INVESTIGATION OF THE IMPACTS OF LINKAGE DISEQUILIBRIUM ON SNP SELECTION STUDIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EKİN KANTAR ÖZÇIRPAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING

JANUARY 2015

Approval of the thesis:

INVESTIGATION OF THE IMPACTS OF LINKAGE DISEQUILIBRIUM ON SNP SELECTION STUDIES

submitted by EKİN KANTAR ÖZÇIRPAN in partial fulfillment of the requirements for the degree of **Master of Science in Biomedical Engineering, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _________________

Prof. Dr. Hakan Işık Tarman
Head of the Department, **Biomedical Engineering** _________________

Prof. Dr. Gerhard Wilhelm Weber
Supervisor, **Biomedical Engineering Dept., METU** _________________

Assoc. Prof. Dr. Cem İyigün
Co-advisor, **Industrial Engineering Dept., METU** _________________

Examining Committee Members:

Prof. Dr. Gerhard Wilhelm Weber
Biomedical Engineering Dept., METU _________________

Assoc. Prof. Dr. Cem İyigün
Industrial Engineering Dept., METU _________________

Assist. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU _________________

Assoc. Prof. Dr. Vilda Purutçuoğlu
Statistics Dept., METU _________________

Assist. Prof. Dr. Öznur Taştan
Computer Engineering Dept., Bilkent University _________________

Date: January 22th, 2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Ekin Kantar Özçırpan

Signature:

ABSTRACT

INVESTIGATION OF THE IMPACTS OF LINKAGE DISEQUILIBRIUM ON SNP SELECTION STUDIES

Kantar Özçırpan, Ekin

M.S., Department of Biyomedical Engineering

Supervisor: Prof. Dr. Gerhard-Wilhelm Weber
Coadvisor: Assist. Prof. Dr. Cem İyigün

January 2015, 90 pages

In many Genome Wide Association Studies (GWAS), the relation between SNPs and complex diseases has being tried to reveal. Moreover it is known that, in GWAS there exist a high amount of data which include relations between SNPs, phenotypes and diseases, etc. Many algorithms have been used to be able to reach the desired information from this huge data. Therefore, in this study, an algorithm one of whose important steps is based on linkage disequilibrium(LD), was constructed to eliminate the redundant information from the high-dimensional data. The algorithm improved in this study has been tested on prostate cancer data set downloaded from dbGaP.

In order to find disease related SNPs in GWAS in a more effective way, we have constructed an algorithm which is based on LD. The web tool called SNAP (SNP Annotation and Proxy Search) was used to obtain the SNPs in the region of LD, which was determined based on the specific threshold value for r^2. This value was selected as 0.5. After obtaining a modified version of original data set based on LD, Using *Fisher's Combination Method*, we have obtained associated combined *p* values for each SNP in this data set. Then using *SNPnexus* database, we tried to achieve disease related SNPs from both data sets which are the original and modified ones. Thus both of the performances being applied on these data sets were evaluated relative to each other. Moreover, after eliminating the redundant data we have applied SNPnexus analysis again and then the results have shown us, by using approximately half of the SNPs, we were able to achieve the desired genes. Besides all of them also *random forest algorithm* was performed on the data set including SNPs with individual *p* values and the modified data set which is including SNPs with combined *p* values. The outputs of both performances were compared.

In addition, one more purpose of this study, being able to reach the most important regulatory SNPs (rSNPs) in GWAS. Based on the data set which was modified using LD, we have focused on the non-coding SNPs, which are located on noncoding regions, through the whole genome. In conclusion, the number of important regulatory SNPs that were found from the modified data set, is much higher than we have found before by using original data set., it is expected from this thesis is that, the studies which have been conducted on prioritization of disease related SNPs are being effected by linkage disequilibrium(LD).

Keywords: SNP, Genome Wide Association Studies, Prostate Cancer, LD, *p* Value, Random Forest.

ÖZ

TEK NÜKLEOTİT POLİMORFİZM (SNP) SEÇİMİ ÇALIŞMALARINDA BAĞLANTI DENGESİZLİĞİNİN ETKİLERİNİN İNCELENMESİ

Kantar Özçırpan, Ekin
Yüksek Lisans, Biyomedikal Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Gerhard-Wilhelm Weber
Yardımcı Tez Yöneticisi: Doç. Dr. Cem İyigün

Ocak 2015, 90 sayfa

Genom ölçeğinde ilişkilendirme çalışmalarında (GWAS), DNA üzerinde tek nükleotid polimorfizmi olarak adlandırılan SNPler ile kompleks hastalıklar arasındaki ilişlki ortaya çıkarılmaya çalışılır. Literatürde, bu amaca yönelik çalışmlardan daha verimli bir şekilde sonuç elde edebimek için çeşitli algoritmlar yer almaktadır. Bizim çalışmamızda da LD' nin, yüksek miktarda, çok boyutlu veri setlerini içeren bu çalışmalar üzerindeki etkisini ölçmek adına yeni bir algortima geliştirilmiştir.

Web tabanlı SNAP (SNP Annotation and Proxy Search) aracı kullanılarak r^2 değeri 0.5 olarak belirlenmiş ve veri setimizde yer alan her bir SNP ile ilgili LD bölgesinde bulunan SNP dizileri elde eilmiştir. Daha sonra elde edilen her bir SNP dizisi için *Fisher's Combination metodu* kullanılarak *combined p value* olarak adlandırdığımız bileşik bir p değeri hesaplanmıştır. Bu değer SNP dizileri içerisinde orjinal p değeri en küçük olan SNP'e atanarak çalışmanın ileriki basamaklarında kullanılacak olan yeni bir veri seti elde edilmiştir. Orjinal veri seti olara dbGAP veritabanından elde edilen prostat kanser verileri kullanılmıştır. LD kullanılarak elde edilen veri ile orjinal veri üzerinde SNPPnexus analizi gerçekleştirilmiş ve bulunan hastalıkla ilişkili SNPler karşılaştırılarak performans değerlendirmeleri yapılmıştır. Ayrıca daha önce de belirttiğimiz gibi GWAS çalışmaları, yüksek miktarda veri üzerinde yürütülmektedir. Bu yüzden karmaşaya ve zaman kaybına neden olan yığınla anlamsız veriden sakınılması gerekmektedir. LD yardımı ile, gereksiz bir takım veriyi filtreleyerek elde ettiğimiz bir diğer veri setinde uygulanan *SNPnexus* analizi gösteriyor ki; yaklaşık yarı yarıya düşen SNP sayısı ile, aynı anlamlı genlere ulaşabiliyoruz. Bu da, çalışmaya başlamadan önce öngördüğümüz bazı sonuçlara ulaştığımızı gösteriyor. Bütün bu çalışmaların yanısıra bu iki veri seti, bir de rastgele orman metoduna girdi olarak verilmiş ve elde edilen çıktılar karşılaştırılarak bu algoritmanın etkisi değerlendirilmiştir.

Bu çalışmanın bir diğer amacı da gen bölgesinde yer almayan ancak bir geni dolaylı olarak etkilebilecek olan düzenleyici SNPleri (rSNPs) de tespit etmek. Bu SNPleri araştırırken de gördük ki; orjinal veri seti ve LD tabanlı veri setinden elde edilen sonuçlar karşılaştırıldığında, LD tabanlı veri setinden daha verimli sonuçlar elde edebiliyoruz. Sonuç olarak anlaşılıyor ki, hastalıkla ilişkili SNPlerin seçimi üzerine yürütülen çalışmlarda LD'nin etkisinin ölçülmesi bu çalışmanın temel amacıdır.

Anahtar Kelimeler: SNP, Genom Ölçeğinde İlişkilendirme Çalışmaları, Prostat Kanseri, LD, *p* Değeri, Rastgele Orman.

Dedicated to my dear mother Solmaz Kantar, my dear father Yüksel Kantar, my dear brother Erdal Kantar, my dear sister Deniz Kantar and my dear husband Hasan Özçırpan.

ACKNOWLEDGEMENTS

First I would like to express my appreciation to my supervisor **Prof. Dr. Gerhard-Wilhelm Weber** and my co-advisor **Assoc. Prof. Dr. Cem İyigün** for their guidance and time devoted to me. In addition, I wish to thank to **Assist. Prof. Dr. Yeşim Aydın Son** for her advices, valuable support and assistance during my thesis study.

I also want to thank my dear uncle and also my friend, **Assist. Prof. Dr. Metin Kantar**, since he has been always supporting my education through my life, since the day I was born.

Further I would like to express my deepest gratitude to **Dr. Serkan Kaygın** for sharing and teaching his own experiences throughout my studentship. I also want to thank to all my friends since, they always support me when I need them.

Finally, I would like to thank to my parents for their endless love and patience through my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

SNP: Single Nucleotide Polymorphism

MDA: Mean Decrease Accuracy

MDG: Mean Decrease Gini

RF: Random Forest

rsID: An accession number used to refer to specific SNPs by researchers and databases

OOB: Out of bag

DNA: Deoxyribonucleic acid

GWAS: Genome-wide association study

SNP-IP: The data set including SNPs with individual *p* values

SNP-CP: The data set including SNPs with combined *p* value

SNP-Filtered: Representative set of SNPs (reduced version of *B* based on LD)

SNP-IP-Ext: Extended version of SNP-IP by adding phenotype information

CHAPTER 1

INTRODUCTION

The aim of Human Genome Project, which was coordinated by the National Institutes of Health and the U.S. Department of Energy, is to determine the sequence of the human genome and to identify the genes consisting of these sequences (Shapiro 1993).This project formally began in 1990 and was completed in 2003. After Human Genome Project was completed, researchers started to understand formation and functionalities of the human DNA. As they learn more about the genes and proteins those form the DNA structure, fields like medicine, biotechnology, or the life sciences, have begun to be developed more qualified than before (Contributors; “Genetic Home Reference” 2014).

The field which has a greatest tendency to be improved was *Genome-wide association studies (GWAS)*. In this method the genome is searched for small variations, called *single nucleotide polymorphisms* or *SNPs* (pronounced “snips”), that will be explained in detail in Subsection 2.1. Since one of the most important medical problems is to obtain the association between complex diseases and SNPs, these methods are mostly developed for this purpose. By looking at hundreds or thousands of SNPs at the same time via GWAS, I would say researchers aim to figure out the gene that may contribute to a person’s risk of developing a certain disease. In addition, one of the main objectives of a GWAS is to develop a prediction model for clinical outcome which is mostly binary (Kim et al. 2013). This can be done by knowing that the SNPs occur more frequently in people with a particular disease than in people without a disease. The outcomes of the study can be used for diagnostic and prognostic purposes in related fields and provides researchers to have better understanding of the relationship between the disease and SNPs

("Genetic Home Reference" 2014). In the literature, many SNP-complex disease relations were studied such as heart diseases (Lettre et al. 2011), diabetes (Reddy et al. 2011), rheumatoid arthritis (Stahl et al. 2010), bipolar disorder (Scott et al. 2009), hypertension (Adeyemo et al. 2009), multiple sclerosis (Jakkula et al. 2010) and cancer types (Yeager et al. 2007; Easton and Eeles 2008; Gerstenblith, Shi, and Landi 2010).

There exist two types of methods using in GWAS: *parametric methods* and *non-parametric methods. Parametric methods* are based on a genetic model. These kinds of models are mostly constructed by statistical calculations such as regression based models. On the other hand *non-parametric methods* do not need any genetic model to achieve the goal of our study. By such methods genetic models are constructed, mostly with using data mining and machine learning techniques (Musani et al. 2007). There are advantages and disadvantages of using these methods. However, the appropriate method should be chosen based on many important criteria such as the problem that is needed to be solved, and the data type in genetic data set that will be used for the associated study. In addition, the aim of the study must be seriously concerned while choosing the group of methods. There are also such kinds of studies that are trying to combine these method to obtain a new *hybrid method* which is expected to improve advantages of the methods while minimizing the disadvantages (Lin 2010; Journal and Computing 2013).

The genetic data used in these studies have a high dimensionality, so that sometimes traditional statistical methods can be inadequate for the analysis. Due to these comments, researchers prefer non-parametric methods over parametric ones (Aguiar, Seoane, and Freire 2010)

As it can be seen in the literature, lots of different machine learning algorithms have been applied in GWAS. The methods using decision trees (Fiaschi, Garibaldi, and Krasnogor 2009; Saangyong Uhmn Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong and Jin Kim 2009; Miyaki et al. 2004; Gomes, Vinga, and Gaspar 2010), artificial neural networks (Tomida et al. 2002; Lucek et al.; Marinov and Weeks 2001; Tomita et al. 2004), Bayesian belief networks (Sinoquet and Leray 2010; Mourad, Sinoquet, and Leray 2011; Jiang, Barmada, and Visweswaran 2010), support vector machines (Zhou and Wang 2007; Chuang et al. 2011; Brown et al.

2000; Waddel et al. 2005) and genetic algorithms (Journal and Computing 2013; Brown et al. 2000), etc., can be given as example of these machine learning methods which have been used in this field. However, since the results of these methods can change due to some variables such as the type of the data as mentioned above, researchers could not prove that one of the methods perform best (Musani et al. 2007).

In this study, we applied the *random forest* algorithm to analyze the data set including SNPs with individual *p* values. In addition, by using linkage disequilibrium (LD) region information which will be explained in detail, we obtained a set of SNPs with combined *p* values. This combined *p* values have been derived by using *Fisher's Combination Method*. After obtaining a set of SNPs with combined *p* values, the *random forest algorithm* was performed again. Since our data set was collected based on prostate cancer information, our aim has been to be able to find SNPs and genes that are associated with prostate cancer. For these purposes, *SNPnexus* analysis has been performed. Moreover, we have shown interest in noncoding regions besides coding regions where the genes are included. We will also explain the regulatory SNPs which may have an effect on prostate cancer. These noncoding SNPs have been analyzed for both data sets including SNPs with individual *p* values and combined *p* values via *HaploReg* website. The explanations of all comparisons done using the obtained results will be discussed.

This thesis is structured as follows. Chapter 1 gives information about the literature related to the SNP selection studies. In Chapter 2, the data sets which are used in this study and *random forest* method will be introduced by giving brief information about the formulation used in this study. In Chapter 3, materials and methods that are applied in this study will be explained in detail. In the following Chapter 4, the results of the data processing part will be given. Chapter 4 will also contain the experimental phase of this study, including an explanation of the way of achieving combined *p* values by using LD. In Chapter 5, a discussion of the results obtained in Chapter 4 is provided. Chapter 5 also contains a comparison of the models with the experimental findings of this study for different *p* values. The closing Chapter 6 contains the conclusion of the study and suggestions for future work.

Firstly the meaning and the usage of SNPs should be understood. Therefore, we will continue by describing SNPs in all aspects in the following chapter.

CHAPTER 2

LITERATURE SURVEY

2.1 The Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation. They represent differences in DNA at nucleotide level. This kind of differences occurs as the replacement of two nucleotides in one of the stretch of DNA. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in that DNA stretch (Figure 2.1).

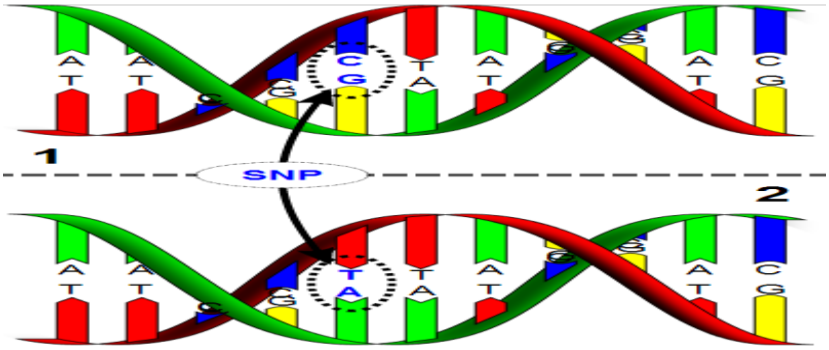


Figure 2.1 A C/T polymorphism. (adapted from http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)

To understand what makes us different from each other, it has to be known those genomic differences in human population. Therefore, this topic is one of the major interest of researchers in genetics science. As mentioned above, SNPs are single base pair differences between individuals and they are important reasons behind the variations that occur in human genome.

According to researches, more than 112 million SNPs have been reported in human genome, which means if we now that there are about 3.2 billion nucleotides through the whole DNA, approximately, it is assumed that in every 28 nucleotides there exist one SNP in the human genome. This information is validated by *dbSNP* database

(Sherry et al. 2001; Eslahchi et al. 2011). There is also another public database called *Ensembl* (T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo et al. 2007) which provides researchers to access identified millions of common SNPs.
SNPs have the important role in genome science to find the disease related genes, since they can be used as biological markers. Therefore, a SNP may have an effect on the gene function directly if it is located in a gene sequence or near a gene sequence in DNA strand. In addition, SNPs are preferred to other genetic markers, such as microsatellites, because of their high abundance, relatively low mutation rate, and easy adaptability to automatic genotyping ("Genetic Home Reference" 2014).

2.2 Linkage Disequilibrium (LD)

An *allele* is an alternative form of same gene or same genetic locus that is located at a specific position on a specific chromosome. This DNA coding determines individual characteristics that can be passed on from parents to offspring. It becomes necessary to deal with the non-random associations of alleles at different loci since allele frequencies could not be able to describe the dynamics of genotypes when genetic variation at two or more loci is considered simultaneously. These non-random associations of alleles at different loci is called *linkage disequilibrium (LD)* (Feingold 1980). LD occurs when genotypes at the one locus is not independent of the genotype at a second locus. However, with not proving the dependency between two genotypes, we cannot exactly say that there exists a LD, since the term is misleading for two reasons. The first one is that non-random associations of alleles at two loci can occur even if the two genes are unlinked, and also just because two loci are linked, this does not mean that they will be in linkage equilibrium (LE). In addition, by proving the independency between two genotypes, a linkage equilibrium exists whose term is the opposite of LD used for this situation.

LD is affected by several factors such as population history, the age and phenotype characteristics of the variants or natural selection, etc. There is a specific area of research studying relative contributions of these factors to LD patterns. Therefore it is expected that LD information in different regions and different populations is used for inferring population histories and localizing genetic variants underlying complex traits (Zhao H. , Pfeiffer R 2003). Biologists and human geneticists are increasingly studying linkage disequilibrium recently, in order to understand past evolutionary and demographic events. In addition, they are using LD information, in order to map genes that are associated with quantitative characters and inherited diseases, so that they can investigate sets of genes that have been linked during evolution (Slatkin 2008).

When it comes how the LD is calculated, in the literature there are some different measurements of LD. Based on the haplotype frequencies, different LD coefficients, that are commonly used ones, can be measured as explained below.
To measure the LD coefficient, it is necessary to know haplotype frequencies as show in Table 2.1.

Table 2.1 Haplotype Frequencies

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

For the distant loci, not the LD but the LE is expected. Therefore, the following equations based on haplotype frequencies becomes as given:

$$p_{AB} = p_A . p_B \tag{2.1}$$

$$p_{Ab} = p_A p_b = p_A(1 - p_B) \tag{2.2}$$

$$p_{aB} = p_a p_B = (1 - p_A) p_B \tag{2.3}$$

$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B) \tag{2.4}$$

For the nearby loci LD is expected and the equations based on haplotype frequencies becomes as follows:

$$p_{AB} \neq p_A p_B \tag{2.5}$$

$$p_{Ab} \neq p_A p_b = p_A(1 - p_B) \tag{2.6}$$

$$p_{aB} \neq p_a p_B = (1 - p_A)\, p_B \tag{2.7}$$

$$p_{ab} \neq p_a p_b = (1 - p_A)(1 - p_B) \tag{2.8}$$

One of the coefficients associated with LD is the disequilibrium coefficient D_{AB}, given below:

$$D_{AB} = p_{AB} - p_A p_B \tag{2.9}$$

$$p_{AB} = p_A p_B + D_{AB} \tag{2.10}$$

$$p_{Ab} = p_A p_b - D_{AB} \tag{2.11}$$

$$p_{aB} = p_a p_B - D_{AB} \tag{2.12}$$

$$p_{ab} = p_a p_B + D_{AB} \tag{2.13}$$

A and *B* are set to be common allele and *a* and *b* are set to be the rare allele mostly found in the literature.

The second coefficient for LD is D'_{AB} which is the scaled version of D. D'_{AB} is given by:

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)}, \text{ if } \quad D_{AB} < 0, \\[2em] \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)}, \text{ if } \quad D_{AB} > 0. \end{cases} \tag{2.14}$$

The value of D'_{AB} varies between -1 and 1. If the value of *D'* is equal to 1 or -1, this means a recombination between the markers cannot be proved. When the allele frequencies are similar, with high *D'* the markers can represent each other. However, in some situations, *D'* can be misleading. For example, in small samples and when one allele is rare in any sample, the value of *D'* can mislead the outcomes of the study.

The third and the last coefficient for LD is r^2 whose calculation based on D_{AB} and the frequencies of haplotypes. This measurement is given by:

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{X^2}{2n} \qquad \textbf{(2.15)}$$

The value of r^2 varies between 0 and 1. When the two markers provide identical information, r^2 becomes 1 and if they are independent from each other, then the value of r^2 becomes 0. In the literature, this measure is mostly used by population geneticists (Hill 1974; Feingold 1980).

2.3 SNP Selection Studies Using LD

It is known that SNPs may be located on the genes related to common and complex diseases, such as cancer. Therefore, by identifying SNPs and using them as markers, researchers have been trying to prove that they may be helpful in personalized medicine, disease risk studies or in researches about inheritance of disease genes within families, etc. They further may also be used in studies about complex diseases such as heart disease, and cancer is a major challenge in current molecular sciences.

SNPs are chosen as markers due to the following reasons: There exist an abundant number of SNPs in DNA, also they have a relatively low mutation rate, and an easy adaptability to automatic genotyping also when compared to properties of microsatellites (Zhang et al. 2004). However, there are some disadvantages of using SNPs. The most common problem which researchers face, is the tremendous number of SNPs on the human genome, which is estimated at more than eleven million. The huge number of SNPs always causes the researchers to slow down, while challenging to obtain and analyze the information of all the SNPs.

It is known that SNPs play an important role in understanding the association between genetic variations and human diseases. Especially, in GWAS thousands of SNPs have to be genotyped to analyze. However, by identifying the correlation between genotypes, LD and SNPs, all of the SNPs does not have to be genotyped necessarily, no longer.

In the literature, there are SNP selection methods which are motivated by the non-random association among SNPs, namely are LD mentioned above (Patil et al. 2001;

Schulze et al. 2004; Jorde 2000; Gabriel 2002; Daly et al. 2001; Carlson et al. 2004). These methods claim that, when high LD exists between SNPs, the nucleotide information of one can usually be inferred from that of the others. Therefore, a relatively small subset of SNPs that still retains most of the nucleotide information of the original set can be selected. In these kinds of studies, the selected SNPs are called *tag* SNPs, while the remaining, unselected SNPs are called *tagged* SNPs. The assumption is that a possible association between a disease phenotype and the unselected tagged SNPs is assumed to be indirectly captured through the selected tag SNPs (Liu, Wang, and Wong 2010). Not for a same but for a similar purpose, we have improved an algorithm in one step of our study. We have obtained the SNPs which are in the same LD region. This region has been obtained by selecting r^2 threshold value as 0.005. By merging equivalent SNPs, we have found a small subset of SNPs which provides us to obtain enough for biologically relevant SNPs on prostate cancer. Therefore, by merging equivalent SNPs, not only computation cost is saved but also the storage can be reduced (Liu, Wang, and Wong 2010).

With the availability of a dense genome-wide map of SNPs, it is now possible to use linkage disequilibrium (LD) to map genes that cause a disease (Reich et al. 2001).
As mentioned, in GWAS hundreds or thousands of SNPs are being examined at the same time, so that researchers are able to figure out the gene that may contribute to a person's risk of developing a certain disease. According to studies, there are two important advantages of using LD that refers to the nonrandom association of alleles at two or more different regions in haplotypes that are inherited from an ancestral chromosomes (Lewontin and July 1964), in association studies. One of the advantages of using LD is: there is no need to genotype the individuals who belong to a pedigree, just need to genotype unrelated individuals. Therefore, it is possible to study a huge number of individuals. A further advantage is that, since too much historical recombination events are reflected by LD information, this may help to develop a map for disease-causing mutations.

In the literature, it is seen that LD have been used in many studies, especially, in those which have the aim of selecting SNPs that represent a region including candidate genes related with diseases (Zhao H. , Pfeiffer R 2003; Byng et al. 2003; Horne and Camp 2004; Ayers and Cordell 2010; Xu, Kaplan, and Taylor 2007).

These studies have been carried out mostly based on the idea that if the LD appears between SNPs, this implies that not all SNPs need to be genotyped in the candidate region (Byng et al. 2003).

According to the previous studies in the literature, there are valuable variation in LD pattern across the human genome (Dunning et al. 2000; Reich et al. 2002; Taillon-Miller et al. 2000; Eisenbarth et al. 2001). There may exist some regions with a high LD and some regions with a low LD across the human genome (Daly et al. 2001; Patil et al. 2001; G. C. Johnson et al. 2001; Dawson et al. 2002; Gabriel 2002). In the literature the regions with high LD are termed as *blocks*. In these blocks, it is commensurate to draw most of haplotype structures by using inconsiderable number of tag SNPs (Patil et al. 2001; G. C. Johnson et al. 2001). Therefore, if the extraction of LD patterns from genotype data is done, may help sufficient number of selected tag SNPs. Available methods which have been developed for haplotype block partitioning and tag SNP selection based on haplotype data or genotype data can be classified into two categories. In one of the categories, firstly, haplotype blocks are obtained by using pairwise LD information of the SNPs (Gabriel 2002) or a four-gamete test (Wang et al. 2002). Then, tag SNPs are selected in each obtained block. In the other categories, haplotype blocks are used as a tool to minimize the total number of tag SNPs over a region of interest or the whole genome (Patil et al. 2001; Zhang et al. 2004). This type of methods can only be used with haplotype data. However, we will not employ haplotype information, we will just use the LD information between SNPs. By using LD information we have obtained lists of SNPs for each SNP in the data set and then given the SNPs that are in the same LD region to *Fisher's Combination Method* as an input. It is assumed that when there are many loci in high LD, this test performs very well (Chapman and Whittaker 2008). There are some other combined *p* value studies which use the LD information indirectly for measuring the combined *p* value (Cui, Li, and Williams 2011). We have selected the correlated SNPs located in the LD region of the representative SNP according to the threshold value for r^2. We have used the threshold value as 0.5 in order not to lose so much information while avoiding redundant information. Finally, the data set including SNPs with combined *p* values was obtained.

CHAPTER 3

MATERIAL AND METHODS

3.1 Data

3.1.1 Prostate Cancer Data

The data set which is consisting of “Multi Ethnic Genome Wide Scan of Prostate Cancer” data was obtained from dbGap database with the study accession number phs000306.v2.p1. Genotyping of the data was done by Broad Institute of MIT and Harvard as a part of GENEVA study. These data are a combination of such case control studies conducted on different ethnicities like African Americans, Latinos and Japanese that live in California and Hawaii. This data set consists of a total of 9415 subjects which 4650 cases and 4795 controls. Each subject has 544408 SNPs in genotyped area of their DNAs, represented by rsID’s. This data set has also phenotype information listed in Table 1; however these phenotype attributes will not be interested much in this study.

Table 3.1 Phenotype variables of prostate cancer data.

Name	Explanation
sex	Gender
Status	Case/Control status
age_cat	Age at entry into cohort
agedx_cat	Age at diagnosis for cancer cases
ageco_cat	Age at blood draw controls
bmi_cat	Body mass index
fh_prca	Family history of prostate cancer (brother or father)

pa_cat	**Hours per day of moderate or vigorous physical activity**
Packyrs_ca	Pack years of smoking cigarettes
ethanol_ca	Alcohol drinks per day
d_lyco_cat	Density of lycopene intake
p_fat_cat	Percentage of calories from fat
d_calc_cat	Density for calcium intake
currsmoke	Currently smoker?
eversmoke	Ever smoked?
severity	Aggressiveness of disease for cases

3.1.2 Data Set Annotations

The different sets of data are labelled as fallows and will be referred to as SNP-IP, SNP-CP, SNP-Filtered and SNP-IP-Ext after this point in the thesis.

SNP-IP: The data set which is preprocessed with Plink and filtered by choosing *p* values threshold as 0.005. Also referred as the data set including SNPs with individual *p* values.

SNP-CP: The data set including SNPs with combined *p* value.

SNP-Filtered: Representative set of SNPs (reduced version of ***SNP-CP*** based on LD).

SNP-IP-Ext: The genotype-phenotype integrated data set (Extended version of ***SNP-IP*** by adding phenotype information).

3.1.3 Preparation of the Data Set Used in the Subsections Based on SNPnexus analysis

We have prepared the data set by using the output of a web tool called *SNAP* which is needed to some operations be performed on, through this study. We have given the data set including 2706 SNPs with individual *p* values to this tool. SNP data set has been selected as "1000 Genomes Pilot" and the population panel has been as CEU. We have decided to choose r^2 threshold as 0.5. Normally, its range is from 0.0

through 1.0 and the default value is 0.8. However, to keep the data to be analyzed as large as needed and not to lose more SNPs, r^2 has been selected as 0.5.

After obtaining the data set from the output of SNAP analysis, it has been shown that many of SNPs yield a warning which has the meanings as the following:

- WARNING No LD data is available for “rs…..” in 1000GenomesPilot1, panel CEU,
- WARNING No matching proxy snps found,
- WARNING Query snp not in 1000GenomesPilot1(A. D. Johnson et al. 2008).

Therefore before trying to find combined *p* values we have to discard these SNPs which have a warning information, from the other SNPs that have LD and proxy SNP information. We have obtained the pure data set by eliminating this kind of redundant information. The pure data set contains 2495 SNPs which means that there are no response values for 211 SNPs of the input data set of this step.

Then, an algorithm has been applied to obtain combined *p* values for each SNPs in the *p* value filtered data set. This algorithm has been constructed by using R language in R Studio. First we have the pure data set which were eliminated from “Warning” information. There are lists of SNPs in LD regions of each SNP which include the response after SNAP analysis. Therefore, we have obtained SNP arrays from this data set and each SNP array represents a list of SNPs which are located in the LD region, where the selected r^2 threshold value is set as 0.5, for one SNP. Each SNP array was given sequentially to *Fisher’s Combination Method* as an input and the output of each step is one of the combined *p* value associated with one SNP array. Then we have identified each combined *p* value with each SNP in the data set which is the input data set of this process. Therefore, we have obtained the most important data set which will be used in further steps of this study.

3.1.4 Selecting Representative Set of SNPs By Using LD

It is assumed that if we use one node SNP instead of other SNPs in the same LD region, we are able to represent the whole SNPs in that region with that node SNP.

We have selected the representative SNP as the SNP with the smallest individual *p* value, which the SNP that has the most statistical meaning in genetics.

By eliminating the SNPs that are located in the same LD region other than the SNP which has the smallest *p* value in that region, we have found a small subset of SNPs which provides us to obtain biologically relevant SNPs for prostate cancer. Therefore, by minimizing the number of SNPs in the data set, not only computation cost is saved but also storage can be reduced (Liu, Wang, and Wong 2010). In addition, when a high LD exists between SNPs, the information of SNPs can be inferred from other SNPs located in the same LD region. Therefore relatively small subset of SNPs can carry most of the biological information of the original set (Friedman, Tibshirani, and Hastie 2009).

After the elimination has been done, the number of SNPs in the data set, which is including SNPs with combined *p* value, was reduced from 2495 to 1758.

3.1.5 Preparation of the Data Set Used in the Subsections Based on RegulomeDB Analysis

It is known that the functional effects of noncoding disease-associated SNPs cannot be determined easily. This is one of the challenging issue in GWAS. We are trying to break the prejudices that the SNPs are considered as unimportant in the noncoding regions through the whole DNA. In the literature, many of these SNPs are likely to be regulatory SNPs which are shown as rSNPs. Their functional ability is known as they are able to effect transcription factor (TF) while binding to DNA (Macintyre et al. 2010).

After *HaploReg* analysis have been done on *SNP-IP* and *SNP-CP* individually while in *SNP-IP*, it has been found that 1538 SNPs in 2706 have no dbSNP function annotation as intronic, 1419 SNPs in *SNP-CP* which includes 2495 SNPs, have no dbSNP function annotation as intronic. As a result, for both data sets, more than half of the SNPs in these data sets are not located on a gene region.

3.2 Data Preprocessing

3.2.1 Plink and P Value Filtering Analyzes

The data set which is “Multi Ethnic Genome Wide Scan of Prostate Cancer” was obtained from dbGap (database of Genotypes and Phenotypes) with the study accession number phs000306.v2.p1 (Mailman et al. 2007). This data set consists of 9457 subjects which 4650 cases and 4795 controls. The data set was given to *Plink* which is a whole genome association analysis tool set, as an input (Purcell 2007). The output of *Plink* analysis was *“association.assoc.adjusted”* file which is consisting of *p* values associated with all of the 544408 SNPs that are the genotype information of the subjects existing in the data set. By using the *p* value information including in this file, due to some threshold value, *p* value filtering has done on the prostate cancer data set. Therefore, these filtering process provides us to eliminate the redundant data which is accepted insignificant according to statistical studies. The threshold for the *p* value was chosen as 0.005 and the data set was filtered using this threshold value in order to find both statistically and biologically relevant SNPs. Since, the *p* value indicates the probability of observing the data by chance, more focused data have been obtained by choosing the *p* value as 0.005.

Due to the large number of single nucleotide polymorphisms (SNPs), it is essential to use only a subset of all SNPs. Therefore, eliminating redundant SNPs in the data set will provide several advantages to the researches while analyzing. After the filtering step, 544408 SNPs which is number of SNPs in *“association.assoc.adjusted”* file, has reduced to 2706 SNPs. This is our main data set that will be studied in more detail. Therefore, almost %99 of the all SNPs has been eliminated in order to simplify the further analysis in this study.

Moreover, an additional data set was obtained by using threshold for *p* value as 0.05 since while looking for the *p* values associated with the SNPs in LD region of each selected SNP in main data set, it was a waste of time searching for those *p* values in the huge data set which has mostly irrelevant SNPs. A *p* value of 0.05 is typically thought to indicate a significance level (Raetz et al. 2001).

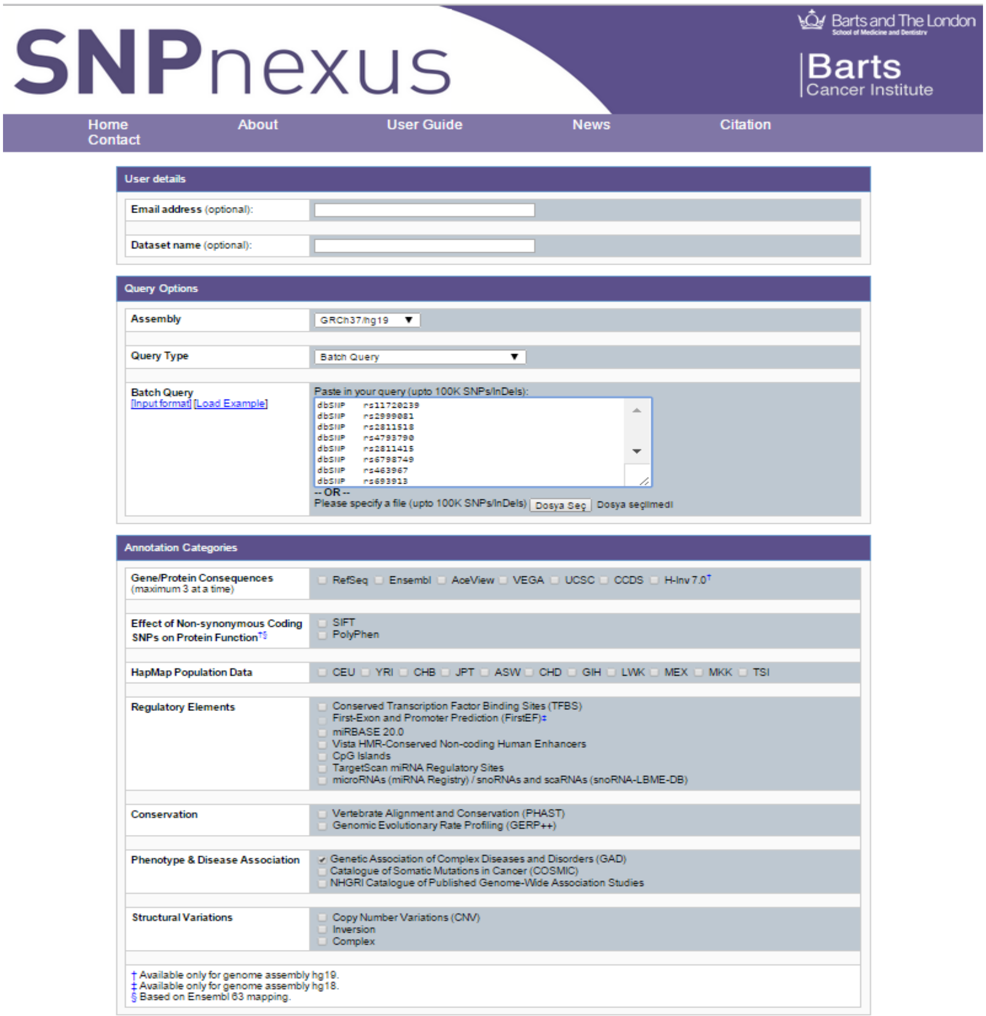
Hence, by choosing threshold value as 0.05 for the *p* value, the additional data set was limited to 26398 SNPs to improve the success of our study.

3.2.2 SNPnexus Analysis

SNPnexus analysis can be accepted as preprocessing for the data which is used as an input to last step of the algorithms improved in this study (Chelala, Khan, and Lemoine 2009; Abu Z. Dayem Ullah, Lemoine, and Chelala 2012; A Z Dayem Ullah, Lemoine, and Chelala 2013).

SNPnexus analysis has done via the website of *SNPnexus* shown in Figure 3.1. The part labeled with *"Paste in your query"* is used for importing our data set which includes the SNPs that we want to obtain disease association information. Finally, by choosing the Genetic Association of Complex Diseases and Disorders (GAD) option in the part labeled with *"Phenotype & Disease Association"* ,the output was set to give the required results (Chelala, Khan, and Lemoine 2009; Abu Z. Dayem Ullah, Lemoine, and Chelala 2012; A Z Dayem Ullah, Lemoine, and Chelala 2013).

Before and after the random forest method is being applied, *SNPnexus* analysis has been done several times in different steps of this study on different number of SNPs chosen in prostate cancer data set in order to compare the number of SNPs associated with prostate cancer. This analysis has also been provided to examine not only the number of SNPs that are related with prostate cancer but also which SNPs are related with prostate cancer with known rsIDs. Therefore, this tool is very useful to show disease association information. In Chapter 4, the tables which are given in Subsection 4.1.1 and 4.2.1, include the SNPs related with prostate cancer found by performing *SNPnexus* analysis.



User details

Email address (optional):

Dataset name (optional):

Query Options

Assembly: GRCh37/hg19 ▼

Query Type: Batch Query ▼

Batch Query [Input format] [Load Example]

Paste in your query (upto 100K SNPs/InDels):

```
dbSNP    rs11720239
dbSNP    rs2999081
dbSNP    rs2811518
dbSNP    rs4793790
dbSNP    rs2811415
dbSNP    rs6798749
dbSNP    rs463967
dbSNP    rs693913
```

-- OR --

Please specify a file (upto 100K SNPs/InDels) Dosya Seç Dosya seçilmedi

Annotation Categories

Gene/Protein Consequences (maximum 3 at a time): □ RefSeq □ Ensembl □ AceView □ VEGA □ UCSC □ CCDS □ H-Inv 7.0†

Effect of Non-synonymous Coding SNPs on Protein Function†§: □ SIFT □ PolyPhen

HapMap Population Data: □ CEU □ YRI □ CHB □ JPT □ ASW □ CHD □ GIH □ LWK □ MEX □ MKK □ TSI

Regulatory Elements:
□ Conserved Transcription Factor Binding Sites (TFBS)
□ First-Exon and Promoter Prediction (FirstEF)‡
□ miRBASE 20.0
□ Vista HMR-Conserved Non-coding Human Enhancers
□ CpG Islands
□ TargetScan miRNA Regulatory Sites
□ microRNAs (miRNA Registry) / snoRNAs and scaRNAs (snoRNA-LBME-DB)

Conservation:
□ Vertebrate Alignment and Conservation (PHAST)
□ Genomic Evolutionary Rate Profiling (GERP++)

Phenotype & Disease Association:
☑ Genetic Association of Complex Diseases and Disorders (GAD)
□ Catalogue of Somatic Mutations in Cancer (COSMIC)
□ NHGRI Catalogue of Published Genome-Wide Association Studies

Structural Variations:
□ Copy Number Variations (CNV)
□ Inversion
□ Complex

† Available only for genome assembly hg19.
‡ Available only for genome assembly hg18.
§ Based on Ensembl 63 mapping.

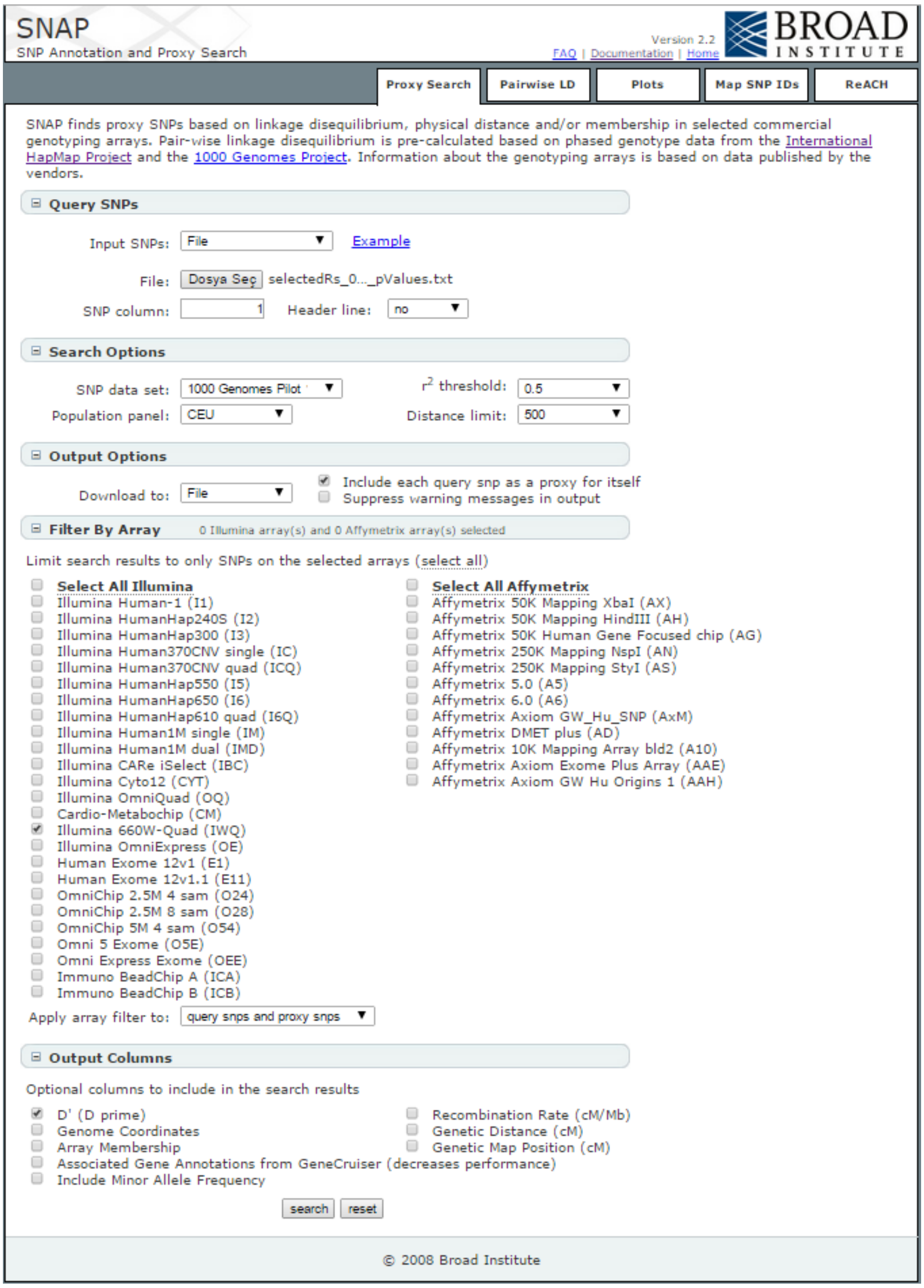
Figure 3.1 *SNPnexus* home page (adapted from Chelala, Khan, and Lemoine 2009; Abu Z. Dayem Ullah, Lemoine, and Chelala 2012; A Z Dayem Ullah, Lemoine, and Chelala 2013).

3.2.3 SNAP Analysis

Based on the ancestral geography of our population and the geographic location where the samples from that population were collected, we have chosen CEU as Population Panel in Figure 3.2. The population named CEU represents Utah residents with ancestry from Northern and Western Europe included in the HapMap. The other populations named like CEU in HapMap is given below (Altshuler et al. 2010):

- YRI : Yoruba in Ibadan, Nigeria,
- JPT : Japanese in Tokyo, Japan,
- CHB : Han Chinese in Beijing, China.

In Figure 3.2, the user interface of SNAP (SNP Annotation and Proxy Search) website is given. Using SNAP we are able to find proxy SNPs which are selected based on linkage disequilibrium, physical distance and/or membership in selected commercial genotyping arrays (A. D. Johnson et al. 2008).



SNAP
SNP Annotation and Proxy Search

Version 2.2
FAQ | Documentation | Home

BROAD INSTITUTE

SNAP finds proxy SNPs based on linkage disequilibrium, physical distance and/or membership in selected commercial genotyping arrays. Pair-wise linkage disequilibrium is pre-calculated based on phased genotype data from the International HapMap Project and the 1000 Genomes Project. Information about the genotyping arrays is based on data published by the vendors.

⊟ Query SNPs

Input SNPs: File ▼ Example

File: Dosya Seç selectedRs_0..._pValues.txt

SNP column: 1 Header line: no ▼

⊟ Search Options

SNP data set: 1000 Genomes Pilot ▼ r^2 threshold: 0.5 ▼

Population panel: CEU ▼ Distance limit: 500 ▼

⊟ Output Options

Download to: File ▼

☑ Include each query snp as a proxy for itself
☐ Suppress warning messages in output

⊟ Filter By Array 0 Illumina array(s) and 0 Affymetrix array(s) selected

Limit search results to only SNPs on the selected arrays (select all)

☐ **Select All Illumina**
☐ Illumina Human-1 (I1)
☐ Illumina HumanHap240S (I2)
☐ Illumina HumanHap300 (I3)
☐ Illumina Human370CNV single (IC)
☐ Illumina Human370CNV quad (ICQ)
☐ Illumina HumanHap550 (I5)
☐ Illumina HumanHap650 (I6)
☐ Illumina HumanHap610 quad (I6Q)
☐ Illumina Human1M single (IM)
☐ Illumina Human1M dual (IMD)
☐ Illumina CARe iSelect (IBC)
☐ Illumina Cyto12 (CYT)
☐ Illumina OmniQuad (OQ)
☐ Cardio-Metabochip (CM)
☑ Illumina 660W-Quad (IWQ)
☐ Illumina OmniExpress (OE)
☐ Human Exome 12v1 (E1)
☐ Human Exome 12v1.1 (E11)
☐ OmniChip 2.5M 4 sam (O24)
☐ OmniChip 2.5M 8 sam (O28)
☐ OmniChip 5M 4 sam (O54)
☐ Omni 5 Exome (O5E)
☐ Omni Express Exome (OEE)
☐ Immuno BeadChip A (ICA)
☐ Immuno BeadChip B (ICB)

☐ **Select All Affymetrix**
☐ Affymetrix 50K Mapping XbaI (AX)
☐ Affymetrix 50K Mapping HindIII (AH)
☐ Affymetrix 50K Human Gene Focused chip (AG)
☐ Affymetrix 250K Mapping NspI (AN)
☐ Affymetrix 250K Mapping StyI (AS)
☐ Affymetrix 5.0 (A5)
☐ Affymetrix 6.0 (A6)
☐ Affymetrix Axiom GW_Hu_SNP (AxM)
☐ Affymetrix DMET plus (AD)
☐ Affymetrix 10K Mapping Array bld2 (A10)
☐ Affymetrix Axiom Exome Plus Array (AAE)
☐ Affymetrix Axiom GW Hu Origins 1 (AAH)

Apply array filter to: query snps and proxy snps ▼

⊟ Output Columns

Optional columns to include in the search results

☑ D' (D prime)
☐ Genome Coordinates
☐ Array Membership
☐ Associated Gene Annotations from GeneCruiser (decreases performance)
☐ Include Minor Allele Frequency

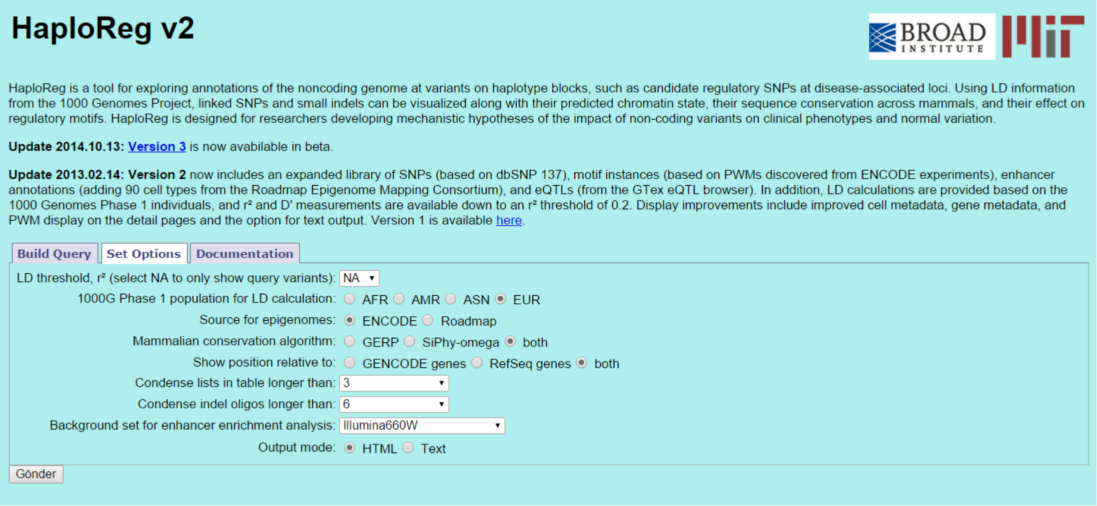
☐ Recombination Rate (cM/Mb)
☐ Genetic Distance (cM)
☐ Genetic Map Position (cM)

© 2008 Broad Institute

Figure 3.2 *SNAP* home page (adapted from A. D. Johnson et al. 2008).

3.2.4 HaploReg Analysis

HaploReg is a web tool for examining the annotations of SNPs in none coding regions of the DNA (Figure 3.3) (Ward and Kellis 2012). The SNPs in the data set including SNPs with individual *p* values and in the data set including SNPs with combined p values were given to this tool as input. While examining the SNPs it is necessary to select an LD threshold as NA (not applicable), since we are just looking among the SNPs that we have given as an input and not all the SNPs in the LD regions. Then, this tool was used in order to eliminate the SNPs which are exactly in the coding regions that have dbSNP functional annotations as intronic.



HaploReg v2

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with their predicted chromatin state, their sequence conservation across mammals, and their effect on regulatory motifs. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2014.10.13: Version 3 is now avabilable in beta.

Update 2013.02.14: Version 2 now includes an expanded library of SNPs (based on dbSNP 137), motif instances (based on PWMs discovered from ENCODE experiments), enhancer annotations (adding 90 cell types from the Roadmap Epigenome Mapping Consortium), and eQTLs (from the GTex eQTL browser). In addition, LD calculations are provided based on the 1000 Genomes Phase 1 individuals, and r² and D' measurements are available down to an r² threshold of 0.2. Display improvements include improved cell metadata, gene metadata, and PWM display on the detail pages and the option for text output. Version 1 is available here.

LD threshold, r² (select NA to only show query variants): NA

1000G Phase 1 population for LD calculation: ○ AFR ○ AMR ○ ASN ● EUR

Source for epigenomes: ● ENCODE ○ Roadmap

Mammalian conservation algorithm: ○ GERP ○ SiPhy-omega ● both

Show position relative to: ○ GENCODE genes ○ RefSeq genes ● both

Condense lists in table longer than: 3

Condense indel oligos longer than: 6

Background set for enhancer enrichment analysis: Illumina660W

Output mode: ● HTML ○ Text

Figure 3.3 *HaploReg* home page (adapted from Ward and Kellis 2012).

After obtaining the SNPs which are not located on the intronic regions by using *HaploReg* web tool, now it is needed to be known which SNPs are more important than others according to their *RegulomeDB* scores. Therefore, *RegulomeDB* is a web tool that provides information based on the regulatory elements in noncoding region of the human genome. The scoring scheme of *RegulomeDB* is given in Table 3.2. As it is shown in the table, the SNPs which have the smallest *RegulomeDB* score, are the most valuable SNPs when being considered as regulatory SNPs. The particular categories from 1a to 1f are more important than other categories and category 1a is the most important one which means a likely to affect binding and linked to expression of a gene target (Boyle et al. 2012).

Table 3.2 The scoring scheme of *RegulomeDB* (Boyle et al. 2012).

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak

1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

3.3 Random Forest Algorithm

To understand the random forest (RF) algorithm used in this study, the meaning of data mining and how the process of supervised learning is performed should be known. In summary, the process of extracting information from a data set and then transforming those data into an understandable structure for further use is called *data mining* (Wikipedia contributors 2014). Moreover, in this study classification type of random forest algorithm was applied and classification is considered as an instance of supervised learning, in which learning based on a training set of correctly identified observations is available, much as in our data set the people whose DNAs were genotyped are grouped as cases and controls. Therefore, the disease status information of the data set was known and can be given as an input into the random forest algorithm.

When different learning models are employed, the accuracy of classification can be increased. This is the main idea of the technique that is called *bagging*. Random Forest Algorithm works as a large collection of decorrelated decision trees. Since lots of decision trees are employed, the algorithm is named as *forest*. Random Forest Algorithm creates a lot of decision trees and uses them to make a classification; that is why it is an algorithm based on bagging technique. From one main sample set, lots of subsets are created with random values. For each one, decision trees are created.

After creating decision trees which are obtained by using subsamples of the entire sample, the number of votes can be accounted for each class (Friedman, Tibshirani, and Hastie 2009).

If we want to describe Random Forests (RF) more general, it is one of the commonly-used data-mining technique which has two main types on the way to achieve the outputs of the algorithm. The first where the response available is continuous is called *regression*. The second is categorical called *classification*, which will be focused on in this study. By building an ensemble of classification trees, this algorithm tries to predict the outcome which is the disease status due to the data set used in our study. Also the prostate cancer data set has a large number of predictors as SNPs those can improve the success of the algorithm (Breiman 2001). In RFs votes are collected from growing an ensemble of trees for selecting the most popular class to improve classification accuracy (Breiman 2001). In the literature RFs, commonly have been proposed for the analysis of genetic data (Goldstein, Polley, and Briggs 2011). Moreover, recently large SNP data sets from GWAS much as in our study have been suggested to be analyzed by using RFs (Goldstein et al. 2010).

With their quality of the prediction in high-dimensional data, RFs can also be used to rank SNPs by giving them variable importance measures (VIM) which is the most important outcome of RFs. These importance values are mean decrease accuracy (MDA) and mean decrease Gini (MDG). In random forest algorithm, with the addition of a single variable, if the accuracy of the algorithm decreases notably, this means that that variable must be taken into account. Therefore, the more the MDA increases, the more important the variable contributes to classification of the data. In addition, homogeneity is one of the important measures of the random forest algorithm, which is represented by MDG. The higher the MDG value grows, the more those variables result in nodes with higher purity (Liaw and Wiener 2002).

Hence, by selecting top ranking SNPs for each ranking based on importance values, one can use the best predictor variables for further studies (Winham et al. 2012). In addition, there is no need for any cross-validation to get an unbiased estimate of the data set error. During the run it is estimated internally which is defined as the out-of-bag (OOB) error rate (Breiman 2001).

3.3.3 Usage in R

The random forest algorithm can be applied in R by using the library of this algorithm which is prepared for R language. In R, Breiman's random forest algorithm based on Breiman and Cutler's original Fortran code is implemented for classification and regression (Breiman 2001). There is a formula which is represented as "randomForest" provides researchers an option to specify the predictors as a matrix or data frame by defining the x argument with defining responses as a vector via y argument. The given formula has too many arguments to be filled; however, in our case, we have only used some of the whole arguments. As shown in this equation, the type of our study is classification. This type is performed when response is a discrete factor. On the other hand, regression is performed when the response is not a factor that is continuous. While using randomForest formula, in order to perform supervised learning, one must specify the responses properly. Therefore in this study the column that is consisting of status information of the people is defined as a response vector via y argument in the Eqn. (3.2) (Liaw and Wiener 2002).

3.4 Fisher's Combination Method

There are a number of ways to combine independent *p* values or some other independent statistical values in different fields. Perhaps the most famous and the most widely used combination method is *Fisher's Combination Method*. In genetics, this method is employed for to combine *p* values of all SNPs in a gene. However, we have used this method to combine the *p* values of list of SNPs that are located in the LD region of one node SNP which is used for as main SNP while detecting SNPs in LD region.

Fisher's Combination Method is based on the fact that the probability of rejecting the global null hypothesis which states that there is no significant difference between the expected and observed result. The global null hypothesis is related to the intersection of the probabilities of each individual test, $\prod_i P_i$. Even if the null hypothesis is true for all partial tests, it is known that $\prod_i P_i$ is not uniformly distributed. Moreover this value cannot be used itself as the joint significance level for the global null hypothesis test. Fisher has rectified this fact by concretizing some interesting properties and relationships among distributions of random variables. These properties are explained below (*The Eugenics Review* 1926).

The *cumulative distribution function* (cdf) of an exponential distribution is:

$F(x) = 1 - e^{-\lambda x}$ **(3.1)**

λ is the rate parameter

The inverse *cdf* is:

$x = -\frac{1}{\lambda} ln(1 - F(x))$ **(3.2)**

$F(x) = P$ **(3.3)**

P is a random variable uniformly distributed in the interval [0,1]

Therefore Eqn (3.5) can be written as:

$x = -\frac{1}{\lambda} \ln(P)$ **(3.4)**

The cdf of a chi-squared distribution with v degrees of freedom, X_v^2, is given by:

$F(x; v) = \frac{\int_0^{\frac{x}{2}} t^{\frac{v}{2}-1} e^{-t} dt}{\left(\frac{v}{2}-1\right)!}$ **(3.5)**

If $v = 2$, and solving the integral we have:

$\mathrm{F}(x; v = 2) = \frac{\int_0^{\frac{x}{2}} t^{\frac{v}{2}-1} e^{-t} dt}{\left(\frac{v}{2}-1\right)!} = \int_0^{\frac{x}{2}} e^{-t} dt = 1 - e^{-\frac{x}{2}}$ **(3.6)**

The X^2 distribution with $v = 2$ is equivalent to an exponential distribution with rate parameter $\lambda = \frac{1}{2}$.

The *moment-generating function* (mgf) of a X_v^2 is:

$M(t) = (1 - 2t)^{-\frac{v}{2}}$ **(3.7)**

The mgf of the sum of k independent variables that follow each a X_2^2 distribution is then given by:

$M_{sum}(t) = \prod_{i=1}^{k} (1 - 2t)^{-\frac{2}{2}} = (1 - 2t)^{-k}$ **(3.8)**

which also defines a X^2 distribution, however, with degrees of freedom $v = 2\,k$.
As a summary of the Eqns. from (3.1) to (3.8), given above by taking the logarithm the product $\prod_{\mathrm{i}} \mathrm{P_i}$ can be converted into a sum. Multiplication of each $\ln(\mathrm{P_i})$ by 2 changes the rate parameter to λ = 1/2 and makes this distribution equivalent to a X^2 distribution with degrees of freedom $v = 2$. If k of these logarithms are summed, the summation also X^2 distribution, but now with $v = 2\,k$ degrees of freedom, i.e. X^2_{2k} .

Therefore the statistic for the Fisher method can be computed as:

$$X = -2\sum_{i=1}^{k} \ln(\mathrm{P_i}) \qquad \textbf{(3.9)}$$

In conclusion, with X following a X^2_{2k} distribution, p value for the global hypothesis can be easily obtained (*The Eugenics Review* 1926; Zaykin et al. 2007). In other words, for combining K independent *p* values Eqn. (3.11) is used as a statistical method (Peng et al. 2010).

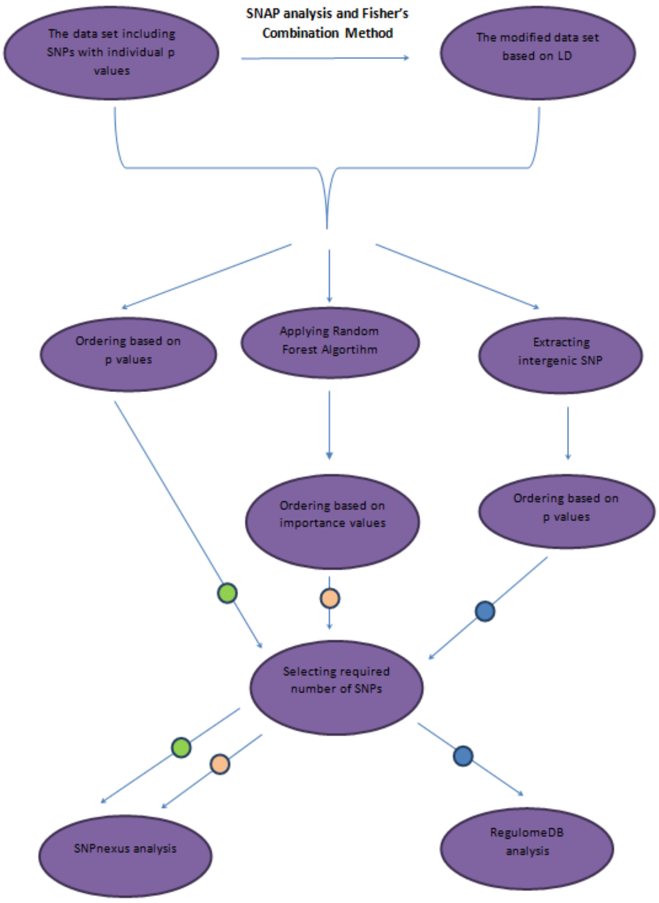
3.4.1 Usage in R

It is very simple to use the *Fisher's Combination Method* in R. It is the code representation of the mathematical expression of this method.
By giving each SNP list that is located in the LD region of one representative SNP, to the method as an input, we are able to calculate the combined *p* value for each group of SNPs. However in our algorithm we have associated the measured combined p value with the representative SNP. Therefore, we have obtained a data set including SNPs with combined *p* value.

CHAPTER 4

RESULTS

The Figure 4.1 given below can be accepted as the main framework of our study. There are also more than one additional steps in our study other than the steps in the following flow chart. These extra steps will also be included in this chapter.



The data set including SNPs with individual p values
SNAP analysis and Fisher's Combination Method
The modified data set based on LD
Ordering based on p values
Applying Random Forest Algortihm
Extracting intergenic SNP
Ordering based on importance values
Ordering based on p values
Selecting required number of SNPs
SNPnexus analysis
RegulomeDB analysis

Figure 4.1 The main flow chart of our study.

4.1 Analyses Performed on the Data Set Including SNPs with Individual *p* Values

After Plink analysis and *p* value filtering, by choosing threshold value as 0.005, the data set including 2706 SNPs with individual *p* values has been obtained. This data set was annotated as *SNP-IP*, throughout this study. In Subsection 4.1, the whole analyses done on *SNP-IP* which is the data set including SNPs with individual *p* values, will be explained.

4.1.1 Outputs of SNPnexus Analysis

Analysis with *SNP-IP*

In this step, *SNPnexus* analysis has been done on all of the SNPs in *SNP-IP* to find out if there are any associated SNPs with prostate cancer or not. The number of SNPs has been found associated with prostate cancer calculated as 57 through the whole of 2706 SNPs in this data set. Those 57 SNPs are shown in Table 4.1 with associated rsIDs. In addition, the Table 4.2 shows the genes where these SNPs are located. It has been observed that 57 SNPs are found to be associated with prostate cancer, located in 29 different gene regions.

Table 4.1 Prostate cancer related SNPs found in *SNP-IP*.

	SNPS		SNPS
1	rs12329598	**30**	rs209998
2	rs531572	**31**	rs4908107
3	rs3912492	**32**	rs927188
4	rs9637471	**33**	rs553371
5	rs6803449	**34**	rs2360995
6	rs12636081	**35**	rs12896434
7	rs17061864	**36**	rs7914154
8	rs8178179	**37**	rs928111
9	rs17775610	**38**	rs4267385
10	rs2999081	**39**	rs7911448
11	rs2811415	**40**	rs13437706
12	rs2811518	**41**	rs1234220

13	rs2811388	**42**	rs4750759
14	rs6798749	**43**	rs1877724
15	rs3764880	**44**	rs2567608
16	rs620359	**45**	rs10776909
17	rs12882037	**46**	rs1266890
18	rs4921943	**47**	rs567700
19	rs7844180	**48**	rs11016862
20	rs4496973	**49**	rs3745540
21	rs4870828	**50**	rs2360999
22	rs16998751	**51**	rs7089141
23	rs6482743	**52**	rs6768256
24	rs9971190	**53**	rs6970999
25	rs7082319	**54**	rs2107280
26	rs213386	**55**	rs8066276
27	rs1380466	**56**	rs10106032
28	rs4665716	**57**	rs11737898
29	rs2622625		

Table 4.2 Genes hosting the SNPs that were found related to prostate cancer in *SNP-IP*.

	GENES	SNPS		GENES	SNPS
1	**EPHX1**	rs1877724	**16**	**KLK12**	rs3745540
2	**BCAS1**	rs12329598 rs16998751	**17**	**FHIT**	rs3912492 rs9637471 rs6803449 rs12636081 rs17061864 rs213386 rs1380466
3	**PTEN**	rs1234220	**18**	**ESRRB**	rs12882037 rs2360995 rs12896434 rs2360999

4	**NCOA1**	rs4665716	**19**	**PIK3AP1**	rs7914154
5	**ABCG2**	rs2622625	**20**	**CUBN**	rs7089141
6	**MGMT**	rs531572 rs6482743 rs9971190 rs7082319 rs553371 rs4750759 rs567700 rs11016862	**21**	**COL4A2**	rs928111
7	**AIFM1**	rs209998	**22**	**NXPH1**	rs6970999 rs2107280
8	**EEFSEC**	rs2999081 rs2811415 rs2811518 rs2811388 rs6798749 rs6768256	**23**	**ACE**	rs4267385 rs8066276
9	**SLC30A7**	rs4908107	**24**	**CAMK1D**	rs7911448
10	**TLR8**	rs3764880	**25**	**PSD3**	rs4921943 rs7844180 rs10106032
11	**SSTR4**	rs2567608	**26**	**ZHX2**	rs4496973 rs4870828
12	**RXRA**	rs10776909	**27**	**CREB5**	rs13437706
13	**PKHD1**	rs927188 rs1266890	**28**	**SORCS2**	rs11737898
14	**PRKDC**	rs8178179	**29**	**PVT1**	rs17775610
15	**C2ORF43**	rs620359			

Analysis with 1000 SNPs of *SNP-IP*

To narrow the region to be analyzed, initially, we have sorted the SNPs according to their individual *p* values. This sorted SNP set can be referred to briefly as sorted

SNP-IP. Then we have selected the first 1000 SNPs of this sorted set. To find the prostate cancer related SNPs in that narrow region, *SNPnexus* analysis has been done on this set. The output SNPs are shown in Table 4.3 with associated rsIDs. It has been observed, 21 SNPs that show a significant association with prostate cancer located in 11 different genes (Table 4.4).

Table 4.3 Prostate cancer related SNPs founded in first 1000 of sorted *SNP-IP*.

	SNPS
1	rs12329598
2	rs531572
3	rs3912492
4	rs9637471
5	rs6803449
6	rs12636081
7	rs17061864
8	rs8178179
9	rs17775610
10	rs2999081
11	rs2811415
12	rs2811518
13	rs2811388
14	rs6798749
15	rs3764880
16	rs620359
17	rs12882037
18	rs4921943
19	rs7844180
20	rs4496973
21	rs4870828

Table 4.4 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 SNPs of sorted *SNP-IP*.

	GENES	SNPS
1	BCAS1	rs12329598
2	MGMT	rs531572
3	FHIT	rs3912492 rs9637471 rs6803449 rs12636081 rs17061864
4	PRKDC	rs8178179
5	PVT1	rs620359
6	EEFSEC	rs2999081 rs2811415 rs2811518 rs2811388 rs6798749
7	TLR8	rs3764880
8	C2ORF43	rs620359
9	ESRRB	rs12882037
10	PSD3	rs4921943 rs7844180
11	ZHX2	rs4496973 rs4870828

Analysis with 1000 SNPs of *SNP-IP* after RF

The data set including 2706 SNPs with individual *p* values which was annotated as *SNP-IP*, was given to random forest as an input. The most important parameters of *random forest* algorithm are "ntree" and "mtry" which should be determined before performing the algorithm. "ntree" which represents the number of trees to grow was determined as 5001 which is large enough to ensure that every input row gets predicted at least a few times. Furthermore, "mtry", which is the number of variables randomly sampled as candidates at each split, was determined as 10. Although the default value of “mtry” for classification is the square root of the number of variables which is approximately 52 in this study, it has been chosen as 10 since based on our group’s experinces the number 10 was found more suitable than the number 52.

The outputs of RF are obtained as importance values which are MDA and MDG values, explained in Subsection 3.3.3. The SNPs are sorted based on both of these values individually. The first 1000 of sorted SNPs based on MDA values were given to *SNPnexus* database in order to figure out the number and the quality of SNPs related with prostate cancer, existing in these SNPs. Then the same procedure was performed on the first 1000 of sorted SNPs based on MDG values. Therefore, this analysis provides the comparison of two different outputs obtained from both individual SNP sets which were sorted according to two distinct importance values. As shown in Table 4.5, while 22 SNPs were obtained from the ranking based on MDA value, 18 SNPs were obtained from the ranking based on MDG value. 14 SNPs are identical in these two sets. The identical SNPs are shown as black and bold in Table 4.5.

Table 4.5 Prostate cancer related SNPs founded in both first 1000 of sorted SNP sets based on MDA values (the list on the left) and MDG values (the list on the right) separately after RF on *SNP-IP*.

	SNPS		SNPS
1	**rs567700**	1	rs553371
2	**rs531572**	2	**rs567700**
3	rs6482743	3	**rs531572**
4	**rs3912492**	4	rs4750759
5	rs6803449	5	rs9971190
6	**rs213386**	6	**rs3912492**
7	rs1380466	7	**rs213386**
8	**rs2622625**	8	**rs2622625**
9	rs6768256	9	rs2567608
10	**rs10776909**	10	**rs10776909**
11	**rs3745540**	11	**rs3745540**
12	rs2360995	12	**rs12882037**
13	**rs12882037**	13	**rs7914154**

14	**rs7914154**	**14**	**rs4267385**
15	rs6970999	**15**	**rs8066276**
16	rs2107280	**16**	**rs4921943**
17	**rs4267385**	**17**	**rs4496973**
18	**rs8066276**	**18**	**rs4870828**
19	**rs4921943**		
20	**rs4496973**		
21	**rs4870828**		
22	rs11737898		

Moreover, the genes hosting these prostate cancer related SNPs can be found easily from the output of SNPnexus analysis. In Table 4.6, the genes that are hosting the prostate cancer related SNPs found in the first 1000 of sorted set of SNPs based on MDA values are shown. In addition the genes that are hosting the prostate cancer related SNPs found in the sorted set of SNPs based on MDG values are given in Table 4.7. It has been observed that, while 22 SNPs that were found to be associated with prostate cancer are located in 13 different gene regions based on the analysis that have been done according to MDA values, 18 SNPs that were found to be associated with prostate cancer are located in 11 different gene regions based on the analysis that has been done according to MDG values. As shown in Table 4.8, 10 genes are shown as black and bold since they are common in both output sets. Based on this observation, we cannot prove certainly that, one of the performances of the analysis that has been done according to both importance values of RF, is better than the other, since there are small differences between the outputs of both analyses. However, we can say that both of the importance values can be used as the main outcome of the *random forest* algorithm.

Table 4.6 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDG values after RF on *SNP-IP*.

	GENES	**SNPS**
1	**MGMT**	rs567700
		rs531572
		rs6482743

2	**FHIT**	rs3912492
		rs6803449
		rs213386
		rs1380466
3	**ABCG2**	rs2622625
4	**EEFSEC**	rs6768256
5	**RXRA**	rs10776909
6	**KLK12**	rs3745540
7	**ESRRB**	rs2360995
		rs12882037
8	**PIK3AP1**	rs7914154
9	**NXPH1**	rs6970999
		rs2107280
10	**ACE**	rs4267385
		rs8066276
		rs4267385
		rs8066276
11	**PSD3**	rs4921943
12	**ZHX2**	rs4496973
		rs4870828
13	**SORCS2**	rs11737898

Table 4.7 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDG values after RF on *SNP-IP*.

	GENES	**SNPS**
1	**MGMT**	rs553371
		rs567700
		rs531572
		rs4750759
		rs9971190
2	**FHIT**	rs3912492
		rs213386
3	**ABCG2**	rs2622625

4	**SSTR4**	rs2567608
5	**RXRA**	rs10776909
6	**KLK12**	rs3745540
7	**ESRRB**	rs12882037
8	**PIK3AP1**	rs7914154
9	**ACE**	rs4267385 rs8066276 rs4267385 rs8066276
10	**PSD3**	rs4921943
11	**ZHX2**	rs4496973 rs4870828

Table 4.8 Comparison of the genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDA values (the list on the left) and MDG values (the list on the right) separately after RF on *SNP-IP*.

	GENES		GENES
1	**MGMT**	**1**	**MGMT**
2	**FHIT**	**2**	**FHIT**
3	**ABCG2**	**3**	**ABCG2**
4	EEFSEC	**4**	SSTR4
5	**RXRA**	**5**	**RXRA**
6	**KLK12**	**6**	**KLK12**
7	**ESRRB**	**7**	**ESRRB**
8	**PIK3AP1**	**8**	**PIK3AP1**
9	NXPH1	**9**	**ACE**
10	**ACE**	**10**	**PSD3**
11	**PSD3**	**11**	**ZHX2**
12	**ZHX2**		
13	SORCS2		

Analysis with 1000 SNPs of *SNP-IP-Ext* after RF

A genotype-phenotype integrated data set was obtained by adding the phenotype information to *SNP-IP* and was annotated as *SNP-IP-Ext*. With the same parameters using in previous Subsection 4.1.3, RF was performed on *SNP-IP-Ext*. As in the previous step two importance values were determined which are MDA and MDG values. After sorting the SNPs based on both of these values individually, first 1000 of both distinct ranking were selected and were given to *SNPnexus* database. We have selected these first 1000 SNPs by not considering the phenotype information which exist at the beginning of the sequence. The reason why we have used phenotype information is that, if these information can affect the out of bag error rate of RF or not and to examine if the output SNPs will be altered with an excessive rate or not. This issue will be discussed in Subsection 5.1.1. The number and the quality of SNPs related with prostate cancer existing in these sets were figured out after the analysis done via SNPnexus. If we want to compare the two rankings, as shown in Table 4.9, while 17 SNPs were obtained from the ranking due to MDA value, 19 SNPs were obtained from the ranking due to MDG value. 10 SNPs are shown as black and bold, since they are common SNPs in these two sets.

Table 4.9 Prostate cancer related SNPs founded in both the first 1000 of sorted SNP sets based on MDA values (the list on the left) and MDG values (the list on the right) separately after RF on *SNP-IP-Ext*.

	SNPS		SNPS
1	rs1877724	1	rs553371
2	**rs531572**	2	rs567700
3	rs9971190	3	**rs531572**
4	**rs3912492**	4	rs4750759
5	rs6803449	5	rs7082319
6	rs12636081	6	**rs3912492**
7	**rs213386**	7	**rs213386**
8	rs620359	8	rs1380466
9	**rs7914154**	9	rs2622625
10	rs6970999	10	rs2567608
11	rs2107280	11	rs10776909

12	**rs4267385**	**12**	rs12882037
13	**rs8066276**	**13**	**rs7914154**
14	**rs4921943**	**14**	**rs4267385**
15	**rs10106032**	**15**	**rs8066276**
16	**rs4496973**	**16**	**rs4921943**
17	**rs4870828**	**17**	**rs10106032**
		18	**rs4496973**
		19	**rs4870828**

Analysis with 100 SNPs after RF on First 1000 SNPs of *SNP-IP*

This part can be accepted as a preliminary study of examining the improvements of performances done on the data set including SNPs with combined *p* values which is annotated as *SNP-CP* throughout this study. These performances will be discussed in detail later.

As described before, the SNPs in *SNP-IP* have been sorted based on associated individual *p* values. Then first 1000 of these SNPs were selected. By performing RF with using the same parameters given in subsection 4.1.3 on these selected 1000 SNPs, MDA and MDG values were obtained for each SNP in this set. Moreover, after RF was performed, the SNPs were sorted based on these two importance values individually. To compare the two rankings, the first 100 of both distinct ordered set were selected and analyzed by using *SNPnexus* database. The output SNPs that are obtained from *SNPnexus* analysis are given in Table 4.10. As shown all of the SNPs are the same in these two sets. Moreover the genes that are hosting these prostate cancer related SNPs are shown in Table 4.11. It has been understood that 3 SNPs are located in 2 different genes. This is the least extensive work through the whole study. Since, this study was assayed that to look for if it will be adequate to be able to do analyses on the smallest data set or not. In the *Discussion* chapter, we will see that the output of this study may mislead the researchers or not.

Table 4.10 Prostate cancer related SNPs founded in both first 100 of sorted SNP sets based on MDA values(the list on the left) and MDG values(the list on the right) separately after RF on the first 1000 SNPs of *SNP-IP*.

	SNPS		SNPS
1	rs12882037	1	rs12882037
2	rs4496973	2	rs4870828
3	rs4870828	3	rs4496973

Table 4.11 The genes hosting the SNPs that were found related to prostate cancer in the first 100 of sorted SNPs based on MDA values(the list on the left) and MDG values(the list on the right) after RF on the first 1000 SNPs of *SNP-IP*.

	GENES	SNPS
1	ESRRB	rs12882037
2	ZHX2	rs4496973
		rs4870828

4.1.2 Outputs of HaploReg and RegulomeDB Analyses

Analysis with 500 non-coding SNPs of *SNP-IP*

After *HaploReg* Analysis have been done on *SNP-IP*, it has found that 1538 SNPs in 2706 have no *dbSNP* function annotation as intronic. This means that, slightly more than half of the SNPs in this data set are not directly located on a gene region, but may affect the genes or proteins indirectly, namely they can be regulatory SNPs. Moreover, it is known that SNPs occur in non-coding regions more frequently than in coding regions. Therefore, our result has supported this statistical reality. As in the previous subsection before selecting desired SNPs, we have sorted those of non-coding SNPs based on associated individual *p* values. Then we have selected the first 500 of this sorted set. In Table 4.12, the SNPs found in the first 500 non-coding SNPS of *SNP-IP*, with *RegulomeDB* scores equal and less than 3, are shown. These 19 SNPs have greater importance compared to other SNPs that have *RegulomeDB* scores higher than 3. We have selected threshold value for this score as 3 since if the *RegulomeDB* score decreases the functional ability of regulatory SNPs is increases.

It can be thought that why we did not use the 1168 SNPs which is the difference between the number of all SNPs and the number of SNPs that have no *dbSNP* function annotation as intronic, to find prostate cancer related genes. Since the aim of this study is not just to improve the performances of the analyses but also to prove that the LD information have exactly effect on the SNP selection procedure.

Table 4.12 The none-coding SNPs with *RegulomeDB* scores equal and less than 3, in the first 500 non-coding SNPs of *SNP-IP*.

	SNPs	RegulomeDB Score
1	rs6708126	2b
2	rs8090231	2b
3	rs12101523	3a
4	rs1762438	1f
5	rs943889	2b
6	rs2063295	2b
7	rs17701543	3a
8	rs10510573	3a
9	rs11751092	2a
10	rs11075236	2b
11	rs888096	2b
12	rs977676	2b
13	rs11253536	2b
14	rs9435409	3a
15	rs1330100	3a
16	rs1379736	2a
17	rs2581717	2b
18	rs9328186	3a
19	rs2062287	3a

4.2 Analyses Performed on the Data Set Including SNPs with Combined *p* Values

From the data set including 2706 SNPs with individual *p* values which was annotated as *SNP-IP*, we have prepared a data set which is consisting of SNPs with combined p values as explained in detail in Subsection 3.1.2. This data set was annotated as *SNP-*

CP, throughout this study. In this part, the whole analyses done on *SNP-CP*, will be explained.

4.2.1 Outputs of SNPnexus Analysis

Analysis with *SNP-CP*

When the data set was organized for obtaining combined *p* value, the number of SNPs in the data set reduced to 2495. Therefore it is expected that the number of SNPs related with prostate cancer will be reduced. The SNPs related with prostate cancer has been found from *SNP-CP* are shown in Table 4.13. After *SNPnexus* analysis has been done on *SNP-CP*, it has proved that, smaller number of prostate cancer related SNPs has been obtained, compared to the number of prostate cancer related SNPs that have been found in *SNP-IP*. However if the success of the analysis is considered as the percentage of the number of SNPs related with prostate cancer, the analysis was accepted as successful as the analysis has done on *SNP-IP*. In addition, the gene regions where these SNPs are located on, are shown in Table 4.13. It has been observed that 52 SNPs that show a significant association with prostate cancer are located on 25 different genes.

Table 4.13 Prostate cancer related SNPs found in *SNP-CP*.

	SNPS		SNPS
1	rs17775610	**27**	rs553371
2	rs8178179	**28**	rs567700
3	rs531572	**29**	rs4750759
4	rs620359	**30**	rs6482743
5	rs3912492	**31**	rs9971190
6	rs9637471	**32**	rs7082319
7	rs6803449	**33**	rs2360995
8	rs12636081	**34**	rs12896434
9	rs17061864	**35**	rs7914154
10	rs12882037	**36**	rs928111
11	rs2999081	**37**	rs2107280
12	rs2811415	**38**	rs4267385
13	rs2811518	**39**	rs7911448

14	rs2811388	**40**	rs13437706
15	rs6798749	**41**	rs1877724
16	rs4921943	**42**	rs10776909
17	rs7844180	**43**	rs11016862
18	rs4496973	**44**	rs3745540
19	rs4870828	**45**	rs2360999
20	rs213386	**46**	rs6768256
21	rs1380466	**47**	rs6970999
22	rs4665716	**48**	rs8066276
23	rs4908107	**49**	rs10106032
24	rs1266890	**50**	rs11737898
25	rs927188	**51**	rs9341218
26	rs1234220	**52**	rs2567608

Table 4.14 Genes hosting the SNPs that were found related to prostate cancer in *SNP-CP*.

	GENES	SNPS		GENES	SNPS
1	**PVT1**	rs17775610	**14**	**PIK3AP1**	rs7914154
2	**PRKDC**	rs8178179	**15**	**COL4A2**	rs928111
3	**MGMT**	rs531572 rs553371 rs567700 rs4750759 rs6482743 rs9971190 rs7082319 rs11016862	**16**	**NXPH1**	rs7914154 rs2107280 rs6970999
4	**C2ORF43**	rs620359	**17**	**ACE**	rs4267385 rs8066276
5	**FHIT**	rs3912492 rs9637471 rs6803449	**18**	**CAMK1D**	rs7911448

		rs12636081 rs17061864 rs213386 rs1380466			
6	**ESRRB**	rs12882037 rs2360995 rs12896434 rs2360999	**19**	**CREB5**	rs13437706
7	**EEFSEC**	rs2999081 rs2811415 rs2811518 rs2811388 rs6798749 rs6768256	**20**	**EPHX1**	rs1877724
8	**PSD3**	rs4921943 rs7844180 rs10106032	**21**	**RXRA**	rs10776909
9	**ZHX2**	rs4496973 rs4870828	**22**	**KLK12**	rs3745540
10	**NCOA1**	rs4665716	**23**	**SORCS2**	rs11737898
11	**SLC30A7**	rs4908107	**24**	**IGFBP2**	rs9341218
12	**PKHD1**	rs1266890 rs927188	**25**	**SSTR4**	rs2567608
13	**PTEN**	rs1234220			

Analysis with 1000 SNPs of *SNP-CP*

According to combined *p* values that we have measured, the SNPs associated with those *p* values were sorted. To narrow the region to be analyzed, we have selected first 1000 SNPs from the ranking which is based on the combined *p* values. *SNPnexus* analysis has been done on first 1000 SNPs of sorted *SNP-CP*. The output SNPs are shown in Table 4.15 with associated rsIDs. It has been observed, there are 23 SNPs that show a significant association with prostate cancer located on 6 different gene regions as given in Table 4.16.

Table 4.15 Prostate cancer related SNPs founded in first 1000 of sorted *SNP-CP*.

	SNPS
1	rs1234220
2	rs553371
3	rs567700
4	rs531572
5	rs4750759
6	rs6482743
7	rs9971190
8	rs7082319
9	rs11016862
10	rs3912492
11	rs9637471
12	rs6803449
13	rs12636081
14	rs17061864
15	rs1380466
16	rs2999081
17	rs2811415
18	rs2811518
19	rs2811388
20	rs6798749
21	rs12896434
22	rs2360999
23	rs10106032

Table 4.16 Genes hosting the SNPs found related to prostate cancer in the first 1000 SNPs of sorted *SNP-CP*.

	GENES	SNPS
1	PTEN	rs1234220
2	MGMT	rs553371 rs567700 rs531572

		rs4750759
		rs6482743
		rs9971190
		rs7082319
		rs11016862
3	**FHIT**	rs3912492
		rs9637471
		rs6803449
		rs1263608
		rs1706186
		rs1380466
4	**EEFSEC**	rs2999081
		rs2811415
		rs2811518
		rs2811388
		rs6798749
5	**ESRRB**	rs12896434
		rs2360999
6	**PSD3**	rs10106032

Analysis with 1000 SNPs of *SNP-CP* after RF

The data set including 2495 SNPs with combined *p* values which was annotated as *SNP-CP*, was given to RF as an input. By using the same parameters that were given in Subsection 4.1.1, the input SNPs have been associated with the importance values which are MDA and MDG values explained before. After sorting the SNPs based on the importance values individually, initially the first 1000 of sorted SNPs based on MDA values were given to *SNPnexus* database in order to figure out the qualifications of SNPs related with prostate cancer existing in this set. Then, the same procedure was performed on the first 1000 of sorted SNPs based on mean MDG values.

As shown in Table 4.17, while 23 SNPs were obtained from the ranking based on MDA values, from the ranking based on MDG values, 18 SNPs were obtained. 13

SNPs are identical in these two sets. These identical SNPs are shown as black and bold in Table 4.17.

Table 4.17 Prostate cancer related SNPs founded in both the first 1000 of sorted SNP sets based on MDA values (the list on the left) and MDG values (the list on the right) separately after RF on *SNP-CP*.

	SNPS		SNPs
1	**rs553371**	1	rs1877724
2	**rs567700**	2	**rs553371**
3	**rs4750759**	3	**rs567700**
4	rs6482743	4	rs531572
5	**rs3912492**	5	**rs4750759**
6	rs12636081	6	rs9971190
7	**rs213386**	7	rs7082319
8	rs1380466	8	**rs3912492**
9	rs2999081	9	**rs213386**
10	rs2811518	10	rs2567608
11	rs6768256	11	**rs12882037**
12	rs10776909	12	**rs7914154**
13	rs3745540	13	**rs4267385**
14	**rs12882037**	14	**rs8066276**
15	**rs7914154**	15	**rs4921943**
16	rs928111	16	**rs10106032**
17	**rs4267385**	17	**rs4496973**
18	**rs8066276**	18	**rs4870828**
19	**rs4921943**		
20	**rs10106032**		
21	**rs4496973**		
22	**rs4870828**		
23	rs11737898		

Moreover the genes that are hosting the prostate cancer related SNPs found in the sorted set of SNPs due to MDA values are shown in Table 4.18. In addition the genes that are hosting the prostate cancer related SNPs found in the sorted set of SNPs due to MDG values are given in Table 4.19. It has viewed that, while 23 SNPs that were found to be associated with prostate cancer are located in 12 different gene regions based on the analysis that have been done according to MDA values, 18 SNPs that were found to be associated with prostate cancer are located in 9 different gene regions based on the analysis that have been done according to MDG values. As shown in Table 4.20, 7 genes are shown as black and bold which means they are common in both output sets. Based on this observation, we cannot prove certainly that, one of the performances of the analysis which have been done according to both importance values of RF, is better than the other.

Table 4.18 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDA values after RF on *SNP-CP*.

	GENES	SNPS
1	**MGMT**	rs553371
		rs567700
		rs4750759
		rs6482743
2	**FHIT**	rs3912492
		rs12636081
		rs213386
		rs1380466
3	**EEFSEC**	rs2999081
		rs2811518
		rs6768256
4	**RXRA**	rs10776909
5	**KLK12**	rs3745540
6	**ESRRB**	rs12882037
7	**PIK3AP1**	rs7914154
8	**COL4A2**	rs928111
9	**ACE**	rs4267385
		rs8066276
		rs4267385
		rs8066276

10	**PSD3**	rs4921943
		rs10106032
11	**ZHX2**	rs4496973
		rs4870828
12	**SORCS2**	rs11737898

Table 4.19 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDG values after RF on *SNP-CP*.

	GENES	**SNPS**
1	**EPHX1**	rs1877724
2	**MGMT**	rs553371
		rs567700
		rs531572
		rs4750759
		rs9971190
		rs7082319
3	**FHIT**	rs3912492
		rs213386
4	**SSTR4**	rs2567608
5	**ESRRB**	rs12882037
6	**PIK3AP1**	rs7914154
7	**ACE**	rs4267385
		rs8066276
		rs4267385
		rs8066276
8	**PSD3**	rs4921943
		rs10106032
9	**ZHX2**	rs4496973
		rs4870828

Table 4.20 Comparison of the genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted SNPs based on MDA values (the list on the left) and MDG values (the list on the right) separately after RF on *SNP-CP*.

	GENES		GENES
1	**MGMT**	1	EPHX1
2	**FHIT**	2	**MGMT**
3	EEFSEC	3	**FHIT**
4	RXRA	4	SSTR4
5	KLK12	5	**ESRRB**
6	**ESRRB**	6	**PIK3AP1**
7	**PIK3AP1**	7	**ACE**
8	COL4A2	8	**PSD3**
9	**ACE**	9	**ZHX2**
10	**PSD3**		
11	**ZHX2**		
12	SORCS2		

Analysis with 100 SNPs after RF on the First 1000 SNPs of *SNP-CP*

In this step of this study, we will further reduce the number of SNPs that will be analyzed. As given in Subsection 4.2.1 the SNPs in *SNP-CP* have been sorted based on associated combined *p* values. Then first 1000 of these SNPs were selected. In the following step, RF was performed on these selected 1000 SNPs, by using the same parameters given in Subsection 4.1.3. MDA and MDG values were obtained for each SNP in this set as in previous sections related with RF. The SNPs were sorted based on these two importance values individually. The first 100 SNPs of both distinct ordered sets were selected and analyzed by using *SNPnexus* database in order to compare two rankings. However we could not find any prostate cancer related SNPs in both of the sorted sets. Therefore analyses done on this small subset of the entire set shows us that, when trying to generalize the results, we need larger subsets of the original data set. Sometimes small subsets may be insufficient to be analyzed and to lead the further studies.

Analysis with *SNP-Filtered*

As described in subsection 3.1.4, we have a small subset of SNPs which provides us to obtain biologically relevant SNPs for prostate cancer. This data set can be described as representative set of the whole SNPs in the original data set. It was obtained by eliminating the SNPs which can be accepted as redundant based on the assumption that, one of the representative SNPs can be able to carry biological information that is also inferred from those redundant SNPs.

SNPnexus analysis has been done on all of the SNPs in *SNP-Filtered*. The number of SNPs has been found associated with prostate cancer calculated as 33 through the whole of 1758 SNPs in this data set. Those 33 SNPs are shown in Table 4.21 with associated rsIDs. In addition the genes where these SNPs are located on are shown in Table 4.22. It has observed that 33 SNPs that show a significant association with prostate cancer are located on 23 different gene regions.

Table 4.21 Prostate cancer related SNPs found in *SNP-Filtered*.

	SNPS		SNPS
1	rs17775610	**18**	rs2567608
2	rs4908107	**19**	rs10776909
3	rs8178179	**20**	rs927188
4	rs531572	**21**	rs1234220
5	rs7082319	**22**	rs11016862
6	rs620359	**23**	rs3745540
7	rs3912492	**24**	rs213386
8	rs9637471	**25**	rs1380466
9	rs12636081	**26**	rs2360995
10	rs12896434	**27**	rs7914154
11	rs12882037	**28**	rs6768256
12	rs928111	**29**	rs2107280
13	rs4267385	**30**	rs7911448
14	rs4921943	**31**	rs10106032
15	rs7844180	**32**	rs13437706
16	rs4870828	**33**	rs11737898
17	rs4665716		

Table 4.22 Genes hosting the SNPs that were found related to prostate cancer in *SNP-Filtered*.

	GENES	SNPS
1	PVT1	rs17775610
2	SLC30A7	rs4908107
3	PRKDC	rs8178179
4	MGMT	rs531572 rs7082319 rs11016862
5	C2ORF43	rs620359
6	FHIT	rs3912492 rs9637471 rs12636081 rs213386 rs1380466
7	ESRRB	rs12896434 rs12882037 rs2360995
8	COL4A2	rs928111
9	ACE	rs4267385
10	PSD3	rs4921943 rs7844180 rs10106032
11	ZHX2	rs4870828
12	NCOA1	rs4665716
13	SSTR4	rs2567608
14	RXRA	rs10776909
15	PKHD1	rs927188
16	PTEN	rs1234220
17	KLK12	rs3745540
18	PIK3AP1	rs7914154
19	EEFSEC	rs6768256
20	NXPH1	rs2107280

21	**CAMK1D**	rs7911448
22	**CREB5**	rs13437706
23	**SORCS2**	rs11737898

4.2.2 Outputs of HaploReg and RegulomeDB Analyses

Analysis with 500 non-coding SNPS of *SNP-CP*

The aim of this step is to focus on SNPs that are not located on gene regions namely, noncoding regions. We have found from the output of *HaploReg* analysis, 1419 SNPs in *SNP-CP*, have no *dbSNP* function annotation as intronic. This means that, more than half of the SNPs in this data set are not located on a gene region, but may have effect on the genes or proteins indirectly. The SNPs with *RegulomeDB* scores equal and less than 3, in the first 500 SNPS of *SNP-CP*, are given in Table 4.23. 27 SNPs have been found with desired *RegulomeDB* scores. These 27 SNPs are more valuable when compared to other SNPs that have *RegulomeDB* scores higher than 3.

Table 4.23 The none-coding SNPs with *RegulomeDB* scores equal and less than 3, in the first 500 non-coding SNPs of *SNP-CP*.

	SNPs	**RegulomeDB Score**
1	rs2832093	3a
2	rs6714287	3a
3	rs304951	1d
4	rs7136770	1f
5	rs10506347	1f
6	rs12998237	2a
7	rs1781079	2b
8	rs12101523	3a
9	rs17701543	3a
10	rs10027556	3a
11	rs9897342	2b
12	rs8090231	2b
13	rs11253536	2b

14	rs2765895	3a
15	rs6894361	3a
16	rs422945	1d
17	rs531805	1f
18	rs6944602	1f
19	rs11696842	1f
20	rs12910685	2b
21	rs6708126	2b
22	rs1636579	3a
23	rs16960555	1f
24	rs1379736	2a
25	rs986046	2b
26	rs977676	2b
27	rs17264915	3a

CHAPTER 5

DISCUSSION

The aim of this thesis is to demonstrate that if there exist any valuable impact of LD on SNP prioritization studies or not. To prove a new algorithm has been constructed and tested on prostate cancer data set downloaded from dbGaP.

In Chapter 5,the result of the analysis, given in the Chapter 4, will be discussed.

5.1 Discussion on SNPnexus Results

5.1.1 Analyses with 1000 SNPs of *SNP-IP* and 1000 SNPs of *SNP-IP-Ext* after RF

In this step, the outputs, which were obtained from the *random forest* algorithm, performed on both *SNP-IP* and *SNP-IP-Ext,* have been compared. In the Subsection 4.1.1, essential *SNPnexus* analysis has been done on the first 1000 SNPs of mentioned SNP sets. A comparison of the results obtained based on MDA value is shown in Table 5.1. 22 SNPs have been obtained from first 1000 of the ordered SNP set based on MDA values which were obtained from the outputs of RF performed on SNP-IP. In addition, 17 SNPs have been obtained from first 1000 of the ordered SNP set based on mean decrease accuracy values which were obtained from the outputs of RF performed on *SNP-IP-Ext*. 12 SNPs are identical in these two sets.

In addition, a comparison of the results obtained based on MDG values is shown in Table 5.2. 18 SNPs have been obtained from the first 1000 of the ordered SNP set based on MDG values which were obtained from the outputs of RF performed on *SNP-IP*. Moreover, 19 SNPs have been obtained from the first 1000 of the ordered SNP set based on MDG values which were obtained from the outputs of RF performed on *SNP-IP-Ext*. 16 SNPs are the same in these two set.

The same parameters have been used for both of the performances done on the data sets *SNP-IP* and *SNP-IP-Ext*. However, error rates of the both performances were

determined very close to each other, as shown in Table 5.3. In Table 5.4 and 5.5, confusion matrices of both performances are given. These matrices list the classes and how the RF classified each one, plus the classification error for each. By using the values from these tables the OOB estimate of error rate can be calculated as shown in Eqn. (5.1). Although the error rate of RF performance *SNP-IP-Ext* is a bit larger than the other, it does not mean that phenotype information is redundant. It can be said that the phenotype information is not necessary for this study. Moreover, the determined rankings were very similar in these two performances. Although phenotype information variables were on the top of the rankings in the output of performance with *SNP-IP-Ext*, when phenotype information was taken out, most of the remaining sorted SNPs are as most of the leading SNPs in the output of performance with *SNP-IP*.

Table 5.1 Prostate cancer related SNPs founded in both the first 1000 of sorted SNP sets based on MDA values after RF on *SNP-IP* (the list on the left) and *SNP-IP-Ext* (the list on the right) individually.

	SNPS		SNPS
1	rs567700	1	rs1877724
2	**rs531572**	2	**rs531572**
3	rs6482743	3	rs9971190
4	**rs3912492**	4	**rs3912492**
5	**rs6803449**	5	**rs6803449**
6	**rs213386**	6	rs12636081
7	rs1380466	7	**rs213386**
8	rs2622625	8	rs620359
9	rs6768256	9	**rs7914154**
10	rs10776909	10	**rs6970999**
11	rs3745540	11	**rs2107280**
12	rs2360995	12	**rs4267385**
13	rs12882037	13	**rs8066276**
14	**rs7914154**	14	**rs4921943**

15	rs6970999	15	rs10106032
16	rs2107280	16	rs4496973
17	rs4267385	17	rs4870828
18	rs8066276		
19	rs4921943		
20	rs4496973		
21	rs4870828		
22	rs11737898		

Table 5.2 Prostate cancer related SNPs founded in both the first 1000 of sorted SNP sets based on MDG values after RF on *SNP-IP* (the list on the left) and *SNP-IP-Ext* (the list on the right) individually.

	SNPS		SNPS
1	rs553371	1	rs553371
2	rs567700	2	rs567700
3	rs531572	3	rs531572
4	rs4750759	4	rs4750759
5	rs9971190	5	rs7082319
6	rs3912492	6	rs3912492
7	rs213386	7	rs213386
8	rs2622625	8	rs1380466
9	rs2567608	9	rs2622625
10	rs10776909	10	rs2567608
11	rs3745540	11	rs10776909
12	rs12882037	12	rs12882037
13	rs7914154	13	rs7914154
14	rs4267385	14	rs4267385
15	rs8066276	15	rs8066276
16	rs4921943	16	rs4921943
17	rs4496973	17	rs10106032
18	rs4870828	18	rs4496973
		19	rs4870828

Table 5.3 OOB estimate of error rate of both performances.

Data Sets	OOB estimate of error rate
SNP-IP	23.81%
SNP-IP-Ext	24.76%

Table 5.4 Confusion matrix of performance with *SNP-IP*.

	1	2	Class error
1	477	151	0.2404459
2	149	483	0.2357595

Table 5.5 Confusion matrix of performance with *SNP-IP-Ext*.

	1	2	Class error
1	481	147	0.2340764
2	165	467	0.2610759

$$\text{OOB estimate error rate} = \frac{\text{\# of wrongly classified SNPs}}{\text{\# of whole SNPs}} \qquad \textbf{(5.1)}$$

5.1.2 Analyses with *SNP-IP*, *SNP-CP* and *SNP-Filtered*

Throughout this study we have obtained three main data sets from the original data set. One of them is the data set including SNPs with individual *p* values which is annotated as *SNP-IP*, the other one is the data set containing SNPs with combined *p* values which is annotated as *SNP-CP* and the last one is the data set that can be accepted as representative set of the two of other SNP sets which is also annotated as *SNP-Filtered*. By performing *SNPnexus* analysis on these data sets we have found the SNPs which are given in Table 5.6 as prostate cancer related SNPs. As expected, while the number of analyzed SNPs decreases, the amount of SNPs found as prostate cancer related, decreases. However, it has proven that when we have calculated the ratio of the prostate cancer related SNPs to the associated data sets, all the three proportion values are close to each other. By considering the proportion values as success rates which are given in Table 5.7, we can say that the results of all the analyses are as successful as each other.

Table 5.6 Prostate cancer related SNPs found in *SNP-IP*, *SNP-CP* and *SNP-Filtered*.

SNPS in *SNP-IP*		**SNPS in *SNP-CP***		**SNPS in *SNP-Filtered***	
1	rs12329598	**1**	rs17775610	**1**	rs17775610
2	rs531572	**2**	rs8178179	**2**	rs4908107
3	rs3912492	**3**	rs531572	**3**	rs8178179
4	rs9637471	**4**	rs620359	**4**	rs531572
5	rs6803449	**5**	rs3912492	**5**	rs7082319
6	rs12636081	**6**	rs9637471	**6**	rs620359
7	rs17061864	**7**	rs6803449	**7**	rs3912492
8	rs8178179	**8**	rs12636081	**8**	rs9637471
9	rs17775610	**9**	rs17061864	**9**	rs12636081
10	rs2999081	**10**	rs12882037	**10**	rs12896434
11	rs2811415	**11**	rs2999081	**11**	rs12882037
12	rs2811518	**12**	rs2811415	**12**	rs928111
13	rs2811388	**13**	rs2811518	**13**	rs4267385
14	rs6798749	**14**	rs2811388	**14**	rs4921943
15	rs3764880	**15**	rs6798749	**15**	rs7844180
16	rs620359	**16**	rs4921943	**16**	rs4870828
17	rs12882037	**17**	rs7844180	**17**	rs4665716
18	rs4921943	**18**	rs4496973	**18**	rs2567608
19	rs7844180	**19**	rs4870828	**19**	rs10776909
20	rs4496973	**20**	rs213386	**20**	rs927188
21	rs4870828	**21**	rs1380466	**21**	rs1234220
22	rs16998751	**22**	rs4665716	**22**	rs11016862
23	rs6482743	**23**	rs4908107	**23**	rs3745540
24	rs9971190	**24**	rs1266890	**24**	rs213386
25	rs7082319	**25**	rs927188	**25**	rs1380466
26	rs213386	**26**	rs1234220	**26**	rs2360995
27	rs1380466	**27**	rs553371	**27**	rs7914154
28	rs4665716	**28**	rs567700	**28**	rs6768256
29	rs2622625	**29**	rs4750759	**29**	rs2107280
30	rs209998	**30**	rs6482743	**30**	rs7911448
31	rs4908107	**31**	rs9971190	**31**	rs10106032

32	rs927188	**32**	rs7082319	**32**	rs13437706
33	rs553371	**33**	rs2360995	**33**	rs11737898
34	rs2360995	**34**	rs12896434		
35	rs12896434	**35**	rs7914154		
36	rs7914154	**36**	rs928111		
37	rs928111	**37**	rs2107280		
38	rs4267385	**38**	rs4267385		
39	rs7911448	**39**	rs7911448		
40	rs13437706	**40**	rs13437706		
41	rs1234220	**41**	rs1877724		
42	rs4750759	**42**	rs10776909		
43	rs1877724	**43**	rs11016862		
44	rs2567608	**44**	rs3745540		
45	rs10776909	**45**	rs2360999		
46	rs1266890	**46**	rs6768256		
47	rs567700	**47**	rs6970999		
48	rs11016862	**48**	rs8066276		
49	rs3745540	**49**	rs10106032		
50	rs2360999	**50**	rs11737898		
51	rs7089141	**51**	rs9341218		
52	rs6768256	**52**	rs2567608		
53	rs6970999				
54	rs2107280				
55	rs8066276				
56	rs10106032				
57	rs11737898				

Table 5.7 Comparison of success rates of *SNP-IP*, *SNP-CP* and *SNP-Filtered*.

	The number of SNPs related	**The number of whole SNPs**	
Data sets	**with prostate cancer**	**Success rate**	
SNP-IP	57	2706	%2.1
SNP-CP	52	2495	%2.08
SNP-Filtered	33	1758	%1.8

In Table 5.8, the genes hosting the prostate cancer related SNPs are given. Common genes are shown as black and bold. As expected all of the genes hosting the prostate cancer related SNPs that were found from *SNP-Filtered*, exist in other two sets. Our experiment has shown that, with using the SNPs which are able to represent other SNPs (*SNP-Filtered*) instead of using *SNP-IP*, while the number of genes hosting the SNPs related to prostate cancer were not changed as to be taken into account, the number of SNPs located on those genes were decreased. Therefore, while validating the biologically relevant results, the experimental cost has been decreased to a relatively significant level. As shown, while the number of SNPs decreased from 57 to 33, the number of genes hosting those SNPs decreased from 29 to 23, so that almost half of the SNPs have been eliminated.

Table 5.8 Genes hosting the SNPs that were found related to prostate cancer in *SNP-IP* (the list on the left), *SNP-CP* (the list in the middle) and *SNP-Filtered* (the list on the right).

GENES in *SNP-IP*		**GENES in *SNP-CP***		**GENES in *SNP-Filtered***	
1	EPHX1	**1**	**PVT1**	**1**	**PVT1**
2	BCAS1	**2**	**PRKDC**	**2**	**SLC30A7**
3	**PTEN**	**3**	**MGMT**	**3**	**PRKDC**
4	**NCOA1**	**4**	**C2ORF43**	**4**	**MGMT**
5	ABCG2	**5**	**FHIT**	**5**	**C2ORF43**
6	**MGMT**	**6**	**ESRRB**	**6**	**FHIT**
7	AIFM1	**7**	**EEFSEC**	**7**	**ESRRB**
8	**EEFSEC**	**8**	**PSD3**	**8**	**COL4A2**
9	**SLC30A7**	**9**	**ZHX2**	**9**	**ACE**
10	TLR8	**10**	**NCOA1**	**10**	**PSD3**
11	**SSTR4**	**11**	**SLC30A7**	**11**	**ZHX2**
12	**RXRA**	**12**	**PKHD1**	**12**	**NCOA1**
13	**PKHD1**	**13**	**PTEN**	**13**	**SSTR4**
14	**PRKDC**	**14**	**PIK3AP1**	**14**	**RXRA**
15	**C2ORF43**	**15**	**COL4A2**	**15**	**PKHD1**
16	**KLK12**	**16**	**NXPH1**	**16**	**PTEN**
17	**FHIT**	**17**	**ACE**	**17**	**KLK12**
18	**ESRRB**	**18**	**CAMK1D**	**18**	**PIK3AP1**

19	**PIK3AP1**	19	**CREB5**	19	**EEFSEC**
20	CUBN	20	**EPHX1**	20	**NXPH1**
21	**COL4A2**	21	**RXRA**	21	**CAMK1D**
22	**NXPH1**	22	**KLK12**	22	**CREB5**
23	**ACE**	23	**SORCS2**	23	**SORCS2**
24	**CAMK1D**	24	**IGFBP2**		
25	**PSD3**	25	**SSTR4**		
26	**ZHX2**				
27	**CREB5**				
28	**SORCS2**				
29	**PVT1**				

5.1.3 Analyses with 1000 SNPs of *SNP-IP* and 1000 SNPs of *SNP-CP*

The main reason why we have started this study is that, the foregoing discussion on the significance level of LD in SNP selection studies. By using the result which were given in the Subsection 4.1.1, we are able to compare the performances of both data set *SNP-IP* and *SNP-CP* which are based on individual p values and combined p values, respectively. As mentioned above, the most important step of the algorithm for obtaining combine p values is that, obtaining a list of SNPs which are expected to be in LD with the associated SNP, according to the given threshold value for r^2. Thus we have claimed that, if we obtain associated combined p values for each SNP in *SNP-IP*, we can be able to demonstrate the impact of LD in such kinds of SNP selection studies. In Table 5.9, prostate cancer related SNPs that were found in both first 1000 of sorted SNP sets based on individual p values and combined p values, are given. As shown, the number of SNPs in two lists are close to each other, however, only 11 SNPs are identical which is about half of one of the set. On the other hand, if we examine the genes hosting the prostate cancer related SNPs in both sets individually, the number of genes decreases, in the second ranking which is based on combined p value. These genes are given in In Table 5.10 and common genes are shown as black and bold. As shown in the table, almost all of the genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted *SNP-CP*, are common in both sets. This result may lead us to the conclusion that, cost effective gene filtration can be done based on our claim which is the reason why we have started to this study.

Table 5.9 Prostate cancer related SNPs that were found in both first 1000 of sorted SNP sets based on individual *p* values (*SNP-IP*) (the list on the left) and combined *p* values (*SNP-CP*) (the list on the right).

	SNPS		SNPS
1	rs12329598	1	rs1234220
2	**rs531572**	2	rs553371
3	**rs3912492**	3	rs567700
4	**rs9637471**	4	**rs531572**
5	**rs6803449**	5	rs4750759
6	**rs12636081**	6	rs6482743
7	**rs17061864**	7	rs9971190
8	rs8178179	8	rs7082319
9	rs17775610	9	rs11016862
10	**rs2999081**	10	**rs3912492**
11	**rs2811415**	11	**rs9637471**
12	**rs2811518**	12	**rs6803449**
13	**rs2811388**	13	**rs12636081**
14	**rs6798749**	14	**rs17061864**
15	rs3764880	15	rs1380466
16	rs620359	16	**rs2999081**
17	rs12882037	17	**rs2811415**
18	rs4921943	18	**rs2811518**
19	rs7844180	19	**rs2811388**
20	rs4496973	20	**rs6798749**
21	rs4870828	21	rs12896434
		22	rs2360999
		23	rs10106032

Table 5.10 Genes hosting the SNPs that were found related to prostate cancer in the first 1000 of sorted *SNP-IP* (the list on the left) and in the first 1000 of sorted *SNP-CP* (the list on the right).

	GENES		GENES
1	BCAS1	1	PTEN
2	**MGMT**	2	**MGMT**
3	**FHIT**	3	**FHIT**
4	PRKDC	4	**EEFSEC**
5	PVT1	5	**ESRRB**
6	**EEFSEC**	6	**PSD3**
7	TLR8		
8	C2ORF43		
9	**ESRRB**		
10	**PSD3**		
11	ZHX2		

5.1.4 Analyses with 1000 SNPs of *SNP-IP* Before and After RF

To figure out the effect of RF on the prostate cancer data set, *SNPnexus* analyzes have been done on this data set before and after RF was performed. After ordering the SNPs based on individual p values, the first 1000 SNPs were selected and analyzed via *SNPnexus* database. In addition, after performing *random forest* algorithm on the whole SNPs, the importance values with associated SNPs have been obtained. Then these SNPs were sorted based on mean decrease accuracy and mean decrease Gini values individually. The first 1000 SNPs were selected and analyzed via *SNPnexus* database as done before RF was performed. The outputs of all these performances are given in the Subsection 4.1.1. We will just bring the all results together to compare and to comment on them.

A comparison between prostate cancer related SNPs that were found in the ordered SNP set based on individual p values and in the ordered SNP set based on MDA values, is shown in Table 5.11. 21 SNPs have been found from the first 1000 of the ordered SNP set due to individual p values, while 22 SNPs have been found from first 1000 of the ordered SNP set based on MDA values. The number of identical SNPs in these two sets is 5 which is considerably low. Moreover, if we compare the genes hosting the SNPs in both SNP lists, it is shown that, while they are numerically similar, one of them has 11 genes and the other one has 13 genes, only 6 genes are common in both (Table 5.13). Therefore, after RF was performed, the SNPs and the genes that will be focused, have altered.

The comparison between prostate cancer related SNPs found in the ordered SNP set based on individual *p* values and in the ordered SNP set based on MDG values, is shown in Table 5.12. 21 SNPs have been found from the first 1000 of the ordered SNP set based on individual *p* values as mentioned above, while 18 SNPs have been found from first 1000 of the ordered SNP set based on MDG values. 6 SNPs are identical in these two sets. As a result, the number of identical SNPs existing in both SNP lists is relatively small. In addition, the number of genes hosting these SNPs in both lists is the same, however again the amount of identical genes is not enough to be analyzed (Table 5.14). Therefore after RF was performed, as in the comparison has done between the lists of SNPs based on individual *p* values and MDA values, the SNPs and the genes that will be focused, have altered, proving that there exist some effects of RF which are needed to be clarified.

SNPs with rsIDs rs531572, rs3912492, rs12882037 and rs4921943 are common in all sets ordered in different three ways which are based on individual *p* values, MDA and MDG values. Moreover, there exist five common genes which are ZHX2, PSD3, ESRRB, FHIT, MGMT in all sets. In further studies, when we will apply all analyses on different data sets, we will be able to make more precise comments about this topic.

Table 5.11 Prostate cancer related SNPs that were found in both first 1000 of sorted SNP sets based on individual *p* values (the list on the left) and MDA values (the list on the right).

	SNPS		**SNPS**
1	rs12329598	**1**	rs567700
2	**rs531572**	**2**	**rs531572**
3	**rs3912492**	**3**	rs6482743

4	rs9637471	4	**rs3912492**
5	**rs6803449**	5	**rs6803449**
6	rs12636081	6	rs213386
7	rs17061864	7	rs1380466
8	rs8178179	8	rs2622625
9	rs17775610	9	rs6768256
10	rs2999081	10	rs10776909
11	rs2811415	11	rs3745540
12	rs2811518	12	rs2360995
13	rs2811388	13	**rs12882037**
14	rs6798749	14	rs7914154
15	rs3764880	15	rs6970999
16	rs620359	16	rs2107280
17	**rs12882037**	17	rs4267385
18	**rs4921943**	18	rs8066276
19	rs7844180	19	**rs4921943**
20	rs4496973	20	rs4496973
21	rs4870828	21	rs4870828
		22	rs11737898

Table 5.12 Prostate cancer related SNPs that were found in both first 1000 of sorted SNP sets based on individual *p* values (the list on the left) and MDG values (the list on the right).

	SNPS		SNPS
1	rs12329598	1	rs553371
2	**rs531572**	2	rs567700
3	**rs3912492**	3	**rs531572**
4	rs9637471	4	rs4750759
5	rs6803449	5	rs9971190
6	rs12636081	6	**rs3912492**
7	rs17061864	7	rs213386
8	rs8178179	8	rs2622625

9	rs17775610	9	rs2567608
10	rs2999081	10	rs10776909
11	rs2811415	11	rs3745540
12	rs2811518	12	**rs12882037**
13	rs2811388	13	rs7914154
14	rs6798749	14	rs4267385
15	rs3764880	15	rs8066276
16	rs620359	16	**rs4921943**
17	**rs12882037**	17	**rs4496973**
18	**rs4921943**	18	**rs4870828**
19	rs7844180		
20	**rs4496973**		
21	**rs4870828**		

Table 5.13 Genes hosting the SNPs that were found related to prostate cancer in the both the first 1000 of sorted SNP sets based on individual *p* values (the list on the left) and MDA values (the list on the right).

	GENES		GENES
1	BCAS1	1	**MGMT**
2	**MGMT**	2	**FHIT**
3	**FHIT**	3	ABCG2
4	PRKDC	4	**EEFSEC**
5	PVT1	5	RXRA
6	**EEFSEC**	6	KLK12
7	TLR8	7	**ESRRB**
8	C2ORF43	8	PIK3AP1
9	**ESRRB**	9	NXPH1
10	**PSD3**	10	ACE
11	**ZHX2**	11	**PSD3**
		12	**ZHX2**
		13	SORCS2

Table 5.14 Genes hosting the SNPs that were found related to prostate cancer in the both first 1000 of sorted SNP sets based on individual p values (the list on the left) and MDG values (the list on the right).

	GENES		GENES
1	BCAS1	1	**MGMT**
2	**MGMT**	2	**FHIT**
3	**FHIT**	3	ABCG2
4	PRKDC	4	SSTR4
5	PVT1	5	RXRA
6	EEFSEC	6	KLK12
7	TLR8	7	**ESRRB**
8	C2ORF43	8	PIK3AP1
9	**ESRRB**	9	ACE
10	**PSD3**	10	**PSD3**
11	**ZHX2**	11	**ZHX2**

5.1.5 Analyses with 1000 SNPs of *SNP-CP* Before and After RF

It is known that the data set including SNPs with combined *p* values which is annotated as *SNP-CP*, has been obtained from the date set including SNPs with individual *p* values which is annotated as *SNP-IP*. Therefore, after RF was performed on *SNP-CP*, we expect to achieve similar results which were obtained in the previous section. To prove this expectation, first, *SNPnexus* analyzes have been done on this data set before and after RF was performed. The outputs of the related analyses are given in the Subsection 4.2. As in the previous section we will bring them together to investigate.

The comparison between prostate cancer related SNPs that were found in the ordered SNP set based on combined *p* values and in the ordered set of SNPs based on MDA values, is shown in Table 5.15. As it can be inferred from the table, in both of the sets, although the number of prostate cancer related SNPs is same, the SNPs in the two lists are not exactly identical. The number of common SNPs in these two sets is 9 which is considerably low again, as expected. Moreover, we compared the genes hosting the SNPs in both SNP lists, as shown in Table 5.17. After RF was performed, we were able to obtain nearly twice the number of genes. 5 genes are identical in these two sets which means, almost all of the genes hosting prostate cancer related SNPs in the first 1000 of the ordered SNP set based on combined *p* values, can be obtained after RF was performed. Thus, although the SNPs that will be focused on, have changed to be taken into consideration after RF was performed, the genes that will be focused in further studies have not altered drastically, but numerically increased.

The comparison between prostate cancer related SNPs found in the ordered SNP set based on combined *p* values and in the ordered SNP set based on MDG values is shown in Table 5.16. 23 SNPs have been found from the first list while 18 SNPs have been found from the second list. 8 SNPs are identical in these two sets. The number of identical SNPs existing in both SNP lists is relatively small. In addition, the hosting genes of these SNPs are shown in Table 5.18. As shown in this table, the similarities between both lists do not provide sufficient information to be able to infer reliable results. So that, we cannot say one of them has better results. However, it can be seen obviously that, after RF was performed, while the SNPs that will be focused have altered as in above analysis, the genes that will be focused, have not altered so much but numerically increased.

SNPs with rsIDs rs553371, rs567700, rs4750759, rs3912492 and rs10106032 are common in all sets ordered in different three ways which are based on combined *p* values, MDA and MDG values. Moreover, there exist five common genes which are MGMT, FHIT, ESRRB, PSD3 in all sets that were analyzed in this Subsection 5.1.5.

Table 5.15 Prostate cancer related SNPs that were found in both the first 1000 of sorted SNP sets based on combined *p* values (the list on the left) and MDA values (the list on the right).

	SNPS		**SNPS**
1	rs1234220	**1**	**rs553371**
2	**rs553371**	**2**	**rs567700**
3	**rs567700**	**3**	**rs4750759**
4	rs531572	**4**	**rs6482743**
5	**rs4750759**	**5**	**rs3912492**

6	**rs6482743**	6	**rs12636081**
7	rs9971190	7	rs213386
8	rs7082319	8	**rs1380466**
9	rs11016862	9	rs2999081
10	**rs3912492**	10	**rs2811518**
11	rs9637471	11	rs6768256
12	rs6803449	12	rs10776909
13	rs12636081	13	rs3745540
14	rs17061864	14	rs12882037
15	**rs1380466**	15	rs7914154
16	**rs2999081**	16	rs928111
17	rs2811415	17	rs4267385
18	**rs2811518**	18	rs8066276
19	rs2811388	19	rs4921943
20	rs6798749	20	**rs10106032**
21	rs12896434	21	rs4496973
22	rs2360999	22	rs4870828
23	**rs10106032**	23	rs11737898

Table 5.16 Prostate cancer related SNPs that were found in both first 1000 of sorted SNP sets based on combined *p* values (the list on the left) and MDG values (the list on the right).

	SNPS		SNPS
1	rs1234220	1	rs1877724
2	**rs553371**	2	**rs553371**
3	**rs567700**	3	**rs567700**
4	**rs531572**	4	**rs531572**
5	**rs4750759**	5	**rs4750759**
6	rs6482743	6	**rs9971190**
7	**rs9971190**	7	**rs7082319**
8	**rs7082319**	8	**rs3912492**
9	rs11016862	9	rs213386
10	**rs3912492**	10	rs2567608

11	rs9637471	**11**	rs12882037
12	rs6803449	**12**	rs7914154
13	rs12636081	**13**	rs4267385
14	rs17061864	**14**	rs8066276
15	rs1380466	**15**	rs4921943
16	rs2999081	**16**	**rs10106032**
17	rs2811415	**17**	rs4496973
18	rs2811518	**18**	rs4870828
19	rs2811388		
20	rs6798749		
21	rs12896434		
22	rs2360999		
23	**rs10106032**		

Table 5.17 Genes hosting the SNPs that were found related to prostate cancer in both the first 1000 of sorted SNP sets based on combined *p* values (the list on the left) and MDA values (the list on the right).

	GENES		GENES
1	PTEN	**1**	**MGMT**
2	**MGMT**	**2**	**FHIT**
3	**FHIT**	**3**	**EEFSEC**
4	**EEFSEC**	**4**	RXRA
5	**ESRRB**	**5**	KLK12
6	**PSD3**	**6**	**ESRRB**
		7	PIK3AP1
		8	COL4A2
		9	ACE
		10	**PSD3**
		11	ZHX2

Table 5.18 Genes hosting the SNPs that were found related to prostate cancer in both the first 1000 of sorted SNP sets based on combined *p* values (the list on the left) and MDG values (the list on the right).

	GENES		GENES
1	PTEN	1	EPHX1
2	**MGMT**	2	**MGMT**
3	**FHIT**	3	**FHIT**
4	EEFSEC	4	SSTR4
5	**ESRRB**	5	**ESRRB**
6	**PSD3**	6	PIK3AP1
		7	ACE
		8	**PSD3**
		9	ZHX2

5.1.6 Analyses with 1000 SNPs of *SNP-IP* and the 1000 SNPs of *SNP-CP* after *RF*

In order to compare the effects of *random forest* algorithm on the data sets *SNP-IP* and *SNP-CP*, we have performed this algorithm on both data sets individually and then the outputs were sorted based on importance values as mentioned above. Finally, *SNPnexus* analyses have been done on first 1000 SNPs in both sorted sets. By using the results which were given in the Subsections 4.1.1 and 4.1.2, we have created the Tables 5.19 and 5.20. The SNPs which are shown as black and bold are identical SNPs in both sets. In Table 5.19, it is shown that more than half of the SNPs of each set identical to each other. 14 SNPs are identical in both sets. Moreover, in Table 5.20, we have obtained similar results as in Table 5.19. Again more than half of the SNPs of each set identical to each other and the number of SNPs which are identical is 14. As expected, since many of the prostate cancer related SNPs in all of the sets are identical, the genes hosting these SNPs are mostly identical as shown in Table 5.24 and 5.25. The number of common genes is as high as we expected. Therefore it can be interpreted that from all of the tables which were created by using the results given before, similar outcomes were obtained.

All of these results point to fact that, the effects of RF to the data set including SNPs with individual *p* values and combined *p* values, are nearly same.

In addition, SNPs with rsIDs rs567700, rs3912492, rs213386, rs12882037, rs7914154, rs4267385, rs8066276, rs4921943, rs4496973, rs4870828 are common in all sets ordered based on different parameters. The number of common SNPs which were given in the previous two subsections is nearly half of the number of common SNPs that we have obtained in this subsection. Moreover, there exist 7 common genes which are MGMT, FHIT, ESRRB, PIK3AP1, ACE, PSD3 and ZHX2 in all sets that were analyzed in this subsection. As expected, the number of common genes that were obtained in this subsection is higher than the number of common genes that were obtained in the previous two Subsections 5.1.4 and 5.1.5.

Moreover if we examine the error rates which are given in Table 5.21, 5.22 and 5.23, it clear that by using both sets *SNP-IP* and *SNP-CP* as an input for RF, any improvement cannot be observed. All of the error rates, OOB estimation error rate and clear error rates are very similar to each other. Therefore, all of these outcomes can be supporting subtractions to the claim that the effects of *random forest* algorithm on the data sets *SNP-IP* and *SNP-CP* are similar. Although by using the SNPs with associated combined *p* values, the SNPs and the genes hosting these SNPs that will be focused on, have altered in a significance level, after RF was performed on these sets, the result have not changed in an acceptable level.

Table 5.19 Prostate cancer related SNPs that were found in both the first 1000 of sorted SNP sets *SNP-IP* (the list on the left) and *SNP-CP* (the list on the right) based on MDA values.

	SNPS		SNPS
1	**rs567700**	1	rs553371
2	rs531572	2	**rs567700**
3	**rs6482743**	3	rs4750759
4	**rs3912492**	4	**rs6482743**
5	rs6803449	5	**rs3912492**
6	**rs213386**	6	rs12636081
7	**rs1380466**	7	**rs213386**
8	rs2622625	8	**rs1380466**
9	**rs6768256**	9	rs2999081
10	**rs10776909**	10	rs2811518

11	**rs3745540**	11	**rs6768256**
12	rs2360995	12	**rs10776909**
13	**rs12882037**	13	**rs3745540**
14	**rs7914154**	14	**rs12882037**
15	rs6970999	15	**rs7914154**
16	rs2107280	16	rs928111
17	**rs4267385**	17	**rs4267385**
18	**rs8066276**	18	**rs8066276**
19	**rs4921943**	19	**rs4921943**
20	**rs4496973**	20	rs10106032
21	**rs4870828**	21	**rs4496973**
22	**rs11737898**	22	**rs4870828**
		23	**rs11737898**

Table 5.20 Prostate cancer related SNPs that were found in both the first 1000 of sorted SNP sets *SNP-IP* (the list on the left) and *SNP-CP* (the list on the right) based on MDG values.

	SNPS		SNPS
1	rs553371	1	rs1877724
2	**rs567700**	2	**rs553371**
3	**rs531572**	3	**rs567700**
4	**rs4750759**	4	**rs531572**
5	**rs9971190**	5	**rs4750759**
6	**rs3912492**	6	**rs9971190**
7	**rs213386**	7	rs7082319
8	rs2622625	8	**rs3912492**
9	**rs2567608**	9	**rs213386**
10	rs10776909	10	**rs2567608**
11	rs3745540	11	**rs12882037**
12	**rs12882037**	12	**rs7914154**
13	**rs7914154**	13	**rs4267385**
14	**rs4267385**	14	**rs8066276**
15	**rs8066276**	15	**rs4921943**

16	rs4921943	16	rs10106032
17	rs4496973	17	rs4496973
18	rs4870828	18	rs4870828

Table 5.21 OOB estimate of error rate of both performances.

Data Sets	OOB estimate of error rate
SNP-IP	23.81%
SNP-CP	25.87%

Table 5.22 Confusion matrix of performance with *SNP-IP*.

	1	2	Class error
1	477	151	0.2404459
2	149	483	0.2357595

Table 5.23 Confusion matrix of performance with *SNP-CP*.

	1	2	Class error
1	469	159	0.2531847
2	167	465	0.2642405

Table 5.24 Genes hosting the SNPs that were found related to prostate cancer in both the first 1000 of sorted SNP sets *SNP-IP* (the list on the left) and *SNP-CP* (the list on the right) based MDA values.

	GENES		GENES
1	MGMT	1	MGMT
2	FHIT	2	FHIT
3	ABCG2	3	EEFSEC
4	EEFSEC	4	RXRA
5	RXRA	5	KLK12
6	KLK12	6	ESRRB
7	ESRRB	7	PIK3AP1

8	**PIK3AP1**	8	COL4A2
9	NXPH1	9	**ACE**
10	**ACE**	10	**PSD3**
11	**PSD3**	11	**ZHX2**
12	**ZHX2**	12	**SORCS2**
13	**SORCS2**		

Table 5.25 Genes hosting the SNPs that were found related to prostate cancer in both the first 1000 of sorted SNP sets *SNP-IP* (the list on the left) and *SNP-CP* (the list on the right) based on MDG values.

	GENES		GENES
1	**MGMT**	1	EPHX1
2	**FHIT**	2	**MGMT**
3	ABCG2	3	**FHIT**
4	**SSTR4**	4	**SSTR4**
5	RXRA	5	**ESRRB**
6	KLK12	6	**PIK3AP1**
7	**ESRRB**	7	**ACE**
8	**PIK3AP1**	8	**PSD3**
9	**ACE**	9	**ZHX2**
10	**PSD3**		
11	**ZHX2**		

In this and the two of previous Subsections, we have analyzed many SNP lists which were sorted based on different important values. Also, we have come up with many prostate cancer related SNPs and the genes hosting those SNPs. Therefore, if a SNP or a gene is common in all of the results, it can have greater importance or may be it means, it is just one of the important SNP or gene that we must obtain for further studies. Although the number of common genes which are MGMT, FHIT, ESRRB and PSD3, is more than one, only one SNP with rsID rs3912492 has been found as common. In addition, this SNP is located on the gene called FHIT which has been found as prostate cancer.

If we focus on this SNP in closer detail by using SNPnexus, we have revealed that many results were related to different types of cancer and other metabolic diseases. Thus can be a proof for it to have many different biological effects.

5.1.7 Analyses with 100 SNPs after RF on the First 1000 SNPs of *SNP-IP* and *SNP-CP*.

The aim of this step is that, if we can be able to generalize our results based on *random forest* algorithm by using less number of SNPs or not. As it is shown in Table 5.26, there are not any prostate cancer related SNPs in the first 100 of the first 1000 SNPs of *SNP-CP*, which were sorted based on importance values. However, in the previous subsection we have proved that the effect of *random forest* on the data sets *SNP-IP* and *SNP-CP* are similar. Thus, minimizing the number of SNPs which are employed in this study, may mislead the outcomes which will be used in further studies. Although the comparison of error rates which are given in Tables 5.27, 5.28 and 5.29 are similar to the comparisons that have been done in previous Subsections, the results that are given in Tables 5.26 and 5.30 are not sufficient for being interpreted.

Table 5.26 Prostate cancer related SNPs that were found in both first 100 of sorted SNP sets, the first 1000 SNPs of *SNP-IP* (the list on the left) and the first 1000 SNPs of *SNP-CP* (the list on the right), based on both MDA and MDG values.

SNPS		SNPS
1	rs12882037	
2	rs4496973	
3	rs4870828	

Table 5.27 OOB estimate of error rate of both performances.

Data Sets	OOB estimate of error rate
The First 1000 SNPs of *SNP-IP*	25.48%
The First 1000 SNPs of *SNP-CP*	23.41%

Table 5.28 Confusion matrix of performance with the first 1000 SNPs of *SNP-IP*.

	1	2	Class error
1	462	166	0.2643312
2	155	477	0.2452532

Table 5.29 Confusion matrix of performance with the first 1000 SNPs of *SNP-CP*.

	1	2	Class error
1	474	154	0.2452229
2	141	491	0.2231013

Table 5.30 Genes hosting the SNPs that were found related to prostate cancer in the both first 100 of sorted SNP sets, the first 1000 SNPs of *SNP-IP* (the list on the left) and the first 1000 SNPs of *SNP-CP* (the list on the right) based on both MDA and MDG values.

	GENES	GENES
1	**ESRRB**	
2	**ZHX2**	

5.2 Discussion on RegulomeDB Results

5.2.1 RegulomeDB Analyzes Done on The First 500 non-coding SNPs of *SNP-IP* and First 500 non-coding SNPs of *SNP-CP*

As an introduction to the studies based on the SNPs which are located on noncoding regions, we have selected the non-coding SNPs in both data sets *SNP-IP* and *SNP-CP*. In order to make a comparison between the results that were obtained in the Subsections 4.1.2 and 4.2.2, fairly , Table 5.31 was created. It is obviously seen that in the data set *SNP-CP* which is including SNPs with combined *p* values, more SNPs were found with *RegulomeDB* scores equal and less than 3. Therefore, this can be very powerful evidence that, using LD is a true insight in order to improve the studies based on selection of desired SNPs.

In addition, in Table 5.32 the non-coding SNPs with *RegulomeDB* score 1 and the genes that these SNPs can effect indirectly, are given. As shown in this table, the number of SNPs in *SNP-CP* is much higher than the number of SNPs in *SNP-IP*. While only one non-coding SNP with *RegulomeDB* score 1 has been found from *SNP-IP*, in *SNP-CP* this number has reached to 8. When everything taken into consideration, it has proved that LD has an importance in other parameters used in SNP selection studies.

Table 5.31 The none-coding SNPs with *RegulomeDB* scores equal and less than 3, in the first 500 non-coding SNPs of *SNP-IP* and *SNP-CP*.

	SNPs		SNPS
1	**rs6708126**	1	rs2832093
2	**rs8090231**	2	rs6714287
3	**rs12101523**	3	rs304951
4	rs1762438	4	rs7136770
5	rs943889	5	rs10506347
6	rs2063295	6	rs12998237
7	**rs17701543**	7	rs1781079
8	rs10510573	8	**rs12101523**
9	rs11751092	9	**rs17701543**
10	rs11075236	10	rs10027556
11	rs888096	11	rs9897342
12	**rs977676**	12	**rs8090231**
13	**rs11253536**	13	**rs11253536**
14	rs9435409	14	rs2765895
15	rs1330100	15	rs6894361
16	**rs1379736**	16	rs422945
17	rs2581717	17	rs531805
18	rs9328186	18	rs6944602
19	rs2062287	19	rs11696842
		20	rs12910685
		21	**rs6708126**
		22	rs1636579
		23	rs16960555

	24	**rs1379736**
	25	rs986046
	26	**rs977676**
	27	rs17264915

Table 5.32 Non-coding SNPs with *RegulomeDB* scores equal to 1 and the genes which they effect indirectly, in both of the first 500 SNPs of the data sets *SNP-IP* and *SNP-CP*.

	SNPS in *SNP-IP*	**RegulomeScore**	**eQTL**
1	rs1762438	1f	MGMT
	SNPS in *SNP-CP*	**RegulomeScore**	**eQTL**
1	rs304951	1d	
2	rs422945	1d	APPBP1
3	rs10506347	1f	STAT6
4	rs7136770	1f	STAT6
5	rs531805	1f	COG7, DCTN5
6	rs16960555	1f	ZNF586
7	rs11696842	1f	PTPNS1L2, SIRPB1
8	rs6944602	1f	ZNFN1A1

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this study, different analysis were performed on different input models which were obtained from the one main original set, in order to show if LD should be considered or not, in such SNP selection studies. In addition, we have examined the SNPs located on noncoding regions, if we can be able to obtain useful information or not. The results and the experimental findings allow us to draw the following conclusions:

1. For the selection of disease related SNPs, obtaining a new modified data set including SNPs with associated combined *p* values, from the original set by using LD.

   We believe that our study serves as a window to an understanding of the process based on contribution of LD in such kinds of SNP selection studies. Although we could not be able to improve the outputs of *random forest* algorithm in an acceptable level, if we consider the improved algorithm alone not as an input to another algorithm, we have proved that cost effective gene filtration can be achieved when analyzing the first 1000 SNPs of data the sets. In addition, if all the SNPs in data set is being analyzed, the results lead us to the conclusion that cost effective SNP filtration can be achieved.

2. Selection of regulatory SNPs by concerning noncoding regions after obtaining a new modified data set including SNPs with associated combined *p* values based on LD.

   It is known that the functional effects of noncoding disease associated SNPs is one of the challenging issue in GWAS. Although we were able to find some regulatory SNPs which have greater importance, by using the data set including SNPs with individual *p* values, we have proved that, after constructed the new data set which is consisting of the SNPs with associated combined *p* values, we

were able to obtain the most important regulatory SNPs in a significant level. Therefore, the results can be the powerful evidence for LD to have positive effects on SNP selection studies, especially on studies based on non-coding SNPs.

As future works;

- By considering the improved algorithm alone, it can be validated after applied on another data set.
- By considering the improved algorithm as an input modification algorithm, after applying it on another set, modified input can be given to another classification algorithm, other than *random forest*.
- After gaining more biological perspective in this area, we will be able to focus on the output SNPs and genes hosting those SNPs in more closer detail.

REFERENCES

Adeyemo, Adebowale, Norman Gerry, Guanjie Chen, Alan Herbert, Ayo Doumatey, Hanxia Huang, Jie Zhou, et al. 2009. “A Genome-Wide Association Study of Hypertension and Blood Pressure in African Americans.” *PLoS Genetics* 5 (7): e1000564. doi:10.1371/journal.pgen.1000564.

Aguiar, Vanessa, Jose A Seoane, and Ana Freire. 2010. “GA-Based Data Mining Applied to Genetic Data for the Diagnosis of Complex Diseases,” 220–40.

Altshuler, David M, Richard a Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Penelope E Bonnen, et al. 2010. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467 (7311): 52–58. doi:10.1038/nature09298.

Ayers, Kristin L, and Heather J Cordell. 2010. “SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression.” *Genetic Epidemiology* 34 (8): 879–91. doi:10.1002/gepi.20543.

Boyle, Alan P, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc a Schaub, Maya Kasowski, Konrad J Karczewski, et al. 2012. “Annotation of Functional Variation in Personal Genomes Using RegulomeDB.” *Genome Research* 22 (9): 1790–97. doi:10.1101/gr.137323.112.

Breiman, Leo. 2001. “Random Forests.” *European Journal of Mathematics* 45 (1): 5–32.

Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. 2000. “Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines.” *Proceedings of the National Academy of Sciences* 97 (1): 262–67. doi:10.1073/pnas.97.1.262.

Byng, M C, J C Whittaker, A P Cuthbert, C G Mathew, and C M Lewis. 2003. “SNP Subset Selection for Genetic Association Studies.” *Ann Hum Genet* 67 (Pt 6): 543–56. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14641242.

Carlson, Christopher S, Michael a Eberle, Mark J Rieder, Qian Yi, Leonid Kruglyak, and Deborah a Nickerson. 2004. “Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage

Disequilibrium.” *American Journal of Human Genetics* 74 (1): 106–20. doi:10.1086/381000.

Chapman, Juliet, and John Whittaker. 2008. “Analysis of Multiple SNPs in a Candidate Gene or Region.” *Genetic Epidemiology* 32 (6): 560–66.

Chelala, Claude, Arshad Khan, and Nicholas R. Lemoine. 2009. “SNPnexus: A Web Database for Functional Annotation of Newly Discovered and Public Domain Single Nucleotide Polymorphisms.” *Bioinformatics* 25 (5): 655–61.

Chuang, Li-yeh, Kuo-chuan Wu, Hsueh-wei Chang, and Cheng-hong Yang. 2011. “Support Vector Machine-Based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes” I: 16–19.

Contributors, Wikipedia. “Human Genome Project.” *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Human_Genome_Project&oldid=637660910.

Cui, Yuehua, Shaoyu Li, and Barry L. Williams. 2011. “A Combined p-Value Approach to Infer Pathway Regulations in eQTL Mapping.” *Statistics and Its Interface* 4 (3): 389–401. doi:10.4310/SII.2011.v4.n3.a13.

Daly, M J, J D Rioux, S F Schaffner, T J Hudson, and E S Lander. 2001. “High-Resolution Haplotype Structure in the Human Genome.” *Nature Genetics* 29 (2): 229–32.

Dawson, Elisabeth, Gonçalo R Abecasis, Suzannah Bumpstead, Yuan Chen, Sarah Hunt, David M Beare, Jagjit Pabial, et al. 2002. “A First-Generation Linkage Disequilibrium Map of Human Chromosome 22.” *Nature* 418 (6897): 544–48.

Dayem Ullah, A Z, N R Lemoine, and C Chelala. 2013. “A Practical Guide for the Functional Annotation of Genetic Variations Using SNPnexus.” *Brief Bioinform* 14 (4): 437–47. doi:10.1093/bib/bbt004.

Dayem Ullah, Abu Z., Nicholas R. Lemoine, and Claude Chelala. 2012. “SNPnexus: A Web Server for Functional Annotation of Novel and Publicly Known Genetic Variants (2012 Update).” *Nucleic Acids Research* 40 (W1).

Dunning, Alison M, Francine Durocher, Catherine S Healey, M Dawn Teare, Simon E Mcbride, Francesca Carlomagno, Chun-fang Xu, et al. 2000. “The Extent of Linkage Disequilibrium in Four Populations with Distinct Demographic Histories,” 1544–54.

Easton, Douglas F, and Rosalind a Eeles. 2008. “Genome-Wide Association Studies in Cancer.” *Human Molecular Genetics* 17 (R2): R109–15. doi:10.1093/hmg/ddn287.

Eisenbarth, I, A M Striebel, E Moschgath, W Vogel, and G Assum. 2001. "Long-Range Sequence Composition Mirrors Linkage Disequilibrium Pattern in a 1.13 Mb Region of Human Chromosome 22." *Human Molecular Genetics* 10 (24): 2833–39.

Eslahchi, Changiz, Ali Katanforoush, Hamid Pezeshk, and Narjes Afzaly. 2011. "Archive of SID Haplotype Block Partitioning and tagSNP Selection under the Perfect Phylogeny Model Archive of SID" 9 (4): 281–89.

Feingold, N. 1980. "Linkage Disequilibrium." *Journal de Genetique Humaine* 28 (2): 105–13. doi:10.1371/journal.pgen.1000147.

Fiaschi, Linda, Jonathan M. Garibaldi, and Natalio Krasnogor. 2009. "A Framework for the Application of Decision Trees to the Analysis of SNPs Data." In *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2009 - Proceedings*, 106–13.

Friedman, J, R Tibshirani, and T Hastie. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York.

Gabriel, S B. 2002. "The Structure of Haplotype Blocks in the Human Genome." *Science* 296: 2225–29.

"Genetic Home Reference." 2014. In *Help Me Understand Genetic*, 9–66. Department of Health & Human Services. http://ghr.nlm.nih.gov/.

Gerstenblith, Meg R, Jianxin Shi, and Maria Teresa Landi. 2010. "Genome-Wide Association Studies of Pigmentation and Skin Cancer: A Review and Meta-Analysis." *Pigment Cell & Melanoma Research* 23 (5): 587–606.

Goldstein, Benjamin A, Alan E Hubbard, Adele Cutler, and Lisa F Barcellos. 2010. "An Application of Random Forests to a Genome-Wide Association Dataset: Methodological Considerations & New Findings." *BMC Genetics* 11: 49.

Goldstein, Benjamin A, Eric C Polley, and Farren B. S. Briggs. 2011. "Random Forests for Genetic Association Studies." *Statistical Applications in Genetics and Molecular Biology*.

Gomes, Bruno C, Susana Vinga, and Jorge Gaspar. 2010. "A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups," 1–8.

Hill, W G. 1974. "Estimation of Linkage Disequilibrium in Randomly Mating Populations." *Heredity* 33 (2): 229–39.

Horne, B D, and N J Camp. 2004. "Principal Component Analysis for Selection of Optimal SNP-Sets That Capture Intragenic Genetic Variation." *Genet Epidemiol* 26 (1): 11–21. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14691953.

Jakkula, Eveliina, Virpi Leppä, Anna-Maija Sulonen, Teppo Varilo, Suvi Kallio, Anu Kemppinen, Shaun Purcell, et al. 2010. “Genome-Wide Association Study in a High-Risk Isolate for Multiple Sclerosis Reveals Associated Variants in STAT3 Gene.” *American Journal of Human Genetics* 86 (2): 285–91. doi:10.1016/j.ajhg.2010.01.017.

Jiang, X, M M Barmada, and S Visweswaran. 2010. “Identifying Genetic Interactions in Genome-Wide Data Using Bayesian Networks.” *Genet Epidemiol* 34 (6): 575–81. doi:10.1002/gepi.20514 [doi].

Johnson, Andrew D., Robert E. Handsaker, Sara L. Pulit, Marcia M. Nizzari, Christopher J. O’Donnell, and Paul I W De Bakker. 2008. “SNAP: A Web-Based Tool for Identification and Annotation of Proxy SNPs Using HapMap.” *Bioinformatics* 24 (24): 2938–39.

Johnson, G C, L Esposito, B J Barratt, A N Smith, J Heward, G Di Genova, H Ueda, et al. 2001. “Haplotype Tagging for the Identification of Common Disease Genes.” *Nature Genetics* 29 (2): 233–37. doi:10.1038/ng1001-233.

Jorde, L.B. 2000. “Linkage Disequilibrium and the Search for Complex Disease Genes.” *Genome Research* 10 (10): 1435–44. doi:10.1101/gr.144500.

Journal, International, and Innovative Computing. 2013. “A GENETIC ALGORITHM – SUPPORT VECTOR MACHINE METHOD FOR SELECTING TAG SINGLE NUCLEOTIDE” 9 (2): 525–41.

Kim, Jinseog, Insuk Sohn, Dennis Dong Hwan Kim, and Sin-Ho Jung. 2013. “SNP Selection in Genome-Wide Association Studies via Penalized Support Vector Machine with MAX Test.” *Computational and Mathematical Methods in Medicine* 2013 (January): 340678. doi:10.1155/2013/340678.

Lettre, Guillaume, Cameron D Palmer, Taylor Young, Kenechi G Ejebe, Hooman Allayee, Emelia J Benjamin, Franklyn Bennett, et al. 2011. “Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project.” *PLoS Genetics* 7 (2): e1001300. doi:10.1371/journal.pgen.1001300.

Lewontin, R C, and Received July. 1964. “The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models’,” 49–67.

Liaw, A, and M Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2: 18–22.

Lin, Min-hui. 2010. “A Hybrid PSO - SVM Approach for Haplotype Tagging SNP Selection Problem” 8 (6): 60–65.

Liu, Guimei, Yue Wang, and Limsoon Wong. 2010. “FastTagger: An Efficient Algorithm for Genome-Wide Tag SNP Selection Using Multi-Marker Linkage Disequilibrium.” *BMC Bioinformatics* 11 (January): 66. doi:10.1186/1471-2105-11-66.

Lucek, P, J Hanke, J Reich, S A Solla, and J Ott. “Multi-Locus Nonparametric Linkage Analysis of Complex Trait Loci with Neural Networks.” *Human Heredity* 48 (5): 275–84.

Macintyre, Geoff, James Bailey, Izhak Haviv, and Adam Kowalczyk. 2010. “Is-rSNP: A Novel Technique for in Silico Regulatory SNP Detection.” *Bioinformatics (Oxford, England)* 26 (18): i524–30. doi:10.1093/bioinformatics/btq378.

Mailman, Matthew D, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, et al. 2007. “The NCBI dbGaP Database of Genotypes and Phenotypes.” *Nature Genetics* 39 (10): 1181–86.

Marinov, M, and D E Weeks. 2001. “The Complexity of Linkage Analysis with Neural Networks.” *Human Heredity* 51 (3): 169–76. doi:53338.

Miyaki, Koichi, Kazuyuki Omae, Mitsuru Murata, Norio Tanahashi, Ikuo Saito, and Kiyoaki Watanabe. 2004. “High Throughput Multiple Combination Extraction from Large Scale Polymorphism Data by Exact Tree Method.” *Journal of Human Genetics* 49 (9): 455–62.

Mourad, Raphaël, Christine Sinoquet, and Philippe Leray. 2011. “A Hierarchical Bayesian Network Approach for Linkage Disequilibrium Modeling and Data-Dimensionality Reduction prior to Genome-Wide Association Studies.” *BMC Bioinformatics* 12 (1). BioMed Central Ltd: 16. doi:10.1186/1471-2105-12-16.

Musani, S K, D Shriner, N Liu, R Feng, C S Coffey, N Yi, H K Tiwari, and D B Allison. 2007. “Detection of Gene X Gene Interactions in Genome-Wide Association Studies of Human Population Data.” *Hum Hered* 63 (2): 67–84. doi:000099179 [pii]\r10.1159/000099179.

Patil, N, A J Berno, D A Hinds, W A Barrett, J M Doshi, C R Hacker, C R Kautzer, et al. 2001. “Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21.” *Science (New York, N.Y.)* 294 (5547): 1719–23.

Peng, Gang, Li Luo, Hoicheong Siu, Yun Zhu, Pengfei Hu, Shengjun Hong, Jinying Zhao, et al. 2010. “Gene and Pathway-Based Second-Wave Analysis of Genome-Wide Association Studies.” *European Journal of Human Genetics : EJHG* 18 (1). Nature Publishing Group: 111–17. doi:10.1038/ejhg.2009.115.

Purcell, Shaun. 2007. “PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis.” *American Journal of Human Genetics,* 81.

Raetz, Elizabeth A., Philip J. Moos, Aniko Szabo, and William L. Carroll. 2001. “GENE EXPRESSION PROFILING.” *Hematology/Oncology Clinics of North America*. doi:10.1016/S0889-8588(05)70257-4.

Reddy, M V Prasad Linga, H Wang, S Liu, B Bode, J C Reed, R D Steed, S W Anderson, L Steed, D Hopkins, and J-X She. 2011. “Association between Type 1 Diabetes and GWAS SNPs in the Southeast US Caucasian Population.” *Genes and Immunity* 12 (3): 208–12.

Reich, David E, Michele Cargill, Stacey Bolk, James Ireland, Pardis C Sabeti, Daniel J Richter, Thomas Lavery, et al. 2001. “Linkage Disequilibrium in the Human Genome” 9 (Table 1): 199–204.

Reich, David E, Stephen F Schaffner, Mark J Daly, Gil McVean, James C Mullikin, John M Higgins, Daniel J Richter, Eric S Lander, and David Altshuler. 2002. “Human Genome Sequence Variation and the Influence of Gene History, Mutation and Recombination.” *Nature Genetics* 32 (1): 135–42.

Saangyong Uhmn Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong and Jin Kim, Dong-Hoi Kim. 2009. “A Study on Application of Single Nucleotide Polymorphism and Machine Learning Techniques to Diagnosis of Chronic Hepatitis.” *Expert Systems* 26 (1): 10. http://www3.interscience.wiley.com/journal/121675322/abstract?CRETRY=1&SRETRY=0.

Schulze, Thomas G, Kui Zhang, Yu-Sheng Chen, Nirmala Akula, Fengzhu Sun, and Francis J McMahon. 2004. “Defining Haplotype Blocks and Tag Single-Nucleotide Polymorphisms in the Human Genome.” *Human Molecular Genetics* 13 (3): 335–42. doi:10.1093/hmg/ddh035.

Scott, Laura J, Pierandrea Muglia, Xiangyang Q Kong, Weihua Guan, Matthew Flickinger, Ruchi Upmanyu, Federica Tozzi, et al. 2009. “Genome-Wide Association and Meta-Analysis of Bipolar Disorder in Individuals of European Ancestry.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (18): 7501–6. doi:10.1073/pnas.0813386106.

Shapiro, L J. 1993. “Human Genome Project.” *The Western Journal of Medicine* 158 (2): 181.

Sherry, S T, M Ward, M Kholodov, J Baker, L Phan, E M Smigielski, K Sirotkin, and K Sirotkin Genome Res. 2001. “dbSNP : The NCBI Database of Genetic Variation” 29 (1): 308–11.

“Single-Nucleotide Polymorphism.” 2013. http://en.wikipedia.org/w/index.php?title=Single-nucleotide_polymorphism&oldid=549487528.

Sinoquet, Christine, and Philippe Leray. 2010. “A Bayesian Network Approach to Model Local Dependencies among SNPs A Bayesian Network Approach to Model Local Dependencies among SNPs.”

Slatkin, Montgomery. 2008. “Linkage Disequilibrium--Understanding the Evolutionary Past and Mapping the Medical Future.” *Nature Reviews. Genetics* 9 (6): 477–85.

Stahl, EA, S Raychaudhuri, EF Remmers, G Xie, S Eyre, BP Thomson, YH Li, et al. 2010. “Genome-Wide Association Study Meta-Analysis Identifies Seven New Rheumatoid Arthritis Risk Loci.” NATURE PUBLISHING GROUP. http://discovery.ucl.ac.uk/140377/.

T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, R. Holland L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, E. Kulesha K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, A. Parker D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, T. Cox A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, G. Proctor V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, and and E. Birney. S. Searle, J. Smith, A. Ureta-Vidal. 2007. “Ensembl.” *Nucleic Acids Research*.

Taillon-Miller, P, I Bauer-Sardiña, N L Saccone, J Putzel, T Laitinen, A Cao, J Kere, G Pilia, J P Rice, and P Y Kwok. 2000. “Juxtaposed Regions of Extensive and Minimal Linkage Disequilibrium in Human Xq25 and Xq28.” *Nature Genetics* 25 (3): 324–28.

The Eugenics Review. 1926. “Statistical Methods for Research Workers” 18 (2): 148–50.

Tomida, S, T Hanai, N Koma, Y Suzuki, T Kobayashi, and H Honda. 2002. “Artificial Neural Network Predictive Model for Allergic Disease Using Single Nucleotide Polymorphisms Data.” *J Biosci Bioeng* 93 (5): 470–78. doi:S1389-1723(02)80094-9 [pii].

Tomita, Yasuyuki, Shuta Tomida, Yuko Hasegawa, Yoichi Suzuki, Taro Shirakawa, Takeshi Kobayashi, and Hiroyuki Honda. 2004. “Artificial Neural Network Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma.” *BMC Bioinformatics* 5 (September): 120. doi:10.1186/1471-2105-5-120.

Waddel, Michael, David Page, Fenghuang Zhan, Bart Barlogie, and John Shaughnessy Jr. 2005. “Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma.” *BIOKDD ’05 Proceedings of the 5th International Workshop on Bioinformatics*, 21–28.

Wang, Ning, Joshua M Akey, Kun Zhang, Ranajit Chakraborty, and Li Jin. 2002. “Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation.” *American Journal of Human Genetics* 71 (5): 1227–34. doi:10.1086/344398.

Ward, Lucas D, and Manolis Kellis. 2012. “HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations within Sets of Genetically Linked Variants.” *Nucleic Acids Research* 40 (Database issue): D930–34. doi:10.1093/nar/gkr917.

Wikipedia contributors. 2014. “Data Mining.” *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Data_mining&oldid=638889170 .

Winham, Stacey J, Colin L Colby, Robert R Freimuth, Xin Wang, Mariza de Andrade, Marianne Huebner, and Joanna M Biernacka. 2012. “SNP Interaction Detection with Random Forests in High-Dimensional Genetic Data.” *BMC Bioinformatics* 13 (1). BMC Bioinformatics: 164. doi:10.1186/1471-2105-13-164.

Xu, Zongli, Norman L Kaplan, and Jack a Taylor. 2007. “Tag SNP Selection for Candidate Gene Association Studies Using HapMap and Gene Resequencing Data.” *European Journal of Human Genetics : EJHG* 15 (10): 1063–70. doi:10.1038/sj.ejhg.5201875.

Yeager, Meredith, Nick Orr, Richard B Hayes, Kevin B Jacobs, Peter Kraft, Sholom Wacholder, Mark J Minichiello, et al. 2007. “Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24.” *Nature Genetics* 39 (5): 645–49.

Zaykin, Dmitri V, Lev A Zhivotovsky, Wendy Czika, Susan Shao, and Russell D Wolfinger. 2007. “Combining P-Values in Large Scale Genomics Experiments.” *Pharmaceutical Statistics* 6 (3): 217–26. doi:10.1002/pst.304.

Zhang, Kui, Zhaohui S Qin, Jun S Liu, Ting Chen, Michael S Waterman, and Fengzhu Sun. 2004. “Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies.” *Genome Research* 14 (5): 908–16. doi:10.1101/gr.1837404.

Zhao H. , Pfeiffer R, Gail M. H. 2003. “Haplotype Analysis in Population Genetics and Association Studies, Pharmacogenomics.”

Zhou, Nina, and Lipo Wang. 2007. “Effective Selection of Informative SNPs and Classification on the HapMap Genotype Data.” *BMC Bioinformatics* 8 (C): 484. doi:10.1186/1471-2105-8-484.