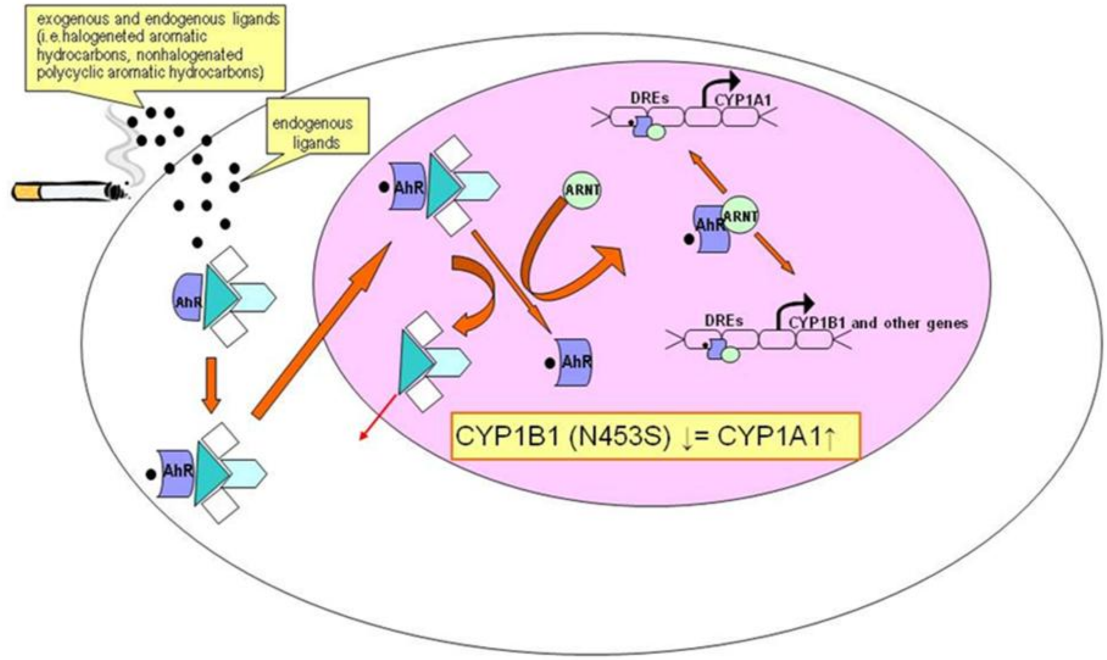

Lung Cancer Risk
1 = case
0 = control
1
2
3
4
5
ln(total NNAL)

Figure 2. Potential molecular mechanism of CYP1B1N453S impact on lung cancer susceptibility. AhR, a ligand-dependent transcription factor, becomes activated upon ligation, to endogenous ligands, inducing chemical binding. The AhR receptor is a part of a multifactor complex consisting of two hsp90 chaperones, XAP2 [35] and p23 [36] (diagramed in white, turquoise and light blue respectively) and it undergoes a conformation change upon activation, resulting in translocation of the complex into the nucleus [37,38]. Release of the ligand-bound AhR from the complex and its subsequent dimerization with ARNT converts the AhR into its high affinity DNA binding form [39,40] that binds to its specific DNA recognition site upstream of CYP1A1, CYP1B1, and other genes, stimulating transcription of those genes [21,24,41,42]. CYP1B1N453S results in downregulation of the cellular protein levels [26], consequently inducing increased AhR receptor production [25] and stimulating CYP1A1 transcription and translation. CYP1A1 enzymatically mediates α-hydroxylation and activates metabolism of NNK and NNAL [6].


exogenous and endogenous ligands (i.e.halogeneted aromatic hydrocarbons, nonhalogenated polycyclic aromatic hydrocarbons)
endogenous ligands
AhR
AhR
AhR
AhR
ARNT
AhR
ARNT
DREs
CYP1A1
DREs
CYP1B1 and other genes
CYP1B1 (N453S) ↓= CYP1A1↑

CHAPTER 5

IDENTIFICATION OF SINGLE NUCLOTIDE POLYMORPHISM INTERACTIONS ASSOCIATED WITH RISK IN LUNG CANCER USING A NOVEL DATA MINING METHOD

Majda Haznadar[1], Gang Fang[2], Wen Wang[2], Vanja Paunic[2], Patrick Day[1], Michael Steinbach[2], Vipin Kumar[2], Timothy Church[3], Stephen Hecht[4], Brian Van Ness[1]

Affiliations:

[1]Institute of Human Genetics, Computer Science and Engineering[2], [3]Environmental Health Sciences

[4]Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota, USA

I processed and genotyped all the samples. The same as in chapter three, I helped guide the process of designing novel combinatorial search algorithms. This is the first application of this novel data mining method developed by our collaborators on lung cancer on genotyping data.

Cigarette smoke has been linked to a range of chronic lung diseases and is a major source of morbidity and mortality. 90% of lung cancers arise as a consequence of cigarette smoking. We applied a novel combinatorial search data mining method to help elucidate genetic interplay that leads to lung cancer. We conducted an association study in 200 smokers, 100 with and a 100 without lung cancer with 3,404 single nucleotide polymorphisms (SNPs), and additional 215 SNPs genotyped using Sequenom. The associations were conducted by constraining groups of SNPs using gene sets collected and publically available at the Molecular Signatures Database (MSigDB). Therefore, we guided this study by focusing on already defined pathways and gene-sets, which also enabled increase in statistical power. We identified two 3-SNP interactions that associate with lung cancer risk, in the following genes: CYP3A4, GSTA5, GSTA4, PLAUR and CYP2C9, polymorphisms in which correlate with an increased risk of lung cancer. This is a novel attempt at identified SNP-SNP interactions that associate with lung cancer risk using novel data mining algorithms.

Introduction:

Cigarette smoke has been connected to a wide array of chronic lung disease and is a major source of morbidity and mortality. In the United States, over 400,000 deaths per year are attributed to smoking, and the number of deaths since 1964 is estimated at $12*10^6$ (1). Active smoking has been linked to 90% of lung cancer cases (2). While the use of tobacco products continues to be a serious public health problem, remarkable progress has been made in the last 20 years, both in understanding tobacco carcinogenesis and in tobacco control. However, 1.3 billion smokers remain in the world, and hundreds of millions of smokeless tobacco users (3).

The "Hoffmann list" of over 60 carcinogens in cigarette smoke is considered the definitive catalogue of its major cancer-causing agents and includes polycyclic aromatic hydrocarbons (PAHs), nitrosamines, aromatic amines, aldehydes, volatile organic compounds, metals, and others (4, 5). Tobacco-specific nitrosamines were first characterized with respect to their presence in tobacco products and their carcinogenicity in the 1970s. They have since emerged as one of the most important groups of carcinogens in tobacco products (6, 7). There have been seven tobacco-specific nitrosamines identified in tobacco products, but two are the most important because of their carcinogenic activities and their abundance in unburned tobacco and its smoke: 4-(methylnitrosamino)-1-(30pyridyl)-1-butanone (NNK) and N'-nitrosonornicotine (NNN) (8). NNK specifically induces mainly lung tumors in laboratory animals, but also causes tumors of the pancreas, nasal mucosa, and liver (9).

It is well-established that dialkylnitrosamines such as NNK require metabolic activation by cytochrome P450 (CYP)-catalyzed α-hydroxylation to exert their carcinogenic effect (10). Multiple CYPs and Phase II enzymes are involved in the metabolic activation and detoxification of NNK and PAHs (9, 11, 12). GSTs act as Phase II enzymes, through glutathione conjugation to electrophilic substances an important role in the protection against oxidative stress induced by carcinogens such as tobacco smoke (13). Thus, the role of CYPs in activation and detoxification of tobacco smoke carcinogens is well-established.

Complexity of genetic interplay in cancer causation and progression has presented association studies with a problem of finding robust methods to study gene and variation interactions. We successfully applied a novel computational approach that efficiently identifies high order SNP interactions and minimizes the false discovery rate (14). We hypothesized that higher levels of tobacco carcinogens and their specific metabolites among long-term current smokers predispose them to higher risks of developing lung cancer. Reversely, long-term smokers who do not develop lung cancer may have genomic variation that protects them against harmful carcinogens from tobacco smoke. Our goal was to identify variations that distinguish the two groups, smokers with lung cancer from smokers without lung cancer.

Materials and Methods:

Parent study

The PLCO is an NCI-funded multi-center, randomized, prospective trial of screening for cancers of the prostate, lung, colorectum and ovaries that began in 1993 and is projected to end in 2011 (15). The screening in the trial includes 77,468 men and women, of whom approximately 25,000 are current or former smokers. In addition to annually screening participants and carefully abstracting cancers from medical records, the PLCO has prospectively collected extensive information from study participants, including smoking history, family history of cancer, and demographic information collected at randomization; and it maintains a bio-repository of blood samples drawn over six annual screening visits starting in 1993. The PLCO trial made available its prospectively collected blood samples from the first screening visit and its extensive baseline and clinical data, thus providing for the direct calculation of lung cancer risks in the groups with different baseline levels of biomarkers. In addition, in the PLCO screening cohort nearly all cases of lung cancer had been screened at least once and so the variability in diagnostic lead times and the potential confounding that variability can produce in unscreened or partially screened cohorts was substantially reduced. At the time our study was initiated, over 800 lung cancer cases had been diagnosed in the screening arm of the PLCO. We randomly selected cases and controls from subjects who reported currently smoking at least 10 cigarettes per day on the baseline questionnaire filled out at the time of randomization. The PLCO was approved by the institutional review boards of each participating institution, and all subjects signed consents permitting the research represented here. In this study, we analyzed all 3,404 SNPs genotyped by BOAC SNP chip, and additional 215 coding-nonsynonymous SNPs genotyped using Sequenom.

Case-control Study

We used a nested case-control approach wherein the source cohort consisted of PLCO participants who at randomization filled out a baseline questionnaire indicating they were free of cancer and currently smoking at least 10 cigarettes per day, and who contributed adequate blood samples to the biorepository. Cases were those smokers subsequently diagnosed with lung cancer and controls were those smokers with no diagnosis of lung cancer before the cut-off date (August 17, 2007). From this cohort, we randomly selected 100 incident lung cancer cases and 100 controls and obtained their demographic and other baseline data from the PLCO database, as well as serum samples adequate to measure total cotinine, total NNAL, and PheT.

Results:

We found two 3-SNP interaction signatures that distinguish the cases from the controls by having a higher frequency of the homozygous-major genotypes in all three SNPs in the controls. Therefore, the homozygous-major genotypes in the respective genes as a group appear to be protective against lung cancer. Reverse can be conceived as true, that those subjects who contain minor alleles in the represented SNPs as a group have an increased risk of lung cancer.

The first 3-SNP signature contains the following SNPs: rs2687105 in cytochrome P450, family 3, subfamily A, polypeptide 4 (CYP3A4), rs2397118 in glutathione S-transferase alpha 5 (GSTA5), and rs669674 in glutathione S-transferase alpha 4 (GSTA4). This 3-SNP signature has a collective odds ratio of 0.2213, thus decreasing the risk of lung cancer, and a p-value of $5.6*10^{-7}$ (Table 1).

The second 3-SNP signature contains the following SNPs: rs182623 in glutathione S-transferase alpha 4 (GSTA4), rs4760 in plasminogen activator, urokinase receptor (PLAUR), and rs1934963 in cytochrome P450, family 2, subfamily C, polypeptide 9 (CYP2C9). The three SNPs collectively have an odds ratio of 0.1514, therefore associated with a decreased risk of lung cancer, with a p-value of $1.3*10^{-7}$ (Table 2).

Figures 1A and 1B graphically represent distributions of the 3-SNP signatures between the cases and the controls, where black blocks represent the subjects with homozygous-major genotypes in the SNPs indicated at the bottom of the graphs (Figure 1). The fourth column in both A and B, denoted as “all three genotypes”, represents a block or a clustering of subjects with homozygous-major genotypes in all three SNPs. It is apparent that there is a significant difference in the frequency of all three genotypes in cases versus controls. As illustrated in Figure 1, the controls have a significantly higher frequency of the 3-SNP signature in both A and B (homozygous-major genotypes). Both graphs represent a total of 200 subjects, all smokers, 100 with lung cancer (cases) and a 100 without (controls).

By constraining the SNPs to pre-identified and publically available gene sets (16), we increased the association power by decreasing the number of SNPs computed by the algorithm, therefore decreasing the possibility of false positives. The computational power allowed interactions of SNPs within either one or two gene sets. Moreover, we had identified three additional 3-SNP signatures, but they contained SNPs in linkage disequilibrium (LD) with the SNPs presented here, as would be expected—all variations within an LD block are expected to be associated with an outcome. We were able to successfully apply a new data mining algorithm for identifying SNP-SNP interactions that associate with lung cancer risk, and identify SNPs in genes biologically relevant to the metabolism of lung cancer causing carcinogens and to other lung cancer pathways.

Discussion:

The amount of data that is generated by genotyping creates possibilities for finding genetic markers that associate with disease risk and outcomes, but also creates a challenge. It is becoming increasingly difficult to analyze all the generated data. We ventured into a collaborative effort to help design a novel combinatorial search data mining algorithm that successfully searches for SNP-SNP interaction signatures, while decreasing a false positive rate (14). This is the first application of the novel computational method on lung cancer.

We were able to discover two 3-SNP signatures that distinguish the cases from the controls. Both 3-SNP signatures (all homozygous-major genotypes) have a significantly higher frequency in the controls; reversely, minor alleles in these SNPs as a group correlate with an increased risk of lung cancer.

CYP3A4, among other CYPs, is an important hepatic enzyme involved in the metabolism of drugs, toxins and other xenobiotics, also expressed to variable degrees in extrahepatic tissues (17, 18). We identified a variation in CYP3A4 as a part of the 3-SNP signature found by constraining the SNPs to one gene set, resulting in interaction with other two SNPs in GSTA5 and GSTA4. There have also been studies implicating CYP3A4, and other CYPs, in the metabolism of NNK, a prominent tobacco smoke carcinogen (9, 19-21). Additionally, a group identified a polymorphism in CYP3A4 that increases risk for small cell lung cancer (22). CYP3A4, therefore, is important to the metabolism of tobacco smoke carcinogens and other xenobiotics.

There have not been any reports of variations in CYP2C9 that significantly associate with lung cancer risk. We found an association of a polymorphism in CYP2C9, which along with variations in GSTA4 and PLAUR associate with lung cancer risk.

Besides detoxifying electrophilic xenobiotics, such as chemical carcinogens, environmental pollutants, and antitumor agents, mammalian glutathione transferase (GST) families inactivate endogenous alpha, beta-unsaturated aldehydes, quinones, epoxides, and hydroperoxides formed as secondary metabolites during oxidative stress. Qian et al. identified polymorphisms in GSTA4 that significantly associate with lung cancer risk (13). Another group discovered variations in GSTA5 in suggestive association with lung cancer survival after platinum-based chemotherapy (23). There have not been any reports of the variations in GSTA4 or GSTA5 reported here in correlation with lung cancer risk.

PLAUR encodes the receptor for urokinase plasminogen activator and, given its role in localizing and promoting plasmin formation, likely influences many normal and pathological processes related to cell-surface plasminogen activation and localized degradation of the extracellular matrix. Almasi et. al. found an increased risk of mortality with increasing levels of liberated domain I of the urokinase plasminogen activator receptor levels in squamous cell lung cancer tumour extracts, which independently predicted overall survival (24). No reports have identified the polymorphism reported here in relation to lung cancer risk.

Given that cancer arises as a consequence of a complex gene and variation interplay, it is of no surprise that we found SNP-SNP interactions in association with lung cancer risk. This effort represents some of the first steps in further uncovering the etiology of smoking-caused lung cancer. We were able to uncover two 3-SNP signatures that as groups associate with lung cancer risk. The mechanism of how these SNP-SNP interactions occur remains to be investigated.

References:

1. Bhalla DK, Hirata F, Rishi AK, Gairola CG. **Cigarette smoke, inflammation, and lung injury: a mechanistic perspective**. *J Toxicol Environ Health B Crit Rev.* 2009; 12: 45-64.
2. Alberg AJ, Samet JM. **Epidemiology of lung cancer**. *Chest*. 2003; 123(1 Suppl): 21S-49S.
3. World Health Organization. **The World Health Report 2003: Shaping the Future**. World Health Organization Geneva, Switzerland. 2003; pp91-94.
4. Hoffman D, and Hecht SS. **Advances in tobacco carcinogenesis**. In *Handbook of Experimental Pharmacology*. (Cooper CS, and Grover PL, Eds.). 1990; pp 63-192, Springer-Verlag. Heidelberg.
5. International Agency for Research on Cancer. **Tobacco smoke and involuntary smoking**. *In IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2004; Vol. 83, pp 81-83, IARC, Lyon, France.
6. International Agency for Research on Cancer. **Smokeless tobacco and tobacco-specific nitrosamines.** *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2003; Vol. 89, IARC, Lyon, France.
7. Hecht SS, and Hoffman D. **Tobacco-specific nitrosamines, an important group of carcinogens in tobacco and tobacco smoke**. *Carcinogenesis*. 1988. 9, 875-884.
8. International Agency for Research on Cancer. **Tobacco smoke and involuntary smoking**. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2004; Vol. 83, pp 59-80, IARC, Lyon, France.
9. Hecht SS. **Biochemistry, biology, and carcinogenicity of tobacco-specific *N*-nitrosamines**. *Chem. Res. Toxicol*. 1998; 11, 559-603.
10. Preussmann R, and Stewart BW. ***N*-Nitroso carcinogens**. In *Chemical Carcinogens*. 2nd ed. (Searle CE, Ed.) ACS Monograph 182, Vol.2, pp 643-828. American Chemical Society, Washington, DC.
11. Jalas JR, Hecht SS, and Myrphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem. Res. Toxicol*. 2005; 18: 95-110.
12. Cooper CS, Grover PL, and Sims P. **The metabolism and activation of benzo[a]pyrene**. *Prog. Drug. Metabo*. 1983; 7: 295-396.
13. Qian J, Jing J, Jin G, Wang H, Wang Y, Liu H, Wang H, Li R, Fan W, An Y, Sun W, Wang Y, Ma H, Miao R, Hu Z, Jin L, Wei Q, Shen H, Huang W, Lu D. **Association between polymorphisms in the GSTA4 gene and risk of lung cancer: a case-control study in a Southeastern Chinese population**. *Mol. Carcinog*. 2009; 48: 253-9.
14. Fang G, Haznadar M, Wang W, Steinbach M, Van Ness B, Kumar V: **A Computationally Efficient and Statistically Powerful Framework for Searching High-order Epistasis with Systematic Pruning and Gene-set Constraints**. Technical Report 013, Department of Computer Science, University of Minnesota, 2010. (submitted)

15. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. **The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status.** *Control. Clin. Trials* 2000; 21(6 Suppl S): 251S-272S.
16. Molecular Signatures Database http://www.broadinstitute.org/gsea/msigdb/index.jsp
17. Ding X, and Kaminsky LS. **Human extrahepatic cytochromes P450: Function in xenobiotic metabolism and tissue-selective chemical toxicity in the respiratory and gastrointestinal tracts**. *Annu. Rev. Pharmacol. Toxicol*. 2003; 43: 149-173.
18. Guengerich FP. **Human cytochrome P450 enzymes**. In *Cytochrome P450: Structure, Mechanism, and Biochemistry* (Ortiz de Montellano PR, Ed.). 1995; pp473-535, Plenum Press, New York.
19. Kushida H, Fujita K, Suzuki A, Yamada M, Endo T, Nohmi T, Kamataki T. **Metabolic activation of N-alkylnitrosamines in genetically engineered Salmonella typhimurium expressing CYP2E1 or CYP2A6 together with human NADPH-cytochrome P450 reductase**. *Carcinogenesis*. 2000; 21: 1227-32.
20. Yamazaki H, Inui Y, Yun CH, Guengerich FP, Shimada T. **Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes**. *Carcinogenesis*. 1992; 13: 1789-94.
21. Crespi CL, Penman BW, Gelboin HV, Gonzalez FJ. **A tobacco smoke-derived nitrosamine, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, is activated by multiple human cytochrome P450s including the polymorphic human cytochrome P4502D6**. *Carcinogenesis*. 1991; 12: 1197-201.
22. Dally H, Edler L, Jäger B, Schmezer P, Spiegelhalder B, Dienemann H, Drings P, Schulz V, Kayser K, Bartsch H, Risch A. **The CYP3A4*1B allele increases risk for small cell lung cancer: effect of gender and smoking dose**. *Pharmacogenetics*. 2003; 13: 607-18.
23. Moyer AM, Sun Z, Batzler AJ, Li L, Schaid DJ, Yang P, Weinshilboum RM. **Glutathione pathway genetic polymorphisms and lung cancer survival after platinum-based chemotherapy**. *Cancer Epidemiol Biomarkers Prev*. 2010; 19: 811-21.
24. Almasi CE, Høyer-Hansen G, Christensen IJ, Danø K, Pappot H. **Prognostic impact of liberated domain I of the urokinase plasminogen activator receptor in squamous cell lung cancer tissue**. *Lung Cancer*. 2005; 48: 349-55.

Table 1. 3-SNP interaction signature constrained to one gene set.

SNP	Gene ID	Function	OR*	p-value	MAF
rs2687105	CYP3A4	intron	0.2213	$5.6*10^{-7}$	2.5%
rs2397118	GSTA5	coding-nonsyn			5.9%
rs669674	GSTA4	intron			8.6%

OR=odds ratio
MAF=minor allele frequency
*odds ratio corresponds to a decreased risk in subjects with major homozygous genotypes in all three SNPs

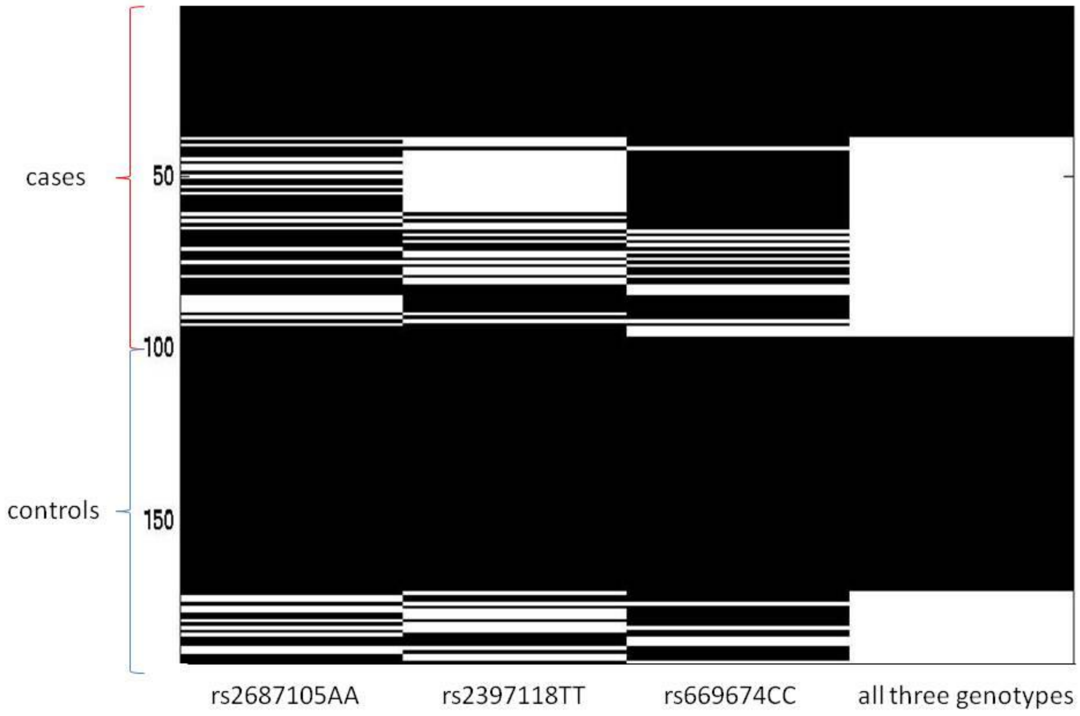
Table 2. 3-SNP interaction signature constrained to two gene sets.

SNP	Gene ID	Function	OR*	p-value	MAF
rs182623	GSTA4	nearGene-5	0.1514	$1.3*10^{-7}$	30.7%
rs4760	PLAUR	coding-nonsyn			12.5%
rs1934963	CYP2C9	intron			20.8%

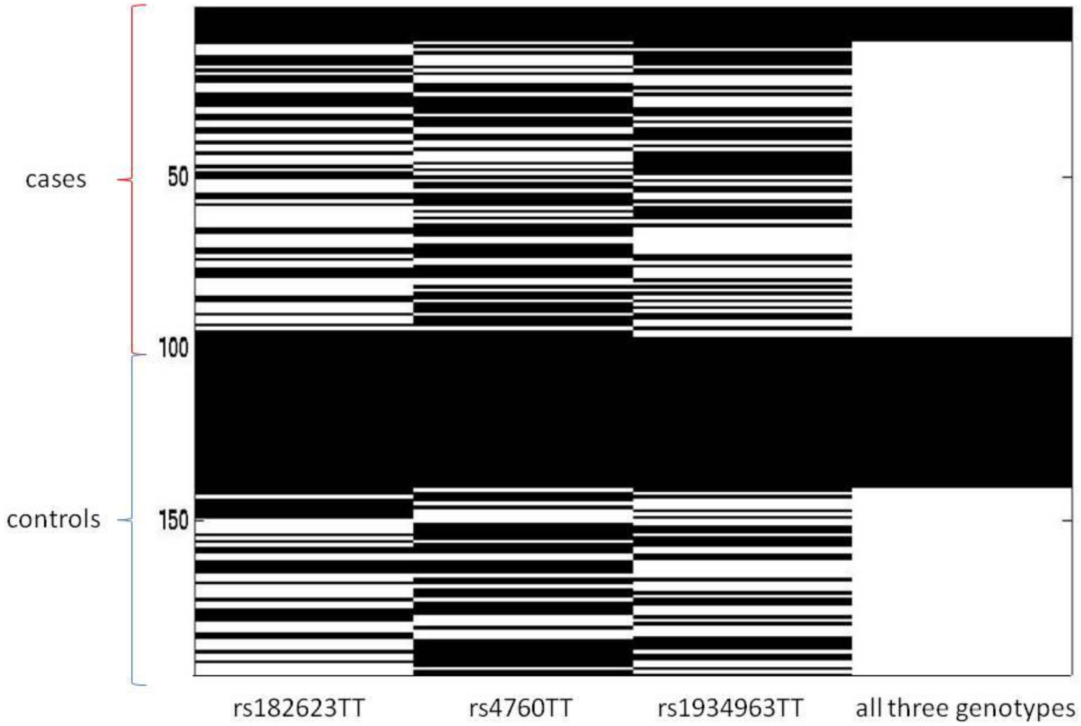
OR=odds ratio
MAF=minor allele frequency
*odds ratio corresponds to a decreased risk in subjects with major homozygous genotypes in all three SNPs

Figure 1. Graphical representation of the 3-SNP signature frequencies in cases vs. controls A) constrained to one gene set B) constrained to two gene sets. The black blocks within the graph represent those subjects that contain homozygous-major genotypes in the three SNPs represented at the bottom. The fourth column to the far right represents only those subjects containing all three homozygous-major genotypes in the represented SNPs. Cases and controls are marked outside the graph. It may be observable that there is a clear difference in the clustering of the subjects containing all three genotypes between the cases and the controls, controls having a higher frequency of the 3 genotypes in both, figures 1A and 1B.

A)


cases
50
100
controls
150
rs2687105AA
rs2397118TT
rs669674CC
all three genotypes

B)


cases
controls
50
100
150
rs182623TT
rs4760TT
rs1934963TT
all three genotypes

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

Multiple myeloma (MM) is an incurable B-cell malignancy characterized by the accumulation of clonal plasma cells in the bone marrow (1). Myeloma is genetically heterogenous and rare, and lack of clinical trials enlisting large subject numbers, as well as lack of repetition of trials with similar treatments, has hindered SNP association studies and validations. However, there are common apoptotic and oncogenic pathways that have enabled us to focus on a biologically driven approach by studying genes and variations relevant in cancer initiation and progression processes—we developed a custom BOAC SNP panel with 3,404 SNP in 983 genes (2).

Active smoking is responsible for about 90% of lung cancer cases. Studying lung cancer has been primarily focused on studying molecular mechanisms of carcinogens in tobacco-smoke; it is one of the rare diseases for which there is a known environmental exposure. Genetic variants can alter lung cancer risk by affecting enzymatic activity of proteins involved in the metabolism of harmful tobacco-smoke carcinogens. The dose to which an individual is exposed is critical and therefore, understanding gene—environment interaction is crucial. Genetic risk variants can produce significant effects at differing levels of exposure (3, 4); we observed a dose-effect in lung cancer, upon discovering an interaction between a variant in *CYP1B1* and 4-(methylnitrosamino)-1-(30pyridyl)-1-butanone (NNK) exposure.

The studies compiled in this thesis focus on genetic variations leading to inter-individual variability to disease risk, clinical outcome and environmental agent metabolism in myeloma and lung cancer. Chapter two of this thesis addresses genetic variant impact on bone disease. Bone disease is a major clinical problem in patients with myeloma and negatively affects their quality of life (5). The novelty of this study is in combining *DKK1* gene expression profile with a newly discovered SNP profile to generate a better predictor of bone disease. Recent studies have emphasized a crucial role of the Wnt-signaling inhibitor *DKK1* in the pathogenesis of the osteolytic bone lesions in myeloma (6). We discovered a SNP profile containing genes *EPHX1*, *IGF1R*, *IL4* and *Gsk3,* which in combination with *DKK1* expression profile, better predicts bone disease, and thus may have potential utility in prognostic models of myeloma bone disease.

Chapter three of this thesis explores the entire 3,404 SNP panel in relation to progression-free survival and disease risk in myeloma. We helped develop novel combinatorial search algorithms to efficiently and robustly search for SNP-SNP interactions in association with risk and outcomes. We explored the entire panel as well as constrained SNPs to defined pathways and gene sets in order to increase the power of the analysis and biologically drive the study. We identified individual and pairs of interacting SNPs in genes involved in drug metabolism and detoxification, immunity, DNA repair and signaling cascades important to MM risk and survival. We were also successful in validating some of our previous findings using these novel computational algorithms. The results in this study will help elucidate genetic complexity of MM by studying interplay of genomic variations.

Chapter four explores gene—environment interaction between variants in genes involved in tobacco-smoke carcinogen metabolism and an established environmental exposure, NNAL. We identified a variant in *CYP1B1*, N453S, which has a functional impact: the variant decreases cellular protein levels and enzyme half-life (7-9). Thus, it is not surprising that this variant alters lung cancer risk upon exposure to NNK, its metabolite 4-(Methylnitrosamino)-1-(3-Pyridyl)-1-Butanol (NNAL) and polycyclic aromatic hydrocarbons (PAHs).

Finally, in chapter five of this thesis we explored the rest of the BOAC SNP panel using a novel combinatorial search computational method, by constraining SNPs to gene sets to decrease the false positive rate and increase the statistical power of the analysis. We were successful in identifying variants in genes *CYP3A4*, *GSTA5*, *GSTA4*, *PLAUR* and *CYP2C9*, correlating with risk of lung cancer. These genes are important in drug metabolism and detoxification, and have been implicated in association with lung cancer risk and metabolism of various exposure agents. In this study, we explored and successfully identified SNP-SNP interactions that alter lung cancer risk. The fact that we obtained non-overlapping results from the two lung cancer studies was to be expected. In chapter four, we uncovered a gene-environment relationship that would have otherwise been hidden had we not included the interaction term (NNAL levels). In chapter five, we focused on SNP main effects, investigating associations between SNPs and lung cancer, without a consideration for the NNAL levels. We did not find any such interactions in chapter four, most likely because chapter five, utilizing new methodology, allowed us to uncover interactions of SNP pairs in association with lung cancer risk.

Since cytochrome P450 and Phase II enzymes are mainly responsible for drug and environmental agent metabolism, it would be appropriate to try to validate our results on a platform that focuses on those genes, decreasing the number of SNPs analyzed, thus increasing the power of the statistical analysis. The validation should be in another cohort of all smokers, half of whom have lung cancer. Measured NNAL levels are required for the validation of chapter four results, needed for the interaction term. Attempting to validate previous results in a larger cohort is essential for decreasing the false positive rate. Validating the results from chapter five should employ a method that can successfully search for SNP-SNP interactions.

The underlining theme of this thesis is that there are genetic variants that affect inter-individual risk and response. With limitations in small sample numbers and lack of additional clinical trials with similar treatments needed for validation, our approaches relied on developing novel computational algorithms to attempt to identify variants and their interactions.

Future directions include a study in collaboration with Dr. Celine Vachon's Lab at Mayo. They conducted a SNP genome-wide scan study (GWAS) and identified top hits. We are collaborating on a validation of their results. Validation will be conducted in 750 mouthwash BOAC samples, patient provided with informed consent. A thousand controls will be provided by Dr. Celine Vachon, subjects without multiple myeloma. Our research has focused on targeted genotyping of predefined pathways rather than GWAS primarily due to small sample numbers. Endeavoring in collaborative research is a crucial step to better understand etiology of this disease, as well as drug metabolism from chemotherapeutic treatments important for prolonging survival and the quality of life of myeloma patients.

The Van Ness Lab is continuing to bank DNA from myeloma patients, BOAC mouthwash, as well as institutional trials. Upon reaching high sample numbers in order to gain power for statistical analysis, there will be a possibility for a myeloma epidemiology study. The first MM etiology study has already been conducted and is presented in Appendix II of this thesis.

There are ongoing studies in the Van Ness Lab, generating drug resistant mouse cell lines. The Van Ness lab has previously developed a series of transgenic animals that target aberrant expression of genes affecting growth and survival pathways in B- and plasma cells (10). Previous research has demonstrated that the genetic background of myeloma patient influences chemotherapeutic response (11). In the future, it will be important to look at human myeloma cell lines and investigate variants in genes whose expression may be correlated with drug resistance. Additionally, some of the variants we identified to correlate with progression-free survival, presented here in Chapter 2, and Appendix I, may be investigated and potentially validated in cell line resistance studies.

There is a new and emerging forefront in SNP research. Multiple studies have highlighted the role that microRNAs have in physiological processes and how their deregulation can lead to cancer. It has been proposed that the presence of single nucleotide polymorphisms in microRNA genes, their processing machinery and target binding sites affects cancer risk, treatment efficacy and patient prognosis (12-16). There is a possibility in investigating SNPs in miRNA genes and their processing machinery in relation to drug resistance in cell lines currently being developed in the Van Ness Lab.

The future of the SNP research field is continuously under review. SNPs are the most frequent type of variation in the human genome, and they occur once every several hundred base pairs (17). They have been studied extensively for defining disease candidate gene regions, evolutionary studies, and establishing functional relationships between genotypic and phenotypic differences. The recent advent of newer technologies has increased the throughput of SNP genotyping and decreased the cost. The latest technologies have allowed the evaluation of SNPs at a single locus, as well as on a genome-wide level at densities that had previously thought to be unobtainable. Such large scale, SNP analysis studies have met with some success in identifying genes and genomic regions involved in the development of various disease phenotypes. There are opposing thoughts as to whether focused, targeted genotyping or the GWAS approaches may be more successful at identifying variants associated with disease phenotypes. While the GWAS approaches, being mapping tools, may successfully identify genomic regions of interest, they require tremendously large sample numbers in order to eliminate the false positives and negatives. Moreover, GWAS studies often identify variants with low associated risk, and poorly capture structural variants, which may have significant functional effects. This could be due to the GWAS technologies missing rare variants, which may be the answer. Targeted genotyping may miss important polymorphic markers, but the potential for type I errors is smaller, and these studies can be conducted on a smaller scale with limited sample numbers, especially relevant to rare diseases sush as myeloma. All of these SNP analysis technologies can be applied towards identifying new cancer genes, but their limitations must be considered and eventually overcome.

Targeted genotyping tools, such as the BOAC SNP chip designed in our laboratory (2), should continuously be updated in order to maintain the value of the information they contain. New polymorphisms are constantly identified and added to increasing databases, many of which are not functionally validated. In order to keep a SNP chip up to date, many of the newly discovered SNPs, relevant to cancer processes and drug metabolism and transport, would have to be added. If we were to redesign our BOAC SNP panel, adding SNPs previously validated in vitro, in addition to those occurring in regulatory regions and expected to have a functional impact, would be crucial. Regulatory-region SNPs are heavily underrepresented on DNA arrays, but may have a large impact on disease processes. Therefore, redesigning BOAC SNP chip would include adding a significant number of regulatory and already functionally validated polymorphisms.

Some may argue that whole-genome sequencing will soon diminish the value of DNA arrays, allowing for a more complete picture of the human genome. However, this technology is still too expensive and requires too long of a turn-around-time to be utilized for commercial purposes. Even though there are efforts being made to increase the speed, decrease the cost and maintain a high accuracy of whole-genome sequencing (18, 19), these relevant parameters are still not attained. Although a sequencing method that produces an error rate of 1 in 100,000 bases may be considered robust, that kind of error rate in the human diploid genome would mean about 60,000 errors per genome, which is a significant number of false positives and negatives, and clinically not applicable.

Moreover, even though whole-genome sequencing would provide significantly more information and a fuller coverage of the human genome than DNA arrays, it would also provide a challenge for the computational world. Whole-genome sequencing does not provide an analysis of what the generated data means or how that data can be utilized in various clinical applications, such as in medicine to help prevent disease. As of now, the companies that are working on providing full genome sequencing do not provide clinical analytical services for the interpretation of the raw genetic data. Therefore, for this data to be useful, researchers and/or companies first need to develop robust analytical methods to conduct data analysis and make it useful to physicians and patients.

Another issue to consider is the protection of human rights to their privacy. Whole-genome sequences may be predictive of an individual's predisposition to diseases, information that may be misused by entitites such as insurance companies. On the other hand, one may speculate that health care professionals will eventually be able to use genomic information to attempt to either minimize the impact of a particular disease or avoid it through the implementation of personalized, preventive medicine. Full genome sequencing will enable the analysis of the entire human genome of an individual, and therefore a detection of all disease-related genetic variants, regardless of the genetic variant's prevalence or frequency. This information will aid a field of Personalized Medicine and will move the clinical genetic revolution forward. Undoubtedly, full genome sequencing is of great importance for research into the basis of genetic disease. However, it should be recognized that despite advancements in genome sequencing technology, incomplete understanding of the significance of individual variants or combinations of variants will limit the widespread usefulness of full genome sequencing in medicine until its clinical utility can be demonstrated.

It is important to consider other variations in human genome and their importance and utility in studying disease risk and outcomes, such as epigenetics, copy number variations, insertions/deletions, etc. In addition, miRNAs add to the complex interplay of epigenetics and create an intricate network. As a consequence of their ability to regulate gene expression, microRNAs are involved in the most crucial cellular processes, from development, differentiation, cell cycle regulation to senescence and metabolism. MicroRNA expression is aberrant in several human diseases, including cancer. The first evidence of microRNA involvement in cancer was reported by Calin et al. (20). They observed knockdown or knockout of miR-15a and miR-16-1 in approximatively 69% of CLL patients. After these initial observations, all the known microRNAgenes were mapped. It was found that miRNA genes are frequently located in cancer associated genomic regions (CAGRs), including minimal regions of amplifications, loss of heterozygosity (LOH), common breakpoint regions in or near oncogenes (OGs) and tumor suppressor genes (TSGs), and fragile sites, which are preferential sites of chromatid exchange, deletion, translocation, amplification, or integration of plasmid DNA and tumor-associated viruses (21). These initial studies have lead to an emergence of miRNA research filed and uncovered their importance in human cancer. They have shown that alterations in miRNA expression in connection to human cancer are not isolated cases, but the rule. Moreover, SNPs in miRNAs, an emerging field and a forefront of SNP research, have become increasingly interesting to the research community. Although there is a rapid shift to whole-genome sequencing, SNPs should remain important players in studying cancer risk, drug metabolism and clinical outcomes in an attempt to generate best predictions.

In summary, we have investigated genetic variations that alter inter-individual response to myeloma and lung cancer risk and clinical outcomes. We demonstrated that using targeted custom SNP chip for genotyping rather than genome-wide scans was a successful approach, due to small sample numbers. We identified variants in genes involved in drug metabolism, DNA repair, inflammation, and bone metabolism in association with myeloma bone disease, prognosis and risk. We also identified a novel gene—environment interaction in lung cancer and additional SNP-SNP interactions. These results will be important in future research investigating genetic background of myeloma and lung cancer patients and how it influences risk and various clinical outcomes.

References:

1. Berg T, Cohen SB, Desharnais J, Sonderegger C, Maslyar DJ, Goldberg J, Boger DL, Vogt PK. **Small-molecule antagonists of Myc/Max dimerization inhibit Myc-induced transformation of chicken embryo fibroblasts**. *Proc Natl Acad Sci U.S.A*. 2002; 99: 3830-5.

2. Van Ness B, Ramos C, Haznadar M, *et.al*. **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med*. 2008; 6:26.

3. Vineis P, Bartsch H, Caporaso N, Harrington AM, Kadlubar FF, Landi MT, Malaveille C, Shields PG, Skipper P, Talaska G, et al. **Genetically based N-acetyltransferase metabolic polymorphism and low-level environmental exposure to carcinogens**. *Nature*. 1994; 369: 154-6.

4. Garte S, Zocchetti C, Taioli E. Garte S, Zocchetti C, Taioli E. **Gene--environment interactions in the application of biomarkers of cancer susceptibility in epidemiology**. *IARC Sci Publ*. 1997; 142: 251-64.

5. Cocks K, Cohen D, Wisløff F, Sezer O, Lee S, Hippe E, Gimsing P, Turesson I, Hajek R, Smith A, Graham L, Phillips A, Stead M, Velikova G, Brown J; EORTC Quality of Life Group. **An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-MY20) in assessing the quality of life of patients with multiple myeloma**. *Eur J Cancer*. 2007; 43: 1670-8.

6. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD Jr. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483-94.

7. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol*. 2005; 67: 435-443.

8. Mammen JS, Pittman GS, Li Y et al. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis.* 2003; 24: 1247-1255.

9. Aklillu E, Ovrebo S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res*. 2005; 65: 5105-5111.

10. Linden M, Kirchhof N, Kvitrud M, Van Ness B. Linden M, Kirchhof N, Kvitrud M, Van Ness B. **ABL-MYC retroviral infection elicits bone marrow plasma cell tumors in Bcl-X(L) transgenic mice**. *Leuk Res*. 2005; 29: 435-44.

11. Rowley M, Liu P, Van Ness B. **Heterogeneity in therapeutic response of genetically altered myeloma cell lines to interleukin 6, dexamethasone, doxorubicin, and melphalan**. *Blood*. 2000; 96: 3175-80.

12. Liu Z, Li G, Wei S, Niu J, El-Naggar AK, Sturgis EM, Wei Q. **Genetic variants in selected pre-microRNA genes and the risk of squamous cell carcinoma of the head and neck**. *Cancer*. 2010. [Epub ahead of print]

13. Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, Wojcik SE, Ferdin J, Kunej T, Xiao L, Manoukian S, Secreto G, Ravagnani F, Wang X, Radice P, Croce CM, Davuluri RV, Calin GA. **Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility**. *Cancer Res*. 2010; 70: 2789-98.

14. Zhou X, Chen X, Hu L, Han S, Qiang F, Wu Y, Pan L, Shen H, Li Y, Hu Z. **Polymorphisms involved in the miR-218-LAMB3 pathway and susceptibility of cervical cancer, a case-control study in Chinese women**. *Gynecol Oncol*. 2010; 117: 287-90.

15. Hu Z, Chen J, Tian T, Zhou X, Gu H, Xu L, Zeng Y, Miao R, Jin G, Ma H, Chen Y, Shen H. **Genetic variants of miRNA sequences and non-small cell lung cancer survival**. *J Clin Invest*. 2008; 118: 2600-8.

16. Despierre E, Lambrechts D, Neven P, Amant F, Lambrechts S, Vergote I. **The molecular genetic basis of ovarian cancer and its roadmap towards a better treatment**. *Gynecol Oncol*. 2010; 117: 358-65.

17. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N et al. **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet*. 1999; 22: 231–238.

18. Carlson, Rob (2007-01-02). "A Few Thoughts on Rapid Genome Sequencing and The Archon Prize - synthesis". Synthesis.cc. http://synthesis.cc/2007/01/a-few-thoughts-on-rapid-genome-sequencing-and-the-archon-prize.html. Retrieved 2009-02-23.

19. PRIZE Overview: Archon X PRIZE for Genomics http://genomics.xprize.org/archon-x-prize-for-genomics/prize-overview

20. G.A. Calin, *et al*. **Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia**. *Proc. Natl. Acad. Sci. U. S. A. 99*. 2002; 15524–15529.

21. G.A. Calin, *et al*. **Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers**. *Proc. Natl. Acad. Sci. U. S. A. 101*. 2004; 2999–3.

BIBLIOGRAPHY

CHAPTER 1

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. **Cancer statistics**. 2009. *CA Cancer J Clin*. 2009; 59: 225-49.

2. Cohen HJ, Crawford J, Rao MK, Pieper CF, Currie MS. **Racial differences in the prevalence of monoclonal gammopathy in a community-based sample of the elderly**. *Am J Med*. 1998; 104: 439-44.

3. Munshi NC. **Plasma cell disorders: an historical perspective.** *Hematology Am Soc Hematol Educ Program*. 2008: 297.

4. Kumar SK, Rajkumar SV, Dispenzieri A, Lacy MQ, Hayman SR, Buadi FK, Zeldenrust SR, Dingli D, Russell SJ, Lust JA, Greipp PR, Kyle RA, Gertz MA. **Improved survival in multiple myeloma and the impact of novel therapies**. *Blood*. 2008; 111: 2516-20.

5. Brenner H, Gondos A, Pulte D. **Recent major improvement in long-term survival of younger patients with multiple myeloma**. *Blood*. 2008; 111: 2521-6.

6. Michigami T, *et al*. **Cell-cell contact between marrow stromal cells andmyeloma cells via VCAM-1 and alpha(4)beta(1)-integrin enhances production ofosteoclast-stimulating activity**. *Blood*. 2000; 96: 1953–1960.

7. Hazlehurst LA, Damiano JS, Buyuksal I, Pledger WJ, Dalton WS. Adhesion **to fibronectin via beta1 integrins regulates p27kip1 levels and contributes to cell adhesion mediated drug resistance (CAM-DR)**. *Oncogene*. 2000; 19: 4319-27.

8. Vacca A, Ria R, Presta M, Ribatti D, Iurlaro M, Merchionne F, Tanghetti E, Dammacco F. **alpha(v)beta(3) integrin engagement modulates cell adhesion, proliferation, and protease secretion in human lymphoid tumor cells**. *Exp Hematol*. 2001; 29: 993-1003.

9. Anderson KC, Lust JA. **Role of cytokines in multiple myeloma**. *Semin Hematol*. 1999; 36(1 Suppl 3): 14-20.

10. Podar K, Hideshima T, Chauhan D, Anderson KC. **Targeting signalling pathways for the treatment of multiple myeloma**. *Expert Opin Ther Targets*. 2005; 9: 359-81.

11. Giuliani N, Rizzoli V, Roodman GD. **Multiple myeloma bone disease: Pathophysiology of osteoblast inhibition.** *Blood*. 2006; 108: 3992-6.

12. Kyle RA, Therneau TM, Rajkumar SV, Offord JR, Larson DR, Plevak MF, Melton LJ 3rd. **A long-term study of prognosis in monoclonal gammopathy of undetermined significance**. *N Engl J Med*. 2002; 346: 564-9.

13. Bataille R, Chappard D, Marcelli C, Dessauw P, Baldet P, Sany J, Alexandre C. **Recruitment of new osteoblasts and osteoclasts is the earliest critical event in the pathogenesis of human multiple myeloma**. *J Clin Invest*. 1991; 88: 62-6.

14. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD Jr. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483-94.

15. Oshima T, Abe M, Asano J, Hara T, Kitazoe K, Sekimoto E, Tanaka Y, Shibata H, Hashimoto T, Ozaki S, Kido S, Inoue D, Matsumoto T. **Myeloma cells suppress bone formation by secreting a soluble Wnt inhibitor, sFRP-2**. *Blood*. 2005; 106: 3160-5.

16. Giuliani N, Colla S, Morandi F, Lazzaretti M, Sala R, Bonomini S, Grano M, Colucci S, Svaldi M, Rizzoli V. **Myeloma cells block RUNX2/CBFA1 activity in human bone marrow osteoblast progenitors and inhibit osteoblast formation and differentiation**. *Blood*. 2005; 106: 2472-83.

17. Standal T, Abildgaard N, Fagerli UM, Stordal B, Hjertner O, Borset M, Sundan A. **HGF inhibits BMP-induced osteoblastogenesis: possible implications for the bone disease of multiple myeloma**. *Blood*. 2007; 109: 3024-30.

18. Lee JW, Chung HY, Ehrlich LA, Jelinek DF, Callander NS, Roodman GD, Choi SJ. **IL-3 expression by myeloma cells increases both osteoclast formation and growth of myeloma cells**. *Blood*. 2004; 103: 2308-15.

19. Ehrlich LA, Chung HY, Ghobrial I, Choi SJ, Morandi F, Colla S, Rizzoli V, Roodman GD, Giuliani N. **IL-3 is a potential inhibitor of osteoblast differentiation in multiple myeloma**. *Blood*. 2005; 106: 1407-14.

20. Clevers H. **Wnt/beta-catenin signaling in development and disease**. *Cell*. 2006; 127: 469-80.

21. Pinzone JJ, Hall BM, Thudi NK, Vonau M, Qiang YW, Rosol TJ, Shaughnessy JD Jr. **The role of Dickkopf-1 in bone development, homeostasis, and disease.** *Blood*. 2009; 113: 517-25.

22. Kaiser M, Mieth M, Liebisch P, Oberländer R, Rademacher J, Jakob C, Kleeberg L, Fleissner C, Braendle E, Peters M, Stover D, Sezer O, Heider U. **Serum concentrations of DKK-1 correlate with the extent of bone disease in patients with multiple myeloma**. *Eur J Haematol*. 2008; 80: 490-4.

23. Drake M, Ng A, Kumar S, et al. **Increases in serum levels of dickkopf 1 are associated with alterations in skeletal microstructure in monoclonal gammopathy of undetermined significance** [abstract]. In: T*he 31st Annual Meeting of the American Society for Bone and Mineral Research*, September 2009, Denver USA.

24. Qiang YW, Barlogie B, Rudikoff S, Shaughnessy JD Jr. **Dkk1-induced inhibition of Wnt signaling in osteoblast differentiation is an underlying mechanism of bone loss in multiple myeloma**. *Bone*. 2008; 42: 669-80.

25. Qiang YW, Chen Y, Stephens O, Brown N, Chen B, Epstein J, Barlogie B, Shaughnessy JD Jr. **Myeloma-derived Dickkopf-1 disrupts Wnt-regulated osteoprotegerin and RANKL production by osteoblasts: a potential mechanism underlying osteolytic bone lesions in multiple myeloma.** *Blood*. 2008; 112: 196-207.

26. Edwards CM, Edwards JR, Lwin ST, Esparza J, Oyajobi BO, McCluskey B, Munoz S, Grubbs B, Mundy GR. **Increasing Wnt signaling in the bone marrow microenvironment inhibits the development of myeloma bone disease and reduces tumor burden in bone in vivo**. *Blood*. 2008; 111: 2833-42.

27. Qiang YW, Shaughnessy JD Jr, Yaccoby S. **Wnt3a signaling within bone inhibits multiple myeloma bone disease and tumor growth**. *Blood*. 2008; 112: 374-82.

28. Raab MS, Breitkreutz I, Anderson KC. **Targeted treatments to improve stem cell outcome: old and new drugs**. *Bone Marrow Transplant*. 2007; 40: 1129-37.

29. Richardson PG, Hideshima T, Mitsiades C, Anderson KC. **The emerging role of novel therapies for the treatment of relapsed myeloma**. *J Natl Compr Canc Netw*. 2007; 5: 149-62.

30. Richardson PG, Barlogie B, Berenson J, Singhal S, Jagannath S, Irwin D, Rajkumar SV, Srkalovic G, Alsina M, Alexanian R, Siegel D, Orlowski RZ, Kuter D, Limentani SA, Lee S, Hideshima T, Esseltine DL, Kauffman M, Adams J, Schenkein DP, Anderson KC. **A phase 2 study of bortezomib in relapsed, refractory myeloma**. *N Engl J Med*. 2003; 348: 2609-17.

31. Garrett IR, Chen D, Gutierrez G, Zhao M, Escobedo A, Rossini G, Harris SE, Gallwitz W, Kim KB, Hu S, Crews CM, Mundy GR. **Selective inhibitors of the osteoblast proteasome stimulate bone formation in vivo and in vitro**. *J Clin Invest*. 2003; 111: 1771-82.

32. Bellido T, Ali AA, Plotkin LI, Fu Q, Gubrij I, Roberson PK, Weinstein RS, O'Brien CA, Manolagas SC, Jilka RL. **Proteasomal degradation of Runx2 shortens parathyroid hormone-induced anti-apoptotic signaling in osteoblasts. A putative explanation for why intermittent administration is needed for bone anabolism**. *J Biol Chem*. 2003; 278: 50259-72.

33. Giuliani N, Morandi F, Tagliaferri S, Lazzaretti M, Bonomini S, Crugnola M, Mancini C, Martella E, Ferrari L, Tabilio A, Rizzoli V. **The proteasome inhibitor bortezomib affects osteoblast differentiation in vitro and in vivo in multiple myeloma patients**. *Blood*. 2007; 110: 334-8.

34. Qiang YW, Hu B, Chen Y, Zhong Y, Shi B, Barlogie B, Shaughnessy JD Jr. **Bortezomib induces osteoblast differentiation via Wnt-independent activation of beta-catenin/TCF signaling**. *Blood*. 2009; 113: 4319-30.

35. Oyajobi BO, Garrett IR, Gupta A, Flores A, Esparza J, Muñoz S, Zhao M, Mundy GR. **Stimulation of new bone formation by the proteasome inhibitor, bortezomib: implications for myeloma bone disease**. *Br J Haematol*. 2007; 139: 434-8.

36. Zavrski I, Krebbel H, Wildemann B, Heider U, Kaiser M, Possinger K, Sezer O. **Proteasome inhibitors abrogate osteoclast differentiation and osteoclast function.** *Biochem Biophys Res Commun.* 2005; 333: 200-5.

37. WHO (February 2006). "Cancer". World Health Organization. http://www.who.int/mediacentre/factsheets/fs297/en/

38. Vaporciyan AA, Nesbitt JC, Lee JS et al. **Cancer Medicine**. *B C Decker*. 2000; pp. 1227–1292.

39. Merck Manual Professional Edition, Online edition. **Lung Carcinoma: Tumors of the Lung**. http://www.merck.com/mmpe/sec05/ch062/ch062b.html#sec05-ch062-ch062b-1405.

40. Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee IM, Marugame T, Palmer JR, Riboli E, Sobue T, Avila-Tang E, Wilkens LR, Samet JM. **Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies**. *PLoS Med*. 2008; 5: e185.

41. Gorlova OY, Weng SF, Zhang Y, Amos CI, Spitz MR. **Aggregation of cancer among relatives of never-smoking lung cancer patients**. *Int J Cancer*. 2007; 121: 111-8.

42. Hackshaw AK, Law MR, Wald NJ. **The accumulated evidence on lung cancer and environmental tobacco smoke**. *BMJ*. 1997; 315: 980-8.

43. Catelinois O, Rogel A, Laurier D, Billon S, Hemon D, Verger P, Tirmarche M. **Lung cancer attributable to indoor radon exposure in france: impact of the risk models and uncertainty analysis**. *Environ Health Perspect*. 2006; 114: 1361-6.

44. O'Reilly KM, Mclaughlin AM, Beckett WS, Sime PJ. **Asbestos-related lung disease**. *Am Fam Physician.* 2007; 75: 683-8.

45. Kabir Z, Bennett K, Clancy L. **Lung cancer and urban air-pollution in Dublin: a temporal association?** *Ir Med J.* 2007; 100: 367-9.

46. Coyle YM, Minahjuddin AT, Hynan LS, Minna JD. **An ecological study of the association of metal air pollutants with lung cancer incidence in Texas**. *J Thorac Oncol*. 2006; 1: 654-61.

47. Chiu HF, Cheng MH, Tsai SS, Wu TN, Kuo HW, Yang CY. **Outdoor air pollution and female lung cancer in Taiwan**. *Inhal Toxicol*. 2006; 18: 1025-31.

48. Carmona, RH. **Secondhand smoke exposure causes disease and premature death in children and adults who do not smoke.** *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General*. U.S. Department of Health and Human Services. 2006. http://www.surgeongeneral.gov/library/secondhandsmoke.

49. WHO International Agency for Research on Cancer. **Tobacco Smoke and Involuntary Smoking**. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans 83*. 2002. http://monographs.iarc.fr/ENG/Monographs/vol83/volume83.pdf.

50. World Health Organization. **The World Health Report 2003: Shaping the Future**. World Health Organization Geneva, Switzerland. 2003; pp91-94.

51. International Union Against Cancer. 2007; www.deathsfromsmoking.net

52. Hecht SS. **Biochemistry, biology, and carcinogenicity of tobacco-specific *N*-nitrosamines**. *Chem. Res. Toxicol*. 1998; 11, 559-603.

53. Jalas JR, Hecht SS, Murphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem Res Toxicol* 2005; 18: 95-110.

54. Cooper CS, P. L. Grover, and P. Sims. The **metabolism and activation of benzo[a]pyrene**. *Prog Drug Metab.* 1983; 7: 295-396.

55. Conney AH. **Induction of microsomal enzymes by foreign chemicals and carcinogenesis by polycyclic aromatic hydrocarbons: G. H. A. Clowes Memorial Lecture**. *Cancer Res*. 1982; 42: 4875-917.

56. Guengerich FP, Shimada T. **Oxidation of toxic and carcinogenic chemicals by human cytochrome P-450 enzymes**. *Chem Res Toxicol*. 1991; 4: 391-407.

57. Shimada T, Hayes CL, Yamazaki H, Amin S, Hecht SS, Guengerich FP, Sutter TR. **Activation of chemically diverse procarcinogens by human cytochrome P-450 1B1**. *Cancer Res*. 1996; 56: 2979-84.

58. Guengerich FP. **Metabolism of chemical carcinogens**. *Carcinogenesis*. 2000; 21: 345-51.

59. Rane A, Sjöqvist F, Orrenius S. **Cytochrome P-450 in human fetal liver microsomes**. *Chem Biol Interact*. 1971; 3 :305.

60. Kim JH, Sherman ME, Curriero FC, Guengerich FP, Strickland PT, Sutter TR. **Expression of cytochromes P450 1A1 and 1B1 in human lung from smokers, non-smokers, and ex-smokers**. *Toxicol Appl Pharmacol*. 2004; 199: 210-9.

61. Buters JT, Sakai S, Richter T, Pineau T, Alexander DL, Savas U, Doehmer J, Ward JM, Jefcoate CR, Gonzalez FJ. **Cytochrome P450 CYP1B1 determines susceptibility to 7, 12-dimethylbenz[a]anthracene-induced lymphomas**. *Proc Natl Acad Sci U.S.A*. 1999; 96: 1977-82.

62. Page TJ, O'Brien S, Holston K, MacWilliams PS, Jefcoate CR, Czuprynski CJ. **7,12-Dimethylbenz[a]anthracene-induced bone marrow toxicity is p53-dependent**. *Toxicol Sci*. 2003; 74: 85-92.

63. Buters JT, Mahadevan B, Quintanilla-Martinez L, Gonzalez FJ, Greim H, Baird WM, Luch A. **Cytochrome P450 1B1 determines susceptibility to dibenzo[a,l]pyrene-induced tumor formation**. *Chem Res Toxicol*. 2002; 15: 1127-35.

64. Han JF, He XY, Herrington JS, White LA, Zhang JF, Hong JY. **Metabolism of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) by human CYP1B1 genetic variants**. *Drug Metab Dispos*. 2008; 36: 745-52.

65. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol*. 2005; 67: 435-43.

66. Aklillu E, Øvrebø S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res*. 2005; 65: 5105-11.

67. Mammen JS, Pittman GS, Li Y, Abou-Zahr F, Bejjani BA, Bell DA, Strickland PT, Sutter TR. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis*. 2003; 24: 1247-55.

68. Vineis P, Bartsch H, Caporaso N, Harrington AM, Kadlubar FF, Landi MT, Malaveille C, Shields PG, Skipper P, Talaska G, et al. **Genetically based N-acetyltransferase metabolic polymorphism and low-level environmental exposure to carcinogens**. *Nature*. 1994; 369: 154-6.

69. Garte S, Zocchetti C, Taioli E. Garte S, Zocchetti C, Taioli E. **Gene--environment interactions in the application of biomarkers of cancer susceptibility in epidemiology**. *IARC Sci Publ*. 1997; 142: 251-64.

70. International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome**. *Nature*. 2001; 409: 860-921.

71. Venter JC, *et.al*. **The sequence of the human genome**. *Science*. 2001; 291: 1304-51.

72. International Human Genome Sequencing Consortium. **Finishing the euchromatic sequence of the human genome**. *Nature*. 2004; 431: 931-45.

73. Van Ness B, Ramos C, Haznadar M, *et.al*. **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med*. 2008; 6:26.

74. The International HapMap Consortium. **The International HapMap Project**. *Nature*. 2003; 426, 789-796. http:/www.hapmap.org

75. Packer, B.R., et al. **SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes**. *Nucleic Acids Res*. 2004; 32 Database issue: D528-32. http://snp500cancer.nci.nih.gov/

CHAPTER 2

1. Hideshima T, Mitsiades C, Tonon G, Richardson PG, Anderson KC. **Understanding multiple myeloma pathogenesis in the bone marrow to identify new therapeutic targets**. *Nat Rev Cancer.* 2007; 7: 585–598.
2. Giuliani N, Rizzoli V, Roodman GD. **Multiple myeloma bone disease: pathophysiology of osteoblast inhibition**. *Blood*. 2006; 108: 3992–3996.

3. Durie BGM, Salmon SE, Mundy GR. **Relation of osteoclast activating factor production to extent of bone disease in multiple myeloma**. *Br J Hematol*. 1981; 47: 21–30.
4. Harada S, Rodan G. **Control of osteoblast function and regulation of bone mass**. *Nature*. 2003; 423: 349–355.
5. Westendorf JJ, Kahler RA, Schroeder TM. **Wnt signaling in osteoblasts and bone diseases.** *Gene.* 2004; 341: 19–39.
6. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B et al. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483–2494.
7. Yaccoby S, Ling W, Zhan F, Walker R, Barlogie B, Shaughnessy Jr JD. **Antibody-based inhibition of DKK1 suppresses tumor-induced bone resorption and multiple myeloma growth in vivo**. *Blood.* 2007; 109: 2106–2111.
8. Colla S, Zhan F, Xiong W, Wu X, Xu H, Stephens O et al. **The oxidative stress response regulates DKK1 expression through the JNK signaling cascade in multiple myeloma plasma cells**. *Blood*. 2007; 109: 4470–4477.
9. Qian J, Xie J, Hong S, Yang J, Zhang L, Han X et al. **Dickkopf-1 (DKK-1) is a widely expressed and potent tumor-associated antigen in multiple myeloma**. *Blood*. 2007; 110: 1587–1594.
10. Choi SJ, Oba Y, Gazitt Y, Alsina M, Cruz J, Anderson J et al. **Antisense inhibition of macrophage inflammatory protein 1-alpha blocks bone destruction in a model of myeloma bone disease**. *J Clin Invest.* 2001; 108: 1833–1841.
11. Lentzsch S, Gries M, Janz M, Bargou R, Dorken B, Mapara MY. **Macrophage inflammatory protein 1-alpha (MIP-1 alpha) triggers migration and signaling cascades mediating survival and proliferation in multiple myeloma (MM) cells**. *Blood*. 2003; 101: 3568–3573.
12. Vallet S, Raje N, Ishitsuka K, Hideshima T, Podar K, Chhetri S et al. **MLN3897, a novel CCR1 inhibitor, impairs osteoclastogenesis and inhibits the interaction of multiple myeloma cells and osteoclasts**. *Blood.* 2007; 110: 3744–3752.
13. Zhan F, Huang Y, Colla S, Stewart JP, Hanamura I, Gupta S et al. **The molecular classification of multiple myeloma**. *Blood.* 2006; 108: 2020–2028.
14. Walker R, Barlogie B, Haessler J, Tricot G, Anaissie E, Shaughnessy Jr JD et al. **Magnetic resonance imaging in multiple myeloma: diagnostic and clinical implications**. *J Clin Oncol*. 2007; 25: 1121–1128.
15. Johnson DC, Corthals S, Ramos C, Hoering A, Cocks K, Dickens NJ et al. **Genetic associations with thalidomide mediated venous thrombotic events in myeloma identified using targeted genotyping**. *Blood.* 2008; 112: 4924–4934.
16. Van Ness B, Ramos C, Haznadar M, Hoering A, Haessler J, Crowley J et al. **Genomic variation in myeloma: design, content, and initial application of the bank on a cure SNP panel to analysis of survival**. *BMC Med.* 2008; 6: 26 [pages not specified].
17. Terry MT, Elizabeth J, Atkinson R. **An introduction to recursive partitioning using the rpart routines.** 1997. *Technical report 61*, Mayo Clinic. 2Available at:

http://mayoresearch.mayo.edu/mayo/research/biostat/techreports.cfm##R package available at: http://cran.r-project.org/src/contrib/Descriptions/rpart.html.

18. Agresti A. **An introduction to categorical data analysis**. *Wiley*: NJ, USA, 1996.
19. Breiman L. **Random forests**. *Machine Learning*. 2001; 45: 5–32.
20. Efron B, Tibshirani R. **An introduction to the Bootstrap**. *Chapman & Hall/CRC*: FL, USA, 1994.
21. Kaplan EL, Meier P. **Nonparametric estimation from incomplete observations**. *J Am Stat Assoc*. 1958; 53: 457–481.
22. Hirschhorn JN, Lohmueller K, Byrne E, Hurschhorn K. **A comprehensive review of genetic association studies**. *Genet Med*. 2002; 4: 45–61.
23. Moghaddam MF, Grant DF, Cheek JM, Green JF, Williamson KC, Hammock BD. **Bioactivation of leukotoxins to their toxic diols by epoxide hydrolase**. *Nature Med*. 1997; 3: 562–566.
24. Burchiel SW, Thompson TA, Lauer FT, Oprea TI. **Activation of dioxin response element (DRE)-associated genes by benzo (A) pyrene3,6-quinone and benzo (A) pyrene1,6-quinone in MCF-10A human mammary epithelial cells**. *Toxicol Appl Pharmacol* 2007; 221: 203–214.
25. Shi C-S, Huang N-N, Harrison K, Han S-B, Kehrl JH. **The mitogenactivated protein kinase kinase kinase kinase GCKR positively regulates canonical and noncanonical Wnt signaling in B lymphocytes**. *Mol Cell Biol* 2006; 26: 6511–6521.
26. Robinson JA, Chatterjee-Kishore M, Yaworsky PJ, Cullen DM, ZhaoW LC et al. **Wnt/b signaling is a normal physiological response to mechanical loading in bone**. *J Biol Chem* 2006; 281: 31720–31728.
27. Shi C-S, Tuscano JM, Witte ON, Kehrl JH. **GCKR links the Bcr-Abl oncogene and Ras to the stress-activated protein kinase pathway**. *Blood* 1999; 93: 1338–1345.
28. Knobloch J, Reimann K, Klotz L-O, Ruther U. **Thalidomide resistance is based upon the capacity of the glutathione-dependent antioxidant defense**. *Mol Pharmaceutics* 2008; 5: 1138–1144.
29. Edwards CM, Edwards JR, Lwin ST, Mundy GR. **Target Wnt signaling in myeloma in vivo; differential effects on tumor burden and myeloma bone disease**. *Blood* 2008; 111: 2833–2842.
30. Caspi M, Zilberberg A, Eldar-Finkelman H, Rosin-Arbesfeld R. **Nuclear GSK-3b inhibits the canonical Wnt signalling pathway in a b-catenin phosphorylation-independent manner**. *Oncogene* 2008; 27: 3546–3555.
31. Staal FJT, Luis TC, Tiemessen MM. **WNT signalling in the immune system: WNT is spreading its wings**. *Immunology* 2008; 8: 581–593.
32. Qiang Y-W, Chen Y, Stephens O, Brown N, Chen B, Epstein J et al. **Myeloma-derived Dixkkopf-1 disrupts Wnt-regulated osteoprotegerin and RANKL production by osteoblasts: a potential mechanism underlying osteolytic bone lesions in multiple myeloma**. *Blood* 2008; 112: 196–207.
33. Qiang Y-W, Shaughnessy JD, Yaccoby S. **Wnt3a signaling within bone inhibits multiple myeloma bone disease and tumor growth**. *Blood* 2008; 112: 374–382.

34. Edward CM. **Wnt signaling: bone's defense against myeloma**. *Blood* 2008; 112: 216–218.
35. Ferlin M, Noraz N, Hertogh C, Brochier J, Taylor N, Klein B. **Insulin-like growth factor induces the survival and proliferation of myeloma cells through an interleukin-6-independent transduction pathway.** *Br J Hematology* 2000; 111: 626–634.
36. Ge N-L, Rudikoff S. **Insulin-like growth factor 1 is a dual effector of multiple myeloma cell growth**. *Blood* 2000; 96: 2856–2861.
37. Mitsiades CS, Mitsiades N, Poulaki V, Schlossman R, Akivama M, Chauhan D et al. **Activation of NF-kB and upregulation of intracellular anti-apoptotic proteins via the IGF-1/Akt signaling in human multiple myeloma calls: therapeutic implications**. *Oncogene* 2002; 21: 5673–5683.
38. Podar K, Tai Y-T, Cole CE, Hideshima T, Sattler M, Hamblin A et al. **Essential role of caveolae in interleukin-6 and insulin-like growth factor I-triggered Akt-1-mediated survival of multiple myeloma cells**. *J Biol Chem* 2003; 278: 5794–5801.
39. DiGirolamo DJ, Mukherjee A, Fulzele K, Gan Y, Cao X, Frank SJ et al. **Mode of growth hormone action in osteoblasts**. *J Biol Chem* 2007; 282: 31666–31674.
40. Brechter AB. **Kinins-important regulators in inflammation induced bone resorption.** *Umea Univ Odontological Dissertation*. Department of Oral Cell Biology, Umea University: Umea, Sweden, 2006 (ISBN 91-7264-195-9).
41. Fujisawa T, Ikegami H, Kawaguchi Y, Ogihata T. **Meta-analysis of the association of Trp Arg polymorphism of b3-adrenergic receptor gene with body mass index**. *J Clin Endocrinol Metab* 1998; 83: 2441–2444.
42. Wang CY, Nguyen ND, Morrison NA, Eisman JA, Center JR, Nguyen TV. **b3-adrenergic receptor gene, body mass index, bone mineral density and fracture risk in elderly men and women: the dubbo osteoporosis epidemiology study (DOES)**. *BMC Med Genet* 2006; 7: 57.
43. Riancho JA, Zarrabeitia MT, Olmos JM, Amado JA, Gonzalez MJ. **Effects of interleukin-4 on human osteoblast-like cells**. *Bone Miner* 1993; 21: 53–61.
44. Durie BGM, Urnovitz HB, Murphy WH. **RT-PCR amplicons in the plasma of multiple myeloma patients, clinical relevance and molecular pathology.** *Acta Oncologica* 2000; 39: 789–796.
45. Hassett C, Robinson KB, Beck NB, Omiecinski CJ. **The human microsomal expoxide hydrolase gene (EPHX1): complete nucleotide sequence and structural characterization**. *Genomics* 1994; 23: 433–442.
46. Fretland AJ, Omiecinski CJ. **Epoxide hydrolases: biochemistry and molecular biology**. *Chem Biol Interact* 2000; 129: 41–59.
47. Omiecinski CJ, Hassett C, Hosagrahara V. **Epoxide hydrolasepolymorphism and role in toxicology**. *Toxicol Lett* 2000; 112–113: 365–370.
48. Berndt SI, Johnson D, Crowley J, Durie BGM, Hoover R, Katz M et al. **Large scale evaluation of genetic variation and the risk of multiple myeloma**. *Blood* 2008; 112, Abstract # 1679.

CHAPTER 3

1. Kyle RA, Rajkumar SV: **Plasma cell disorders**. In *Cecil textbook of medicine* 22nd edition. Edited by: Goldman L, Ausiello DA. Philadelphia: W.B. Saunders; 2004:1184-1195.
2. Venter JC, *et al*: **The sequence of the human genome**. *Science*. 2001, 291:1304-1351.
3. Lander ES, from the International Human Genome Sequencing Consortium *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
4. International Human Genome Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
5. Marchini J, Donnelly P, Cardon LR. **Genome-wide strategies for detecting multiple loci that influence complex diseases**. *Nat Genet* 2005; 37(4):413-7.
6. http://myeloma.org/ArticlePage.action?articleId=2099
7. Van Ness B, Ramos CR, Haznadar M, Hoering A, Haessler J, Crowley J, Jacobus S, Oken M, Rajkumar V, Greipp P, Barlogie B, Durie B, Katz M, Atluri G, Fang G, Gupta R, Steinbach M, Kumar V, Mushlin R, Johnson, D, Morgan G: **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med* 2008, **8**:6:26.
8. Fang G, Haznadar M, Wang W, Steinbach M, Van Ness B, Kumar V: **A Computationally Efficient and Statistically Powerful Framework for Searching High-order Epistasis with Systematic Pruning and Gene-set Constraints**. Technical Report 013, Department of Computer Science, University of Minnesota, 2010. (submitted)
9. Oken MM, Leong T, Lenhard RE, Greipp PR, Kay NE, Van Ness B, KeimowitzRM, Kyle RA: **The addition of interferon or high dosecyclophosphamide to standard chemotherapy in the treatmentof patients with multiple myeloma: Phase III EasternCooperative Oncology Group Clinical Trial EST 9486.** *Cancer* 1999, **86(6):**957-968.
10. Barlogie B, Kyle RA, Anderson KC, Greipp PR, Lazarus HM, HurdDD, McCoy J, Moore DF Jr, Sakhil SR, Lanier KS, Chapman RA,Cromer JN, Salmon SE, Durie B, Crowley JC: **Standard chemotherapycompared with high-dose chemoradiotherapy formultiple myeloma: final results of phase III US IntergroupTrial S9321.** *J Clin Oncol* 2006, **24(6):**929-936.

11. Hardenbol P, Baner J, Jain M *et al*. **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nat Biotechnol* 2003; 21(6):673-678.

12. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, Kumar V. **Subspace differential coexpression analysis: problem definition and a general approach**. *Pac Symp Biocomput*. 2010, 145-56.

13. Fang G, Pandey G, Gupta M, Steinbach M, Kumar V. **Mining Low-support Discriminative Patterns from Dense and High-dimensional Data**. Tech report; Department of Computer Science, University of Minnesota. 2009: 011. (Minor revision, IEEE Transactions on Data and Knowledge Engineering, 2010).

14. Hahn LW, Ritchie MD, Moore JH. **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions**. *Bioinformatics*. 2003; 12;19(3):376-82.

15. The Molecular Signature Database http://www.broadinstitute.org/gsea/msigdb/index.jsp

16. Yang JJ, Bhojwani D, Yang W, Cai X, Stocco G, Crews K, Wang J, Morrison D, Devidas M, Hunger SP, Willman CL, Raetz EA, Pui CH, Evans WE, Relling MV, Carroll WL. **Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia**. *Blood*. 2008; 112:4178-83.

17. Gross E, Busse B, Riemenschneider M, Neubauer S, Seck K, Klein HG, Kiechle M, Lordick F, Meindl A. **Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients**. *PLoS One*. 2008; 3: e4003.

18. Kleibl Z, Fidlerova J, Kleiblova P, Kormunda S, Bilek M, Bouskova K, Sevcik J, Novotny J. **Influence of dihydropyrimidine dehydrogenase gene (DPYD) coding sequence variants on the development of fluoropyrimidine-related toxicity in patients with high-grade toxicity and patients with excellent tolerance of fluoropyrimidine-based chemotherapy**. *Neoplasma*. 2009; 56:303-16.

19. Ofverholm A, Arkblad E, Skrtic S, Albertsson P, Shubbar E, Enerbäck C. **Two cases of 5-fluorouracil toxicity linked with gene variants in the DPYD gene**. *Clin Biochem*. 2010; 43:331-4.

20. Hayden PJ, Tewari P, Morris DW, Staines A, Crowley D, Nieters A, Becker N, de Sanjosé S, Foretova L, Maynadié M, Cocco PL, Boffetta P, Brennan P, Chanock SJ, Browne PV, Lawler M. **Variation in DNA repair genes XRCC3, XRCC4, XRCC5 and susceptibility to myeloma**. *Hum Mol Genet*. 2007; 15;16: 3117-27.

21. Roddam PL, Allan JM, Dring AM, Worrillow LJ, Davies FE, Morgan GJ. **Non-homologous end-joining gene profiling reveals distinct expression patterns associated with lymphoma and multiple myeloma**. *Br J Haematol*. 2010; 149:258-62.

22. Alexandrakis MG, Passam FH, Pappa CA, Dambaki C, Sfakiotaki G, Alegakis AK, Kyriakou DS, Stathopoulos E. **Expression of proliferating cell nuclear antigen (PCNA) in multiple myeloma: its relationship to bone marrow microvessel density and other factors of disease activity**. *Int J Immunopathol Pharmacol*. 2004; 17: 49-56.

23. Dumontet C, Landi S, Reiman T, Perry T, Plesa A, Bellini I, Barale R, Pilarski LM, Troncy J, Tavtigian S, Gemignani F. **Genetic polymorphisms associated with outcome in multiple myeloma patients receiving high-dose melphalan**. *Bone Marrow Transplant*. 2009. [Epub ahead of print]

24. Lincz LF, Kerridge I, Scorgie FE, Bailey M, Enno A, Spencer A**. Xenobiotic gene polymorphisms and susceptibility to multiple myeloma**. *Haematologica*. 2004; 89:628-9.

25. Liang J, Wang H, Xiao H, Li N, Cheng C, Zhao Y, Ma Y, Gao J, Bai R, Zheng H. **Relationship and prognostic significance of SPARC and VEGF protein expression in colon cancer**. J *Exp Clin Cancer Res*. 2010; 29(1):71. [Epub ahead of print]

26. Zhou Y, Li G, Wu J, Zhang Z, Wu Z, Fan P, Hao T, Zhang X, Li M, Zhang F, Li Q, Lu B, Qiao L. **Clinicopathological significance of E-cadherin, VEGF, and MMPs in gastric cancer**. *Tumour Biol*. 2010 [Epub ahead of print]

27. Donnem T, Al-Shibli K, Andersen S, Al-Saad S, Busund LT, Bremnes RM. **Combination of low vascular endothelial growth factor A (VEGF-A)/VEGF receptor 2 expression and high lymphocyte infiltration is a strong and independent favorable prognostic factor in patients with nonsmall cell lung cancer**. *Cancer*. 2010 [Epub ahead of print]

28. Zhang L, Hu Y, Sun CY, Li J, Guo T, Huang J, Chu ZB. **Lentiviral shRNA silencing of BDNF inhibits in vivo multiple myeloma growth and angiogenesis via down-regulated stroma-derived VEGF expression in the bone marrow milieu**. *Cancer Sci*. 2010; 101:1117-24.

29. Chen Q, Van der Sluis PC, Boulware D, Hazlehurst LA, Dalton WS. **The FA/BRCA pathway is involved in melphalan-induced DNA interstrand cross-link repair and accounts for melphalan resistance in multiple myeloma cells**. *Blood*. 2005; 106:698-705.

30. Wang M, Liu S, Liu P. **Gene expression profile of multiple myeloma cell line treated by arsenic trioxide**. *J Huazhong Univ Sci Technolog Med Sci*. 2007; 27:646-9.

31. Chiarle R, Simmons WJ, Cai H, Dhall G, Zamo A, Raz R, Karras JG, Levy DE, Inghirami G. **Stat3 is required for ALK-mediated lymphomagenesis and provides a possible therapeutic target**. *Nat Med*. 2005 Jun;11(6):623-9. Epub 2005 May 15.

32. Du ZX, Meng X, Zhang HY, Guan Y, Wang HQ. **Caspase-dependent cleavage of BAG3 in proteasome inhibitors-induced apoptosis in thyroid cancer cells**. *Biochem Biophys Res Commun*. 2008; 369:894-8.

33. Zhang H, Vakil V, Braunstein M, Smith EL, Maroney J, Chen L, Dai K, Berenson JR, Hussain MM, Klueppelberg U, Norin AJ, Akman HO, Ozçelik T, Batuman OA. **Circulating endothelial progenitor cells in multiple myeloma: implications and significance**. *Blood*. 2005; 105:3286-94.

34. Martin SK, Diamond P, Williams SA, To LB, Peet DJ, Fujii N, Gronthos S, Harris AL, Zannettino AC. **Hypoxia-inducible factor-2 is a novel regulator of aberrant CXCL12 expression in multiple myeloma plasma cells**. *Haematologica*. 2010; 95:776-84.

CHAPTER 4

1. International Agency for Research on Cancer. **Tobacco Smoke and Involuntary Smoking**. Lyon, FR: *IARC*; 2004. p 33-1187.

2. International Agency for Research on Cancer. **Smokeless Tobacco and Tobacco-Specific Nitrosamines**. Lyon, FR: *IARC*; 2007. p 548-553.

3. Straif K, Baan R, Grosse Y, Secretan B, El Ghissassi F, Cogliano V. **Carcinogenicity of polycyclic aromatic hydrocarbons**. *Lancet Oncol* 2005;6(12):931-932.

4. Hecht SS. **Tobacco carcinogens, their biomarkers and tobacco-induced cancer**. *Nat Rev Cancer* 2003;3(10):733-744.

5. Jalas JR, Hecht SS, Murphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem Res Toxicol* 2005;18(2):95-110.

6. Hecht SS. B**iochemistry, biology, and carcinogenicity of tobacco-specific N-nitrosamines**. *Chem Res Toxicol* 1998;11(6):559-603.

7. Cooper CS, P. L. Grover, and P. Sims. **The metabolism and activation of benzo[a]pyrene**. *ProgDrug Metab* 1983;7:295-396.

8. Church TR, Anderson KE, Caporaso NE et al. **A prospectively measured serum biomarker for a tobacco-specific carcinogen and lung cancer in smokers**. *Cancer Epidemiol Biomarkers Prev* 2009;18(1):260-266.

9. Hecht SS, Chen M, Yagi H, Jerina DM, Carmella SG. **r-1,t-2,3,c-4-Tetrahydroxy-1,2,3,4-tetrahydrophenanthrene in human urine: a potential biomarker for assessing polycyclic aromatic hydrocarbon metabolic activation**. *Cancer Epidemiol Biomarkers Prev* 2003;12(12):1501-1508.

10. Hecht SS, Carmella SG, Yoder A et al. **Comparison of polymorphisms in genes involved in polycyclic aromatic hydrocarbon metabolism with urinary phenanthrene metabolite ratios in smokers**. *Cancer Epidemiol Biomarkers Prev* 2006;15(10):1805-1811.

11. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. **The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status**. *Control Clin Trials* 2000;21(6 Suppl S):251S-272S.

12. Rothman KJ, Greenland S, editors. **Modern Epidemiology**. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998. xiii, 737 p.

13. Van Ness B, Ramos C, Haznadar M et al**. Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med* 2008;6:26.

14. Carmella SG, Yoder A, Hecht SS. C**ombined analysis of r-1,t-2,3,c-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol in smokers' plasm**a. *Cancer Epidemiol Biomarkers Prev* 2006;15(8):1490-1494.

15. Hecht SS. **Tobacco smoke carcinogens and lung cancer**. *J Natl Cancer Inst* 1999;91(14):1194-1210.

16. Hardenbol P, Baner J, Jain M et al. **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nat Biotechnol* 2003;21(6):673-678.

17. Greenland S, Pearl J, Robins JM. **Causal diagrams for epidemiologic research**. *Epidemiology* 1999;10(1):37-48.

18. Peto RLA, Boreham J et al. **Mortality from smoking in developed countries 1950-2000: Indirect estimates from National Vital Statistics**: Oxford University Press; 2006.

19. Biesalski HK, Bueno de Mesquita B, Chesson A et al. **European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel**. *CA Cancer J Clin* 1998;48(3):167-176; discussion 164-166.

20. Hecht SS. **Recent studies on mechanisms of bioactivation and detoxification of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), a tobacco-specific lung carcinogen**. *Crit Rev Toxicol* 1996;26(2):163-181.

21. Lin P, Hu SW, Chang TH. **Correlation between gene expression of aryl hydrocarbon receptor (AhR), hydrocarbon receptor nuclear translocator (Arnt), cytochromes P4501A1 (CYP1A1) and 1B1 (CYP1B1), and inducibility of CYP1A1 and CYP1B1 in human lymphocytes**. *Toxicol Sci* 2003;71(1):20-26.

22. Hoffman EC, Reyes H, Chu FF et al. **Cloning of a factor required for activity of the Ah (dioxin) receptor**. *Science* 1991;252(5008):954-958.

23. Jiang BH, Rue E, Wang GL, Roe R, Semenza GL. **Dimerization, DNA binding, and transactivation properties of hypoxia-inducible factor 1**. *J Biol Chem* 1996;271(30):17771-17778.

24. Whitlock JP, Jr**. Induction of cytochrome P4501A1**. *Annu Rev Pharmacol Toxicol* 1999;39:103-125.

25. Alexander DL, Zhang L, Foroozesh M, Alworth WL, Jefcoate CR. **Metabolism-based polycyclic aromatic acetylene inhibition of CYP1B1 in 10T1/2 cells potentiates aryl hydrocarbon receptor activity**. *Toxicol Appl Pharmacol* 1999;161(2):123-139.

26. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol* 2005;67(2):435-443.

27. Mammen JS, Pittman GS, Li Y et al. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis* 2003;24(7):1247-1255.

28. Aklillu E, Ovrebo S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res* 2005;65(12):5105-5111.

29. Nebert DW, Roe AL, Dieter MZ, Solis WA, Yang Y, Dalton TP. **Role of the aromatic hydrocarbon receptor and [Ah] gene battery in the oxidative stress response, cell cycle control, and apoptosis**. *Biochem Pharmacol* 2000;59(1):65-85.

30. Nebert DW, Benedict, W. F., and Kouri, R. E. *Chemical Carcinogenesis*. In: Ts'o POP, and DiPaolo, J. A, editor. New York: Marcel Dekker, Inc.; 1974. p 271-288.

31. Kellermann G, Shaw CR, Luyten-Kellerman M. **Aryl hydrocarbon hydroxylase inducibility and bronchogenic carcinoma**. *N Engl J Med* 1973;289(18):934-937.

32. Nebert DW, Dalton TP, Okey AB, Gonzalez FJ. **Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer**. *J Biol Chem* 2004;279(23):23847-23850.

33. Chang JT, Chang H, Chen PH, Lin SL, Lin P. **Requirement of aryl hydrocarbon receptor overexpression for CYP1B1 up-regulation and cell growth in human lung adenocarcinomas**. *Clin Cancer Res* 2007;13(1):38-45.

34. Shimada T, Hayes CL, Yamazaki H et al. **Activation of chemically diverse procarcinogens by human cytochrome P-450 1B1**. *Cancer Res* 1996;56(13):2979-2984.

35. Meyer BK, Pray-Grant MG, Vanden Heuvel JP, Perdew GH. **Hepatitis B virus X-associated protein 2 is a subunit of the unliganded aryl hydrocarbon receptor core complex and exhibits transcriptional enhancer activity**. *Mol Cell Biol* 1998;18(2):978-988.

36. Kazlauskas A, Poellinger L, Pongratz I. **Evidence that the co-chaperone p23 regulates ligand responsiveness of the dioxin (Aryl hydrocarbon) receptor**. *J Biol Chem* 1999;274(19):13519-13524.

37. Hord NG, Perdew GH. **Physicochemical and immunocytochemical analysis of the aryl hydrocarbon receptor nuclear translocator: characterization of two monoclonal antibodies to the aryl hydrocarbon receptor nuclear translocator**. *Mol Pharmacol* 1994;46(4):618-626.

38. Pollenz RS, Sattler CA, Poland A. **The aryl hydrocarbon receptor and aryl hydrocarbon receptor nuclear translocator protein show distinct subcellular localizations in Hepa 1c1c7 cells by immunofluorescence microscopy**. *Mol Pharmacol* 1994;45(3):428-438.

39. Hankinson O. **The aryl hydrocarbon receptor complex**. *Annu Rev Pharmacol Toxicol* 1995;35:307-340.

40. Probst MR, Reisz-Porszasz S, Agbunag RV, Ong MS, Hankinson O. **Role of the aryl hydrocarbon receptor nuclear translocator protein in aryl hydrocarbon (dioxin) receptor action**. *Mol Pharmacol* 1993;44(3):511-518.

41. Denison MS, Elferink CF, Phelan D. **The Ah receptor signal transduction pathway**. In: *Denison* MS, W.G. H, editors. Toxicant-Receptor Interactions in the Modulation of Signal Transduction and Gene Expression. Philadelphia: Taylor & Francis; 1998. p 3-33.

42. Denison MS, Fisher JM, Whitlock JP, Jr. **The DNA recognition site for the dioxin-Ah receptor complex. Nucleotide sequence and functional analysis**. *J Biol Chem* 1988;263(33):17221-17224.

CHAPTER 5

1. Bhalla DK, Hirata F, Rishi AK, Gairola CG. **Cigarette smoke, inflammation, and lung injury: a mechanistic perspective**. *J Toxicol Environ Health B Crit Rev.* 2009; 12: 45-64.

2. Alberg AJ, Samet JM. **Epidemiology of lung cancer**. *Chest*. 2003; 123(1 Suppl): 21S-49S.
3. World Health Organization. **The World Health Report 2003: Shaping the Future**. World Health Organization Geneva, Switzerland. 2003; pp91-94.
4. Hoffman D, and Hecht SS. **Advances in tobacco carcinogenesis**. In *Handbook of Experimental Pharmacology*. (Cooper CS, and Grover PL, Eds.). 1990; pp 63-192, Springer-Verlag. Heidelberg.
5. International Agency for Research on Cancer. **Tobacco smoke and involuntary smoking**. *In IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2004; Vol. 83, pp 81-83, IARC, Lyon, France.
6. International Agency for Research on Cancer. **Smokeless tobacco and tobacco-specific nitrosamines.** *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2003; Vol. 89, IARC, Lyon, France.
7. Hecht SS, and Hoffman D. **Tobacco-specific nitrosamines, an important group of carcinogens in tobacco and tobacco smoke**. *Carcinogenesis*. 1988. 9, 875-884.
8. International Agency for Research on Cancer. **Tobacco smoke and involuntary smoking**. IARC *Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2004; Vol. 83, pp 59-80, IARC, Lyon, France.
9. Hecht SS. **Biochemistry, biology, and carcinogenicity of tobacco-specific *N*-nitrosamines**. *Chem. Res. Toxicol*. 1998; 11, 559-603.
10. Preussmann R, and Stewart BW. ***N*-Nitroso carcinogens**. In *Chemical Carcinogens*. 2nd ed. (Searle CE, Ed.) ACS Monograph 182, Vol.2, pp 643-828. American Chemical Society, Washington, DC.
11. Jalas JR, Hecht SS, and Myrphy SE. **Cytochrome P450 enzymes as catalysts of metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a tobacco specific carcinogen**. *Chem. Res. Toxicol*. 2005; 18: 95-110.
12. Cooper CS, Grover PL, and Sims P. **The metabolism and activation of benzo[a]pyrene**. *Prog. Drug. Metabo*. 1983; 7: 295-396.
13. Qian J, Jing J, Jin G, Wang H, Wang Y, Liu H, Wang H, Li R, Fan W, An Y, Sun W, Wang Y, Ma H, Miao R, Hu Z, Jin L, Wei Q, Shen H, Huang W, Lu D. **Association between polymorphisms in the GSTA4 gene and risk of lung cancer: a case-control study in a Southeastern Chinese population**. *Mol. Carcinog*. 2009; 48: 253-9.
14. Fang G, Haznadar M, Wang W, Steinbach M, Van Ness B, Kumar V: **A Computationally Efficient and Statistically Powerful Framework for Searching High-order Epistasis with Systematic Pruning and Gene-set Constraints**. Technical Report 013, Department of Computer Science, University of Minnesota, 2010. (submitted)
15. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. **The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status.** *Control. Clin. Trials* 2000; 21(6 Suppl S): 251S-272S.
16. Molecular Signatures Database http://www.broadinstitute.org/gsea/msigdb/index.jsp
17. Ding X, and Kaminsky LS. **Human extrahepatic cytochromes P450: Function in xenobiotic metabolism and tissue-selective chemical toxicity in the respiratory and gastrointestinal tracts**. *Annu. Rev. Pharmacol. Toxicol*. 2003; 43: 149-173.

18. Guengerich FP. **Human cytochrome P450 enzymes**. In *Cytochrome P450: Structure, Mechanism, and Biochemistry* (Ortiz de Montellano PR, Ed.). 1995; pp473-535, Plenum Press, New York.
19. Kushida H, Fujita K, Suzuki A, Yamada M, Endo T, Nohmi T, Kamataki T. **Metabolic activation of N-alkylnitrosamines in genetically engineered Salmonella typhimurium expressing CYP2E1 or CYP2A6 together with human NADPH-cytochrome P450 reductase**. *Carcinogenesis*. 2000; 21: 1227-32.
20. Yamazaki H, Inui Y, Yun CH, Guengerich FP, Shimada T. **Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes**. *Carcinogenesis*. 1992; 13: 1789-94.
21. Crespi CL, Penman BW, Gelboin HV, Gonzalez FJ. **A tobacco smoke-derived nitrosamine, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, is activated by multiple human cytochrome P450s including the polymorphic human cytochrome P4502D6**. *Carcinogenesis*. 1991; 12: 1197-201.
22. Dally H, Edler L, Jäger B, Schmezer P, Spiegelhalder B, Dienemann H, Drings P, Schulz V, Kayser K, Bartsch H, Risch A. **The CYP3A4*1B allele increases risk for small cell lung cancer: effect of gender and smoking dose**. *Pharmacogenetics*. 2003; 13: 607-18.
23. Moyer AM, Sun Z, Batzler AJ, Li L, Schaid DJ, Yang P, Weinshilboum RM. **Glutathione pathway genetic polymorphisms and lung cancer survival after platinum-based chemotherapy**. *Cancer Epidemiol Biomarkers Prev*. 2010; 19: 811-21.
24. Almasi CE, Høyer-Hansen G, Christensen IJ, Danø K, Pappot H. **Prognostic impact of liberated domain I of the urokinase plasminogen activator receptor in squamous cell lung cancer tissue**. *Lung Cancer*. 2005; 48: 349-55.

CHAPTER 6

1. Berg T, Cohen SB, Desharnais J, Sonderegger C, Maslyar DJ, Goldberg J, Boger DL, Vogt PK. **Small-molecule antagonists of Myc/Max dimerization inhibit Myc-induced transformation of chicken embryo fibroblasts**. *Proc Natl Acad Sci U.S.A*. 2002; 99: 3830-5.

2. Van Ness B, Ramos C, Haznadar M, *et.al*. **Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival**. *BMC Med*. 2008; 6:26.

3. Vineis P, Bartsch H, Caporaso N, Harrington AM, Kadlubar FF, Landi MT, Malaveille C, Shields PG, Skipper P, Talaska G, et al. **Genetically based N-acetyltransferase metabolic polymorphism and low-level environmental exposure to carcinogens**. *Nature*. 1994; 369: 154-6.

4. Garte S, Zocchetti C, Taioli E. Garte S, Zocchetti C, Taioli E. **Gene--environment interactions in the application of biomarkers of cancer susceptibility in epidemiology**. *IARC Sci Publ*. 1997; 142: 251-64.

5. Cocks K, Cohen D, Wisløff F, Sezer O, Lee S, Hippe E, Gimsing P, Turesson I, Hajek R, Smith A, Graham L, Phillips A, Stead M, Velikova G, Brown J; EORTC Quality of Life Group. **An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-MY20) in assessing the quality of life of patients with multiple myeloma**. *Eur J Cancer*. 2007; 43: 1670-8.

6. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD Jr. **The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma**. *N Engl J Med*. 2003; 349: 2483-94.

7. Bandiera S, Weidlich S, Harth V, Broede P, Ko Y, Friedberg T. **Proteasomal degradation of human CYP1B1: effect of the Asn453Ser polymorphism on the post-translational regulation of CYP1B1 expression**. *Mol Pharmacol*. 2005; 67: 435-443.

8. Mammen JS, Pittman GS, Li Y et al. **Single amino acid mutations, but not common polymorphisms, decrease the activity of CYP1B1 against (-)benzo[a]pyrene-7R-trans-7,8-dihydrodiol**. *Carcinogenesis.* 2003; 24: 1247-1255.

9. Aklillu E, Ovrebo S, Botnen IV, Otter C, Ingelman-Sundberg M. **Characterization of common CYP1B1 variants with different capacity for benzo[a]pyrene-7,8-dihydrodiol epoxide formation from benzo[a]pyrene**. *Cancer Res*. 2005; 65: 5105-5111.

10. Linden M, Kirchhof N, Kvitrud M, Van Ness B. Linden M, Kirchhof N, Kvitrud M, Van Ness B. **ABL-MYC retroviral infection elicits bone marrow plasma cell tumors in Bcl-X(L) transgenic mice**. *Leuk Res*. 2005; 29: 435-44.

11. Rowley M, Liu P, Van Ness B. **Heterogeneity in therapeutic response of genetically altered myeloma cell lines to interleukin 6, dexamethasone, doxorubicin, and melphalan**. *Blood*. 2000; 96: 3175-80.

12. Liu Z, Li G, Wei S, Niu J, El-Naggar AK, Sturgis EM, Wei Q. **Genetic variants in selected pre-microRNA genes and the risk of squamous cell carcinoma of the head and neck**. *Cancer*. 2010. [Epub ahead of print]

13. Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, Wojcik SE, Ferdin J, Kunej T, Xiao L, Manoukian S, Secreto G, Ravagnani F, Wang X, Radice P, Croce CM, Davuluri RV, Calin GA. **Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility**. *Cancer Res*. 2010; 70: 2789-98.

14. Zhou X, Chen X, Hu L, Han S, Qiang F, Wu Y, Pan L, Shen H, Li Y, Hu Z. **Polymorphisms involved in the miR-218-LAMB3 pathway and susceptibility of cervical cancer, a case-control study in Chinese women**. *Gynecol Oncol*. 2010; 117: 287-90.

15. Hu Z, Chen J, Tian T, Zhou X, Gu H, Xu L, Zeng Y, Miao R, Jin G, Ma H, Chen Y, Shen H. **Genetic variants of miRNA sequences and non-small cell lung cancer survival**. *J Clin Invest*. 2008; 118: 2600-8.

16. Despierre E, Lambrechts D, Neven P, Amant F, Lambrechts S, Vergote I. **The molecular genetic basis of ovarian cancer and its roadmap towards a better treatment**. *Gynecol Oncol*. 2010; 117: 358-65.

17. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N et al. **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet*. 1999; 22: 231–238.

18. Carlson, Rob (2007-01-02). "A Few Thoughts on Rapid Genome Sequencing and The Archon Prize - synthesis". Synthesis.cc. http://synthesis.cc/2007/01/a-few-thoughts-on-rapid-genome-sequencing-and-the-archon-prize.html. Retrieved 2009-02-23.

19. PRIZE Overview: Archon X PRIZE for Genomics http://genomics.xprize.org/archon-x-prize-for-genomics/prize-overview

20. G.A. Calin, *et al*. **Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia**. *Proc. Natl. Acad. Sci. U. S. A. 99*. 2002; 15524–15529.

21. G.A. Calin, *et al*. **Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers**. *Proc. Natl. Acad. Sci. U. S. A. 101*. 2004; 2999–3.

APPENDIX I

GENOMIC VARIATION IN MYELOMA: DESIGN, CONTENT, AND INITIAL APPICATION OF THE BANK ON A CURE SNP PANEL TO DETECT ASSOCIATIONS WITH PROGRESSION-FREE SURVIVAL

Brian Van Ness[*1], Christine Ramos[1], Majda Haznadar[1], Antje Hoering[2], Jeff Haessler[2], John Crowley[2], Susanna Jacobus[3], Martin Oken[4], Vincent Rajkumar[5], Philip Greipp[5], Bart Barlogie[6], Brian Durie[7], Michael Katz[8], Gowtham Atluri[9], Gang Fang[9], Rohit Gupta[9], Michael Steinbach[9], Vipin Kumar[9], Richard Mushlin[10], David Johnson[11] and Gareth Morgan[11]

Address: [1]Cancer Center, University of Minnesota, Minneapolis, MN, USA, [2]Cancer Research and Biostatistics, Seattle, WA, USA, [3]Dana Farber Cancer Institute, Boston, MA, USA, [4]North Memorial Hospital, Minneapolis, MN, USA, [5]Hematology, Mayo Clinic, Rochester, MN, USA, [6]University of Arkansas Medical Sciences Center, Little Rock, AK, USA, [7]Cedar Sinai Medical Center, Los Angeles, CA, USA, 8International Myeloma Foundation, Hollywood, CA, USA, [9]Electrical Engineering & Computer Science, University of Minnesota, Minneapolis, MN, USA, [10]IBM Research, TJ Watson Research Center, Yorktown Heights, NY, USA and [11]Royal Marsden Hospital, London, UK

Reprinted with permission (license # 00334241) from the BioMed Central, BMC Medicine 2008, 6:26 doi:10.1186/1741-7015-6-26

I helped genotype some of the samples represented in the study, and was marginally involved in writing the manuscript by contributing with some data analysis and editing. I provided the analysis of linkage disequilibrium variations, which served as an internal control (all variations in the same LD block are expected to associate together).

Background: We have engaged in an international program designated the Bank On A Cure, which has established DNA banks from multiple cooperative and institutional clinical trials, and a platform for examining the association of genetic variations with disease risk and outcomes in multiple myeloma. We describe the development and content of a novel custom SNP panel that contains 3404 SNPs in 983 genes, representing cellular functions and pathways that may influence disease severity at diagnosis, toxicity, progression or other treatment outcomes. A systematic search of national databases was used to identify non-synonymous coding SNPs and SNPs within transcriptional regulatory regions. To explore SNP associations with PFS we compared SNP profiles of short term (less than 1 year, n = 70) versus long term progression-free survivors (greater than 3 years, n = 73) in two phase III clinical trials.

Results: Quality controls were established, demonstrating an accurate and robust screening panel for genetic variations, and some initial racial comparisons of allelic variation were done. A variety of analytical approaches, including machine learning tools for data mining and recursive partitioning analyses, demonstrated predictive value of the SNP panel in survival. While the entire SNP panel showed genotype predictive association with PFS, some SNP subsets were identified within drug response, cellular signaling and cell cycle genes.

Conclusion: A targeted gene approach was undertaken to develop an SNP panel that can test for associations with clinical outcomes in myeloma. The initial analysis provided some predictive power, demonstrating that genetic variations in the myeloma patient population may influence PFS.

Introduction:

The draft sequence of the human genome published in 2001 [1,2], followed by the more recent improved sequence release of the International Human Genome Consortium [3], have shown that there are large genetic variations in the human genome (polymorphisms). Unlike somatic mutations, polymorphisms are stable and heritable. Polymorphisms include single nucleotide polymorphisms (SNPs), and micro- and minisatellites, and may include heritable insertions and deletions (indels). Significantly, SNPs account for over 90% of genetic variation in the human genome [2]. An important principle that has emerged from the consideration of genetic variation is that disease risk and clinical outcomes can be influenced by individual genetic backgrounds. Thus, while many diseases may have their unique genetic signatures, individual patient outcomes are dependent on heritable variations in a wide variety of genes and pathways affecting cellular functions and drug responses. Moreover, genetic variations in such global functions as inflammation, immunity and cellular signaling in the tumor microenvironment can have an impact on diverse clinical responses.

Multiple myeloma (MM) is a universally fatal disease characterized by the accumulation of malignant plasma cells in the bone marrow [4]. It accounts for 2% of all cancer deaths and 15% of all hematologic malignancies, with about 13,000 deaths per year in the USA [4]. While there are certain common clinical features such as anemia, bone lesions, hypercalcemia, immunodeficiency and renal failure, the disease shows significant heterogeneity with regard to morphology, disease progression, response to therapy and incidence of secondary malignancies. This heterogeneity likely is due, in part, to differences in genetic abnormalities within the malignant clone, as shown in many studies on chromosomal abnormalities [5] and gene expression profiles [6-8].

The growth of MM plasma cells is dependent on a complex interplay among various growth factors, adhesion molecules and other factors in the tumor microenvironment. Thus it might be expected that genetic variations in this interplay could have a profound influence on disease initiation, progression, associated bone complications, and response. Moreover, genetic variation in immunity and inflammation is an important consideration, as are variations in genes coding for drug metabolism and transport. Indeed, death from MM commonly results from infections associated with a severely compromised immune system resulting, in part, from therapeutic toxicities that may be related to variable rates of drug metabolism [9].

In order to address these issues we have engaged in an international program designated as the Bank On A Cure (BOAC). A cooperative program was established to bank DNA from multiple cooperative groups and institutional trials, and to develop a platform for examining the association of genetic variation with disease risk and outcomes. BOAC receives samples through Material Transfer Agreements, and clinical outcomes are provided through agreements with the Cancer Research and Biostatistics Group (Seattle) and the University of Minnesota (with Institutional Review Board, IRB, approval). Currently, the bank has over 2100 samples from the USA, representing six different clinical trials, patient-provided BOAC buccal cell kit samples, and unaffected controls accumulated since 1987.

In this report we describe the development of a novel custom SNP panel based on the Affymetrix/Gene Chip Targeted Genotyping Platform, which contains 3404 SNPs representing variations in a variety of cellular functions and networks, and its initial application to myeloma DNA samples collected in the BOAC bank. We examined population frequencies in affected and unaffected individuals among different ethnic groups, and we developed some novel early approaches in using the SNP panel to determine whether genomic variations in the patient population influence survival.

Materials and Methods:

Control and patient samples

DNA was prepared from 102 Coriell cell lines [10], representing 31 Caucasian, 24 African American, 23 Hispanic, and 24 Asian racial groups (unaffected by myeloma). DNA samples were also prepared from 143 myeloma patients enrolled in phase III clinical trials: E9486, n = 52 [11] and S9321, n = 91 [12], with informed consent; and 34 unaffected, spousal controls (all Caucasian). E9486 patients ranging in age from 55 to 70 years were treated with Vincristine, Busulfan, Melphalan, Cyclophossphamide, Prednisone (VBMCP) followed by randomization143 to no further treatment, IFN-a, or cylcophosphamide; and, although there was variation in survival among all patients, no significant differences in survival were noted among the three arms of the trial [11]. Patients included in this study from S9321 were in the same age range, and received Vincristine, Adriamycin, Dexamethasone (VAD) induction followed by VBMCP [12]. S9321 patients in the trial arm randomized to high dose melphalan+TBI followed by transplant were not included. Patients for this analysis were selected based on progression-free survival (PFS) of less than 1 year (n = 70) or greater than 3 years (n = 73).

Custom BOAC SNP chip design and content

A directed, custom BOAC SNP chip design was developed with specific criteria from public and commercial databases. Rather than a total genome wide screen, a plan was undertaken to develop a custom SNP chip, focusing on functionally relevant polymorphisms playing a role in normal and abnormal cellular functions, inflammation and immunity, as well as drug responses. Candidate gene lists were created and each gene in the candidate list was systematically investigated with a selection of SNP databases to harvest SNPs that may have a functional effect on gene action. Figure 1 outlines the approach. Searches for genes were developed, using public and commercial software programs in PubMed, iHOP [13], as well as pathway databases, such as

PharmGKB Pathways [14], BioCarta [15], KEGG [16], Ingenuity, and Pathway Assist (Stratagene, Inc.). The Human Gene Mutation Database [17] contains a searchable database of polymorphisms associated with diseases cited in the literature. This database was used in conjunction with SNP500, SNPper, and MutDB to obtain the SNP id (rs number) of polymorphisms in the gene lists. A systematic search for all non-synonymous SNPs (ie, resulting in coding change) with a validated, minor allele frequency greater than 2% in all of the candidate genes was completed using SNP 500, dbSNP, and Affymetrix databases. SNPs failing to meet a 2% population frequency were included if the frequency was higher than 5% in selected racial subgroups (eg, Asian, African American, Caucasian). A systematic search of the promoter regions in all the candidate genes for SNPs present in homologous regions between human and mouse with a minor allele frequency greater than 2% were identified using the PromoLign Database [18]. Many of the SNPs selected in this method were seen to lie in or adjacent to transcription binding sites. Some additional promoter SNPs that may affect transcription binding sites were also identified using the FESD [19] database. Affymetrix provided several in-house validated SNP lists, including: inflammation and immunity, drug metabolism, and cancer lists. Three groups of admixture SNPs, which differ in frequency between Asian, African and European groups, were added to allow corrections in data analyses for racial specific variations [20]. TagSNPs in genes influencing drug metabolism and transport were added from the supplementary data from Ahmadi et al. [21]. The full SNP panel includes 3404 SNPs in 983 genes. Genotyping was performed using the Affymetrix® Gene- Chip® Scanner 3000 Targeted Genotyping System (GCS 3000 TG System), which utilizes molecular inversion probes to simultaneously identify the 3404 pre-selected SNPs. The protocol has previously been described [22]. All genotyping experiments were performed in strict adherence to the manufacturer's protocol.

Statistical methods

Patients from the Eastern Cooperative Oncology Group (ECOG) and Southwest Oncology Group (SWOG) trials were selected using the following criteria: they were all Caucasian and between 55 and 70 years of age at diagnosis; patients with IgA subtype were excluded (as this is an independent, poor prognosis variable). Patients with the longest progression-free survival (PFS > 3 years) and patients with the shortest progression-free survival (PFS < 1 year) were selected. Two approaches were used to determine whether there was true discrimination of SNP genotypes in the PFS analysis, when analyzed as a conglomerate data set.

1) Leave-one-out cross-validation [23]. In this approach, the original data set of 143 patients was divided into two groups: one consisting of a single patient and one consisting of the remaining 142 patients. A classification model was built using the 142 patients as a training set and then this classification model was used to classify the single 'left out' patient. For this study, as well as the class label study below, we used a support vector machine (SVM) classifier, as implemented by the Weka package [24], and specified a liner kernel.

2) Randomization of class labels [25]. For the original data and labels, we followed the standard practice for building and evaluating a classifier [25], that is, compare the performance of a classifier using the original and randomly shuffled class labels (permutations). There were 143 subjects, consisting of 73 cases and 70 controls. The training set was created by randomly selecting 50 cases and 50 controls and using the remaining subjects as a test set. One hundred runs were performed for the original data and class labels. We also analyzed each clinical trial data set separately and used the other clinical trial data set as a validation set.

Fisher's exact test was used as a univariate screening tool to rank the SNPs by how strongly they are associated with PFS. The top 50 SNPs of each trial with the smallest pvalue were selected and used in a recursive partitioning analysis. For this recursive partitioning analysis we used RPART from the R software package, a language and environment for statistical computing. The tree-based library RPART was developed as described [26]. The regression tree resulting from the analysis was subsequently pruned in order to avoid over-fitting. This regression tree was used on both the trial it was developed on as well as the other trial for validation purposes. Specificity and sensitivity were determined for each data set. Finally, we attempted univariate ranking and recursive partitioning of the conglomerate data set (both trials combined) using random forests [27]. Validation was examined by randomly mixing survival data sets and determining and comparing the predictive accuracy of true survival subsets and random subsets.

Results:

SNP chip panel design

A final custom SNP chip panel of 3404 SNPs from 983 genes meeting the above criteria was produced for the Affymetrix/Gene Chip Targeted Genotyping Platform. A full documented list of SNPs is found in [28] and includes rs assignments, gene identifiers (Entrez), functional grouping, and SNP effect (coding, non-coding, regulatory, haptag). Table 1 summarizes a variety of functional categories represented on the chip. Figure 1S (found in [28]) shows the chromosomal distribution of the SNPs included. Although the SNP chip was not designed to serve as a genome linkage panel, the chromosomal distribution is quite broad (see additional File 1), and may provide functional targets for higher density linkage chips or regions identified by other approaches such as comparative genomic hybridizations or genome wide screens.

We also examined representation among a variety of define metabolic and signaling pathways. Because of the filter criteria, it was found that most pathways did not have a high degree of representation, suggesting SNPs for many of the genes not included may not have coding or regulatory impact. This becomes an important consideration in attempts to associate outcomes with specific pathways. Instead, common functional groupings turned out to be a better analytical target (see below). It is noteworthy to compare the content of the BOAC SNP chip to the SNPs represented on the Affymetrix 500K Array genome wide scan. The 500K Array panel is primarily derived from two restriction enzyme cleavage fragmentations, with SNP representation for each fragment, providing a comprehensive, global SNP panel. Well over 95% of the panel is intragenic, non-coding; and thus, its primarily use is to identify copy number, chromosomal regions, and linkage. Indeed, of the 3404 SNPs on the BOAC SNP chip, only 401 are present on the 500K Array panel. Thus, while the BOAC SNP chip does not have gene wide coverage, it does have a higher density of coding and regulatory content.

Samples and quality control assessments

For this study, a total of 279 DNA samples were profiled by the BOAC SNP chip. One hundred and thirty-six unaffected controls from the Coriell panel and spouses of myeloma patients were profiled. The Coriell panel included 31 Caucasian, 24 Asian, 23 Hispanic, and 24 African American samples of unaffected individuals. One hundred and forty-three myeloma samples were profiled, from the phase III clinical trials, ECOG E9486 (n = 52) and the chemotherapy arm of the ECOG-SWOG intergroup trial S9321 (n = 91). Treatment protocols are given in the Methods section. This study was in compliance with the Helsinki Declaration, and approved by the IRB at the University of Minnesota (approval # 0311M53428), with patient consents collected by the clinical cooperative groups' trial offices. Among all samples profiled, we had an average SNP call rate of 96%. The profiles of the Coriell panel allowed us to determine allelic frequencies in racial groups and unaffected populations. Of the 3404 SNPs on the BOAC panel, 786 were contained in the SNP500 cancer database, allowing us to determine concordance between the two Coriell data sets.

We found very good agreement between our data set and the national database, with an average of > 97% concordance. We also duplicated the profile of a number of samples (n = 10), and found better than 99.7% reproducibility between duplicate samples. This concordance and duplication rate was also equivalent when comparing the BOAC SNP panel run in the USA and UK facilities, providing a cross validation between BOAC laboratory sites. Finally, for every batch run of 24 samples, the Affymetrix platform includes a control DNA sample, and this provided continuous monitoring and quality assurance across the study.

Allelic variations by race

It has been well established that there are significant allelic frequency differences by race, or ethnic and regional origins [29]. Part of the SNP panel design included the admixture SNP panel that shows significant racial variation. Figure 2A shows a diagonal plot in which each SNP minor allelic frequency is plotted by frequency in the Caucasian (n = 92) versus the African American (n = 27) myeloma populations. Equivalent frequencies would be expected to cluster on the 45 degree angle; an it is readily apparent that frequencies of many of the SNPs vary widely between races. Indeed, the racial disparities in allelic frequencies were far more significant than could be assessed in case-control or outcome studies, so that subsequent initial survival analyses were done only on a single racial group. Moreover, given the inclusion criteria of SNPs (included if greater than 5% in one racial group), it is noteworthy that 401 SNPs show allelic variation only in the African Americans (ie, no variations seen in Caucasian). In contrast, in a comparison of unaffected samples with affected samples, restricted to Caucasians, there is high concordance across the total panel of SNPs (Figure 2B). This provides an opportunity to examine smaller clusters or functional associations with disease that may be masked by the larger multi-racial pools. However, the object of this study was not to compare variations within different ethnic patient populations.

Allelic variations associated with progression-free survival

Although genetic deregulation within the tumor population has been shown to stratify clinical outcomes [5-8], a significant impact on therapeutic outcomes may result from genetic variations in germline DNA affecting a number of important functions, including drug metabolism, transport, DNA repair, immune response, growth factors, angiogenesis, etc. To explore the SNP associations on the BOAC SNP chip we chose to examine an extreme phenotype comparison in two phase III clinical trials with similar chemotherapeutic treatments. E9486 patients ranging in age from 55 to 70 years were treated with VBMCP followed by randomization to no further treatment, IFN-a, or cylcophosphamide; and, although there was variation in survival among all patients, no significant differences in survival were noted among the three arms of the trial [11]. Patients included in this study from S9321 were in the same age range, and received VAD induction followed by VBMCP [12]. S9321 patients in the trial arm receiving high dose melphalan+TBI, going on to transplant were not included. The goal was to identify SNPs that may distinguish short term (less than 1 year) versus long term progression-free survivors (greater than 3 years).

While our banking represents one of the largest collections of myeloma specimens, one of the difficulties encountered in this data analysis still results from relatively small sample sizes of patients with similar treatment protocols. With the data sets we had, we used a variety of approaches to determine whether there was true discrimination of SNP associations between the two PFS groups. The 'leave-one-out' approach is a standard approach in classification [23], but is not typically used, due to computational cost, when data sets are large. However, in this case, only 143 classification runs were necessary.

The results of those runs performed using an SVM classifier from Weka [24] are summarized in Table 2. SVM is a supervised method used for classification and regression. It belongs to a family of generalized linear classifiers. A special property of SVM is that it simultaneously minimizes the empirical classification error and maximizes the predictive separation.

The classification accuracy of the leave-one-out approach is (50+45)/143 = 0.66. If there was no true discriminating signal in the data, then the classifiers built by the leaveone-out procedure should produce a table with a relatively evenly distributed number of entries among the four cells, since the classes are of roughly the same size and the predictions should be random. However, the observed table is far from that random distribution. By using Fisher's exact test it is possible to compute the probability (p-value) for obtaining a table with the same or better accuracy of prediction by random chance. Specifically, the p-value is 7.7 × 10-5, which strongly indicates that the result is not due to random chance. The calculated odds ratio (OR) for survival is 3.9 CI (2.0, 7.8). We subsequently focused on SNP subsets that might provide more directed functional associations and found that the best predictor of survival was achieved when just the non-synonymous SNPs and the promolign SNP subset in introns was used. The accuracy of prediction increased to 75.5% OR = 9.6 CI (4.5, 20.5).

To determine whether genotypes in the SNP panel had true discriminatory power we randomly permuted the outcome across the two groups and calculated the classification accuracy. A total of 10,000 random group comparisons were performed (ie, survival groups were randomly mixed) with the distribution of accuracy shown in Figure 3. As expected, the most common accuracy was close to 50%, with a random distribution around the mean. Notably, no random grouping achieved the accuracy of the original survival classification of 66%, nor the 75% sub set, indicating that, as a group, the SNPs are providing a measure of true discrimination of survival.

Another approach that is commonly used for classification is the generation of random subsets for training and validation. A classification model is built on the training subset, and is evaluated on a separate test set. The process is repeated and yields a distribution of accuracy on the data set labels (eg. short versus long PFS). This is then repeated with random shuffling of the survival data sets to determine whether there is true signal accuracy. For this analysis, we again used a support vector machine classifier, as implemented by the Weka package [25], using a preset linear kernel option). The training set consisted of repeated samples of 50 short term and 50 long term survivors, with the remaining patient samples used for test validation. One hundred runs were performed for the short and long term classifiers, and the average accuracy on test set analysis was 61.4% +/-7.1%. One hundred runs of random mixed set comparisons generated an average accuracy of 47.5% +/-7.3%. This further suggests that there are true differences in the genotypes that impact survival classification. A t-test was performed to evaluate the difference between the classification results based on the original and randomized class labels, resulting in a pvalue for survival classification of less than 0.0001. This further indicates that, as a group, the SNPs are providing a measure of true discrimination of survival.

Each of the above approaches demonstrated that the SNP panel provided discrimination; however, we attempted to explore possible subsets of SNPs that may drive the association. We used Fisher's exact test as a univariate screening tool to determine the association of SNPs with each trial separately (Tables 3 and 4), then using the top 50 rank ordered SNPs, performed recursive partitioning to identify the combination of SNPs that best distinguish PFS groups. In recursive partitioning each genotype is evaluated on its ability to make a correct prediction, creating a decision node. A pruned decision tree is created in which the minimum number of the strongest nodes creates a group prediction. From the results of each trial, we then validated on the other trial. In the univariate ranking, we did not correct for multiple comparisons; that would certainly reduce the p-value significance, but would not alter the rank order comparison. This approach does examine the association of each individual SNP; it is more likely that complex interactions may drive association of groups of SNPs not revealed by univariate ranking. Nevertheless, among the top ranked individual SNP variations in both trials were those associated with drug metabolism/detoxification/transport, including: cyp genes, multiple variants of GSTA4, SLCO, UGT1, NAT2, ABCB genes; as well as genes impacting cellular response, including: BMP2 (inducing myeloma apoptosis) [30], cathepsin B (inducing IL-8 dependent cellular migration and angiogenesis [31,32], XRCC5 (DNA repair); and genes associated with proliferative responses (PCNA, MAPK, cyclin kinase). The association of multiple alleles of GSTA4 is particularly compelling, suggesting consistency in its impact across several variant alleles. In addition, several alleles are in linkage disequilibrium, appearing as a cluster in the list – providing quality controls (as linked genes would be expected to show the same association).

The first survival separation was analyzed for clinical trial S9321, and the top 20 rank ordered SNPs are presented inTable 3 (more extended rank order presented in Table 2S of [28]). Figure 4 shows the pruned recursive partitioning tree, resulting in two SNPs with the highest classification prediction of survival groups. One SNP is in catechol methyl transferase (COMT) and one is in Ghrelin precursor (GHRL). The potential significance of these SNPs on outcomes is discussed below. The correct classification rate (survival prediction) was 71%, which dropped to 58% on validation testing with the E9486 trial. The specificity and sensitivities are also presented. The converse analysis was done (E9497 training set; S9321 validation); and the rank order of SNPs was determined (Table 4 and an expanded Table 3S in [28]). A recursive partitioning tree of two SNPs showed 79% classification on the training E9487 set, and 56% on the S9321 validation (Figure 4). The SNPs identified in this trial were farnesyl transferase (FDFT) and ABCC1 (in the family of ATP transporters). The potential significance is also provided in the Discussion.

As an exploratory approach, we combined the data sets from both trials, then used Fisher's exact test as a univariate screening tool to determine the association of each SNP with survival. When we treated the top 165 SNPs (univariate $p < 0.02$) as a set for short versus long PFS classification prediction using a random forest multiple sampling approach [27], we found a 79% correct classification rate. However, similar classification accuracy could be achieved with random class labels, demonstrating the potential of false positive associations in such complex data sets.

Discussion:

We have designed a novel SNP panel, containing 3404 genetic variations associated with 983 genes involved in a variety of cellular functions that could impact population variations in tumor progression and response (Table 1 and [28]). This approach is distinct from using genomewide SNP arrays of 500,000 SNPs. The Affymetrix 500K SNP Panel is based on restriction enzyme cleavage sites and representative spacing on the chromosomes. While having significantly greater content, over 90% of the SNPs on the whole genome array are intragenic; and the chip is most often analyzed for linkage associations. The multiple comparison false positive error rate is large, and the technology considerably more expensive. Indeed, of the 3404 gene-associated SNPs on the BOAC SNP panel, only 401 are contained on the 500K SNP panel.

There are limitations to the BOAC SNP panel as well. The public and Affymetrix databases used to construct the chip content are constantly updating, so that missing elements may be noted. While we targeted SNPs in non-synonymous coding sequences or highly conserved regulatory sequences, many of the SNPs have not yet been functionally documented for effects. As such, SNP associations in the BOAC panel represent a first step in exploring the genome for clinically relevant genetic variations that will require both extensive validations as well as functional assays to confirm their effect.

We made a considerable effort to ensure that quality controls were in place. The Affymetrix platform provided a high call rate (96%) as well as very high concordance in replicate samples, even those run at different facilities. The concordance extended to 786 SNPs on the panel that were documented for the Coriell cell lines we have included [10]. All of the samples we analyzed had high quality DNA (A260/280 ratios > 1.7, and little DNA degradation).

In subsequent unrelated studies, we found that even highly degraded DNA provides robust, high call rates and reproducibility (not shown); probably because the initial amplifications are across 100–150 bp of DNA. The most likely source for quality control error may come from sample misidentification or placement in multi-well plates. To control for this, we routinely incorporate randomly positioned controls and replicates. Within the Coriell cell line panel is a distribution of racial groups. It is striking how much allelic frequencies differ in the African American vs. Caucasian racial groups. It is likely there is more refinement of allelic variations associated with more geographical based lineages [33], as racial definitions are somewhat subjective and often self reported. Importantly, as the BOAC database increases, multiple comparisons can be done with appropriate corrections for allelic variations among races. It will be important to include the full spectrum of patients as the database expands.

Disease progression, response and survival vary widely among patients. There are a number of studies that have examined variations in tumor cell chromosomal abnormalities [5] and gene expression profiles [6-8]. The evidence strongly suggests that patient outcomes are impacted by these tumor cell variations. However, patient populations show considerable germline variation that could influence the microenvironment, immune status, and drug metabolism or transport. For example, the authors (DJ, GM) have presented evidence that germline variations in GSTP1 show alterations in melphalan metabolism, and have been associated with different outcomes in patients receiving high dose melphalan therapies [34]. Numerous examples of variations in drug metabolism, transport, and DNA repair have been documented, with emerging associations on therapeutic outcomes.

Our approach was to provide a more global germline analysis that was driven by bioinformatic searches for potentially relevant variations in multiple genes and gene functions. This is still an exploratory approach to identify potential variations of functions that impact upon therapeutic responses and disease progression that may result in differences in survival outcomes. Rather than a linear progression of survival, we chose to examine two extreme ends of the PFS spectrum, to maximize the first steps in identifying potential functional variations. Patients were stratified by short (< 1 year) versus longer (> 3 years) PFS groups. Nevertheless, it is likely that survival is a complex endpoint resulting from both tumor progression and therapeutic failure that may impact upon multiple organ systems. Moreover, we recognized that a) tumor variation among patients may have dominant effects that are associated with survival; b) the trials we examined used multidrug regimens, and each drug response may be impacted upon by complex genetic variations in transport, metabolism, and export; and c) sample number is still limiting statistical power. Thus, our initial approaches in this study were to determine whether germline variations had any measurable influence on survival.

We felt it was important to determine, first, if there were any true discrimination of the SNP panel in the two PFS groups, when the complete SNP profile was considered. Using a variety of methods that were tested against randomly mixed sample analysis, we found the SNP panel had true signal to discriminate the short and long progression-free survivors, although the accuracy did not reach the level of prediction that would allow clinical application. Notably, a smaller subset increased the predictive power. Significantly, no individual genetic variation provided a strong, independent prediction of survival. This likely reflects the fact that individual germline variations may impact upon response, but are not solely responsible; and it is likely that such variations are the result of complex interactions. Indeed, genetic variations in the tumor cell may play a dominant role in response and survival. Thus, patient responses are likely to involve interactions affecting multiple functions within the tumor cell as well as external factors affecting tumor progression and drug response. Nevertheless, our analysis of the SNP panel as a group suggests it is likely that germline variations impact upon patient survival and deserve further attention.

Recognizing the limited statistical power to detect single SNPs associated with PFS, we did perform a univariate analysis to rank order the SNPs that individually best discriminated the groups in the two similar phase III clinical trials. We did not correct for multiple comparisons, which would certainly reduce the p-value significance but would not alter the rank order comparison. This approach also assumes association for the individual SNP. It is more likely that complex multi-SNP groupings influence response. Nevertheless, among the top SNP variations in both trials were those associated with drug metabolism/ detoxification/transport, including: cyp genes, multiple variants of GSTA4, SLCO, UGT1, NAT2, ABCB genes; as well as genes impacting cellular response, including: BMP2 (inducing myeloma apoptosis), cathepsin B (inducing IL-8 dependent cellular migration and angiogenesis [31,32], XRCC5 (DNA repair); and genes associated with proliferative responses (PCNA, MAPK, cyclin kinase). The association of multiple alleles of GSTA4 is particularly compelling, suggesting consistency in its impact across several variant alleles. In addition, several alleles are in linkage disequilibrium, appearing as a cluster in the list – providing quality controls (as linked genes would be expected to show the same association). Surprisingly absent from the SNP association lists are cytokines, growth factors and receptors that might be expected to cause variations in disease progression and resistance, with the exception of IL-10, which has been reported in previous studies [35].

While still an exploratory analysis, the paired SNPs identified by recursive partitioning in each trial have some intriguing possible connections to PFS. COMT (catechol-O-methyltransferase) metabolizes catechol drugs, and has been linked to breast cancer risk and survival [36]; GHRL has been shown to stimulate angiogenesis [37] and regulate bone formation through osteoblasts [38,39]; FDFT is the farnesyl transferase that may regulate important signaling (eg, ras) [40,41]; and ABCC is among a class of transporters that may influence multi-drug resistance [42]. It is noted that strong association in one trial was significantly reduced in the validation trial. Nevertheless, the functional impact of these genetic variations may warrant further investigation.

Conclusion

The exploratory analyses provide some of the first attempts to use larger, targeted SNP panels to develop models of genomic variations that may influence treatment outcomes, and that may deserve further analysis of functional significance. Not surprisingly, among the most significant variations correlating with survival were genes that could be functionally categorized as pharmacologic. However, the group analysis suggests various functions may interplay in disease progression and response. It is important to consider the fact that we could not identify a small driver set of SNPs that strongly associated with survival, particularly with the limited sample size. However, we note that, as a group, germline genomic variations do have impact on event-free survival. As the Bank On A Cure data set is expanding, SNP associations are being analyzed for more specific phenotypes in response, disease complications (eg, bone disease), and adverse or toxic drug effects (eg, thrombolytic events associated with thalidomide). Heterogeneity in tumor gene deregulation certainly contributes to variation in disease outcome. It would seem appropriate to consider combining an understanding of tumor heterogeneity (chromosomal and expression profiles) with germline variations (eg, SNP variations associated with pharmacologic functions or disease complications) that can lead to development of more individualized therapies that take into account both tumor and population variations.

Abbreviations

BOAC: Bank On A Cure; ECOG: Eastern Cooperative Oncology Group; IRB: Institutional Review Board; MM: multiple myeloma; OR: odds ratio; PFS: progression-free survival; SNP: single nucleotide polymorphism; SVM: support vector machine; SWOG: Southwest Oncology Group; VAD: Vincristine, Adriamycin, Dexamethasone; VBMCP: Vincristine, Busulfan, Melphalan, Cyclophossphamide, Prednisone.

Competing interests

The authors declare that they have no competing interests. Authors' contributions BVN is the principal investigator of study, involved in study design and writing the manuscript. CR and MH developed genotyping assays and generated genotype data. AH, JH, and JC provided input on study design and statistical analysis. SJ provided clinical outcome data from the office of ECOG Statistical Center. MO was the clinical chair of the ECOG phase III trial. VR and PG have shared chairmanship of the ECOG Myeloma Committee and participated in clinical trial study design and clinical evaluations. BB chairs the SWOG myeloma clinical design and evaluations. BD is the Clinical Director of the Bank On A Cure project described. MK supervizes integration of laboratory efforts between the USA and the UK. VR, MS, GA, GF, RG, and RM developed statistical approaches in vector machine applications. DJ participated in SNP panel design and genotyping assembly. GM co-directs Bank On A Cure with BVN, and has been involved in study design and data evaluation.

References:

1. Lander ES, from the International Human Genome Sequencing Consortium, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

2. Venter JC, et al.: The sequence of the human genome. Science 2001, 291:1304-1351.

3. International Human Genome Consortium: Finishing the euchromatic sequence of the human genome. Nature 2004, 431:931-945.

4. Kyle RA, Rajkumar SV: Plasma cell disorders. In Cecil textbook of medicine 22nd edition. Edited by: Goldman L, Ausiello DA. Philadelphia: W.B. Saunders; 2004:1184-1195.

5. Bergsagel PL, Kuehl WM: Molecular pathogenesis and consequent classification of multiple myeloma. J Clin Oncol 2005, 23(26):6333-6338.

6. Shaughnessy JD, Zhan F, Burington BE, Huang Y, Colla S, Hanamura I, Stewart JP, Kordsmeier B, Randolph C, Williams DR, Xiao Y, Xu H, Epstein J, Anaissie E, Krishna SG, Cottler-Fox M, Hollmig K, Mohiuddin A, Pineda-Roman M, Tricot G, van Rhee F, Sawyer J, Alsayed Y, Walker R, Zangari M, Crowley J, Barlogie B: A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood 2007, 109(6):2276-2284.

7. Jenner MW, Leone PE, Walker BA, Ross FM, Johnson DC, Gonzalez D, Chiecchio L, Dachs Cabanas E, Dagrada GP, Nightingale M, Protheroe RK, Stockley D, Else M, Dickens NJ, Cross NC, Davies FE, Morgan GJ: Gene mapping and expression analysis of 16q loss of heterozygosity identifies WWOX and CYLD as being important in determining clinical outcome in multiple myeloma. Blood 2007, 110(9):3291-3300.

8. Chng WJ, Kumar S, Vanwier S, Ahmann G, Price-Troska T, Henderson K, Chung TH, Kim S, Mulligan G, Bryant B, Carpten J, Gertz M, Rajkumar SV, Lacy M, Dispenzieri A, Kyle R, Greipp P, Bergsagel PL, Fonseca R: Molecular dissection of hyperdiploid multiple myeloma by gene expression profiling. Cancer Res 2007, 67(7):2982-2989.

9. Kay NE, Leong TL, Bone N, Vescole DH, Greipp PR, Van Ness B, Oken MM, Kyle RA: Blood levels of immune cells predict survival in myeloma patients: results of an Eastern Cooperative Oncology Group phase 3 trial for newly diagnosed multiple myeloma patients. Blood 2001, 98(1):23-28.

10. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, Eckert A, Puri V, Gerhard DS, Chanock SJ: SNP500 Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res 2006:D617-621.

11. Oken MM, Leong T, Lenhard RE, Greipp PR, Kay NE, Van Ness B, Keimowitz RM, Kyle RA: The addition of interferon or high dose cyclophosphamide to standard

chemotherapy in the treatment of patients with multiple myeloma: Phase III Eastern Cooperative Oncology Group Clinical Trial EST 9486. Cancer 1999, 86(6):957-968.

12. Barlogie B, Kyle RA, Anderson KC, Greipp PR, Lazarus HM, Hurd DD, McCoy J, Moore DF Jr, Sakhil SR, Lanier KS, Chapman RA, Cromer JN, Salmon SE, Durie B, Crowley JC: Standard chemotherapy compared with high-dose chemoradiotherapy for multiple myeloma: final results of phase III US Intergroup Trial S9321. J Clin Oncol 2006, 24(6):929-936.

13. Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A: Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE 2005:pe21.

14. Pharmgkb [http://www.pharmgkb.org/search/pathway/path way.jsp]

15. BioCarta [http://www.biocarta.com]

16. KEG [http://cgap.nci.nih.gov/Pathways/Kegg_Standard_Pathways]

17. Stenson PD, Ball EV, Mort M, Phillips AD, Sheil JA, Thomas NS, Abeysinghe S, Krawszak M, Cooper DN: The Human Gene Mutation Database (HGMD®): 2003 Update. Hum Mutat 2003, 21:577-581.

18. Zhao T, Chang LW, McLeod HL, Stormo GD: PromoLign: A Database for Upstream Region Analysis and SNPs. Human Mutation 2004, 23:534-539.

19. Kang HJ, Choi KO, Kim BD, Kim S, Kim YJ: FESD: a Functional Element SNPs Database in human. Nucleic Acids Res 2005, 33(1):D518-522.

20. Miller RD, Phillips MS, Jo I, et al.: High-density single-nucleotide polymorphism maps of the human genome. Genomics 2005, 86(2):117-126.

21. Ahmadi KR, Weale ME, Xue ZY, Soranzo N, Yarnall DP, Briley JD, Maruyama Y, Kobayashi M, Wood NW, Spurr NK, Burns DK, Roses AD, Saunders AM, Goldstein DB: A single-nucleotide polymorphism tagging set for human drug metabolism and transport. Nat Genet 2005, 37(1):84-89.

22. Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, Davis RW: Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol 2003, 21(6):673-678.

23. Kearns M, Ron D: Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In Proceedings of the Tenth Annual Conference on Computational Learning theory (Nashville, Tennessee, United States, July 06 – 09, 1997). COLT '97 ACM, New York, NY; 1997:152-162.

24. Dinu Valentin, Zhao Hongyu, Miller Perry L: Integrating domain knowledge with statistical and data mining methods for highdensity genomic SNP disease association analysis. Biomedical Informatics 2007, 40:750-760.

25. Witten Ian H, Frank Eibe: "Data Mining: Practical machine learning tools and techniques". 2nd edition. Morgan Kaufmann, San Francisco; 2005.

26. Therneau Terry M, Atkinson Elizabeth J: An introduction to recursive partitioning using the rpart routines. Technical report 61, Mayo Clinic 1997. http://mayoresearch.mayo.edu/mayo/research/biostat/techreports.cfm, R package available at http://cran.r-project.org/src/contrib/Descriptions/rpart.html

27. 27. Agreti A: An introduction to categorical data analysis. Wiley, NJ; 1996.

28. Breiman L: Random forests. Machine Learning 2001, 45:5-32.

29. Efron B, Tibshirani R: An introduction to the Bootstrap. Chapman & Hall/CRC, FL; 1994.

30. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958, 53:457-481.

31. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, Jones KW: Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2005, 2(2):81-89.

32. Kawamura C, Kizaki M, Ikeda Y: Bone morphogenetic protein (BMP)-2 induces apoptosis in human myeloma cells. Leuk Lymphoma 2002, 43(3):635-9.

33. Schraufstatter IU, Trieu K, Zhao M, Rose DM, Terkeltaub RA, Burger M: IL-8-mediated cell migration in endothelial cells depends on cathepsin B activity and transactivation of the epidermal growth factor receptor. J Immunol 2003, 171:6714-6722.

34. Yanamandra N, Gumidyala KV, Waldron KG, Gujrati M, Olivero WC, Dinh DH, Rao JS, Mohanam S: Blockade of cathepsin B expression in human glioblastoma cells is associated with suppression of angiogenesis. Oncogene 2004, 23:2224-2230.

35. Bamshad M: Genetic influence on health: does race matter? JAMA 2005, 294:937-946.

36. Dasgupta RK, Adamson PJ, Davies FE, Rollinson S, Roddam PL, Ashcroft AJ, Dring AM, Fenton JA, Child JA, Allan JM, Morgan GJ: Polymorphic variation in GSTP1 modulates outcome following therapy for multiple myeloma. Blood 2003, 102(7):2345-2350.

37. Zheng C, Huang D, Liu L, Wu R, Bergenbrant Glas S, Osterborg A, Bjorkholm M, Holm G, Yi Q, Sundblad A: Interleukin-10 gene promoter polymorphisms in multiple myeloma. Int J Cancer 2001, 95(3):184-188.

38. Long JR, Cai Q, Shu XO, Cai H, Gao YT, Zheng W: Genetic polymorphisms in estrogen-metabolizing genes and breast cancer survival. Pharmacogenet Genomics 2007, 17(5):331-338.

39. Li A, Cheng G, Zh GH, Tarnawaski AS: Ghrelin stimulates angiogenesis in human microvascular endothelial cells: Implications beyond GH release. Biochem Biophys Res Commun 2007, 353(2):238-2343.

40. Kim SW, Her SJ, Park SJ, Kim D: Ghrelin stimulates proliferation and differentiation and inhibits apoptosis in osteoblastic MC3T3-E1 cells. Bone 2005, 37(3):359-69.

41. Fukushima N, Hanada R, Teranishi H, Fuku Y: Ghrelin directly regulates bone formation. J Bone Miner Res 2005, 20(5):790-798.

42. Alvarado Y, Giles FJ: Ras as a therapeutic target in hematologic malignancies. Expert Opin Emerg Drugs 2007, 2:271-284.

43. Hu L, Shi Y, Hsu JH, Gera J, Van Ness B, Lichtenstein A: Downstream effects of oncogenic ras in multiple myeloma cells. Blood 2003, 101(8):3126-3135.

44. Pol MA van der, Broxterman HJ, Pater JM, Feller N, Mass M van der, Weijers GW, Scheffer GL, Allen JD, Scheper RJ, van Loevezijn A, Ossenkoppele GJ, Schuurhuis GJ: Function of the ABC transporters, P-glycoprotein, multidrug resistance protein and breast cancer resistance protein, in minimal residual disease in acute myeloid leukemia. Haematologica 2003, 88(2):134-147.

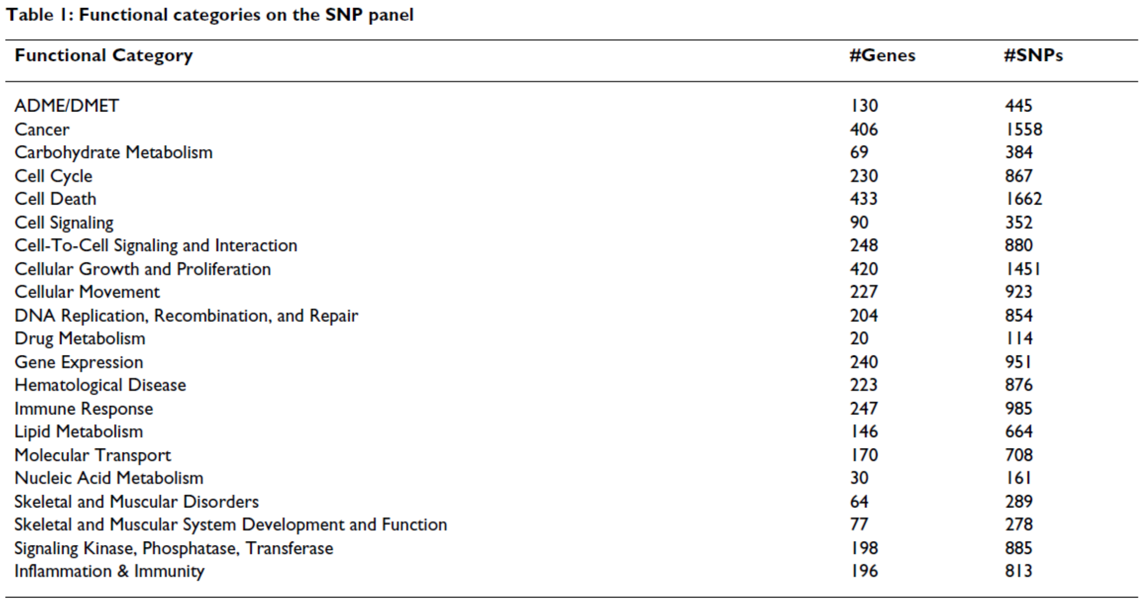


Table 1: Functional categories on the SNP panel

Functional Category	#Genes	#SNPs
ADME/DMET	130	445
Cancer	406	1558
Carbohydrate Metabolism	69	384
Cell Cycle	230	867
Cell Death	433	1662
Cell Signaling	90	352
Cell-To-Cell Signaling and Interaction	248	880
Cellular Growth and Proliferation	420	1451
Cellular Movement	227	923
DNA Replication, Recombination, and Repair	204	854
Drug Metabolism	20	114
Gene Expression	240	951
Hematological Disease	223	876
Immune Response	247	985
Lipid Metabolism	146	664
Molecular Transport	170	708
Nucleic Acid Metabolism	30	161
Skeletal and Muscular Disorders	64	289
Skeletal and Muscular System Development and Function	77	278
Signaling Kinase, Phosphatase, Transferase	198	885
Inflammation & Immunity	196	813

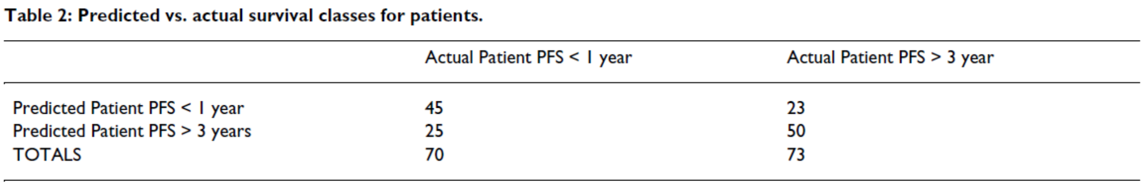


Table 2: Predicted vs. actual survival classes for patients.

	Actual Patient PFS < 1 year	Actual Patient PFS > 3 year
Predicted Patient PFS < 1 year	45	23
Predicted Patient PFS > 3 years	25	50
TOTALS	70	73

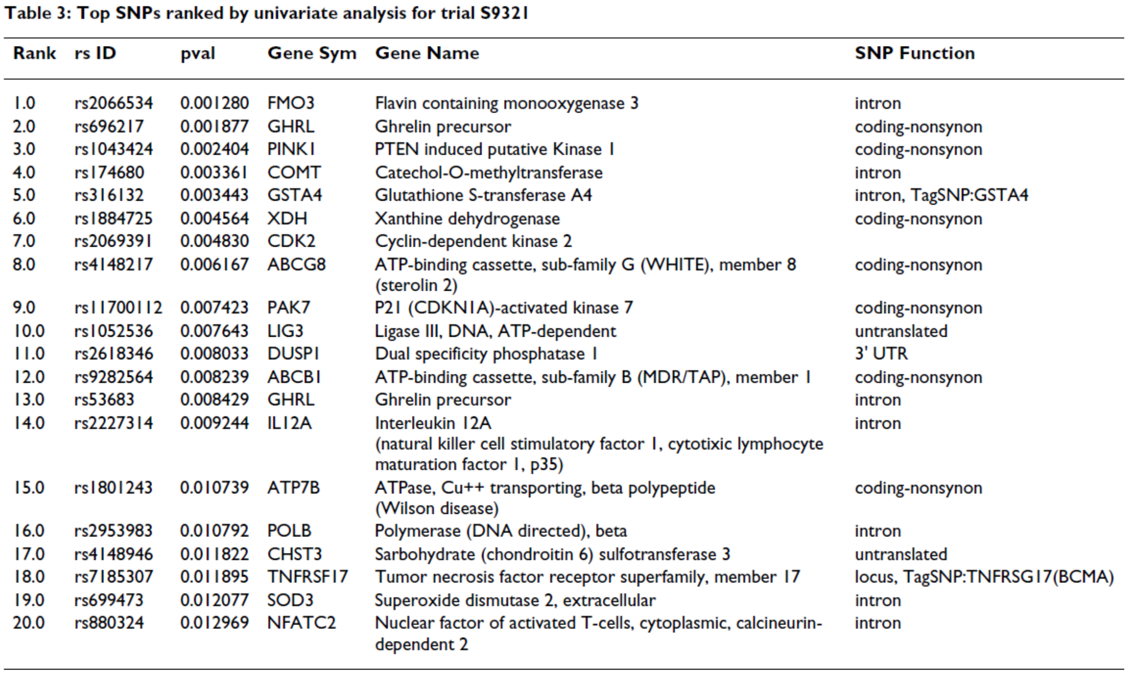


Table 3: Top SNPs ranked by univariate analysis for trial S9321

Rank	rs ID	pval	Gene Sym	Gene Name	SNP Function
1.0	rs2066534	0.001280	FMO3	Flavin containing monooxygenase 3	intron
2.0	rs696217	0.001877	GHRL	Ghrelin precursor	coding-nonsynon
3.0	rs1043424	0.002404	PINK1	PTEN induced putative Kinase 1	coding-nonsynon
4.0	rs174680	0.003361	COMT	Catechol-O-methyltransferase	intron
5.0	rs316132	0.003443	GSTA4	Glutathione S-transferase A4	intron, TagSNP:GSTA4
6.0	rs1884725	0.004564	XDH	Xanthine dehydrogenase	coding-nonsynon
7.0	rs2069391	0.004830	CDK2	Cyclin-dependent kinase 2	
8.0	rs4148217	0.006167	ABCG8	ATP-binding cassette, sub-family G (WHITE), member 8 (sterolin 2)	coding-nonsynon
9.0	rs11700112	0.007423	PAK7	P21 (CDKN1A)-activated kinase 7	coding-nonsynon
10.0	rs1052536	0.007643	LIG3	Ligase III, DNA, ATP-dependent	untranslated
11.0	rs2618346	0.008033	DUSP1	Dual specificity phosphatase 1	3' UTR
12.0	rs9282564	0.008239	ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	coding-nonsynon
13.0	rs53683	0.008429	GHRL	Ghrelin precursor	intron
14.0	rs2227314	0.009244	IL12A	Interleukin 12A (natural killer cell stimulatory factor 1, cytotixic lymphocyte maturation factor 1, p35)	intron
15.0	rs1801243	0.010739	ATP7B	ATPase, Cu++ transporting, beta polypeptide (Wilson disease)	coding-nonsynon
16.0	rs2953983	0.010792	POLB	Polymerase (DNA directed), beta	intron
17.0	rs4148946	0.011822	CHST3	Sarbohydrate (chondroitin 6) sulfotransferase 3	untranslated
18.0	rs7185307	0.011895	TNFRSF17	Tumor necrosis factor receptor superfamily, member 17	locus, TagSNP:TNFRSG17(BCMA)
19.0	rs699473	0.012077	SOD3	Superoxide dismutase 2, extracellular	intron
20.0	rs880324	0.012969	NFATC2	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2	intron

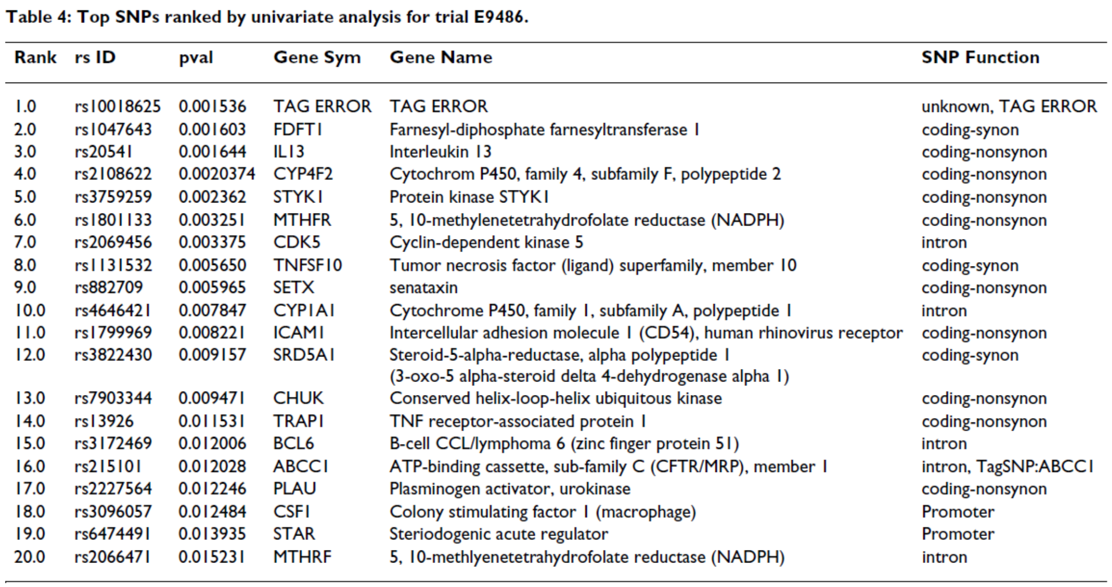
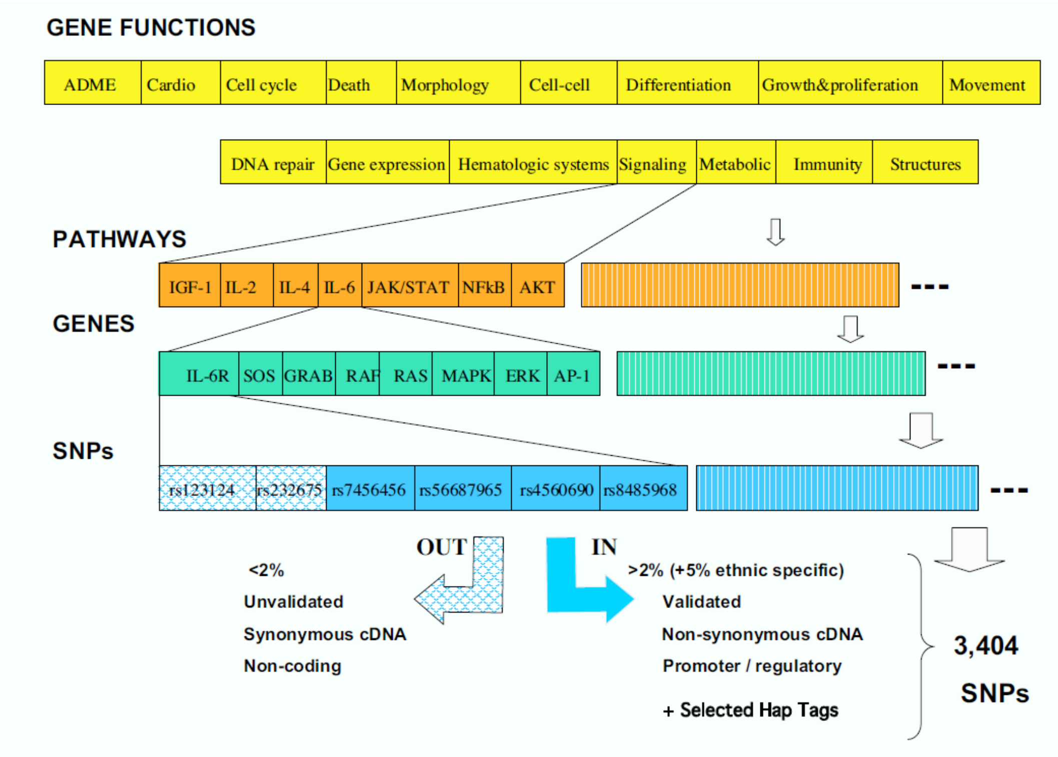


Table 4: Top SNPs ranked by univariate analysis for trial E9486.

Rank	rs ID	pval	Gene Sym	Gene Name	SNP Function
1.0	rs10018625	0.001536	TAG ERROR	TAG ERROR	unknown, TAG ERROR
2.0	rs1047643	0.001603	FDFT1	Farnesyl-diphosphate farnesyltransferase 1	coding-synon
3.0	rs20541	0.001644	IL13	Interleukin 13	coding-nonsynon
4.0	rs2108622	0.0020374	CYP4F2	Cytochrom P450, family 4, subfamily F, polypeptide 2	coding-nonsynon
5.0	rs3759259	0.002362	STYK1	Protein kinase STYK1	coding-nonsynon
6.0	rs1801133	0.003251	MTHFR	5, 10-methylenetetrahydrofolate reductase (NADPH)	coding-nonsynon
7.0	rs2069456	0.003375	CDK5	Cyclin-dependent kinase 5	intron
8.0	rs1131532	0.005650	TNFSF10	Tumor necrosis factor (ligand) superfamily, member 10	coding-synon
9.0	rs882709	0.005965	SETX	senataxin	coding-nonsynon
10.0	rs4646421	0.007847	CYP1A1	Cytochrome P450, family 1, subfamily A, polypeptide 1	intron
11.0	rs1799969	0.008221	ICAM1	Intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	coding-nonsynon
12.0	rs3822430	0.009157	SRD5A1	Steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)	coding-synon
13.0	rs7903344	0.009471	CHUK	Conserved helix-loop-helix ubiquitous kinase	coding-nonsynon
14.0	rs13926	0.011531	TRAP1	TNF receptor-associated protein 1	coding-nonsynon
15.0	rs3172469	0.012006	BCL6	B-cell CCL/lymphoma 6 (zinc finger protein 51)	intron
16.0	rs215101	0.012028	ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	intron, TagSNP:ABCC1
17.0	rs2227564	0.012246	PLAU	Plasminogen activator, urokinase	coding-nonsynon
18.0	rs3096057	0.012484	CSF1	Colony stimulating factor 1 (macrophage)	Promoter
19.0	rs6474491	0.013935	STAR	Steriodogenic acute regulator	Promoter
20.0	rs2066471	0.015231	MTHRF	5, 10-methlyenetetrahydrofolate reductase (NADPH)	intron

Figure 1. SNP selection strategy for the BOAC SNP panel. For full description, see Methods and Results. Numbers under the cell functions indicate the final number of SNPs on the chip in each category.


GENE FUNCTIONS
ADME
Cardio
Cell cycle
Death
Morphology
Cell-cell
Differentiation
Growth&proliferation
Movement
DNA repair
Gene expression
Hematologic systems
Signaling
Metabolic
Immunity
Structures
PATHWAYS
IGF-1
IL-2
IL-4
IL-6
JAK/STAT
NFkB
AKT

GENES
IL-6R
SOS
GRAB
RAF
RAS
MAPK
ERK
AP-1

SNPs
rs123124
rs232675
rs7456456
rs56687965
rs4560690
rs8485968

OUT
<2%
Unvalidated
Synonymous cDNA
Non-coding
IN
>2% (+5% ethnic specific)
Validated
Non-synonymous cDNA
Promoter / regulatory
+ Selected Hap Tags
3,404
SNPs

Figure 2. Racial allelic frequency patterns. A) Diagonal plot comparing minor allele frequencies between BOAC SNPs of Caucasian versus African American myeloma patients. Note high rate of allelic variation. B) Diagonal plot comparing minor allele frequencies between BOAC SNPs of Caucasian myeloma patients versus unaffected Caucasians.

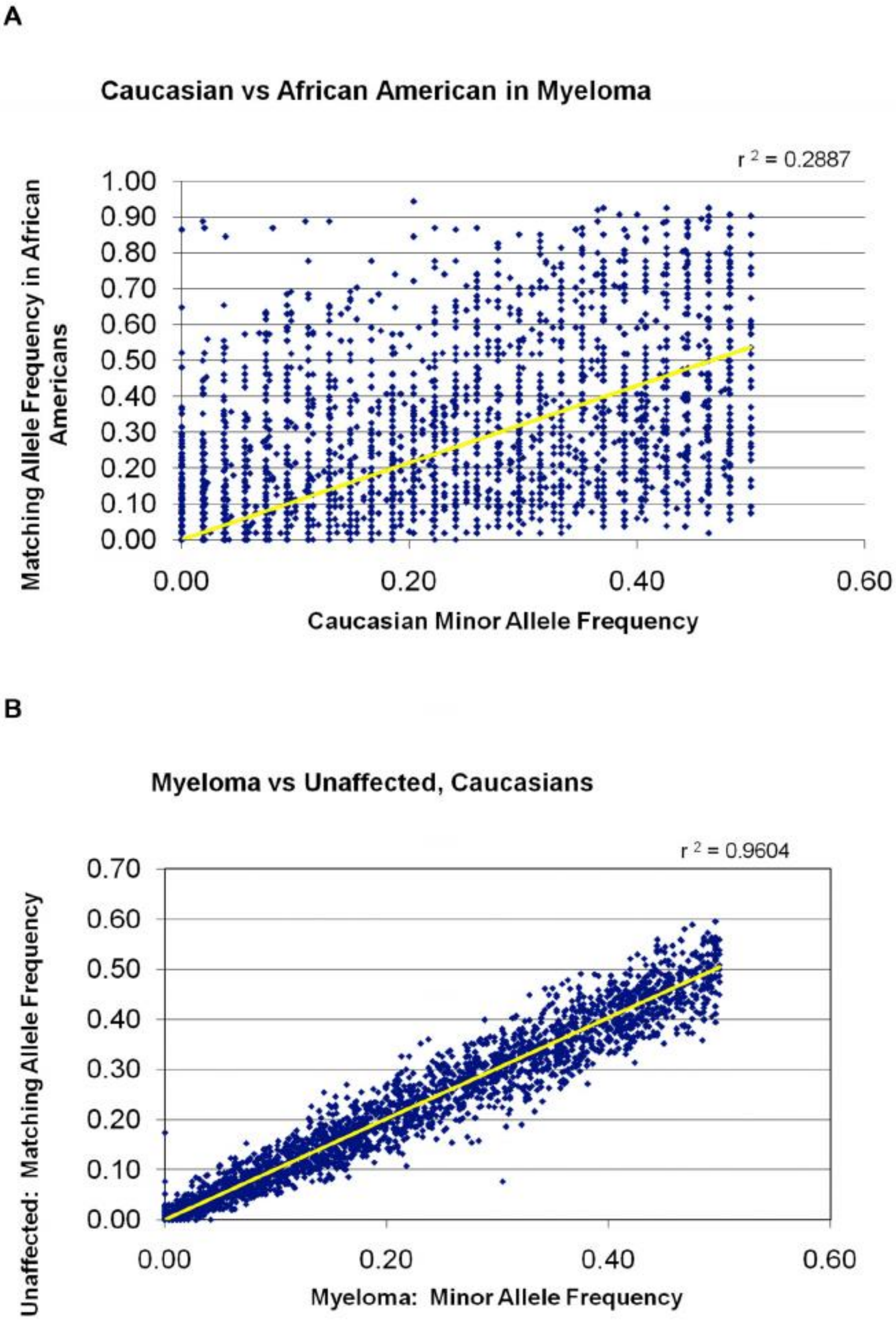

A
Caucasian vs African American in Myeloma
r 2 = 0.2887
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Matching Allele Frequency in African Americans
0.00
0.20
0.40
0.60
Caucasian Minor Allele Frequency
B
Myeloma vs Unaffected, Caucasians
r 2 = 0.9604
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Unaffected: Matching Allele Frequency
0.00
0.20
0.40
0.60
Myeloma: Minor Allele Frequency

Figure 3. Survival prediction accuracy versus distribution of random subsets from the BOAC SNP panel. The 173 SNP profiles were randomly paired 10,000 times, and the accuracy of the SNP prediction was determined, resulting in a distribution of accuracy, centered around 50%. This is compared with survival group prediction accuracy of the full SNP panel (66%) and the subset of SNPs (76%) described in the text. Odds ratios and confidence intervals are given for each. In both cases the pvalue for predictive power is less than 0.0001.

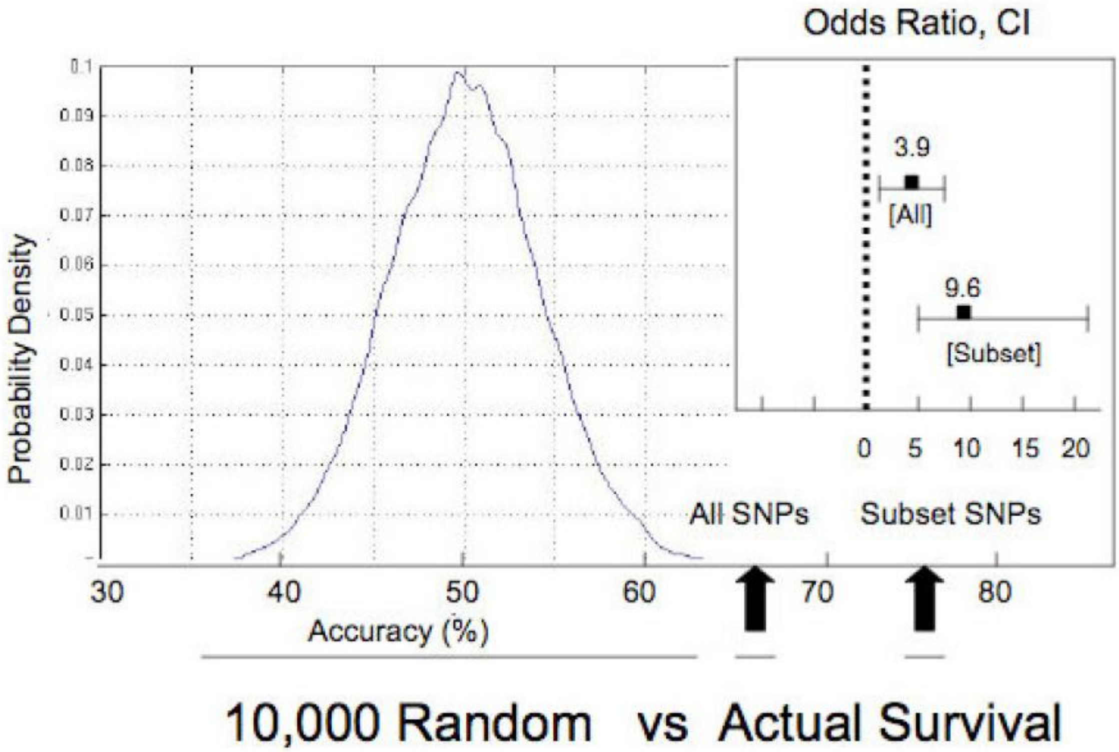
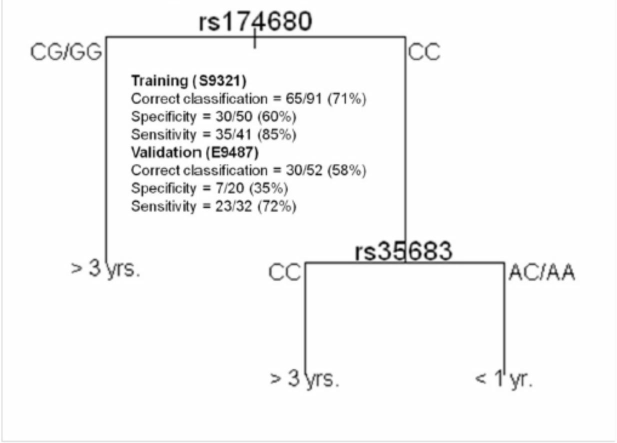

Odds Ratio, CI
3.9
[All]
9.6
[Subset]
0
5
10
15
20
All SNPs
Subset SNPs
Probability Density
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
30
40
50
60
70
80
Accuracy (%)
10,000 Random vs Actual Survival

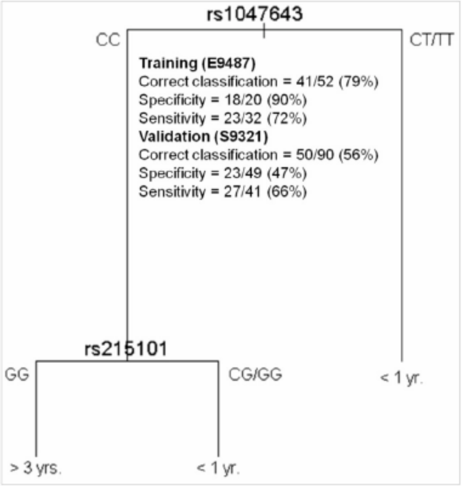
Figure 4. Recursive partitioning tree from S9321 and E9486. The classification prediction was calculated for one trial and tested on the other as validation.

Recursive partitioning tree from S9321



Training (S9321)
Correct classification = 65/91 (71%)
Specificity = 30/50 (60%)
Sensitivity = 35/41 (85%)
Validation (E9487)
Correct classification = 30/52 (58%)
Specificity = 7/20 (35%)
Sensitivity = 23/32 (72%)

Recursive partitioning tree from E9487



Training (E9487)
Correct classification = 41/52 (79%)
Specificity = 18/20 (90%)
Sensitivity = 23/32 (72%)
Validation (S9321)
Correct classification = 50/90 (56%)
Specificity = 23/49 (47%)
Sensitivity = 27/41 (66%)

APPENDIX II

INHERITED GENETIC VARIATION AND THE RISK OF DEVELOPING MULTIPLE MYELOMA

Johnson D.C[1], Sonneveld P[2], Corthals S.L[2], Davies F.E[1], Ramos C3, Shaughnessy JD Jr[4], Walker B.A[1], Gregory W.M[5], Haznadar M[3], Gonzalez D[1], A.G. Uitterlinden[6], Lokhorst H.M[7], Durie B.G[8], Barlogie B[5], Van Ness B[3], Dalsu Baris[9], Morgan G.J[1].

[1] Section of Haemato-Oncology, Institute of Cancer Research, London, UK.[2] Department of Hematology, Erasmus Medical Center, Rotterdam, the Netherlands. [3] Department of Genetics, Cell Biology, and Development, University of Minnesota. Minneapolis, USA. [4] Myeloma Institute for Research and Therapy, University of Arkansas, Little Rock, AR, USA. [5] Clinical Trials Research Unit, University of Leeds, Leeds, UK. [6] Department of Internal Medicine, Erasmus Medical Center, Rotterdam, the Netherlands [7] Department of Hematology, University Medical Center, Utrecht the Netherlands. [8] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [9] International Myeloma Foundation and Cedars Sinai Comprehensive Cancer Center, Los Angeles, USA.

This investigation was supported by International Myeloma Foundation (Bank on a Cure project (BOAC)), Myeloma UK and the National Institute for Health and Biological Research at the Royal Marsden Hospital. We would like to thank members of the MRC Myeloma-IX Trial Management Group and HOVON-50/GMMG-HD3 trials offices.

I contributed to genotyping of some clinical trials represented in the study, as well as with editing of the manuscript. This study was a large collaborative effort.

We have performed an expanded candidate gene study by combining genotyping data from 2595 presenting myeloma cases with 8974 matched population controls across three cross-validating cohorts of European origin from the UK, US and the Netherlands. Strong associations were found with non-synonymous SNPs in FCRL5 (rs6679793, odds ratio (OR) =1.54, P=4.2 x 10-14); CYP2C19 (rs3758581, OR=2.02, P=3.08 x 10-9); CAMKK2 (rs1132780, OR=1.56, P= 8.58 x 10-8) and SELP (rs6127, OR=1.26; P=9.5 x 10-7). We also see associations with tagging-SNPs across potential regulatory regions in NFATC1 (rs4799055, OR=0.77, P=6.93 x10-13); CYP1A2 (rs2472304, OR=1.26, P=3.40 x 10-8); IFNGR2, rs8131980, OR=1.24, P=4.23 x 10-7); GSTA4 (rs7496, OR=1.59, P=6.33 x 10-6); and TNFRSF17 (rs3743591, OR=2.92 P=1.72 x 10-11). These data provide evidence of an underlying genetic susceptibility to multiple myeloma disease.

Introduction:

Approaches aimed at identifying the causative factors associated with the risk of developing Multiple Myeloma (MM), have largely been inconclusive (1-4) and genetic epidemiology approaches offer an important new avenue of research. Evidence supporting a role for genetic variation in the aetiology of MM comes from a variety of sources, where a greater risk is seen in first degree relatives of patients with MM 5 and MGUS (6-7), there are reported familial clusters of MM (8-11), and an increased risk is seen in African Americans (12-13). Although studies to date using genetic epidemiological methods have lacked power and have used only a limited numbers of SNPs (14-18), they have suggested possible associations with genetic variation in a number of pathways, including growth factors, DNA repair, one carbon metabolism, cell cycle progression and apoptosis pathways (19-24).

In this work, we have taken a hypothesis driven approach to examine the role of genetic variation associated with MM, using a custom genotyping array to study 2595 presenting MM cases of European origin derived from the UK, US and the Netherlands. The custom array consisted of 3404 SNPs in 964 genes, focussing on SNP variation in molecular pathways involved in the pathogenesis and treatment response of MM (25-27). In order to understand the distribution of these variants within the normal unaffected population, we accessed 8974 population control genotypes generated by the Wellcome Trust Case Controls Consortium (WTCCC) and the Erasmus Rotterdam Health and the Elderly Study Group (ERGO). A comparison of the genotypic distribution between the case and control populations has allowed us to identify genetic variation that associates with MM predisposition.

Materials and Methods:

A hypothesis driven custom (BOAC) array (26-28) was built by generating a candidate list, comprising of genes with experimental evidence of a functional role linked to the risk and progression of MM from the literature. The candidate gene list was systematically interrogated by SNP databases including dbSNP, SNP500 (29), Promolign (30) focusing on common coding and regulatory SNPs with minor allele frequency (maf) >2%. We sought to validate previously published work in myeloma epidemiology and pharmacogenetics in general by including variants with previously described associations or putative functional effects in myeloma. We also had a focus on ADME genes (absorption, distribution, metabolism and excretion) (31). The final array comprised of 3404 SNPs, selected in predicted functional regions within 964 genes spanning 67 molecular pathways.

SNPs were genotyped using DNA extracted from the peripheral blood of 2595 presenting MM cases derived from a number of randomised clinical trials from the UK (MRC Myeloma-IX study), US (ECOGA4/A03, ECOG9486, SWOG9321) and the Netherlands

(HOVON-50, HOVON-65, GMMG-HD3). The MM cases were genotyped on Affymetrix® Targeted arrays using True-Tag protocol version 1.5. Genotype calling involves background subtraction and spectral overlap correction, followed by expectation-maximisation (E/M) clustering across all processed arrays. Samples were filtered using a >90% call rate cut-off and SNPs were filtered for call rates >95% and for departures from Hardy Weinberg equilibrium P >10-5 across each SNP cluster. To understand and protect the analysis from potential platform effects, 58 DNA samples from Coriell CEPH Hapmap individuals were assayed on the custom assay to validate the call performance of the panel. A total of 2606 SNPs were present on both the custom array and HAPMAP, with a SNP call correlation of 96.1%; SNPs falling below this were removed from the analysis (132 SNPs). The Coriell sample genotype validation was replicated in each of the three labs to ensure there was no differential bias in genotyping scoring between sites.

The control population sets consisted of the WTCCC2 study (32-33) with 3000 individuals from the 1958 British Birth Cohort (58C) and the UK Blood Service collections (UKBS), genotyped on both the Illumina 1.2M Duo (Human1-2M-DuoCustom_v1) and the Affymetrix v6.0 chips; and the 5974 individuals from the Dutch >55yrs old population controls from the ERGO study (34-35), genotyped on the Illumina 550K array. 2352 autosomal SNPs typed on the targeted array with MAF >1% were also available by direct genotyping in the publicly accessible WTCCC population datasets. 2164 SNPs were available from the WTCCC2 data genotyped on the Illumina 1.2M Duo array (Human1-2M-DuoCustom_v1). 719 SNPs from the targeted array were available on the Affymetrix v6.0 chip, 601 SNPs of these were not seen on the Illumina array. An additional 230 SNPs was provided by imputation methods. Indirectly population genotypes were generated by imputation with IMPUTE2 as described (36), utilising both 1000 genomes and HapMap3 data sets. We also selected a set of proxy SNPs in complete LD (D' and R2 =1) with the target SNP using SNAP (37). The WTCCC BS dataset were paired with the US cases and the WTCCC 58C were used to provide population controls for the UK cases. In the Dutch series there was 992 SNPs available by direct genotyping on the Illumina 550K array from the ERGO study, a further 255 SNPs were available following imputation and 630SNPs by identification of proxies. Quality control measures were applied with a >90% call rate for the controls samples. A call rate of >95% filter was applied for each SNP assay; departures from Hardy-Weinberg equilibrium (HWE) with a P>10-5 were also excluded.

Statistical Analysis

Analysis was performed in the program PLINK 1.07 (38). Subjects with evidence of cryptic relatedness and non-European background were excluded from the analysis. A Genomic inflation factor λ was evaluated based on median chi-squared for each set of analyses and showed little evidence for an inflation of test statistics (Figure 2). A logistic regression based additive model was fitted to each SNP with MM risk as the outcome measure and was adjusted for the covariates age and sex. There were 1240 SNPs in 582 genes fulfilling quality control criteria across all three case-controls sets available for meta-analysis which was carried out to examine associations across the discovery and validation datasets. The PLINK --epistasis option was used for case and controls analysis, and the --fast-epistasis option was used for case-only analysis. We performed search of pair-wise interactions using epistasis analysis (39). Combinations were restricted to SNPs more than 5 MB apart or on different chromosomes. To correct for multiple-testing in the epistasis analysis, analyses were restricted to SNPs producing nominally significant values of P<10-4 (40).

The relationship of expression with genotype in SNPs associated with risk was examined in a set of 212 UK (41) and 264 Arkansas cases (42-43) with expression data derived from the U133 Plus 2.0 expression array (Affymetrix). Utilising publicly available data through NCBI GEO datasets and Hapmap, we also correlated mRNA expression level with genotype in the same SNPs from 86 Epstein-Barr virus (EBV) transformed lymphoblastoid cell lines in

European Hapmap individuals using Sentrix Human-6 Expression BeadChips (Illumina) (44). A Wilcoxon-type test for trend (45) was used to compare differences in the distribution of levels of mRNA expression between SNP genotypes in these sets.

Results:

The study design and comparator groups are shown in Figure 1 and the strongest positive associations seen are summarised in Table 1. The non-synonymous variants significantly associated with myeloma risk across each of the studies are shown as a Forest plot, Figure 3. Further haplotypic structures around the individual SNPs are described in more detail in the supplementary data. The strongest association is seen with the non-synonymous SNP rs6679793 in the gene FCRL5 (Fc receptor-like 5). Under a random-effects model, the A-allele in the SNP rs6679793 is associated with an increased myeloma risk with a P=4.2 x 10-14, OR=1.54, CI 1.38-1.72. An association is seen with an intronic SNP rs4799055 in the gene NFATC1 (Nuclear factor of activated T-cells, cytoplasmic 1), under a random-effects model the T-allele being protective, P=6.93 x10-13, OR=0.77 CI 0.66-0.87. We see an association with a SNP in the untranslated region of the gene TNFRSF17 (Tumour Necrosis Factor Receptor Superfamily, member 17), the G-allele of SNP rs3743591 being associated with risk, P=1.72 x 10-11, OR=2.92 CI 2.14-3.99, in a random-effect model. An association is seen with a non-synonymous SNP rs3758581 in CYP2C19 (Cytochrome P450 2C19), the A–allele being associated with increased risk, P=3.08 x 10-9 OR=2.02 CI 1.60-2.55 under a random-effect model. We found an association with a non-synonymous SNP rs1132780 in the gene CAMKK2 (Calcium/calmodulin-dependent protein kinase kinase), the A-allele being associated with risk P=8.58 x 10-8 OR=1.56 CI 1.32-1.83. An association is found with the intronic SNP rs2472304 in the gene CYP1A2 (Cytochrome P450 1A2) the G-risk allele being associated with an increased risk P=3.40 x 10-8 OR=1.26 CI 1.16-1.37. We see an association with the gene IFNGR2 (interferon gamma receptor 2), SNP rs8131980, the A-allele being associated with increased risk P=4.23 x 10-7 OR=1.24 CI 1.14-1.35.

We see two further associations that did not achieve genome wide significance, but are close enough to be reported. There is an association with a non-synonymous SNP rs6127 in the gene SELP (Selectin P), the C-risk allele being associated with risk, P=3.09 x 10-6, OR=1.18 CI 1.09-1.26; and also with a SNP rs7496 in the untranslated region of GSTA4 (Glutathione S-transferase alpha 4), the T-allele being associated with increased risk, P=6.33 x 10-6 OR=1.59 CI 1.07-2.20. It is important to note, given that imputation techniques can bias genotype frequencies that SNP rs37458581 was imputed across all three control cohorts, whilst rs6127 was imputed in the Dutch cohort only; all other case and controls genotypes across the associated MM risk variants were directly genotyped.

In order to examine potential functionality of significantly associated SNPs present within regulatory regions, we correlated the presence of the risk variant with the expression level of the gene probeset in two separate myeloma datasets and a lymphoblastoid cell line set (HAPMAP). All associated genes were shown to be expressed highly in myeloma tumour cells and the results of this analysis are given in full in the supplementary data. We see evidence of a significant trend in expression level with the risk associated genotypes in SELP, NFATC1, CAMKK2, FCRL5, GSTA4 in one of the two myeloma expression sets, but no SNP showed a significant correlation with expression across both myeloma sets.

There is increasing evidence that gene–gene interactions (epistasis) could play a role in susceptibility to complex diseases.46-48 To investigate statistical evidence of epistasis, we tested whether the observed counts for SNP-SNP combinations between associated variants, reflected the expected values within the null hypothesis of no interaction. Examining SNP x SNP interactions in both cases and controls we see that rs6127 (SELP) has a significant interaction with a number of SNPs including rs3743591 (TNFRSF17) ORinteraction=0.66, P=1.6x10-9, rs3758581 (CYP2C19) OR interaction=1.54, and rs7496 (GSTA4) P=6.5x10-6, ORinteraction=1.33, P=1.6x10-5. It is possible to increase the power to detect such interactions by performing a case-only analysis.49 In the case-only method we see a major interaction between rs6127 (SELP) and rs3743591 (TNFRSF17), with seven further interactions exceeding P>10-5. (Table 2).

Discussion:

In this study we describe a number of inherited genetic variants that associate with the risk of developing MM that are informative of both the biological and environmental contributions in this respect. It is the largest genetic epidemiological study to date addressing the genetic contribution to the risk of developing MM with the study design laying somewhere between a classical candidate gene study and a whole genome scan. We have assayed 1284 SNPs across 524 genes, in genes that are potentially involved in the pathogenesis of MM and specifically asked how they may modulate the risk of developing MM. The associations seen are strong, across three cohorts of European origin with a number achieving genome wide significance P<10-7. In addition there is experimental support for the functional role of the variants identified. However, it should be noted, this hypothesis driven approach cannot detect associations outside the candidate panel and will not, therefore, observe associations potentially detectable by a genome wide approach.

Micro-environmental interactions and B cell signalling pathways are relevant to the development of myeloma possibly by mediating cell survival following genetic damage. In this context, we find a number of relevant associations including with 2 SNPs derived from chromosome 1, a region which is frequently associated with myeloma progression and poor clinical outcome (42, 50-52). The strongest association was seen with the SNP rs6679793 in

FCRL5, which is situated at bp 157514097 and codes for an amino acid change of Tyrosine to Histidine at position 267. FCRL5, also known as IRTA2, is an immunoglobulin-like cell surface receptor, expressed on the plasma membrane of the majority of MM patients (53), is involved in B-cell differentiation and is thought to play an immuno-regulatory role in the marginal zone (54-55). The second associated SNP from chromosome 1, rs6127, lies at bp 169566313 in the SELP gene and results in an amino acid change of Arginine to Asparagine at position 603. SELP is a cell adhesion molecule on the surfaces of activated endothelial cells, with the P-selectin ligand being highly expressed in MM cells compared to normal plasma cells (56).

We see an association with, the SNP rs8131980 in IFNGR2 located on chromosome 21 at bp 34810007; IFNGR2 is an integral part of the IFNγ signal transduction pathway and interacts with GAF, JAK1, and/or JAK2 to deliver survival signals. A further associated SNP rs3743591, located in TNFRSF17 at bp 12059032 on chromosome 16, is a receptor that is preferentially expressed in mature B lymphocytes, and is important for B cell development and autoimmune response. TNFRSF17 can bind to TNFRSF13b also known as BAFF, leading to NFκB and MAPK8/JNK activation. TNFRSF17 also binds to various TRAF family members, and thus may transduce signals for cell survival and proliferation.

The intronic SNP rs4799055, in NFATC1, at bp 77182003 on chromosome 18, represents another associated MM risk variant in a gene involved in mediating signalling via cytokine signalling pathways. NFATC1 can influence the expression of cytokine genes, such as IL-2 or IL-4 and may not only regulate B cell activation and proliferation, but also the differentiation and programmed death of T-lymphocytes. NFATC1 is also a transcription factor that plays a central role in osteoclast formation An associated risk variant in a gene mediating similar effects is rs1132780 in CAMKK2 leading to an amino acid change at position 363 from an Arginine to a Serine, located at bp 121691096 on chromosome 12. CAMKK2 plays a key role in autophagy and cell survival and it is also known that CAMKK-dependent Akt activation inhibits IL-1β-induced NFκB activation through an interference with the coupling of IRAK1 to MyD88 (57). The potential importance of the gene in myeloma was highlighted in an RNAi scan for genes critical in myeloma cell function (58). Further evidence for variation in immune system related genes impacting on risk comes from the observation of associations with a SNP in the untranslated region of IL-8RB (rs1126579), P < 5.27 x 10-5 and NFκB intron SNP rs4648133 at P< 9.32 x 10-7. These analyses were only performed across two studies, as the data was not available for the Dutch set.

The associations with absorption, distribution, metabolism, excretion (ADME) genes seen in this study suggest that there are potential environmental exposures and opens the way for further validation in studies designed to investigate gene environment interactions. We see associations with a number of ADME genes including the CYP2C19 SNP rs3758581 present in poor metaboliser haplotypes CYP2C19*2 and CYP2C19*3, found on chromosome 10 at bp 96602623, resulting in an amino change from Isoleucine to Valine at position 331. CYP2C19 is a key metabolising gene residing in the endoplasmic reticulum and can be induced to high levels in liver and other tissues by various relevant environmental exposures (59). The CYP1A2 SNP rs2472304 is associated with myeloma risk, is located on chromosome 15 at bp 75044238. CYP1A2 is found in the endoplasmic reticulum and is induced by some polycyclic aromatic hydrocarbons (PAHs). We could not demonstrate a relationship between the presence of the risk variant and CYP1A2 expression levels in the myeloma plasma cells, but this may be due to its impact being mediated at the level of the liver. A further association in an ADME gene with observed with SNP rs7496 in the GSTA4 gene from chromosome 6 at bp 52842839.

An observation of potential importance from the analysis of the SNP x SNP interactions is the suggestion that the two major pathways outlined above seem to mediate risk in tandem. In the strongest interaction discovered in this analysis we see two risk alleles from the immune response pathway, rs6157 (SELP) and rs3743591 (TNFRSF17). This interaction was negative, and the risk alleles were seen to be partially exclusive of each other, indicating that may be serving a similar functional role in governing risk. We also then see a number of synergistic interactions between members from the two major pathway, an example of this is interaction between rs6157 (SELP) from immune response and rs3758581 (CYP2C19) from an ADME pathway.

In addition to describing novel associations we sought to validate previously reported associations with MM risk in our series of patients. While DNA damage and repair has been linked to myeloma risk, we saw no evidence of an association with the DNA repair genes XRCC5, MRE11A, BRCA1, BRCA2 and FANCA. However, we did see a strong association with a variant within ERCC4, but we are unsure of the significance of this finding due to the heterogeneity seen between studies. While prior studies have suggested associations with one carbon metabolism, we did not see such an association (20, 60-63). We also did not see an association with genetic variation with growth factor signalling pathways as has been reported previously (18). Thus while classic epidemiological association studies have identified potential environmental exposures as well as chronic immune stimulation, as being relevant associations with the risk of developing MM and a number of groups have suggested that inherited variation in IL-6, IL-1B, TNFα and NFκB (23, 64-72) may play a role in MM risk, none of these associations have been either adequately replicated or reached genomewide significance.

Our findings provide the first evidence of common genetic variants linked with the risk of developing MM at the level of genome-wide significance. The nine positive associations fall into two broad groups mediating immune response and the response to environmentally encountered carcinogens by ADME genes. We provide evidence that individuals at greatest risk of developing MM carry risk alleles in both the immune response and ADME pathways. Further evaluation of these pathways potentially by sequencing approaches in both MM and MGUS cohorts will provide a greater understanding of the mechanisms driving the transition from a plasma cell to the myeloma tumour cell and potentially enable biomarker discovery to allow anticipation of an individual’s MM risk.

References:

1. Morgan GJ, Davies FE, Linet M. Myeloma aetiology and epidemiology. Biomed Pharmacother. 2002;56:223-234.

2. Morgan GJ, Adamson PJ, Mensah FK, et al. Haplotypes in the tumour necrosis factor region and myeloma. Br J Haematol. 2005;129:358-365.

3. Cartwright RA, Gilman EA, Nicholson P, Allon D. Epidemiology of multiple myeloma in parts of England, 1984-1993. Hematol Oncol. 1999;17:31-38.

4. Alexander DD, Mink PJ, Adami HO, et al. Multiple myeloma: a review of the epidemiologic literature. Int J Cancer. 2007;120 Suppl 12:40-61.

5. Kristinsson SY, Bjorkholm M, Goldin LR, McMaster ML, Turesson I, Landgren O. Risk of lymphoproliferative disorders among first-degree relatives of lymphoplasmacytic lymphoma/Waldenstrom macroglobulinemia patients: a population-based study in Sweden. Blood. 2008;112:3052-3056.

6. Landgren O, Kristinsson SY, Goldin LR, et al. Risk of plasma cell and lymphoproliferative disorders among 14621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. Blood. 2009;114:791-795.

7. Vachon CM, Kyle RA, Therneau TM, et al. Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance. Blood. 2009;114:785-790.

8. Lynch HT, Thome SD. Familial multiple myeloma. Blood. 2009;114:749-750.

9. Lynch HT, Watson P, Tarantolo S, et al. Phenotypic heterogeneity in multiple myeloma families. J Clin Oncol. 2005;23:685-693.

10. Ogmundsdottir HM, Haraldsdottirm V, Johannesson GM, et al. Familiality of benign and malignant paraproteinemias. A population-based cancer-registry study of multiple myeloma families. Haematologica. 2005;90:66-71.

11. Jain M, Ascensao J, Schechter GP. Familial myeloma and monoclonal gammopathy: a report of eight African American families. Am J Hematol. 2009;84:34-38.

12. Brown LM, Gridley G, Check D, Landgren O. Risk of multiple myeloma and monoclonal gammopathy of undetermined significance among white and black male United States veterans with prior autoimmune, infectious, inflammatory, and allergic disorders. Blood. 2008;111:3388-3394.

13. Landgren O, Weiss BM. Patterns of monoclonal gammopathy of undetermined significance and multiple myeloma in various ethnic/racial groups: support for genetic factors in pathogenesis. Leukemia. 2009;23:1691-1697.

14. Brown EE, Lan Q, Zheng T, et al. Common variants in genes that mediate immunity and risk of multiple myeloma. Int J Cancer. 2007;120:2715-2722.

15. Hayden PJ, Tewari P, Morris DW, et al. Variation in DNA repair genes XRCC3, XRCC4, XRCC5 and susceptibility to myeloma. Hum Mol Genet. 2007;16:3117-3127.

16. Hosgood HD, 3rd, Baris D, Zhang Y, et al. Genetic variation in cell cycle and apoptosis related genes and multiple myeloma risk. Leuk Res. 2009;33:1609-1614.

17. Hosgood HD, 3rd, Baris D, Zhang Y, et al. Caspase polymorphisms and genetic susceptibility to multiple myeloma. Hematol Oncol. 2008;26:148-151.

18. Birmann BM, Tamimi RM, Giovannucci E, et al. Insulin-like growth factor-1- and interleukin-6-related gene variation and risk of multiple myeloma. Cancer Epidemiol Biomarkers Prev. 2009;18:282-288.

19. Maggini V, Buda G, Galimberti S, et al. Lack of association of NQO1 and GSTP1 polymorphisms with multiple myeloma risk. Leuk Res. 2008;32:988-990.

20. Zintzaras E, Giannouli S, Rodopoulou P, Voulgarelis M. The role of MTHFR gene in multiple myeloma. J Hum Genet. 2008;53:499-507.

21. Roddam PL, Rollinson S, O'Driscoll M, Jeggo PA, Jack A, Morgan GJ. Genetic variants of NHEJ DNA ligase IV can affect the risk of developing multiple myeloma, a tumour characterised by aberrant class switch recombination. J Med Genet. 2002;39:900-905.

22. Cozen W, Gebregziabher M, Conti DV, et al. Interleukin-6-related genotypes, body mass index, and risk of multiple myeloma and plasmacytoma. Cancer Epidemiol Biomarkers Prev. 2006;15:2285-2291.

23. Dring AM, Davies FE, Rollinson SJ, et al. Interleukin 6, tumour necrosis factor alpha and lymphotoxin alpha polymorphisms in monoclonal gammopathy of uncertain significance and multiple myeloma. Br J Haematol. 2001;112:249-251.

24. Davies FE, Rollinson SJ, Rawstron AC, et al. High-producer haplotypes of tumor necrosis factor alpha and lymphotoxin alpha are associated with an increased risk of myeloma and have an improved progression-free survival after treatment. J Clin Oncol. 2000;18:2843-2851.

25. Van Ness B, Ramos C, Haznadar M, et al. Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival. BMC Med. 2008;6:26.

26. Johnson DC, Corthals S, Ramos C, et al. Genetic associations with thalidomide mediated venous thrombotic events in myeloma identified using targeted genotyping. Blood. 2008;112:4924-4934.

27. Durie BG, Van Ness B, Ramos C, et al. Genetic polymorphisms of EPHX1, Gsk3beta, TNFSF8 and myeloma cell DKK-1 expression linked to bone disease in myeloma. Leukemia. 2009;23:1913-1919.

28. Van Ness BG, Crowley JC, Ramos C, et al. SNP Associations with Event Free Survival in Myeloma from Two Phase III Clinical Trials Using the Bank On A Cure Chip. ASH Annual Meeting Abstracts. 2006;108:131-.

29. Packer BR, Yeager M, Burdett L, et al. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res. 2006;34:D617-621.

30. Zhao T, Chang LW, McLeod HL, Stormo GD. PromoLign: a database for upstream region analysis and SNPs. Hum Mutat. 2004;23:534-539.

31. Ahmadi KR, Weale ME, Xue ZY, et al. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. Nat Genet. 2005;37:84-89.

32. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464:713-720.

33. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661-678.

34. Uitterlinden AG, Ralston SH, Brandi ML, et al. The association between common vitamin D receptor gene variations and osteoporosis: a participant-level meta-analysis. Ann Intern Med. 2006;145:255-264.

35. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet. 2009;41:25-34.

36. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.

37. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24:2938-2939.

38. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-575.

39. Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. Am J Hum Genet. 2006;79:831-845.

40. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004;74:765-769.

41. Walker BA, Leone PE, Jenner MW, et al. Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. Blood. 2006;108:1733-1743.

42. Shaughnessy JD, Jr., Haessler J, van Rhee F, et al. Testing standard and genetic parameters in 220 patients with multiple myeloma with complete data sets: superiority of molecular genetics. Br J Haematol. 2007;137:530-536.

43. Shaughnessy JD, Jr., Zhan F, Burington BE, et al. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood. 2007;109:2276-2284.

44. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315:848-853.

45. Cuzick J. A Wilcoxon-type test for trend. Stat Med. 1985;4:87-90.

46. Schupbach T, Xenarios I, Bergmann S, Kapur K. FastEpistasis: A high performance computing solution for quantitative trait epistasis. Bioinformatics. 2010.

47. Nicodemus KK, Callicott JH, Higier RG, et al. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. Hum Genet. 2010.

48. Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet. 2009;17:1231-1240.

49. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392-404.

50. Leone PE, Walker BA, Jenner MW, et al. Deletions of CDKN2C in multiple myeloma: biological and clinical implications. Clin Cancer Res. 2008;14:6033-6041.

51. Decaux O, Lode L, Minvielle S, Avet-Loiseau H. [Genetic abnormalities in multiple myeloma: role in oncogenesis and impact on survival]. Rev Med Interne. 2007;28:677-681.

52. Fonseca R, Bergsagel PL, Drach J, et al. International Myeloma Working Group molecular classification of multiple myeloma: spotlight review. Leukemia. 2009;23:2210-2221.

53. Ise T, Nagata S, Kreitman RJ, et al. Elevation of soluble CD307 (IRTA2/FcRH5) protein in the blood and expression on malignant cells of patients with multiple myeloma, chronic lymphocytic leukemia, and mantle cell lymphoma. Leukemia. 2007;21:169-174.

54. Miller I, Hatzivassiliou G, Cattoretti G, Mendelsohn C, Dalla-Favera R. IRTAs: a new family of immunoglobulinlike receptors differentially expressed in B cells. Blood. 2002;99:2662-2669.

55. Hatzivassiliou G, Miller I, Takizawa J, et al. IRTA1 and IRTA2, novel immunoglobulin superfamily receptors expressed in B cells and involved in chromosome 1q21 abnormalities in B cell malignancy. Immunity. 2001;14:277-289.

56. Azab AK, Quang P, Azab F, et al. Role of Selectins in the Pathogenesis of Multiple Myeloma. ASH Annual Meeting Abstracts. 2009;114:951-.

57. Chen BC, Wu WT, Ho FM, Lin WW. Inhibition of interleukin-1beta -induced NF-kappa B activation by calcium/calmodulin-dependent protein kinase kinase occurs through Akt activation associated with interleukin-1 receptor-associated kinase phosphorylation and uncoupling of MyD88. J Biol Chem. 2002;277:24169-24179.

58. Tiedemann RE, Zhu YX, Schmidt J, et al. Kinome-wide RNAi studies in human multiple myeloma identify vulnerable kinase targets, including a lymphoid-restricted kinase, GRK6. Blood. 2010;115:1594-1604.

59. Brown LM, Pottern LM, Silverman DT, et al. Multiple myeloma among Blacks and Whites in the United States: role of cigarettes and alcoholic beverages. Cancer Causes Control. 1997;8:610-614.

60. Gonzalez-Fraile MI, Garcia-Sanz R, Mateos MV, et al. Methylenetetrahydrofolate reductase genotype does not play a role in multiple myeloma pathogenesis. Br J Haematol. 2002;117:890-892.

61. Kim HN, Kim YK, Lee IK, et al. Polymorphisms involved in the folate metabolizing pathway and risk of multiple myeloma. Am J Hematol. 2007;82:798-801.

62. Chiusolo P, Farina G, Putzulu R, et al. Analysis of MTHFR polymorphisms and P16 methylation and their correlation with clinical-biological features of multiple myeloma. Ann Hematol. 2006;85:474-477.

63. Ortega MM, Honma HN, Zambon L, et al. GSTM1 and codon 72 P53 polymorphism in multiple myeloma. Ann Hematol. 2007;86:815-819.

64. Abazis-Stamboulieh D, Oikonomou P, Papadoulis N, Panayiotidis P, Vrakidou E, Tsezou A. Association of interleukin-1A, interleukin-1B and interleukin-1 receptor antagonist gene polymorphisms with multiple myeloma. Leuk Lymphoma. 2007;48:2196-2203.

65. Duch CR, Figueiredo MS, Ribas C, Almeida MS, Colleoni GW, Bordin JO. Analysis of polymorphism at site -174 G/C of interleukin-6 promoter region in multiple myeloma. Braz J Med Biol Res. 2007;40:265-267.

66. Mazur G, Bogunia-Kubik K, Wrobel T, et al. IL-6 and IL-10 promoter gene polymorphisms do not associate with the susceptibility for multiple myeloma. Immunol Lett. 2005;96:241-246.

67. Aladzsity I, Kovacs M, Semsei A, et al. Comparative analysis of IL6 promoter and receptor polymorphisms in myelodysplasia and multiple myeloma. Leuk Res. 2009;33:1570-1573.

68. Du J, Yuan Z, Zhang C, et al. Role of the TNF-alpha promoter polymorphisms for development of multiple myeloma and clinical outcome in thalidomide plus dexamethasone. Leuk Res. 2010.

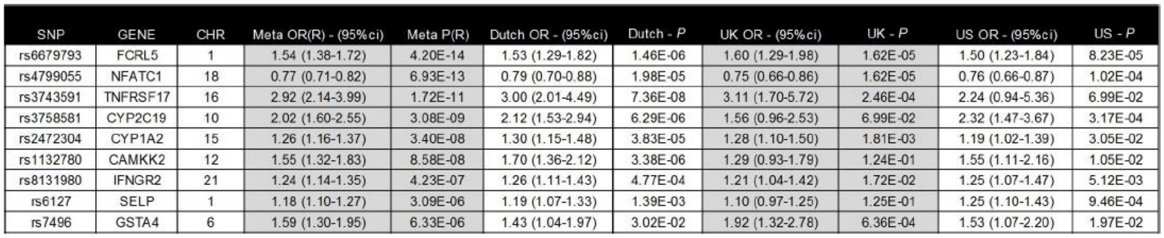
69. Kadar K, Kovacs M, Karadi I, et al. Polymorphisms of TNF-alpha and LT-alpha genes in multiple myeloma. Leuk Res. 2008;32:1499-1504.

70. Spink CF, Gray LC, Davies FE, Morgan GJ, Bidwell JL. Haplotypic structure across the I kappa B alpha gene (NFKBIA) and association with multiple myeloma. Cancer Lett. 2007;246:92-99.

71. Zheng C, Huang D, Liu L, et al. Interleukin-10 gene promoter polymorphisms in multiple myeloma. Int J Cancer. 2001;95:184-188.

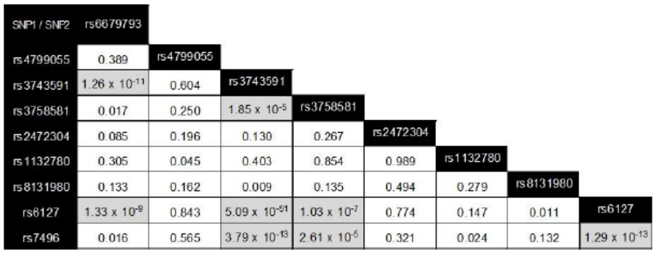
72. Zheng C, Huang DR, Bergenbrant S, et al. Interleukin 6, tumour necrosis factor alpha, interleukin 1beta and interleukin 1 receptor antagonist promoter or coding gene polymorphisms in multiple myeloma. Br J Haematol. 2000;109:39-45.

Table 1: Associated myeloma risk variants reaching genomic significance, following meta-analysis for random-effects in UK, US and Dutch populations.



SNP	GENE	CHR	Meta OR(R) - (95%ci)	Meta P(R)	Dutch OR - (95%ci)	Dutch - *P*	UK OR - (95%ci)	UK - *P*	US OR - (95%ci)	US - *P*
rs6679793	FCRL5	1	1.54 (1.38-1.72)	4.20E-14	1.53 (1.29-1.82)	1.46E-06	1.60 (1.29-1.98)	1.62E-05	1.50 (1.23-1.84)	8.23E-05
rs4799055	NFATC1	18	0.77 (0.71-0.82)	6.93E-13	0.79 (0.70-0.88)	1.98E-05	0.75 (0.66-0.86)	1.62E-05	0.76 (0.66-0.87)	1.02E-04
rs3743591	TNFRSF17	16	2.92 (2.14-3.99)	1.72E-11	3.00 (2.01-4.49)	7.36E-08	3.11 (1.70-5.72)	2.46E-04	2.24 (0.94-5.36)	6.99E-02
rs3758581	CYP2C19	10	2.02 (1.60-2.55)	3.08E-09	2.12 (1.53-2.94)	6.29E-06	1.56 (0.96-2.53)	6.99E-02	2.32 (1.47-3.67)	3.17E-04
rs2472304	CYP1A2	15	1.26 (1.16-1.37)	3.40E-08	1.30 (1.15-1.48)	3.83E-05	1.28 (1.10-1.50)	1.81E-03	1.19 (1.02-1.39)	3.05E-02
rs1132780	CAMKK2	12	1.55 (1.32-1.83)	8.58E-08	1.70 (1.36-2.12)	3.38E-06	1.29 (0.93-1.79)	1.24E-01	1.55 (1.11-2.16)	1.05E-02
rs8131980	IFNGR2	21	1.24 (1.14-1.35)	4.23E-07	1.26 (1.11-1.43)	4.77E-04	1.21 (1.04-1.42)	1.72E-02	1.25 (1.07-1.47)	5.12E-03
rs6127	SELP	1	1.18 (1.10-1.27)	3.09E-06	1.19 (1.07-1.33)	1.39E-03	1.10 (0.97-1.25)	1.25E-01	1.25 (1.10-1.43)	9.46E-04
rs7496	GSTA4	6	1.59 (1.30-1.95)	6.33E-06	1.43 (1.04-1.97)	3.02E-02	1.92 (1.32-2.78)	6.36E-04	1.53 (1.07-2.20)	1.97E-02

Table 2: *P*-values generated by SNP x SNP interaction analysis within top SNP associations with MM risk following meta-analysis. (Cases-only)



SNP1 / SNP2	rs6679793	rs4799055	rs3743591	rs3758581	rs2472304	rs1132780	rs8131980	rs6127
rs4799055	0.389							
rs3743591	1.26×10^{-11}	0.604						
rs3758581	0.017	0.250	1.85×10^{-5}					
rs2472304	0.085	0.196	0.130	0.267				
rs1132780	0.305	0.045	0.403	0.854	0.989			
rs8131980	0.133	0.162	0.009	0.135	0.494	0.279		
rs6127	1.33×10^{-9}	0.843	5.09×10^{-51}	1.03×10^{-7}	0.774	0.147	0.011	
rs7496	0.016	0.565	3.79×10^{-13}	2.61×10^{-5}	0.321	0.024	0.132	1.29×10^{-13}

Figure 1. Analysis scheme: Quality controls measures were applied across each of the three cases and three control datasets; cases and controls were combined; odds ratios were calculated for single SNP by Logistic regression; Meta-analysis of overlapping SNPs.

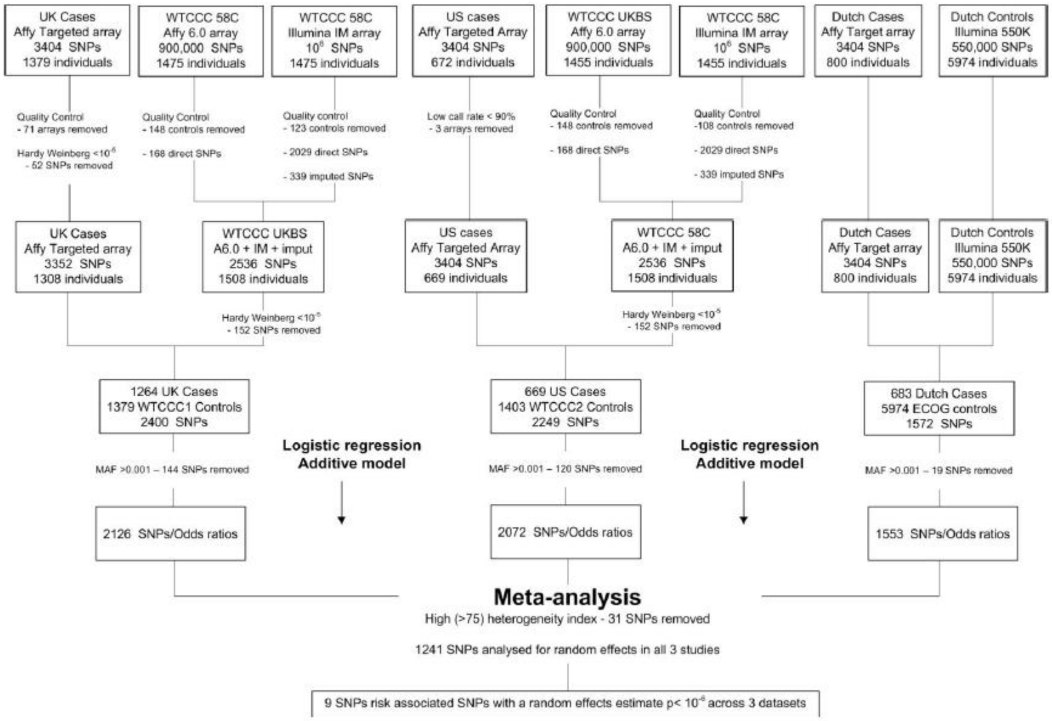
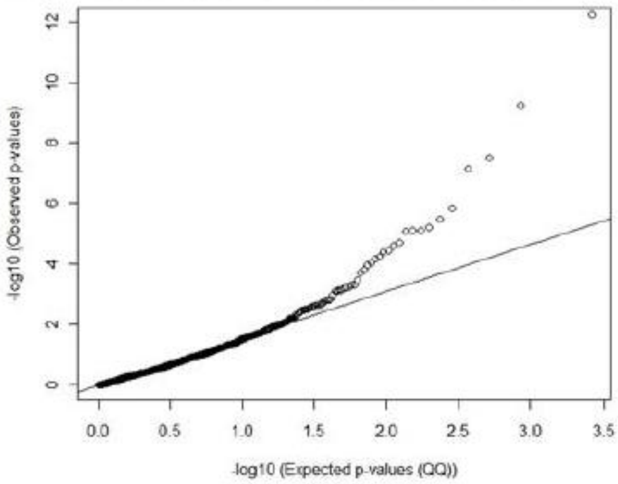

UK Cases
Affy Targeted array
3404 SNPs
1379 individuals
WTCCC 58C
Affy 6.0 array
900,000 SNPs
1475 individuals
WTCCC 58C
Illumina IM array
10^6 SNPs
1475 individuals
US cases
Affy Targeted Array
3404 SNPs
672 individuals
WTCCC UKBS
Affy 6.0 array
900,000 SNPs
1455 individuals
WTCCC 58C
Illumina IM array
10^6 SNPs
1455 individuals
Dutch Cases
Affy Target array
3404 SNPs
800 individuals
Dutch Controls
Illumina 550K
550,000 SNPs
5974 individuals
Quality Control
- 71 arrays removed
Hardy Weinberg $<10^{-6}$
- 52 SNPs removed
Quality Control
- 148 controls removed
- 168 direct SNPs
Quality control
- 123 controls removed
- 2029 direct SNPs
- 339 imputed SNPs
Low call rate < 90%
- 3 arrays removed
Quality Control
- 148 controls removed
- 168 direct SNPs
Quality Control
-108 controls removed
- 2029 direct SNPs
- 339 imputed SNPs
UK Cases
Affy Targeted array
3352 SNPs
1308 individuals
WTCCC UKBS
A6.0 + IM + imput
2536 SNPs
1508 individuals
US cases
Affy Targeted Array
3404 SNPs
669 individuals
WTCCC 58C
A6.0 + IM + imput
2536 SNPs
1508 individuals
Dutch Cases
Affy Target array
3404 SNPs
800 individuals
Dutch Controls
Illumina 550K
550,000 SNPs
5974 individuals
Hardy Weinberg $<10^{-5}$
- 152 SNPs removed
Hardy Weinberg $<10^{-5}$
- 152 SNPs removed
1264 UK Cases
1379 WTCCC1 Controls
2400 SNPs
669 US Cases
1403 WTCCC2 Controls
2249 SNPs
683 Dutch Cases
5974 ECOG controls
1572 SNPs
Logistic regression
Additive model
Logistic regression
Additive model
MAF >0.001 – 144 SNPs removed
MAF >0.001 – 120 SNPs removed
MAF >0.001 – 19 SNPs removed
2126 SNPs/Odds ratios
2072 SNPs/Odds ratios
1553 SNPs/Odds ratios
Meta-analysis
High (>75) heterogeneity index - 31 SNPs removed
1241 SNPs analysed for random effects in all 3 studies
9 SNPs risk associated SNPs with a random effects estimate p< 10^{-6} across 3 datasets

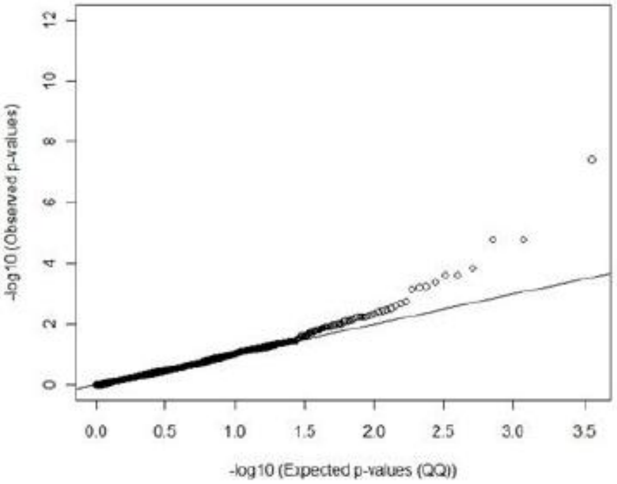
Figure 2. Quantile-quantile plots of test statistics for MM risk across the Dutch, UK and US datasets.

Quantile-quantile plot of the test statistics - Dutch - ERGO



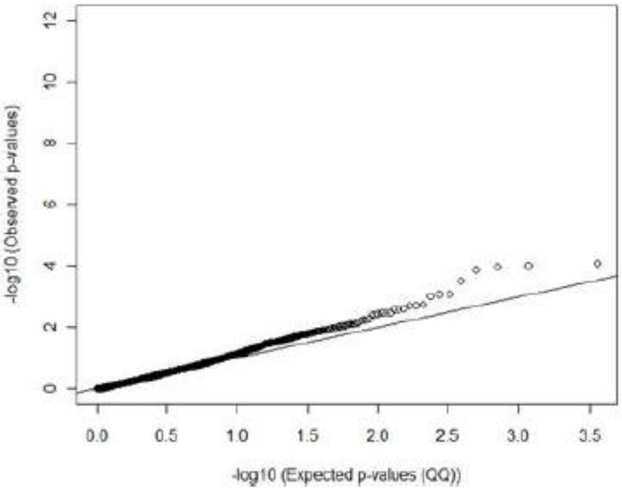
Genome inflation =1.55

Quantile-quantile plot of the test statistics - UK - WTCCC BC



Genome inflation = 1

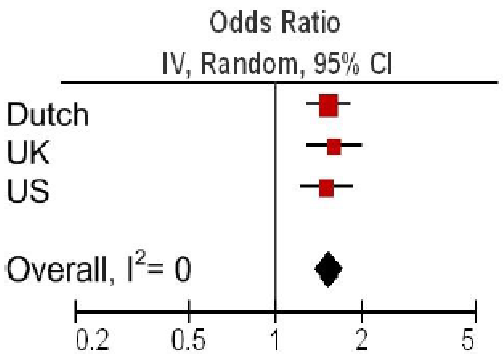
Quantile-quantile plot of the test statistics - US - WTCCC 58C



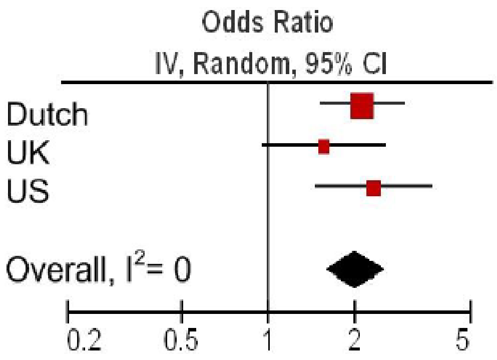
Genome inflation = 1

Figure 3. Forest plots of an inverse variance (IV) meta-analysis for random-effects.

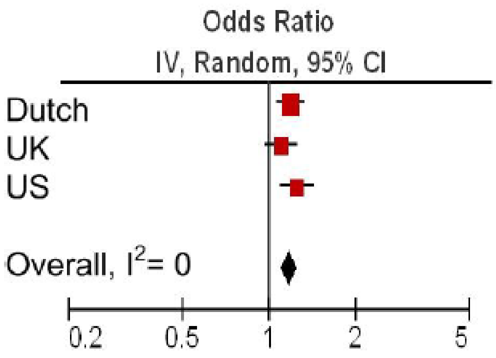
FCRL5 - coding-nonsynonymous - rs6679793



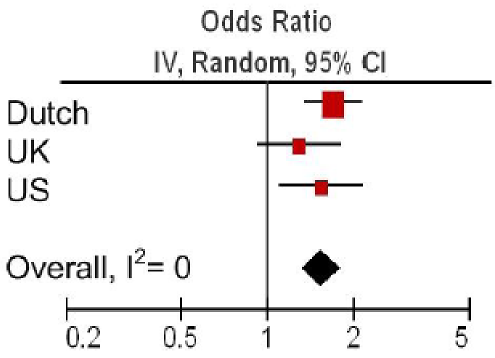
CYP2C19- coding-nonsynonymous- rs3758581



SELP- coding-nonsynonymous- rs6127



CAMKK2 - coding-nonsynonymous - rs1132780



APPENDIX III

A COMPUTATIONALLY EFFICIENT AND STATISTICALLY POWERFUL FRAMEWORK FOR SEARCHING HIGH-ORDER EPISTASIS WITH SYSTEMATIC PRUNING AND GENE-SET CONSTRAINTS

Gang Fang[1], Majda Haznadar[2], Wen Wang[1], Michael Steinbach[1], Brian Van Ness[2], Vipin Kumar[1]

[1]Department of Computer Science, [2]Department of Genetic, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA.

I helped guide the design of novel data mining algorithms represented here. I regularly met with our collaborators and provided them with biological interpretation. I conducted the data analysis following the production of the results by the algorithm.

Motivation: Genotype-phenotype association study from both local and genome-wide single-nucleotide polymorphism data have revolutionized our ability to identify genetic variants that are associated with complex diseases. Recently, discovering epistasis (gene-gene interactions with strong disease association but weak marginal effect) is receiving more attention over single-locus studies since many diseases, such as cancer and diabetes, are believed to be complex traits. Several efficient approaches have been proposed recently for searching two-locus epistasis from genome-wide datasets. However, in searching for higher order epistasis, a much more computationally expensive method, existing methods adopt either brute-force search, which can only handle a smaller number of SNPs, or greedy search that may miss interesting epistasis with weak marginal effect. In addition, due to large numbers of hypotheses tested, existing approaches searching for epistasis also lacks statistical power.

Results: In this paper, we propose an efficient framework for searching high-order epistasis from SNP datasets with thousands of SNPs. The framework is extended from the association analysis techniques developed in the data mining community, which leverage the anti-monotonicity of the objective functions to systematically prune the exponential search space of high-order epistasis while guaranteeing the completeness with respect to the problem formulation. This framework can also take gene set (or pathway) constraints to reduce false positive rates while maintaining the anti-monotonicity of the object function. Both synthetic and real datasets are used to demonstrate the efficiency of the proposed framework for searching high-order epistasis and the enhanced statistical power by using gene set constraints. (Availability: Source code and datasets: http://vk.cs.umn.edu/PCD/).

Introduction:

Genotype-phenotype association studies from both local and genome-wide publically available data, have revolutionized our ability to identify causal genetic variants that are associated with complex diseases. Many loci have been linked to diseases such as Crohn's disease [6], type 1 diabetes [6, 27], type 2 diabetes [29] etc. Most existing approaches are based on the univariate tests, and thus may ignore the interactions among SNPs, i.e. epistasis. Epistasis is defined as gene-gene interactions that may be used to represent combinations of loci that are strongly associated with a phenotype, even though they have weak or close-to-zero marginal effects [11]. Indeed, many complex diseases such as cancer, diabetes and obesity are believed to be driven by multiple SNP interactions, and there have already been several such discoveries [24]. The clear challenge of searching for epistasis from a large number of SNPs is the exponentially increasing number of combinations. This is even more challenging given that a large number of permutation tests [36] (e.g. frequently 1000 times) are needed to correct for multiple hypothesis testing.

As discussed in [36], given a GWAS dataset with millions of SNPs, it is computationally challenging even if we only intend to compute all the pair-wise combinations of SNPs. Several efficient approaches have been proposed recently for searching two-locus epistasis from genome-wide datasets [38, 39, 37], by reusing the repeated computations in the permutation tests, which reduce the overall run-time by several orders of times. With the substantially improved efficiency, the complete search of pair-wise epistasis can handle 120,000 SNPs [35], which is still far from the entire set of SNPs generally available in existing GWAS datasets, as high as 500,000 [4] or even 900,000 [23]. To further address the computational challenges, several approaches have proposed to use existing biological knowledge such as biological gene sets and protein interaction networks to reduce the search space of epistasis [22, 13]. From a different perspective, recent developments on GPUs have also been utilized to enable the complete search of 2-locus epistasis discoveries from complete GWAS dataset [18].

As discussed above, a variety of efficient computational approaches have been proposed for two-locus epistasis discoveries. However, to search for higher order epistasis, which is much more computationally expensive, existing methods adopt either brute-force or greedy search, both of which have advantages and limitations. On one hand, the brute-force search-based approaches such as MDR [28] can guarantee the completeness of the search, which is important for detecting epistasis with weak marginal effects [21]. However, brute-force approaches can handle relatively small numbers of SNPs [28, 17]. On the other hand, greedy approaches such as [34, 14, 31] are computationally efficient but may miss interesting epistasis with weak marginal effects [36]. In addition to the computational challenges, existing approaches for epistasis discoveries also lacks statistical power; due to large numbers of hypothesis tested and limited sample sizes, some combinations of SNPs can be associated with a disease phenotype by random chance. As will be shown in our experiments, in some datasets high-order epistasis can be discovered even when we randomize the phenotypic class labels. In this paper, we aim to address both the computational and statistical challenges of searching high-order epistasis.

1. To improve the computational efficiency, we propose a framework for searching high-order epistasis from SNP datasets for focused studies (thousands of SNPs). The framework is extended from the association analysis techniques [3, 16] developed in the data mining community, which leverage the anti-monotonicity of the objective functions to systematically prune the exponential search space of high-order epistasis while guaranteeing the completeness with respect to the problem formulation. A unique advantage of Apriori over brute-force search is that it can avoid exploring the whole search space (all sets of SNP genotypes) by pruning a large number of candidates that are guaranteed to disqualify the threshold based on the anti-monotonicity of the object function.

2. To improve the statistical power of high-order epistasis search, the proposed framework can take gene sets (or pathways) as constraints for high-order epistasis search. Specifically, the gene set constraints only consider epistasis composing SNPs which are on the genes that span k gene sets (e.g. pathways). Such a constraint formulation naturally fits in the association analysis framework, because it maintains the anti-monotonicity of the object function, and can reduce the false positive rate while further improving the computational efficiency. Carefully designed experiments with synthetic data demonstrate that the proposed framework is more scalable for searching high-order epistasis from SNP datasets for focused studies (thousands of SNPs), and that the gene set constraints can enable the discovery of significant high order epistasis that would otherwise be missed. We also applied the proposed framework on four real SNP datasets for focused studies (thousands of SNPs), in which we discovered many significant epistasis.

Methods:

In this section, we describe the proposed framework for efficient search of high-order epistasis and how it can naturally incorporate gene set constraints. Firstly, we will briefly introduce the traditional association analysis and the anti-monotonicity based pruning of combinatorial search space that is originated for the analysis of marker-basket data. Secondly, we describe one specific branch of association analysis named discriminative pattern mining, which is motivated by discoveries of combinations of entities (e.g. combinations of genotypes) that have substantially different occurrence between different classes (e.g. disease vs. control populations). Then, we formulate high-order epistasis search as a discriminative pattern mining problem via a lossless transformation of the categorical genotype data, and relate the formulation to MDR. Finally, we show how the framework can naturally incorporate gene set constraints while maintaining the anti-monotonicity of the object function.

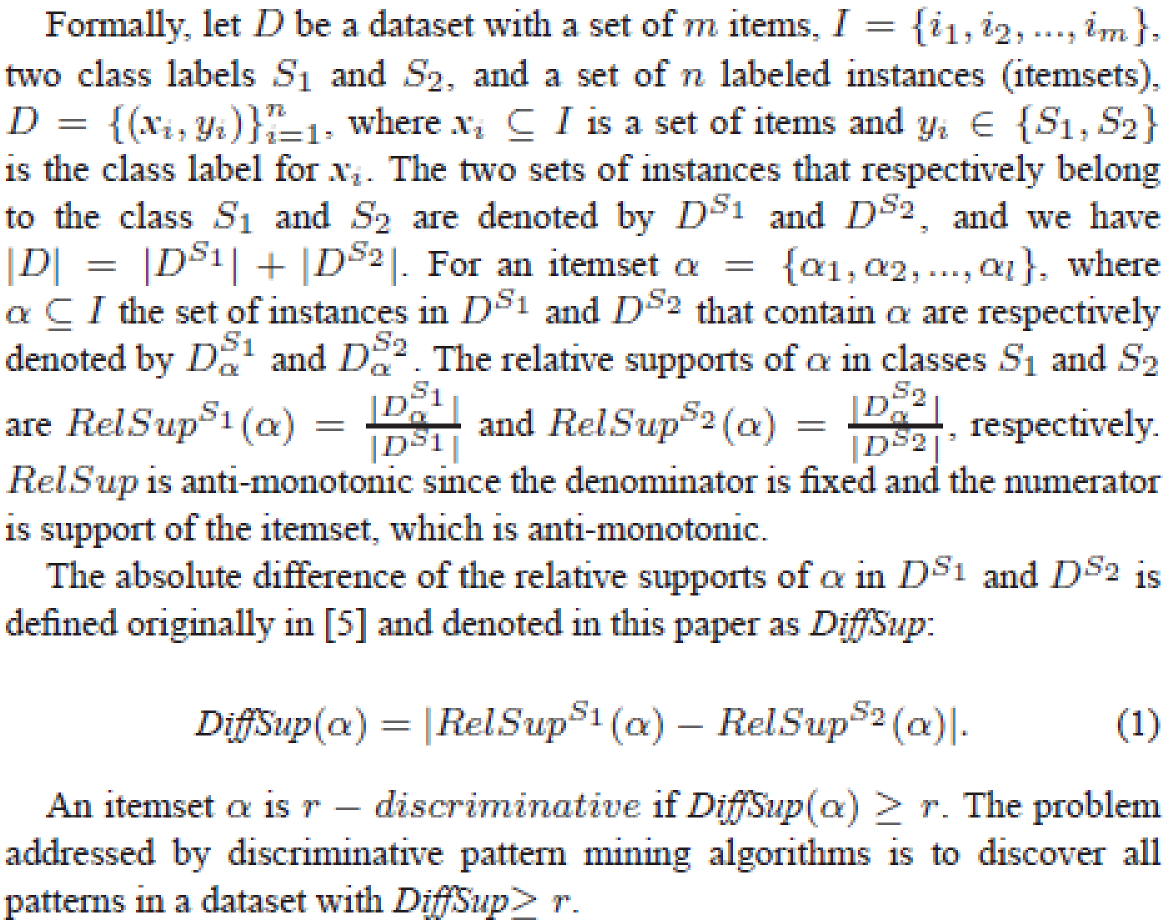
A brief introduction about association analysis

In this section, we briefly describe the association analysis techniques and its key concept, i.e. antimonotonicity based pruning of combinatorial search space. The concept of frequent itemset was first introduced for mining market basket databases [2] and recently surveyed in [19]. Let I = {i1, i2, . . . , in} be a set of all items, which in market basket data can be coke, bread, milk etc. A transaction t is a subset of items, i.e. t I, and a transaction dataset D = ˘t1, t2, . . . , t|D|¯ is a collection of transactions. A transaction dataset D can be represented as a |D|×n binary matrix in which the (i, j) entry is 1 if the jth item appears in the ith transaction. For example, Figure 1 displays a transaction dataset that has 20 transactions and 15 items, in which a blue entry indicates 1, and white entry indicates 0. A k − itemset, which consists of k items from I, is frequent if it occurs in a transaction database D no lower than |D| times, where is a userspecified minimum support threshold (called minsup), and |D| is the total number of transactions in D. In the matrix representation, for a frequent k − itemset, there are no less than |D| rows on which all the k columns have 1s. For instance, in Figure 1, itemsets P2 = {i5, i6, i7} and P4 = {i12, i13, i14} are frequent given a minsup of 10. The straightforward way to find all the frequent itemsets can be a bruteforce search, in which all the combinations of items from size − 1 to size − n are checked and those that satisfy the minsup are selected as frequent itemsets. However, since there are usually large numbers of items in a real transaction database (e.g. Walmart database), the number of total combinations of items can be enormous. Therefore, it is necessary to develop scalable methods for mining frequent itemsets in a large transaction database. Agrawal and Srikant [3] observed an interesting anti-monotonic property of support, called Apriori: A k − itemset is frequent only if all of its sub-itemsets are frequent. Based on this observation, whenever a k − itemset is found to be infrequent, all its supersets can be pruned in the search space without missing any frequent itemsets. Through such antimonotonicity based pruning, the Apriori framework can search for frequent itemset much faster then brute-force search. Indeed, beyond the specific frequent itemset mining problem, the Apriori framework can essentially serve as an exhaustive yet efficient (compared to brute-force) framework for any type of combinatorial search, for which the objective function has the anti-monotonic property. Next, we describe an extended branch of association analysis named discriminative pattern mining, where the transactions in a dataset are associated with class labels, e.g. disease vs. non-disease.

A specific branch in association analysis:

Discriminative pattern mining

For data sets with class labels, association patterns [3, 33] that occur with disproportionate frequency in some classes versus others can be of considerable value in many applications. Such applications include census data analysis that aims at identifying differences among demographic groups [12, 5] and biomarker discovery, which searches for groups of genes or related entities, that are associated with diseases [7, 30, 1]. We will refer to these patterns as discriminative patterns 1 in this manuscript, although they have also been investigated under other names [26], such as emerging patterns [12] and contrast sets [5]. In this manuscript, we focus on 2-class problems, which can be generalized to multi-class problems as described in [5]. To introduce some key ideas relevant to discriminative patterns and make the following discussion easier to follow, consider Figure 2, which displays a sample dataset 2 containing 15 items (columns) and 2 classes, each with 10 instances (rows). In the figure, four patterns (sets of binary variables) can be observed: P1 = {i1, i2, i3}, P2 = {i5, i6, i7}, P3 = {i9, i10} and P4 = {i12, i13, i14}. P1 and P4 are interesting discriminative patterns that occur with different frequencies in the two classes, for which DiffSup is 0.6 and 0.7, respectively. In contrast, P2 and P3 are uninteresting patterns with a relatively uniform occurrence across the classes, both having a DiffSup of 0. Furthermore, P4 is a discriminative pattern where individual items are also highly discriminative, while those of P1 are not. Based on the support in the entire dataset, P2 is a frequent non-discriminative pattern, while P3 is a relatively infrequent non-discriminative pattern.



Formally, let D be a dataset with a set of m items, $I = \{i_1, i_2, ..., i_m\}$, two class labels S_1 and S_2, and a set of n labeled instances (itemsets), $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \subseteq I$ is a set of items and $y_i \in \{S_1, S_2\}$ is the class label for x_i. The two sets of instances that respectively belong to the class S_1 and S_2 are denoted by D^{S_1} and D^{S_2}, and we have $|D| = |D^{S_1}| + |D^{S_2}|$. For an itemset $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$, where $\alpha \subseteq I$ the set of instances in D^{S_1} and D^{S_2} that contain α are respectively denoted by $D_{\alpha}^{S_1}$ and $D_{\alpha}^{S_2}$. The relative supports of α in classes S_1 and S_2 are $RelSup^{S_1}(\alpha) = \frac{|D_{\alpha}^{S_1}|}{|D^{S_1}|}$ and $RelSup^{S_2}(\alpha) = \frac{|D_{\alpha}^{S_2}|}{|D^{S_2}|}$, respectively. $RelSup$ is anti-monotonic since the denominator is fixed and the numerator is support of the itemset, which is anti-monotonic.

The absolute difference of the relative supports of α in D^{S_1} and D^{S_2} is defined originally in [5] and denoted in this paper as *DiffSup*:

$$\mathit{DiffSup}(\alpha) = |RelSup^{S_1}(\alpha) - RelSup^{S_2}(\alpha)|. \qquad (1)$$

An itemset α is $r - discriminative$ if $\mathit{DiffSup}(\alpha) \geq r$. The problem addressed by discriminative pattern mining algorithms is to discover all patterns in a dataset with $\mathit{DiffSup} \geq r$.

Without any loss of generality, we only consider discriminative patterns for the binary-class problems. A possible strategy for improving the performance of the two-step approaches is to directly utilize the support of a pattern in the two classes for pruning some non-discriminative patterns in the pattern mining stage. Indeed, several approaches have been proposed [5, 8], where the anti-monotonic upper bounds of discriminative measures, such as DiffSup, are used for pruning some non-discriminative patterns in an Apriori-like framework [3].

Nominal categorical data can be converted to binary data without any loss of information, while ordinal categorical data and continuous data can be binarized, although with some loss of magnitude and order information. This strategy, like the two-step approaches, also guarantees to find the complete set of discriminative patterns with respect to a threshold, although in a more efficient manner. Next, we will show how the search of high-order epistasis can be formulated as a discriminative pattern mining problem.

High-order epistasis search formulated as discriminative pattern mining

As mentioned previously, a transaction dataset can be converted into a binary matrix with rows being samples and columns being items. Traditional association analysis approaches including discriminative pattern mining can be performed and are designed to search for patterns from such binary matrixes.

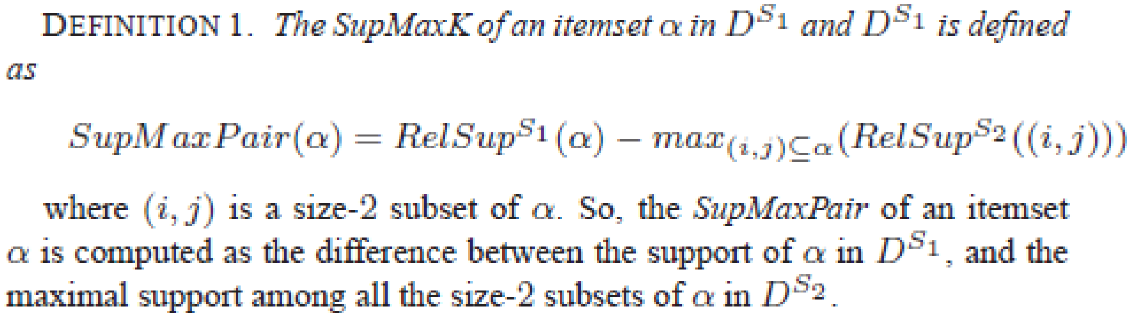
However, because the genotype datasets are categorical (i.e. a SNP has three possible genotypes: homozygous minor (mm), heterozygous (Mm) and homozygous major (MM)), existing discriminative pattern mining algorithms can not be directly applied on SNP datasets. In order to formulate high-order epistasis search into a discriminative pattern mining problem, we need to first convert a categorical genotype data into a binary matrix data. This can be done in a lossless way by creating, for each SNP, three columns corresponding to the three genotypes. In the converted matrix, an entry (e.g. SNPk genotype Mm) is 1 if a sample has a Mm genotype for SNPk. Figure 3 illustrates an example in which a dataset with three SNPs is converted from categorical format to a binary matrix

[1] The terms ”pattern” and ”itemset” are used interchangeably in this paper. [2] The discussion in this paper assumes that the data is binary.

After the conversion, the SNP dataset can be input into a discriminative pattern mining algorithm, which can discover combinations of SNP genotypes that occur with disproportionate frequency between cases and controls, e.g. such as P1 and P4 shown in Figure 2. However, it is worth noting that, there are two challenges in applying traditional association analysis techniques on the converted matrix format. Firstly, the conversion results in a binary matrix with guaranteed 33.33% density 3, which is a very high number for traditional association analysis. Secondly, the conversion will result in a three times of variables 4, i.e. a high dimensional binary matrix. Discriminative pattern mining from dense and high dimensional dataset is computationally challenging because existing techniques can not effectively prune frequent non-discriminative patterns [16] such as P2 in Figure 2, which is expected to be common in dense and high-dimensional datasets.

[3] This is because for each sample and each SNP, one of the three columns has 1. [4] This is because, each SNP in the categorical format is converted into three binary columns in the binary format.

This is especially challenging if the disease-associated epistasis covers only a small fraction of samples, which further adds more computational expense to current approaches for discriminative pattern mining [16]. Recently, Fang et al. [16] proposed an antimonotonic measure named SupMaxPair for efficiently mining low support discriminative patterns from dense and high-dimensional gene expression data [16, 15]. We build up on this approach for discovering high-order epistasis from SNP datasets in this study, with the data represented here described above.



DEFINITION 1. *The SupMaxK of an itemset α in D^{S_1} and D^{S_1} is defined as*

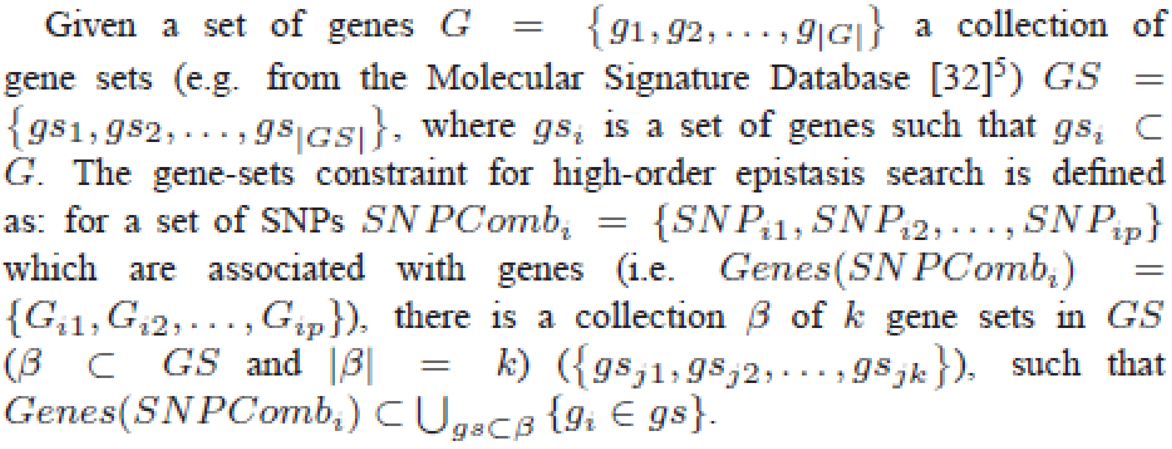
$$SupMaxPair(\alpha) = RelSup^{S_1}(\alpha) - max_{(i,j)\subseteq\alpha}(RelSup^{S_2}((i,j)))$$

where (i,j) is a size-2 subset of α. So, the *SupMaxPair* of an itemset α is computed as the difference between the support of α in D^{S_1}, and the maximal support among all the size-2 subsets of α in D^{S_2}.

SupMaxPair has the anti-monotonicity [16], and thus can be used in the Apriori framework for the systematic yet efficient discovery of discriminative patterns. It has been demonstrated, on gene expression datasets, that SupMaxPair is effective for pruning frequent nondiscriminative patterns, and scalable to discover discriminative patterns (in this case high-order epistasis) from dense and high dimensional datasets. Next, we further describe how the framework can naturally incorporate gene set constraints while maintaining the anti-monotonicity of the object function.

Incorporating gene set constraints in the search of high-order epistasis

Recently, several approaches have proposed to use existing biological knowledge such as biological gene sets and protein interaction networks to reduce the search space of low-order epistasis (mostly size − 2 and some size − 3) [22, 13] for better scalability on genome-wide datasets. Such an incorporation of gene set constraints can, on one hand, further improve the computational efficiency, and on the other hand, control false positive rates since the genes that are physically or functionally related to each other are more likely to host epistasis. In this section, we show that similar constraints can be naturally incorporated into the Apriori framework for discriminative mining.



Given a set of genes $G = \{g_1, g_2, \ldots, g_{|G|}\}$ a collection of gene sets (e.g. from the Molecular Signature Database [32][5]) $GS = \{gs_1, gs_2, \ldots, gs_{|GS|}\}$, where gs_i is a set of genes such that $gs_i \subset G$. The gene-sets constraint for high-order epistasis search is defined as: for a set of SNPs $SNPComb_i = \{SNP_{i1}, SNP_{i2}, \ldots, SNP_{ip}\}$ which are associated with genes (i.e. $Genes(SNPComb_i) = \{G_{i1}, G_{i2}, \ldots, G_{ip}\}$), there is a collection β of k gene sets in GS ($\beta \subset GS$ and $|\beta| = k$) ($\{gs_{j1}, gs_{j2}, \ldots, gs_{jk}\}$), such that $Genes(SNPComb_i) \subset \bigcup_{gs \subset \beta} \{g_i \in gs\}$.

Conceptually, a combination of SNPs is considered a candidate for epistasis if and only if these SNPs are on the genes that only span no more than k gene sets. Such gene set constraints also have the anti-monotonic property, i.e. if a size − k itemset spans no more than k gene sets, all of its subsets must also span no more than k gene sets. From a different perspective, whenever a size − k SNP combination is found to span more than k gene sets, all its supersets can be pruned in the Apriori framework. Since two anti-monotonic measures can be used together in the Apriori framework and still maintain the anti-monotonicity, SupMaxPair and the gene set constraints will be used in the Apriori framework together for searching high-order epistasis. Such an incorporation of gene set constraints can, on one hand effectively control false positive rates since the genes that are physically or functionally related to each other are more likely to host epistasis.

[5] http://www.broadinstitute.org/gsea/msigdb/index.jsp

On the other hand, this method can also further improve the computational efficiency of the overall framework, because the use of gene-set constraints reduces the combinatorial search space for SNP-combination discoveries. Such additional computational benefit allows the use of lower SupMaxPair thresholds for discriminative pattern mining which are otherwise computationally challenging due to the high density and dimensionality of SNP datasets [16]. Consequently, the use of lower SupMaxPair thresholds may enable the discovery of low-support but statistically significant SNP-combinations, which are especially interesting in biomarker discovery, because highorder epistasis may only be observable over a small fraction of samples in a given dataset, due to a heterogeneity of complex diseases. It is worth noting that adding such gene set constraints may also miss true epistasis that are on some unknown gene sets, which is a limitation of this approach. However, as will be shown in the experiments, although some highly differential SNP combinations are indeed missed due to the gene set constraints, the overall statistical power is gained. Specifically, there are many statistically significant epistasis discoveries with gene set constraints which are otherwise considered insignificant. This indicates that adding gene set constraints does provide an effective control of false positive rates.

Experiments and Results:

In this section, we use both synthetic and real SNP datasets to study the effectiveness of the proposed framework for discovering statistically significant discriminative SNP combinations. We first describe the datasets used in the experiments, and then present the experimental design and finally discuss the experimental results.

Datasets

We used one synthetic and four real SNP datasets in our experiments.We first describe how the synthetic dataset is simulated and then present the details of the real SNP datasets.

Synthetic Datasets

We first use Hap-Sample simulator (6) to simulate genotype data. Specifically, we used the 3404 SNPs used in a recent study on multiple myeloma [25] as input, out of which 2172 SNPs are included in the Hap-Sample. The synthetic dataset contains 70 cases and 70 controls. Note that, this genotype dataset by itself does not contain disease-associated loci. Therefore, we further embed a high-order epistasis of size 4 (denoted as EP (embedded pattern)), as a proof of concept. Figure 4 provides a visualization of the pattern. This combination of four SNP genotypes illustrates a high order epistasis, in which the four SNP have a DiffSup of 20/70 = 0.29 (chi-square statistic 23.33), while the best DiffSup and corresponding chi-square statistic for size-3, size-2 and size- 1 patterns are 14/70 = 0.20 (9.25), 8/70 = 0.11 (2.59) and 2/70 = 0.03 (0.14), respectively. Given this dataset, our goal is to discover the embedded size-4 pattern (rather than its subsets) and access its statistical significance. In addition to the dataset, we also need to generate a collection of gene sets that can be used to demonstrate the effectiveness of constraints on improving the statistical power of the proposed framework. Specifically, the pathways were constructed in the following steps:

[6] http://www.hapsample.org

1. We created 100 randomized class labels (permutation) from the original 70 cases and 70 controls [32, 35]. For each permutation, we used our algorithm to identify a list of discriminative patterns that meet a *SupMaxPair* threshold (0.1). We collect all the discriminative patterns in the 100 permutations together and treat each pattern as a gene-set candidate.
2. We compute chi-square value for each pattern and separate the gene-set candidates into two groups by their Chi-square levels: those with chi-square values equal to or greater than θ are called group I, while those with chi-square values lower than θ are called group II.
3. We select the gene sets $\widehat{GS}$ in group II that have at most one gene overlapping with any gene set in group I as the final gene sets. The motivation is to use these gene sets to disqualify all the discriminative patterns with chi-square above θ, discovered in the 100 permutations. To select this subset of gene sets from group II. We use the following approach: Each gene set has multiple genes and each gene can have multiple SNPs, which can be represented by a gene-set by SNP matrix, in which the rows correspond to gene sets and the columns correspond to SNPs, and the value of each element of the matrix could either be 1 or 0, with 1 meaning that the corresponding SNP is on the corresponding pathway and 0 otherwise. If we use the notation M_I and M_{II} for the geneset-SNP matrix for the gene sets in group I and II respectively. Then, $M_{II} \times M_I^T$ is a gene-set (group II) by gene-set (group I) matrix. In this matrix, the rows with all entries no more than 1 correspond to those gene sets in group II that overlap with any gene set in group I by at most one gene.
4. Finally, $\widehat{GS} \cup EP$ is the final collection of gene sets to be used in the experiment.

Real Datasets

In addition to the synthetic dataset, we selected four SNP datasets designed for studying four different types of diseases using the same BOAC chip with 3404 SNPs [25]. The four phenotypes studied are respectively: (i) short vs. long survival of multiple myeloma patients [25] (denoted as Survival), (ii) multiple myeloma patients vs. their normal spouse [20] (denoted as EPI), (iii) lung cancer vs. non-long cancer, all heavy smokers. For these four real datasets, there is a collection of 1892 gene sets from the Molecular Signature Database (MSigDB, C2) [32].

Experiment Design

On the synthetic dataset, we compare the discriminative pattern mining approach under two different setups: (i) one without gene-set constraints and (ii) one with gene-set constraints. Note that we make SupMaxPair the same (0.1) in the comparison. Such a design aims to demonstrate the effectiveness of gene-set constraints for reducing false positive SNP combinations. For each of the real datasets, we also compare the two discriminative pattern mining setups, i.e. without and with gene-set constraints, with the same SupMaxPair threshold (0.5). In addition, these two setups are also compared with a third one, in which a lower SupMaxPair threshold (0.2) is used in the gene-set constrained approach. This additional setup aims to demonstrate the computational benefits from using gene set constraints, in addition to the statistical benefits as described earlier. The criterion for the above comparisons is based on the number of statistically significant (w.r.t. false discovery rate (FDR) threshold 0.25) SNP combinations discovered with a specific setup. We adopt an empirical approach to estimate the FDR for each discovered SNP combination as used in [32]. This approach has zero assumptions on the nature and structure of each real dataset. In more detail, after a set of patterns are discovered with a particular set of parameters, the statistical significance of this set of patterns is estimated via permutation test. Specifically, we first apply the proposed algorithm to the data with original class labels to get a list of discriminative patterns which are called the original patterns.

Then we randomly shuffle the original class labels 100 times to get 100 lists of discriminative patterns which are called the random patterns. Then, false discovery rate (FDR) of each original pattern is calculated using the chi-square values as follows: for an original pattern with Chi-square value of c, if there are m patterns from original class labela that have Chi-square value greater than c and with n patterns discovered in the permutation test that have Chi-square value greater than c, then the false discovery rate is (n/100)/m. The gene-set sizes in the MSigDB C2 range from 1 to 1839. The larger the gene set is, the less it constrains the search of SNP combinations. To study how the sizes of gene sets affect the efficiency of the proposed framework, we use a parameter (maxpathsize) to filter out the gene sets with too many genes. Specifically, for each real dataset, we conducted experiments with maxpathsize = 20, 40, 60, 80 and 100 respectively. Note that, we only vary maxpathsize no more than 100 because we observed that when the gene sets with more than 100 genes are used, there are mostly no statistically significant SNP combinations in the permutation test.

Experimental results: Next, we present the results of the experiments as presented previously.

Synthetic Dataset

From the computational perspective, MDR (7) is still computing size-3 SNP combinations after 5 hours, thus failed to identify the embedded size-4 pattern. In contrast, the run time of proposed framework for searching high-order combinations (of size 2 to 5, and thus including the embedded size-4 pattern) on the synthetic dataset ranges from 15 to 30 minutes depending on whether or not there are gene set constraints and how many gene sets are used as constraints. Figure 5 illustrates the probability distribution of the chi-square statistics of the true patterns (discovered from the original case-control dataset) and the random patterns (discovered in permutation test) on the synthetic data, without the pathway setup (left) and with the pathway setup (right).

[7] http://www.epistasis.org/software.html

A couple of observations can be made:

1. When there are no gene-set constraints, the patterns discovered with randomized class labels have similar chi-square distribution as the patterns discovered from the true class labels. This indicates that all the true patterns discovered from the true class labels have FDR values close to 1. This serves as an illustration that, on dense and high-dimensional SNP datasets, the statistical significance of truly differentiating SNP combinations (e.g the size-4 combination that we embedded with a chi-square statistic of 17) is not observable. As discussed in the introduction, such situations are often encountered on SNP datasets with low sample sizes, but high dimensionality.

2. In contrast, after the gene-set constraints are added (as described earlier), most of the random patterns with high chi-square values are eliminated (because they do not qualify the gene-set constraints) (Figure 5). Meanwhile, the real line indicates that there are many highly discriminative patterns (those with high chisquare values) existing outside the random patterns discovered in the permutation test (the dashed line), and thus having a FDR of 0. These two contrasting observations on the synthetic dataset demonstrate the key concept of the proposed gene-set constrained approach, that is, gene set constraints can improve the statistical power of higher-order SNP-combination discovery by effectively controlling false positives. Note that in order to illustrate the key concept in the experiment on the synthetic dataset, the gene sets are constructed in such a way that it will reduce all the random patterns. Next, we further study the effectiveness of real gene sets for reducing false positives on real datasets.

Real Datasets

After demonstrating the the effectiveness of gene-set constrains on controlling false positive SNP combinations on synthetic data, we further applied our method to the four real datasets as described earlier. We applied gene-set constrains with different maxpathsize (20, 40, 60, 80, 100). Table 2 displays, for each dataset, the detailed information about the parameters we used for each experiment and the number of significant epistasis discovered for each value of maxpathsize. The highest number of statistically significant SNP-combinations for each parameter setting on each dataset is indicated by bold font. Figure 6 provides further detailed chi-square statistics and FDRs for each of the patterns discovered by the three experimental setups (A, B and C) on the four datasets.

Several observations can be made from Table 2 and Figure 6:

1. Gene-set constraints are generally effective for improving the statistical power of SNP combination discovery: On three of the four datasets (Kidney, Survival, Lungcancer), there are no statistically significant SNP combinations discovered in the “without-constraint” setup. In contrast, with gene-set constrains, there are many significant SNP combinations discovered. Specifically, Figure 6 shows that, the patterns discovered with gene-set constraints have smaller chi-square statistics but higher FDRs than those discovered without geneset constrains. This indicates that, although several SNP combinations with high chi-square statistics are not discovered in the “with-constraint” setup (because they do not qualify for the gene-set constraints), the remaining patterns have better statistical significance due to the control of false positives by the gene-set constraints.

2. Gene-set constraints sometime can miss statistically significant SNP combinations: In the EPI dataset, “without-constraint” setup discovered 459 statistically significant SNP combinations, while the with-constraint setup with SupMaxPair = 0.2 only discovered less than 200 significant combinations. Furthermore, the “with-constraint” setup with SupMaxPair = 0.5 does not identify any significant combinations. The possible explanation is that the gene sets in MSigDBC2 do not describe the interaction associated with the phenotype in the EPI dataset (multiple myeloma vs. non-multiple myeloma controls). This observation indicates that the effectiveness of gene-set constraints depends on the gene sets used and vary from phenotype to phenotype. This serves as an example to show the potential limitation of the proposed framework.

3. “With-constraint” setup with the lower SupMaxPair is generally more effective: for three out of the four datasets (Kidney, Survival and EPI), more significant SNP-combinations are

discovered when the lower SupMaxPair (0.2) is used (Table 2). This demonstrates the existence of low-support yet statistically significant SNP combinations and thus the benefits of the ability to search low-support SNP-combinations, enabled by using gene-set constraints.

4. Maxpathsize affects the effieciency of the proposed framework: on the kidney dataset, smaller maxpathsize values (20 and 40) discovered significant SNP combinations while larger values (60, 80 and 100) did not. This seems to indicate that the smaller the gene sets that are used, the more constraints are added and the better the statistical power. However, this observation does not hold for the other three datasets. This is because that the effectiveness of the framework essentially depends on both gene sets and the datasets.

5. Most statistically significant SNP combinations are of size-2: Although the proposed framework can search for high-order SNP combinations up to size-7 on these datasets, most of the discovered statistically significant SNP-combinations are of size-2, except for a small number of size-3 combinations. This may be due to the insufficient number of samples in the four real datasets. However, the highlights of the results are, on one hand, that the proposed discriminative pattern mining approach can search for high-order SNP combinations much more efficiently than existing approaches such as MDR; on the other hand, gene-set constraints can generally improve the statistically power of SNP combination discoveries. Among the discovered statistically significant SNP-combinations, we specifically identify those combinations with weak marginal effect (association with the phenotypes), but strong association with the phenotypes as a combination. For this purpose, we use a measure calculated as the difference between the chi-square statistic of the SNP-combination and the best chi-square statistic from all the subsets of the SNP combination, i.e. chi-square jump. These SNP combinations may indicate phenotype-associated interactions between the genes that the SNPs affect. Detailed biological interpretations of the discovered epistasis are presented in a separate paper [20] focusing on the biological insights obtained from the discovered SNP combinations in multiple myeloma.

Discussion:

In this paper, we targeted the efficient search of statistically significant high-order epistasis. We first discussed the two specific goals of the study: the computational efficiency and the statistical power. To improve the computational efficiency, we proposed a framework for searching high-order epistasis from SNP datasets for focused studies (with thousands of SNPs). The framework was extended from the association analysis framework, which leverages the anti-monotonicity of the objective functions (frequency of a SNP-genotype combination and its differentiation) to systematically prune the exponential search space of high-order epistasis while still guaranteeing the completeness of the search. To improve the statistical power of high-order epistasis search, we proposed to utilize the known biological gene sets as constraints for high-order epistasis search. Specifically, the gene set constraints only consider epistasis composing of SNPs which are on the genes that span k gene sets (e.g. pathways). Such a formulation aims to reduce false positive rates and to further improve the computational efficiency. Both synthetic and real datasets were used to demonstrate that the proposed framework is much more efficient for searching high-order SNP-combinations from SNP case-control datasets, and that the gene set constraints improve the statistical power of high-order epistasis search, and they also discover significant high order epistases that would otherwise be missed.

References:

[1] Mental Health Services Administration (2006). The role of biomarkers in the treatment of alcohol use disorders. *Substance Abuse Treatment Advisory*, 5(4):4206–4223.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases*, pages 487–499, 1994.

[4] J. Barrett and L. Cardon. Evaluating coverage of genome-wide association studies. *Nature genetics*, 38(6):659–662, 2006.

[5] S. Bay and M. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[6] L. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. Kwiatkowski, M. McCarthy, W. Ouwehand, N. Samani, J. Todd, P. Donnelly, et al. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[7] C. Carlson, M. Eberle, L. Kruglyak, and D. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452, 2004.

[8] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *Proceedings of International Conference on Data Engineering*, pages 716–725, 2007.

[9] T. Church, K. Anderson, N. Caporaso, M. Geisser, C. Le, Y. Zhang, A. Benoit, S. Carmella, and S. Hecht. A prospectively measured serum biomarker for a tobacco-specific carcinogen and lung cancer in smokers. *Cancer Epidemiology Biomarkers & Prevention*, 18(1):260, 2009.

[10] T. Church, M. Haznadar, M. Geisser, K. Anderson, N. Caporaso, C. Le, S. Abdullah, S. Hecht, M. Oken, and B. Van Ness. Interaction of CYP1B1, cigarette-smoke carcinogen metabolism, and lung cancer risk. *submitted*, 2010.

[11] H. Cordell. Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463, 2002.

[12] G. Dong and J. Li. Efficient mining of emerging paterns: Discovering trends and differences. In *Proceedings of the 2001 ACM SIGKDD international conference on knowledge discovery in databases*, pages 43–52, 1999.

[13] M. Emily, T. Mailund, J. Hein, L. Schauser, and M. Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 2009.

[14] D. Evans, J. Marchini, A. Morris, and L. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, 2006.

[15] G. Fang, R. Kuang, G. Pandey, M. STEINBACH, C. MYERS, and V. KUMAR. Subspace differential coexpression analysis: problem definition and a general approach. volume 15, pages 145–56, 2010.

[16] G. Fang, G. Pandey, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. Technical Report 011, Department of Computer Science, University of Minnesota, 2009.

[17] C. Greene, D. HIMMELSTEIN, H. NELSON, K. KELSEY, S. WILLIAMS, A. ANDREW, M. KARAGAS, and J. MOORE. Enabling personal genomics with an explicit test of epistasis. In *Pac. Symp. Biocomput*, pages 327–336, 2010.

[18] C. Greene, N. Sinnott-Armstrong, D. Himmelstein, P. Park, J. Moore, and B. Harris. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics*, 26(5):694, 2010.

[19] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15:55–86, 2007.

[20] M. Haznadar, G. Fang, W. Wang, V. Paunic, M. Steinbach, V. Kumar, and B. Van Ness. Single nucleotide polymorphism (SNP) interactions and associations with survival and disease outcome in Multiple Myeloma through a novel data mining method. *submitted*, 2010.

[21] H. He, W. Oetting, M. Brott, and S. Basu. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC Medical Genetics*, 10(1):127, 2009.

[22] C. Herold, M. Steffens, F. Brockschmidt, M. Baur, and T. Becker. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 25(24):3275, 2009.

[23] A. Jankowska, H. Szpurka, R. Tiu, H. Makishima, M. Afable, J. Huh, C. O'Keefe, R. Ganetzky, M. McDevitt, and J. Maciejewski. Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood*, 113(25):6403, 2009.

[24] J. Marchini, P. Donnelly, and L. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature genetics*, 37(4):413–417, 2005.

[25] B. V. Ness, C. Ramos, M. Haznadar, A. Hoering, J. C. Jeff Haessler, S. Jacobus, M. Oken, V. Rajkumar, P. Greipp, B. Barlogie, B. Durie, M. Katz, G. Atluri, G. Fang, R. Gupta, M. Steinbach, V. Kumar, R. Mushlin, D. Johnson, and G. Morgan. Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival. *BMC Medicine,* 6:66, 2008.

[26] P. Novak, N. Lavrac, and G. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and SubgroupMining. *Journal of Machine Learning Research,* 10:377–403, 2009.

[27] N. Pub. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, 39(7):857, 2007.

[28] M. Ritchie and et. al. Multifactordimensionality reduction reveals high-order iteractions among estrogen- metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1):1245–1250, 2001.

[29] R. Saxena, B. Voight, V. Lyssenko, N. Burtt, P. de Bakker, H. Chen, J. Roix, S. Kathiresan, J. Hirschhorn, M. Daly, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336, 2007.

[30] D. Segre, A. DeLuna, G. Church, and R. Kishony. Modular epistasis in yeast metabolism. *Nature Genetics*, 37:77–83, 2004.

[31] J. Storey, J. Akey, L. Kruglyak, et al. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3(8):1380, 2005.

[32] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545– 15550, 2005.

[33] P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. *Addison- Wesley*, 2005.

[34] N. Yosef, Z. Yakhini, A. Tsalenko, V. Kristensen, A. Borresen-Dale, E. Ruppin, and R. Sharan. A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics*, 23(2):91–98, 2007.

[35] X. Zhang, S. Huan, F. Zou, , and W. Wang. Team: Efficient two-locus epistasis tests in human genome-wide association study. In *Annual International Conference on Intelligent Systems for Molecular Biology*, page in press, 2010.

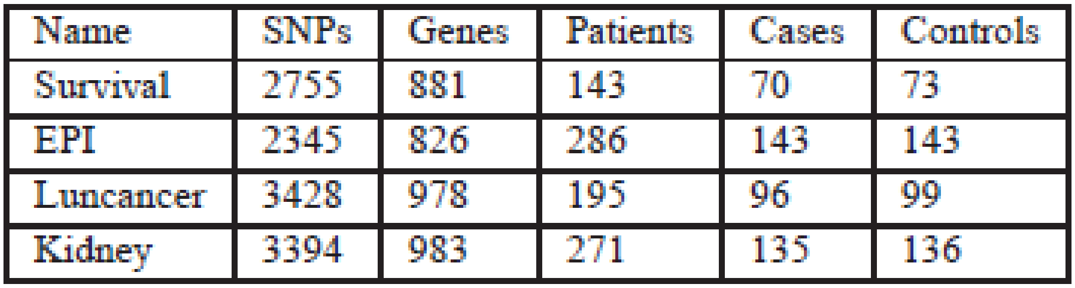
[36] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang. COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study. In *Research in Computational Molecular Biology*, pages 253–269. Springer.

[37] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang. COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study.
In *Research in Computational Molecular Biology*, volume 5541, pages 253–269, 2009.

[38] X. Zhang, F. Zou, and W. Wang. Fastanova: an efficient algorithm for genomewide association study. In *Proceeding of the ACM SIGKDD international conference on knowledge discovery in databases*, pages 109–118, 2008.

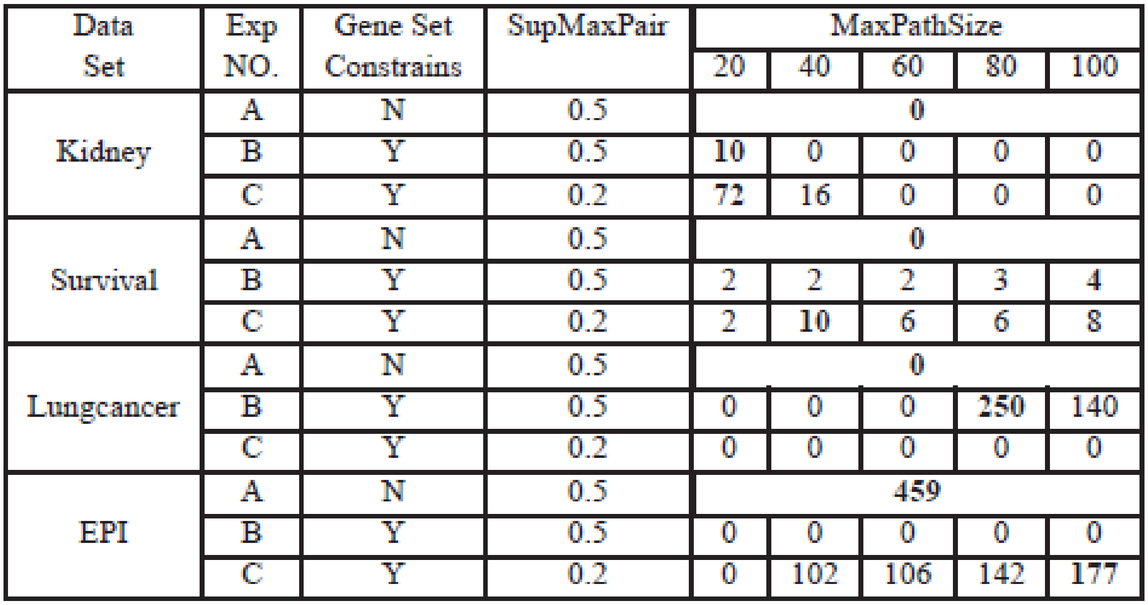
[39] X. ZHANG, F. ZOU, andW.WANG. FASTCHI: AN EFFICIENT ALGORITHM FOR ANALYZING GENE-GENE INTERACTIONS. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 528, 2009.

Table 1. Information about the real datasets.



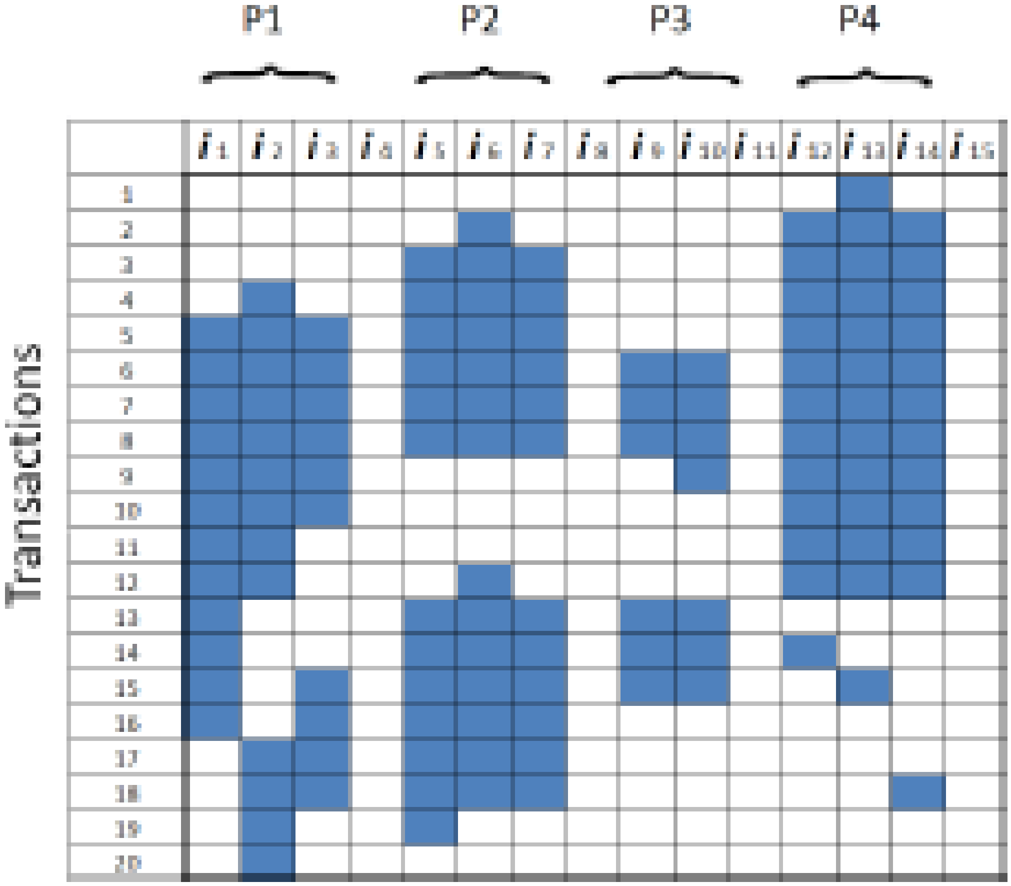
Name	SNPs	Genes	Patients	Cases	Controls
Survival	2755	881	143	70	73
EPI	2345	826	286	143	143
Luncancer	3428	978	195	96	99
Kidney	3394	983	271	135	136

Table 2. Parameters used and number of significant patterns discovered w.r.t. FDR 0.25 for each of the three approaches (A: without constraint, *SupMaxPair* = 0.5; B: with constraints, *SupMaxPair* = 0.5 and C: with constraints, *SupMaxPair* = 0.2) on each of the four real datasets. The highest number of statistically significant SNP-combinations for each parameter setting on each dataset, is indicated by bold font.



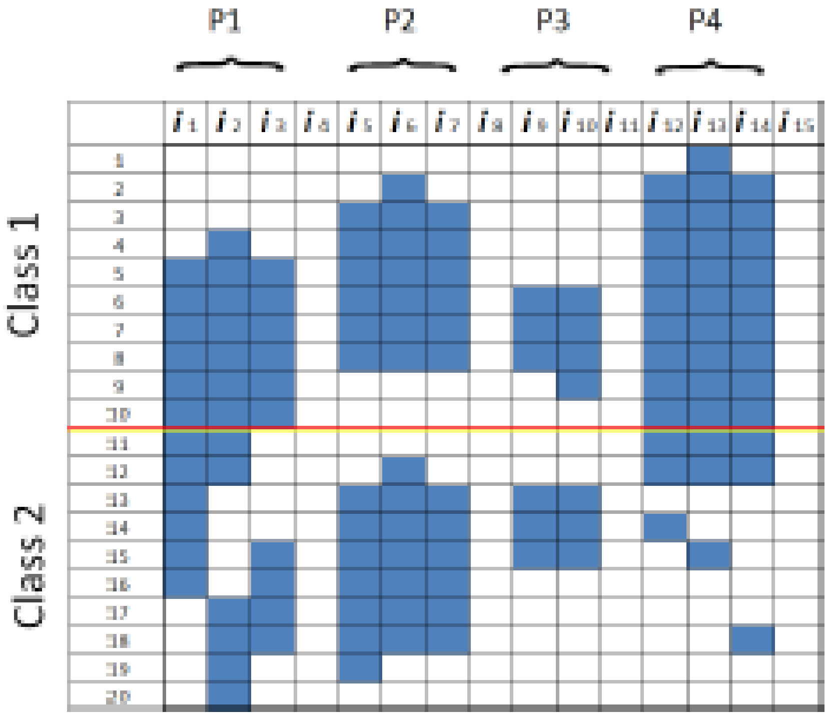
Data Set	Exp NO.	Gene Set Constrains	SupMaxPair	MaxPathSize				
				20	40	60	80	100
Kidney	A	N	0.5	**0**				
	B	Y	0.5	**10**	0	0	0	0
	C	Y	0.2	**72**	16	0	0	0
Survival	A	N	0.5	**0**				
	B	Y	0.5	2	2	2	3	**4**
	C	Y	0.2	2	**10**	6	6	**8**
Lungcancer	A	N	0.5	**0**				
	B	Y	0.5	0	0	0	**250**	140
	C	Y	0.2	0	0	0	0	0
EPI	A	N	0.5	**459**				
	B	Y	0.5	0	0	0	0	0
	C	Y	0.2	0	102	106	142	177

Figure 1. An illustration of the matrix represented for a transaction dataset.


P1
P2
P3
P4
Transactions

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}
1													■		
2						■						■	■	■	
3					■	■	■					■	■	■	
4		■			■	■	■					■	■	■	
5	■	■	■		■	■	■					■	■	■	
6	■	■	■		■	■	■		■	■		■	■	■	
7	■	■	■		■	■	■		■	■		■	■	■	
8	■	■	■		■	■	■		■	■		■	■	■	
9	■	■	■							■		■	■	■	
10	■	■	■									■	■	■	
11	■	■										■	■	■	
12	■	■				■						■	■	■	
13	■				■	■	■		■	■					
14	■				■	■	■		■	■		■			
15	■		■		■	■	■		■	■			■		
16	■		■		■	■	■								
17		■	■		■	■	■								
18		■	■		■	■	■							■	
19		■			■										
20		■													

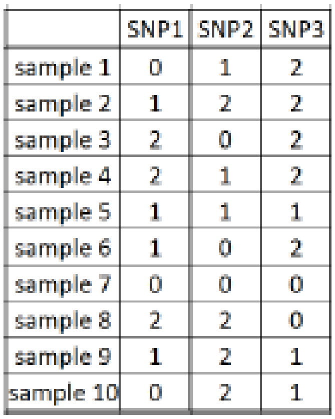
Figure 2. Matrix of sample data sets with interesting discriminative patterns (P1, P4) and uninteresting patterns (P2, P3).



	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}
1													■		
2						■						■	■	■	
3					■	■	■					■	■	■	
4		■			■	■	■					■	■	■	
5	■	■	■		■	■	■					■	■	■	
6	■	■	■		■	■	■		■	■		■	■	■	
7	■	■	■		■	■	■		■	■		■	■	■	
8	■	■	■		■	■	■		■	■		■	■	■	
9	■	■	■							■		■	■	■	
10	■	■	■									■	■	■	
11	■	■										■	■	■	
12	■	■				■						■	■	■	
13	■				■	■	■		■	■					
14	■				■	■	■		■	■		■			
15	■		■		■	■	■		■	■			■		
16	■		■		■	■	■								
17		■	■		■	■	■								
18		■	■		■	■	■							■	
19		■			■										
20		■													

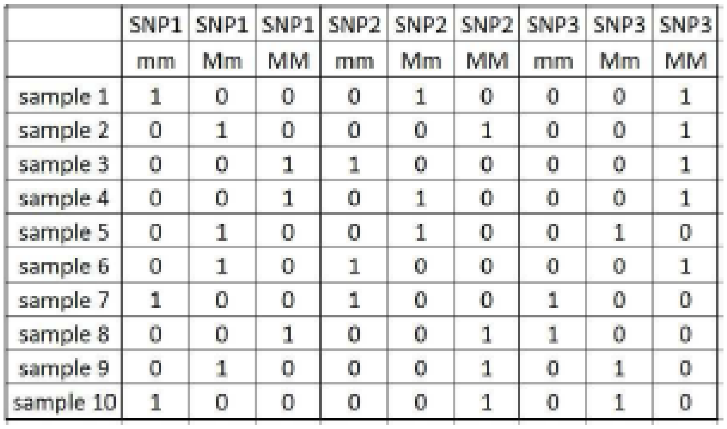
Figure 3. (a) SNP genotype matrix (categorical) (b) the corresponding matrix representation (binary)

a)



	SNP1	SNP2	SNP3
sample 1	0	1	2
sample 2	1	2	2
sample 3	2	0	2
sample 4	2	1	2
sample 5	1	1	1
sample 6	1	0	2
sample 7	0	0	0
sample 8	2	2	0
sample 9	1	2	1
sample 10	0	2	1

b)



	SNP1	SNP1	SNP1	SNP2	SNP2	SNP2	SNP3	SNP3	SNP3
	mm	Mm	MM	mm	Mm	MM	mm	Mm	MM
sample 1	1	0	0	0	1	0	0	0	1
sample 2	0	1	0	0	0	1	0	0	1
sample 3	0	0	1	1	0	0	0	0	1
sample 4	0	0	1	0	1	0	0	0	1
sample 5	0	1	0	0	1	0	0	1	0
sample 6	0	1	0	1	0	0	0	0	1
sample 7	1	0	0	1	0	0	1	0	0
sample 8	0	0	1	0	0	1	1	0	0
sample 9	0	1	0	0	0	1	0	1	0
sample 10	1	0	0	0	0	1	0	1	0

Figure 4. Visualization of the embedded size-4 combinations of SNP genotypes. The four columns A, B, C and D are respectively the minor-minor genotype of four SNPs. The pink color indicates 1s and the light blue indicates 0s, in the binary format. This combination of four SNP genotypes illustrates a high order epistasis.

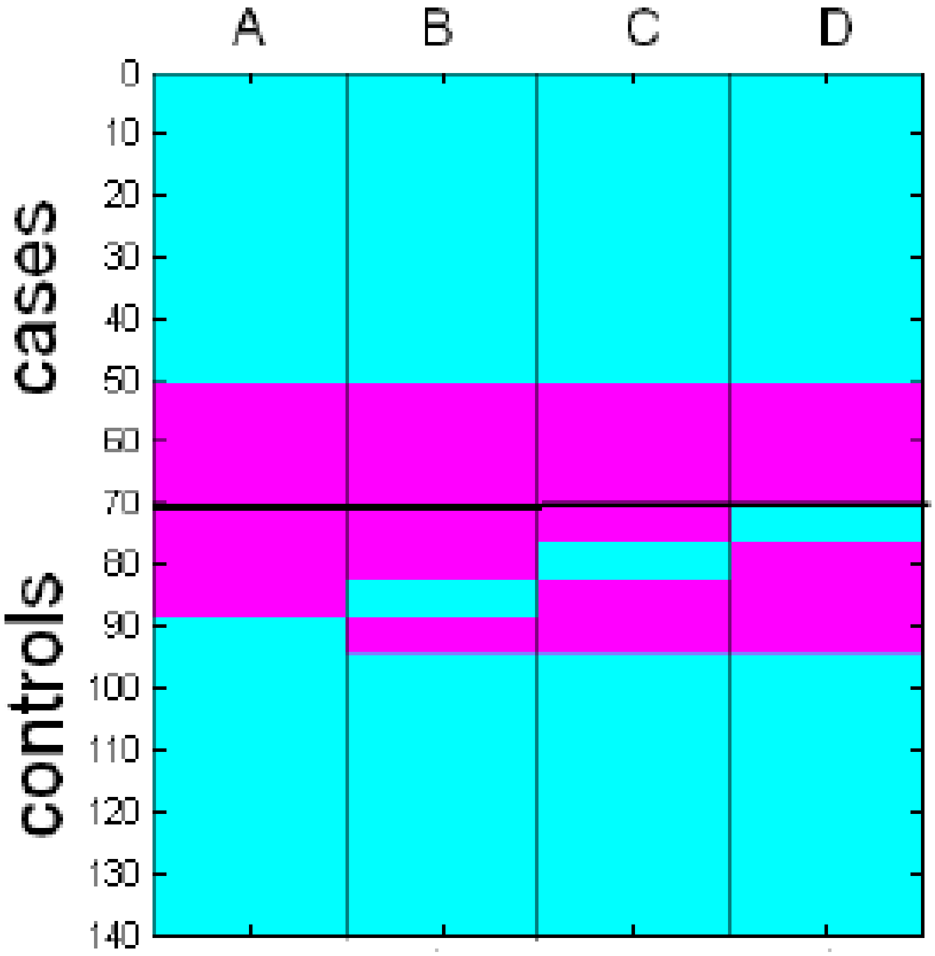

A
B
C
D
cases
controls
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140

Figure 5. Probability distribution of the the chi-square statistics of the true patterns (discovered from the original case-control dataset) and the random patterns (discovered in permutation test) on the synthetic data for the "without pathway" setup (left) and the "with-pathway" setup (right).

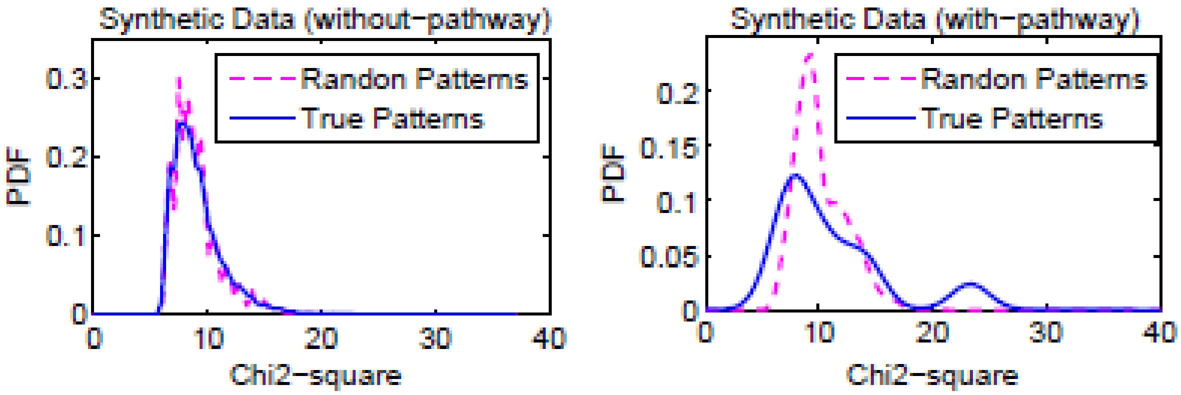
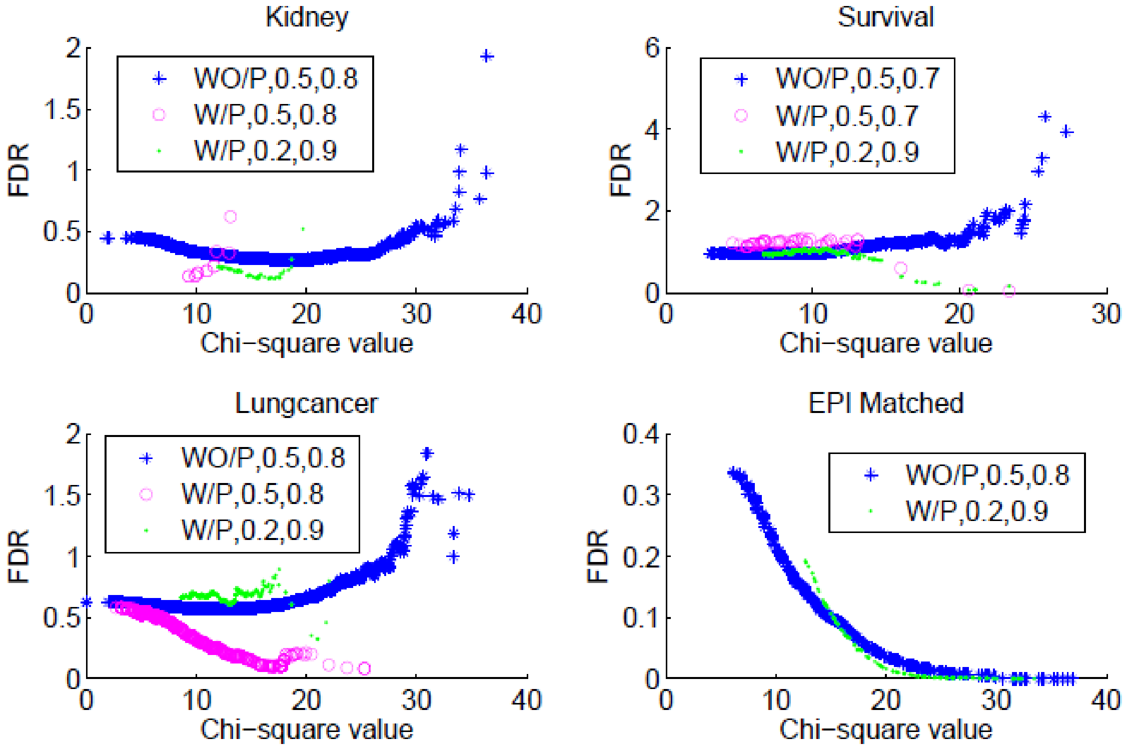

Synthetic Data (without−pathway)
PDF
0.3
0.2
0.1
0
0
10
20
30
40
Chi2−square
Randon Patterns
True Patterns
Synthetic Data (with−pathway)
PDF
0.2
0.15
0.1
0.05
0
0
10
20
30
40
Chi2−square
Randon Patterns
True Patterns

Figure 6. Detailed chi-square statistics and FDRs for each of the patterns discovered by the three experimental setups (A, B and C from Table 2) on the four datasets. WO/P is short for without gene-set constraint, and W/P is short for with gene-set constraints.


Kidney
FDR
Chi−square value
WO/P,0.5,0.8
W/P,0.5,0.8
W/P,0.2,0.9
0
0.5
1
1.5
2
0
10
20
30
40
Survival
FDR
Chi−square value
WO/P,0.5,0.7
W/P,0.5,0.7
W/P,0.2,0.9
0
2
4
6
0
10
20
30
Lungcancer
FDR
Chi−square value
WO/P,0.5,0.8
W/P,0.5,0.8
W/P,0.2,0.9
0
0.5
1
1.5
2
0
10
20
30
40
EPI Matched
FDR
Chi−square value
WO/P,0.5,0.8
W/P,0.2,0.9
0
0.1
0.2
0.3
0.4
0
10
20
30
40