# Genome wide scan for prostate cancer susceptibility genes

Master's thesis
Institute of Medical Technology
University of Tampere
Ha,Nati
June 2008

# ACKNOLEDGEMENTS

The practical and written parts of this thesis were done in Institute of the Medical technology (IMT) at University of Tampere. First, I owe my special thanks to Professor Johanna Schleutker, who has given me a great possibility to work and learn in her group. I also direct my deep appreciation to Professor Johanna Schleutker, for her professional guidance and support throughout this project. Major compliments belong to the other member of the Prostate cancer Investigator Group (PIG), especially to Tiina Wahlfors, PhD, and other personnel of IMT for their valuable advice and help.

I am the most grateful to my parents and for both financial and mental support. Without their encouragement I had not been able to fulfill my dreams and achieve my goals. I also want to express my gratitude to Professor Mauno Vihinen for discussion and important advice. In addition, I want to thank my friends for their powerful words of encouragement.

Tampere, June 2008

Ha,Nati

# MASTER'S THESIS


| | |
|---|---|
| Place: | UNIVERSITY OF TAMPERE<br>Faculty of Medicine<br>Institute of Medical Technology |
| Author: | Ha, Nati |
| Title: | Genome wide scan for Prostate Cancer<br>Susceptibility genes |
| Pages: | 76 pp + appendices 8 pp. |
| Supervisors: | Tiina Wahlfors, PhD; Professor Johanna Schleutker |
| Reviewer: | Professor Mauno Vihinen, Professor Johanna Schleutker |
| Time: | June 2008 |


## Abstract


**Background and aims:** Prostate cancer is the leading cancer type in Finnish men. It was suggested that 5% to 10% of incident cases are attributed to rare, highly penetrated alleles in single gene forms of disease. The aim of the study was to use genome wide linkage scan in 56 Finnish Families with multiple prostate cancer cases to detect possible prostate cancer susceptibility genes.

**Methods:** Genotyping data of 490 microsatellite markers was analyzed with two point, and multi-point parametric and non parametric linkage analyses. Using Prostate cancer genotype data as input, Mendelian errors were checked by Pedcheck program. Two Point linkage analyses were carried out by FastLink program. Single-point nonparametric, multi-point parametric and nonparametric analyses were carried out by GENEHUNTER program.

**Results:** The most significant results are obtained from chromosome 13 which gave the best two point LOD score 2.67 with marker D13S173 at 13q33, chromosome 17 at 17q21, where the best two point LOD score was 2.46, The third significant LOD score come from chromosome 3 at 3q26, where the best two point LOD score was 2.38 (theta=0.1) with marker D3S1565, and chromosome X at Xq27, where the best two point LOD score 2.03 (theta =0.2) with marker DXS1227.

**Conclusion:** Results from chromosome 3, 8, 12, and X, support previous linkage analyses. In addition, the present linkage analyses also reveal areas which are not presented in the previous linkage analyses, such as chromosome 17 at 17q21, chromosome 13 at 13q33, chromosome 2 at 2q37 and chromosome 6 at 6p21.

# CONTENTS

## Abbreviations

| | |
|---|---|
| AD | Autosomal Dominant |
| APM | Affected-Pedigree Member method |
| AR | Autosomal Recessive |
| CI | Confidence Interval |
| CNVs | Copy Number Variations |
| FRR | Familial Relative Risk |
| GAS | Genetic Analysis System |
| HMM | Hidden Markov chain Model |
| HPC | Hereditary Prostate Cancer |
| IBD | Identity/Identical By Decent |
| IBS | Identity/Identical By State |
| ICPCG | International Consortium for Prostate Cancer hereditary Genetics |
| LD | Linkage Disequilibrium |
| LOD | maximum Logarithm of Odds |
| HLOD | Heterogeneity Logarithm of Odds |
| LOH | Loss Of Heterozygosity |
| MLE | Maximum Likelihood |
| MCMC | Markov Chain Monte Carlo |
| NPL | Non Parametric LOD |
| NMM | No Male-to-Male |
| OMIM | Online Mendelian Inheritance in Man |
| PIC | Polymorphic Information Content |
| PSA | Prostate Specific Antigen |
| RFLPS | Restriction fragment length of polymorphisms |
| RAPD | Random Amplification of Polymorphic |
| SSLP | Single Sequence Length polymorphism |
| SSR | Simple sequence Repeats |
| STMS | Sequence Tagged Microsatellites |
| STRP | show tandem repeat polymorphisms |
| SNP | single nucleotide polymorphisms |
| XD | X-linked Dominant |
| XR | X-linked Recessive |

# 1. Introduction

Prostate cancer is the most common cancer among Finnish men, and the incidence of prostate cancer has increased markedly in recent decades (http://www.cancerregistry.fi). Except age and ethnicity, the most significant risk factors are the presence of several affected first-degree relatives and an affected that had an uncommon early age onset (Keetch et al. 1995). Evidence from twin studies support that this familiar risk has an inherited basis (Page et al.1997). In 1993, it was suggested that around 5% to 10% of prostate cancer cases are attributed to rare, highly penetrant alleles in single gene forms of disease (Carter et al. 1993). Since then, multiple linkage analyses were performed during 1996-2003 (Schaid 2004) based on this epidemiological evidence. Early linkage studies and consequent fine mapping has revealed three high-penetrance candidate genes, *ELAC2*, *RNASEL* and *MSR1* (Shaid 2004). However, mutation in these genes seems to be extremely rare explaining only a small population of familial prostate cancer cases, also in Finland (Rökman et al. 2001; Rökman et al. 2002; Seppälä et al. 2003). Finnish population is one of the genetically homogeneous founder populations of the World, where linkage analysis should be most powerful (de la Chapella 1993; Peltonen 2000). To date, only 10 Finnish prostate cancer families have been analyzed with genome wide linkage analysis (Schleutker et al, 2003). Since none of the known candidate genes seems to explain the familial aggregation in Finland and on the other hand, two novel loci specific for Finns have been identified, a second analysis with more extensive material was needed.

The parts of DNA that are responsible for coding protein structures are called genes. They are inherited according to the Mendelian laws. In human, DNA, containing both coding sequences and non-coding sequences, is divided into 23 segments called chromosomes. Every human produces germ cells (sperm or egg); which contain one copy of each chromosome. During meiosis, when a germ cell is formed, the two homologous copies of each chromosome pair up, each member of pair goes into one or the other

daughter cell. When the pairs of homologous chromosomes line up side by side, they undergo a process called crossing over, which is referred as recombination. Recombination happens frequently, and it seems that at least one chiasma must happen on each chromosome in each meiosis (Sturt, 1976). The basis of linkage analysis is that recombination events happen between two genetic loci (genes, DNA markers, chromosomal abbreviations, etc) at a rate related to the distance between them on the same chromosome. The goal of linkage analysis is to determine whether two loci tend to co-segregate more often than they should if they are not physically close together on same chromosome.

Linkage analysis approaches can be classified into two main classes: parametric and nonparametric methods. Methods in the first class require specification of genetic parameters, such as penetrance, disease-allele frequency, phenocopy and mutation rates describing the mode of disease inheritance. In contrast, methods in second class, was model free.

The purposes of this study were to extend the finding of the previous genome wide scan of Hereditary Finnish Prostate families, and to locate other possible prostate cancer susceptibility genes.

# 2. Review of literature

# 2. 1 Genetic Epidemiology

## 2.1.1 General Genetic epidemiology

There are two main studies in genetic epidemiology: 1) the study of the etiology of disease among groups of relatives to reveal the causes of family resemblance and 2) the study of inherited causes of disease in population (Morton and Chung, 1978). Other studies seem to focus on genetic epidemiology mainly to the analysis of familial aggregation. Alternatively, it was pointed out by Roberts (1985) that the underlying genetic structure of a population is important in determining disease and other physiologic processes that could be considered within the range of normal human variation.

Considering disease etiology, geneticists have viewed disease etiology as ranging from totally genetic causal events to totally environmental causal events, epidemiologists' ideas of disease etiology have also been expressed in terms of complicated interaction among the agent, the host, and the environment, the trio known as the "epidemiology triangle" (Mausner and Kramer, 1985). Disease is defined as the result of a chain of events that comprises a delicate interaction of external causal events and internal pathogenetic mechanisms (MacMahon and Pugh, 1970). Despite the appeal of any simple classification of disease, it is becoming more apparent that most diseases are not purely genetic or environmental in etiology, but depend on a complex interaction of these two factors.

As mentioned above, it is clear that genetic epidemiology is generalized as the study of the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations; the environmental factors are defined as

exogenous factors, such as chemical, physical, infectious and nutritional factors. Genetic epidemiology aims to explain the role of genetic factors in the etiology of disease in human populations with the objective of disease control and prevention. From that perspective, available study designs (e.g., family studies, inbreeding studies, population surveys), and statistical techniques (e.g., segregation analysis, linkage analysis) are considered. Family studies are one of several available study methods to explore the role of genetic factors in disease. Studies design should follow sound epidemiologic guidelines as to case definition, sample representativeness, ascertainment methods, and data collection pertaining to disease and exposures, and appropriate methods of analysis (Dorman et al., 1988), or it would be difficult to make broad inferences, based on the segregation analysis performed on pedigrees of patients, if pedigrees were not collected in a systematic and unbiased fashion. And if the study sample was restricted to high-risk families or to the referred patients from one specialty clinic, it may also be difficult to generalize findings to the population at large. Furthermore, statistical evidence to a particular model from pedigree data remains uncertain until specific and measurable genetic factors are documented, using molecular or biochemical methods.

Two broad but overlapping groups of studies, which provide parallel methodologies for studying the role of genetic factors in disease, are: 1) descriptive studies, focusing on the distribution of genetic traits and diseases in populations or family, and 2) analytic studies, which focus on the determinations of the distribution of genetic traits and their correspondent role in health and disease in population and families.

Genetic epidemiology mainly focused on population and family studies. Population studies generally include the study of the distribution and determinants of observable genetic traits in populations, and the study of the role of genetic factors, many of which are not directly observable, in disease processes and other physiologic variations.

However, the central theme in genetic epidemiology is family studies (King et al., 1984), and family studies have been the focus of most of methodological and statistical efforts. Genetic epidemiology faces three main questions in family studies (King et al. 1984).

The first question is about disease clustering in families, this can be approached either by comparing disease frequency in relatives of cases and controls, or by comparing disease frequency in relatives of cases with that in the general population and calculating some relative risk. The occurrence of a high degree of familial aggregation, however, does not prove the existence of a genetic mechanism nor does a low relative risk preclude a genetic mechanism, because infectious diseases frequently cluster in families (Susser, 1985; Susser and Susser1987a, 1987b). On the other hand, single-gene disorder can also give low relative risks at certain gene frequencies (Weiss et al., 1982). Incomplete penetrance in a genetic form of the disease can further undermine the use of the case control approach to measure familiar aggregation (Majumder et al., 1983), because unaffected controls may carry the disease genotype without expressing it, and therefore have relatives at high risk.

The second question is about checking if familial clustering related to environmental exposure, or inherited susceptibility, which is approached by using both genetic and epidemiologic methods that could be applied to quantitative and qualitative traits. Epidemiologic methods can be used to determine whether the familial aggregation has a non-Mendelian basis (Susser, 1985; Susser and Susser 1987a). If approached by using genetic methods, the multi factorial model of inheritance is the basic frame structure for these analyses. This model uses statistical methods of analysis of variance and path analysis to infer the degree of genetic control in either quantitative or qualitative traits.

If both evidence of familial aggregation and genetic control are suggested for a disease or quantitative phenotype, the third question is about identifying the responsible genetic mechanism. Segregation analyses are useful to test for Mendelian transmission of the phenotypes (discrete or quantitative) in pedigree data. Theses methodologies yield maximum likelihood estimators for major genetic parameters (Elston and Stewart, 1971; Cannings et al., 1978; Hasstedt, 1982). The parameters include transmission probabilities, gene frequencies, and penetrance parameters in Mendelian models; heritability, sample means and variances in polygenic models; and both types of parameters in the “mixed model” (Morton, 1982, 1984). The value of this latter mixed model is that it allows

specific test of hypotheses to discriminate between a single-locus model and polygenic inheritance.

If a single locus model is proved to explain the distribution of a disease or trait in families' best, there are still possibilities remains that a non-genetic mechanism may be the true etiologic agent, and it is merely mimicking a genetic mechanism as illustrated by Lilienfeld (1959) and McGuffin and Huckle (1990). This is true when the families are small. The final bit of statistical evidence for Mendelian inheritance can come from genetic linkage with a known genetic marker, using statistical methods and other mapping strategies.

## 2.1.2 Fundamental Genetic Concepts

Human nuclear DNA is divided into 22 pairs of autosomal chromosomes and 2 sex specific chromosomes X and Y, which become visible microscopically only during cell division. Humans are diploid, so there are two copies of each autosome in the cell nuclear; each parent gives one of each pair. Each chromosome is composed of two arms separated by a centromere; the shorter arm is denoted as *p* and the longer arm is denoted as *q*.

In most cell divisions, the duplicated nuclear material divides into daughter cells by a process called mitosis, the only cell division process for all somatic cells. In the reproductive or germinal cells, a different process, called meiosis, occurs. The behavior of a chromosomes pair in mitosis and meiosis is illustrated in Figures 2-1 and Figure 2-2, respectively. The basic difference between these two types of cell division is that mitosis exactly replicates the entire genetic complement in each daughter cell, with no changes in chromosome number or arrangement, while meiosis results in a systematic reduction of the usual diploid number into haploid daughter cells with 23 chromosomes, each having one member of each pair. Thus, when two haploid gametes fuse to form a new zygote, the original diploid complement of 46 chromosomes is restored. Furthermore, in meiosis, there is independent assortment of chromosomes of maternal and paternal origin within

each of 22 homologous pairs of autosomes, as well as random assortment of the sex chromosomes into the resulting haploid daughter cell that will go on to develop into gametes. Along with genetic recombination because of crossing over between loci located along the length of the chromosomes, this random assortment or independent segregation of homologous chromosomes guarantees a large number of different combinations of genetic traits at each generation.

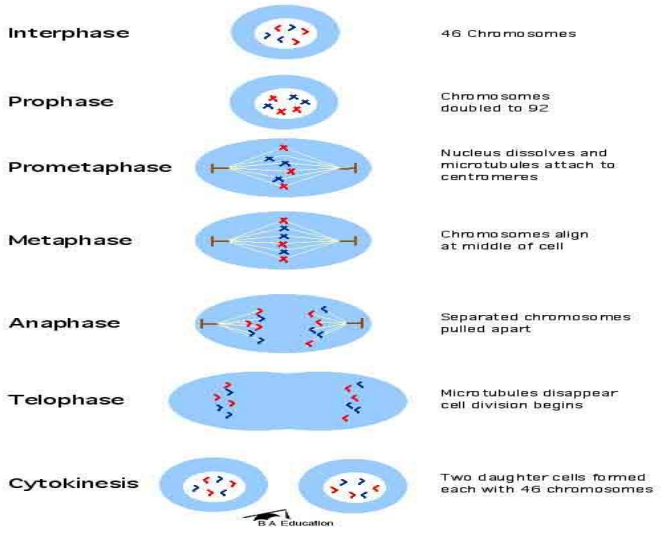


**Figure 2-1 Phases of mitosis(http://www.ba-education.demon.co.uk/for/science/dnabiology1.html)**

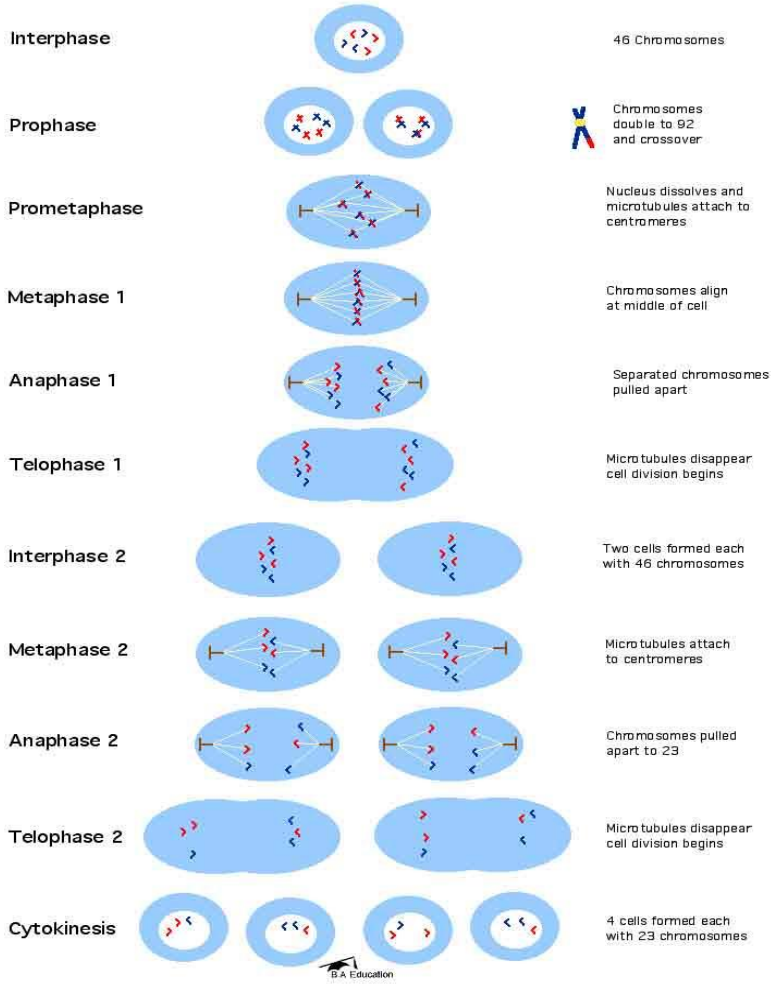


**Figure 2-2 Phases of meiosis(http://www.ba-education.demon.co.uk/for/science/dnabiology1.html)**

While nuclear DNA is organized into these distinct chromosomes, the basic unit of hereditary information is the gene or locus, which codes for some gene product (protein). Many different forms of a gene representing individual mutations may exist at a given locus, and these are called alleles. Each person carries two copies of every autosomal gene, and these alleles may or may not be different. Heterozygous individual have two different alleles, while homozygous individual have two copies of the same allele. Females can carry two alleles at all loci on the X chromosome, but male are hemizygous for every locus on both the X and Y chromosomes.

Because genes specify the coding of protein that in turn construct the structural and functional building segments of the human organism, any alteration (mutation) in the

genetic material that leads to a disturbance in the structure or/and function of a protein can result in disease. Mutation is broadly defined as any changes in the genetic material, and thus many mutations will disrupt the structure and function of gene products. Mutations can occur in both somatic and germinal cells, but only germinal mutations are heritable and transmitted to subsequent generations.

Single base substitutions do not always causes changes in the amino acid sequence of the gene product because the genetic code is degenerate. However, a single base substitution which causes premature termination of transcription would result in the absence of a functional gene product. Depending on the role of the gene product, this might lead to disease. Insertion and deletion of one or two nucleotides (any number divided by 3, which gives reminder not equal to 0) results in frame shift mutations, Frame shift mutations can also cause premature termination of translation if they result in converting a codon into the stop signal for translation. When considering the overall effect of mutations on the occurrence of disease, it is important to establish the mode of expression of a mutant allele. If the phenotype is altered by a mutant allele, in both the homozygous and the heterozygous states, the disease or trait is said to be dominant. If the phenotype is altered only in the homozygous state the disease or trait is said to be recessive. When both alleles in heterozygote are fully expressed, that is, the heterozygote is phenotypically distinct from the two homozygotes, the trait is said to be co-dominant. The variety of known single-gene disorders (autosomal dominant, autosomal recessive, and X-linked) has been cataloged and updated in OMIM (http://www.ncbi.nlm.nih.gov/omim/). The number of such diseases, both confirmed and suspected, has grown remarkably over time. These diseases are referred to as Mendelian disorders because they follow Mendel's law for single gene transmission in families.

Chromosomal mutations include abnormalities of chromosome number and aberrations of chromosome structure, and they represent a form of genetic diseases distinctly different from the Mendelian Disease. Gene and chromosome mutations can also occur in somatic cells. Although these are not heritable, large evidence suggests that they play an important role in the pathogenesis of human disease, notably cancer.

DNA copy number variation has long been associated with specific chromosomal rearrangements and genomic disorders. In humans, copy number variants (CNVs) account for a substantial amount of genetic variation. Since many CNVs include genes that result in differential levels of gene expression, CNVs may account for a significant proportion of normal phenotypic variation.

Recent advance in human genetics have improved the concept of the structure of genes. However, the principles of Mendelian transmission, both in families and in populations, are still holding, and were served as the cornerstone of genetic analysis. Mendelian principles have been successfully applied to the study of transmission of single-gene traits and diseases in families. There are four major types of single locus Mendelian transmission: autosomal dominant (AD), autosomal recessive (AR), X-linked recessive (XR), and X-linked dominant (XD). Table 2-2 lists several prominent features of these four mechanisms that can be used to discriminate among competing hypotheses for individual pedigrees (diagrams of extend families).

**Table 2-2.General features of disease distribution in families and populations under the four major forms of Mendelian inheritance (chapter 2, Fundamentals of Genetic Epidemiology by Muin J. Khoury, Terri H. Beaty, Bernice H. cohen)**

| | Autosomal Dominant | Autosomal Recessive | X-linked Dominant | X-linked Recessive |
|---|---|---|---|---|
| Both males and females affected in equal frequency of population | Yes | Yes | No<br>For rare diseases, approximately 2/3 affected will be female | No<br>For rare diseases, female frequency equals square of male frequency |
| Transmission by both sexes | Yes | Yes | No<br>Father-son Transmission not Possible | NO<br>Father-son Transmission not possible |
| Status of parents of an affected child | At least 1 parent of affected child must be affected | For rare diseases, typically, both parents normal; consanguinity more common than general pops | At least 1 parent affected; affected males must havae affected mothers except for new mutation | Usually both parents normal, although maternal male relatives frequently affected |
| Most common at risk mating type( for a rare disorder) | Aa x aa<br>Affected x normal | Aa x Aa<br>Normal x Normal | XY x XX-<br>Normal x affected heterozygote | XY x XX-<br>Normal x normal heterozygote |
| Segregation ratio in offspring (normal:affected) from | 1:1 from mating of affected x normal; no risk | 3:1 segregation | 1:1 for both sons and daughters | 1:1 in males only; all daughters phenotypically |

| | | | | |
|---|---|---|---|---|
| most common mating | to children of 2 normal, barring incomplete penetrance and new mutation | | | normal, but 50% are carriers |
| Other prominent mating types to be considered | Aa x Aa<br>Affected x affected | aa x AA<br>Affected x affected | X-Y x XX<br>Affected male x normal female | X-Y x XX affected male x normal female |
| Segregation Ratio | 3: 1 affected to normal, frequently homozygous individuals are more severely affected | All offspring normal, but all are carriers | All daughters affected, no sons affected | X-Y x XX affected male x normal female<br>No affected offspring all daughters carries |
| Variations on pattern | Late onset; Incomplete penetrance variable expressively | Complementation between genetically distinct forms of disease; incomplete penetrance | Variable Expression in heterozygous females possibly mimicking incomplete penetrance; more severe in males | Heterozygous females may show decreased levels of gene product, but with substantial variation |

# 2.2 Statistical Methods

## 2.2.1 Linkage analysis

### 2.2.1.1 Genome wide approach

There are two general strategies for identifying complex trait loci depending on what is known about the trait biologically. If not reasonable hypothesis-based candidate genes can directly be tested, the second strategy, a hypothesis generating approach is considered instead. In this case, anonymous polymorphisms uniformly distributed throughout the genome are tested for presence of linked trait locus at each of the loci. This is so called positional cloning or genome wide scan strategy, which represents a unique tool for detecting previously unknown trait. Positional cloning begins with the identification of a chromosomal region that is transmitted within families, along with the disease phenotype of interest (genetic linkage). Positional cloning has been extremely useful in the identification of genes responsible for diseases with simple Mendelian inheritance. The

ultimate goal of positional cloning is to identify sequence variants within the gene associated with the phenotype.

Sometimes, Linkage and association studies are occasionally mixed up. They aim to answer different questions and provide different answers. Linkage is a phenomenon of cosegregating loci, within families. Linkage studies are used for coarse mapping as they have a limited genetic resolution of about 1 cM. If two markers are close, there will not be much recombination between them and they will cosegregate. Association studies at the population level are the next step for fine mapping. Association may result from direct involvement of the gene or linkage disequilibrium (LD) with the disease gene at the population level. Linkage always leads to an association but this is usually intrafamilial with no association at the population level (linkage of genotype for a genetic marker to disease may be unique to the particular family). In other words, linkage does not necessarily mean a consistent association with a particular allele. Allelic association, on the other hand, may or may not be due to linkage. While recombination fraction is what linkage studies rely on, LD is the foundation of association studies. The assumption is that the genetic marker studied is close enough to the actual disease gene and this will result in an allelic association at the population level (Jorde et al. 2000, Weiss et al. 2002, Carlson et al. 2004, Morton et al. 2005). Association studies focus on population frequencies, whereas linkage studies focus on concordant inheritance. One may be able to detect linkage without association when there are many independent trait-causing chromosomes in a population; or association without linkage when an allele explains only a minor proportion of the variance for a trait, so that the allele may occur more often in affected individuals but does a poor job of predicting disease status within a pedigree (Lander & Schork, 1994).

### 2.2.1.2 Genetic markers

Alleles are alternative forms of a gene and are present at the specific locus, which is at specific position on the genome. Variation in sequence at this position (the position of allele) or locus will lead to a phenomenon called polymorphism. Polymorphism is a

change in the DNA sequence or repeat element at a specific location, these are called markers. Every individual might change at this location of allele and since markers are spanned all over the genome, and they are extremely useful in mapping human disease genes as they are close to the disease gene. Many such markers have been identified of which some are RFLP (Restriction Fragment Length Polymorphism), RAPD (Random Amplification of Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), Microsatellites and recently SNPs (Single Nucleotide Polymorphism).

Difference in the genome between individuals has the potential to effect the function of the gene and hence the gene product, which might lead to diseases. Most commonly used genetic markers these days are microsatellites and SNPs because they are more advantageous over first generation DNA markers (RFLPs, RAPDs etc.).

Microsatellites are DNA regions with variable number of short tandem repeats flanked by a unique sequence (Queller et al. 1993). The repeats are usually simple dinucleotides with dinucleotide repeated about ten times. Human genome has highly polymorphic mono, tri and tetra or bigger repeat elements and the high degree of polymorphism in the repeats make them marker of choice for mapping studies. Some of the advantages of using microsatellites are that they are locus specific, codominance of alleles, PCR based, random distribution throughout the genome, and they are often quite informative. Sometime it can be misleading, for example the heterozygotes can be misclassified as homozygotes when null alleles occur due to mutation in primer annealing sites. Microsatellites have many synonyms, like SSLP (Singel Sequence Length polymorphism), SSR (Simple sequence Repeats), STMS (Sequence Tagged Microsatellites).

SNP (Single Nucleotide Polymorphism) is a small genetic variation in the DNA sequence. SNP variation occurs when a single nucleotide, for example **A**, replaced by **T**, one of the three nucleotides (**T, G, C)** (illustrated in following two sequences).

Seq 1 ATT **A** AATCCA

Seq 2 ATT **T** AATCCA

**(A ß à T)**

SNPs are bi-allelic, and they may occur in non-coding regions, or in the coding regions, which would be more interesting since they cause variation in the function of the protein. Human genome contains about 10-30 million SNPs with an average of SNP every 100-300 bases, more than 4 million SNPs have been identified and the information is publicly available (http://www.ncbi.nlm.nih.gov/projects/SNP/). They are stable from evolutionary stand point by not changing much from generation to generation making them easier for genetic studies.

### 2.2.1.3 Introduction of linkage analysis

Linkage analysis plays an important role in genetic epidemiology because it identifies a biological mechanism for transmission of a trait or disease. The term “linkage” has been used to denote the situation where alleles from two loci segregate together in a family. The most obvious biologic explanation for such an observation is that the two loci are physically located near one another on the same chromosome. Elston (1981) argues that demonstrating linkage is the highest level of statistical “proof” that a disease is due to a genetic mechanism. While proof of genetic control must be identification of the gene product and a biologic explanation for pathogenesis, confirmation of reported linkage in multiple studies can establish genetic transmission of a complex disease. The position of markers linked to a disease on the genetic or physical map of the human genome automatically uncovers further areas for research at the molecular level.

When there is linkage between two loci, it might be possible that the specific alleles segregating together in one family may differ from alleles at these same loci segregating together in other families. Family studies are always necessary to measure genetic linkage. While population studies can be used to detect general association between a given allele at a marker locus and a disease, they cannot test for genetic linkage or estimate the recombination fraction between different loci. Association is property of alleles, while linkage is a property of loci and must involve all alleles at the marker locus.

Linkage is widely used to map markers on each chromosome in the human genome, to map genetic diseases, and further to identify genetic forms of common diseases. There are several critical questions need to be considered before carrying linkage analysis, because of the large number of linkage studies now being undertaken and linkage analyses are being extended to diseases of complex etiology where genetic transmission of the disease is very complex. Thus, issues like how to incorporate genetic marker information and linkage analysis into studies of complex diseases, where both genetic and nongenetic factors may jointly distribute to the disease, represent a major challenge in genetic epidemiology (Risch, 1990a).

The probability of a marker being informative for linkage analysis is a function of the frequency of heterozygotes, which in turn is a function of the number of marker alleles and their frequencies. The polymorphic information content (PIC) is used to summarize the probability of a marker locus being informative (Botstein et al., 1980), and highest PIC score are attained by markers with many equally frequent alleles.

Linkage analysis is straightforward when recombinant offspring can be counted directly. It is equally straightforward in experimental genetic where mating can be arranged. In these situations, questions of sample size and statistical power are addressed by relying on the familiar binomial distribution, and counts of recombinant versus nonrecombinant children can be tallied over all families of a single mating type. However, in human genetics, there are several reasons why this process is generally more complicated. Firstly, not all mating are “phase-known” (some family member’s data might be unavailable or not informative). Secondly, diseases with incomplete penetrance or age-dependent penetrance make it impossible to identify all carriers of the disease allele accurately. Thirdly, if carrier detection is impossible, for strictly recessive diseases, linkage analysis becomes uniformly difficult due to the limited information on genotypes of critical individuals.

### 2.2.1.4 Maximum likelihood methods for linkage analysis

The maximum likelihood approach to linkage analysis dates back to Haldane and Smith (1947), but did not come widely used until Morton (1955) published tables of log-odds (or LOD) scores that could be used in the sequential analysis of family data. Several widely available computer software packages for two point or multipoint analyses now exist (Lathrop et al., 1985; Cottingham et al. 1993; Schäffer et al. 1994; Kruglyak et al. 1996; Kruglyak and Lander 1998).

As seen in Table 2-3, the probability of co-segregation is dependent on the genetic distance between these loci, usually measured by the recombination fraction $\theta$. Also, the probability of a child inheriting the disease allele and the "-"allele at the marker or inheriting the normal allele and the "+" allele should be less than ¼, again dependent on the actual recombination fraction between the loci.

**Table 2-3. Probability of receiving alleles at two loci (a dominant disease locus with alleles D and d, a marker locus with allele + and -, under independent linkage)**

| | Linkage Disease | |
|---|---|---|
| Marker | D | D |
| + | $\frac{1-\theta}{2}$ | $\frac{\theta}{2}$ |
| - | $\frac{\theta}{2}$ | $\frac{1-\theta}{2}$ |

The method to calculate linkage is developed by Newton E Morton (1955). The probabilities effectively determine the likelihood function on a family where $r$ is the number of recombinant children out of total of $n$ children and $\theta$ is the recombination fraction, that is,

$$L(\theta) = \left(\frac{\theta}{2}\right)^{r}\left(\frac{1-\theta}{2}\right)^{n-r} .$$

So this likelihood is merely proportional to the actual probability of observing any one family. The null hypothesis of no linkage corresponds to a recombination value $\theta$=0.5.

The log-odds or LOD score serves as a useful summary of all information on linkage, that is,

*LOD*=log(Probability with linkage( $\theta$ ) / Probability with no-linkage (0.5)).

$$LOD = \log \frac{L(\theta)}{L(\theta = 0.5)} = \log\left( \frac{(\theta/2)^r [(1-\theta)/2]^{n-r}}{(0.25)^n} \right) = \log 2^n \theta^r (1-\theta)^{n-r} \quad .$$

Since the likelihoods of independent families are multiplied to accumulate a total likelihood for any one sample, those log-likelihoods or LOD scores are simply summon over all independent mating. A LOD of 3.0 or more has been considered strong evidence for linkage, while a LOD score of -2.0 or less has been taken as evidence against linkage. These critical values correspond to 1000:1 odds for linkage and 100:1 odds against linkage, respectively, at some specified value of $\theta$ . The prior probability of two loci being linked has been estimated at approximately 5% based on the relative length of all autosomes (Renwick, 1971). The approach used in linkage analysis has evolved as a compromise between statistical principles and recognized biologic constraints. Morton (1955) originally developed these critical values in the context of sequential testing for linkage, where families were sampled until conclusive evidence either for or against linkage was accumulated. Even though the framework of sequential testing has not been strictly followed, and often estimation of $\theta$ is a primary goal, most test of significance in linkage analysis still rely on this critical value of 3.0 (Ott, 1985). The probability of a type I error (i.e., falsely identifying two loci as linked) must always consider the low prior probability of linkage , and from theoretical grounds this probability of non-linkage at an LOD score of 3.0 seems to be 3 to 4% (Smith 1986; Conneally and Rivas,1980). Empirical evidence suggests that less than 2% of linkage giving LOD>=3.0 are spurious (Rao et al., 1978). Maximum likelihood approaches to linkage analysis have very little power to detect loose linkage ($0.25 < \theta < 0.45$).

### 2.2.1.5 Multipoint analysis

Multipoint mapping refers to linkage analysis of more than two loci at a time. Considering multipoint loci simultaneously gives substantial increase in information for both estimating the recombination fraction and establishing the order of linked loci.
Recombination fractions can be converted to map distances by the use of a mapping function, for close linkage, map distances and recombination fraction can be assumed to be equal, although truly comprehensive genetic maps must be based on some mapping function.

Multipoint analysis has been implemented in software packages, e.g. FASTLINK (Cottingham et al. 1993; Schäffer et al. 1994), GENEHUNTER (Kruglyak et al. 1996; Kruglyak and Lander 1998). Multiple markers are particularly useful when the IBD (identical/identity by decent) relations of family members at the loci of interest are ambiguous, because multipoint analysis can use haplotype information from several markers to infer the IBD relations. However, it has its own problem. The specification of inter-marker distances is subject to error, particularly in small regions, and such misspecification can adversely affect the power of a multipoint analysis. The genetic distance between two markers is estimated empirically by observing the frequency of recombinant events in human meiosis, and then using a mapping function to convert the frequency to distance in centimorgans (cM). However these distance estimates are subject to statistical error, particularly in small regions containing markers so closely spaced that recombination between them is rare.

### 2.2.1.6 Parametric and nonparametric linkage analysis

Linkage analysis aims to retrieve all available inheritance information from pedigrees and to test for coinheritance of chromosomal regions with a trait. Basically, one can use either parametric method, which is testing whether the inheritance pattern fits a specific model, or use nonparametric method, which is testing if the inheritance pattern deviates from expectation under independent assortment.

In a pedigree, *nonfounders (n)* are those individuals whose has parents in the pedigree. Individuals whose don't have parents in the pedigree are defined as *founders (f)*. Founders will be assumed to be unrelated to each other; they carry *2f* alleles that are distinct by descent. First, one starts to infer information about the inheritance pattern of a pedigree, and then decide if the inheritance information indicates the presence of a trait-causing gene. The inheritance pattern at each point *x* (genetic location) is completely described by a binary inheritance vector *v(x)=(P1,M1,P2,M2,…Pn,Mn)* whose coordinates describe the outcome of paternal and maternal meioses giving rise to the *n nonfounders* in the pedigree (Lander and Green 1987). So, the inheritance vector specifies which of the *2f* distinct *founder* alleles are inherited by each *nonfounder*. The set of all $2^{2n}$ possible inheritance vectors will be denoted *V*.

In the practical situation, it is not possible to determine the true inheritance vector at every point in the genome, because not all of the genotyping are phase known due to lots of reasons. Partial information extracted from a pedigree can be used to compute a probability distribution over the possible inheritance vector at each locus in the genome, that is $P(v(x)=w)$ for all inheritance vectors $w$ ($w$ from $V$,$2^{2n}$ possible inheritance vectors). In the absence of any genotype information, all inheritance vectors are equally likely according to Mendel's first law, and the probability distribution is uniform (*P*-uniform). As genotype information is added, the *P*-uniform is concentrated on certain inheritance vectors.

In parametric analysis, one assumes a model describing the probability of phenotype given genotype at diseases locus and calculates the likelihood ratio under the hypothesis that a disease gene is at *x*, versus the hypothesis that is unlinked to *x*. In the special case when the inheritance vector is known, the scoring function *S* is the likelihood ratio,

$$S = LR(v) = \frac{P(\Phi \mid v)}{\sum_{\omega \in v} P(\Phi \mid w) P_{uniform}(w)}$$ (Kruglyak et al. 1996).

$P(\Phi \mid v)$ is the likelihood of observed phenotypes $\Phi$, conditioned on the particular inheritance vector *v*; it depends only on the penetrance values and allele frequencies at the disease locus. For each *v*, one can efficiently compute $P(\Phi \mid v)$ by a simple adaptation of standard peeling methods for pedigrees without loops (Elston and Stewart 1971; Lange and Elston 1975; Cannings et al. 1978; Whittemore and Halpern 1994b) and by a combination of peeling, loop breaking, and enumeration of founder genotypes for pedigrees with loops. Calculating the likelihood for each of the $2^{2n-f}$ equivalence classes of inheritance vectors is very quick for moderate-sized pedigrees, both with and without loops.

In the general case, one takes the expectation of the scoring function over the inheritance distribution, as in equation (2):

$$\overline{LR}(\chi) = \sum_{w \in v} LR(w) P(v(\chi) = w) = \frac{\sum_{w \in v} P(\Phi \mid w) P_{complete}(w)}{\sum_{w \in v} P(\Phi \mid w) P_{uniform}(w)}$$ (Kruglyak et al. 1996).

This expression is seen to be equivalent to the traditional definition of the likelihood ratio; the numerator is proportional to the multipoint likelihood when the disease gene is at *x*, whereas the denominator is proportional to the unlinked likelihood. According to long-standing tradition, one reports the LOD score, $\log_{10} \overline{LR}$.

Because parametric linkage analysis can be highly sensitive to misspecification of the linkage model ( Clerget-Darpoux et al 1986), nonparametric analysis is a key tool for all but the simplest of traits. Nonparametric analysis has primarily two methods. The first approach is to break pedigrees into nuclear families and apply sib-pair analysis; it wastes a great deal of inheritance information contained in pedigree structure. To partly utilize pedigree information, Weeks and Lange (1998, 1992) developed the affected-pedigree-member method (APM). APM solves the issue of tracing the inheritance pattern in a pedigree by focusing on whether affected relatives happen to show the same alleles at a locus (i.e., identity/identical by state (IBS), regardless of whether the allele is actually inherited from a common ancestor (i.e., identity/identical by decent (IBD)). The extent of IBS sharing among all pairs of affected members of the pedigree is compared with

Mendelian expectation under the hypothesis of no linkage. There are two suitable scoring functions for non parametric analysis, which are S-pair and S-all. In S-pair scoring function; IBD sharing in pairs, one possible approach is to count pair wise allele sharing among affected relatives. Given the inheritance vector *v*, $S_{pairs}(v)$ is defined to be the number of pairs of alleles from distinct affected pedigree members that are IBD. The traditional APM statistic also counts pair wise allele sharing, but it based on sharing IBS rather than on sharing IBD; the two statistics will coincide only at markers for which IBS unambiguously determines IBD.

In S-all scoring function; IBD sharing in larger sets, one can often increase statistical power by considering larger sets of affected relatives. Whittemore and Halpern (1994a) proposed a statistic to capture the allele sharing associated with a given inheritance vector *v*. Let *a* denote the number of affected individuals in the pedigree, let *h* be a collection of alleles obtained by choosing one allele from each of these affected individuals, and let $b_i(h)$ denote the number of times that *i*-th founder allele appears in *h* (for *i*=1,…,2*f*). The score $S_{all}$ is defined as

$S_{all}(v) = 2^{-a}\sum_{h}\left[\prod_{i=1}^{2f} b_i(h)!\right]$ ( Kruglyak et al. 1996),

where the sum is taken over the $2^a$ possible ways to choose *h*. In effects, the score is the average number of permutations that preserve a collection obtained by choosing one allele from each affected person. It gives sharply increasing weight as the number of affected individuals sharing a particular allele increases.

For either approach, a normalized score was defined

$Z(v) = [S(v) - \mu]/\sigma$ ,

Where μ and σ are the mean and SD (Standard Deviation) of $S$ (scoring function) under P-uniform (the uniform distribution over the possible inheritance vectors). Under the null hypothesis of no linkage, the normalized score Z has mean 0 and variance 1. To combine scores among *m* pedigrees, one can take a linear combination

$$Z = \sum_{i=1}^{m} \gamma_i Z_i ,$$

where $m$ is the number of pedigrees, $Zi$ denotes the normalized score for $i$-th pedigree, and the $\gamma i$ are weighting factors. Now this $Z$ is referred as NPL score for the collection of the pedigree.

# 2.3 Prostate Cancer

## 2.3.1 Inheritable factors in common cancers

A small part of all cancers is associated with inherited predisposition to cancer (Ponder 2001). There are two kinds of mechanisms that associate cancer risk with genetic status. These two mechanisms are shown in table 2-4. First, genetic predisposition associated with a very high risk can be used to explain inherited cancer syndromes. Second, genetic mechanism associated with familial cancers may be caused by genetic susceptibility via individual or ethnic polymorphisms.

**Table 2-4 inherited predisposition to cancer (Ponder 2001)**

| | Contribution to overall cancer incidence | Clinical features | Frequency of predisposing alleles | Effect on individual risk |
|---|---|---|---|---|
| Inherited cancer syndromes | 1-2% | Rare cancers or combination of cancers. Mendelian dominant inheritance | Rare (1:1000 or less) | Strong Lifetime risk up to 50-80% |
| Familial cancers | Up to 10% depending on definition | Families with several cases of common cancers. Generally dominant inheritance | Uncommon to common | Moderate to weak |
| Predisposition without evident family clustering | No precise figure possible substantial fraction of cancer incidence within predisposed population | Single cases of cancer at any site, some with one or two affected relatives. | Multiple common alleles | Weak |

The estimated values for heritability of four common cancers obtained from cohort or twin studies were shown in table 2-5. In study 1, Familial relative risk (FRR) from the Utah population database was estimated by detecting all cases of cancer in first degree relatives of 35,228 cancer probands (Goldgar et al. 1994). In study 2, FRR was estimated

by using Swedish Family can database (Dong and Hemminki 2001). In both studies, Moderate risk ratios were used to characterize the most common cancers. In 2000, data on 44,788 pairs of twins listed in the Swedish, Danish, and Finnish twin registries were gathered in order to assess the risks of cancer (Lichtenstein et al. 2000). Comparing by heritable factors, prostate cancer was placed first among the common cancers with over 40% of heritability, colorectal cancer was second and breast cancer third. Of interest, both for breast and colorectal cancer, major risk genes have been identified, making them in that respect different from prostate cancer.

**Table 2-5 Heritability of four common cancers**

| Cancer type | Study 1<br>Family risk ratio | Study 2<br>Family risk ratio | Proportion of variance due to heritable factors |
|---|---|---|---|
| Lung | 2.55 | 1.68 | 0.26 |
| Colorectal | 2.54 | 1.86 | 0.35 |
| Prostate | 2.21 | 2.82 | 0.42 |
| Breast | 1.83 | 1.86 | 0.27 |

## 2.3.2 Prostate Cance

Prostate cancer is a cancer disease that develops in the prostate (a gland in the male reproductive system). It occurs when cells of the prostate start mutation and mutated cell begin to increase out of control. These cells may even spread from the prostate to other organs. The symptoms are difficulty in urinating, erectile dysfunction, and even causing pain.

The rates of prostate cancer vary widely between countries; it is not so common in South and East Asia, more common in Europe, and most common in the United States (Parkin et al. 1997). According to the American Cancer Society (www.cancer.org), prostate cancer is least common among Asian men and most common among black men, with figures for white men in-between. However, these high rates can be reasoned to the increasing rates of detection.

Prostate cancer occurs mostly in men over fifty years of age. This cancer can occur only in men, since the prostate is only in the male reproductive system. It is the most common type of cancer in men in the Finland; take current incidence from the Finnish Cancer Registry (www.cancerregistry.fi) where it is responsible for more male deaths than any other cancer, except lung cancer. However, there are cases like many men who develop prostate cancer never noticed, undergo no therapy, and eventually die of other reasons. In most cases, most of the patients are very old, so they often has other diseases, which makes their causes of death unrelated to the prostate cancer, such as heart/circulatory disease, pneumonia, other unconnected cancers or old age. Many other factors, including genetics and diet, have been implicated in the development of prostate cancer (Steinberg et al. 1990; Gann et al, 2005)

The methods used in detecting prostate cancer is mostly physical examination or by screening blood tests, such as the PSA (prostate specific antigen) test. Suspected case of prostate cancer is confirmed by examining a piece of the prostate (biopsy) under a microscope. Further tests are used to test for spreading of cancer, such as X-rays and bone scans.

The specific causes of prostate cancer remain unclear. A man's risk of developing prostate cancer may be related to his age, ethnicity, diet habit, medications, and other possible factors. Result from segregation analyses points out that familial clustering of prostate cancer can be best explained by transmission of a rare hereditary factor accounting for 5-10% of total prostate cancer cases (Carter et al. 1993). Beside that, two large twin studies reported higher prostate cancer concordance rate for monozygotic twins versus dizygotic twins suggesting a strong genetic factor on risk (Page el al. 1997; Lichtenstein et al. 2000). Early hope that searching of susceptibility genes would be as straightforward as it was for breast and colorectal cancer. However, this hope has not been fulfilled by the difficulty of replicating promising regions of linkage (Nupponen and Carpten 2001; Schaid 2004). A major problem in prostate cancer genetics is the correctness of modes of inheritance for familial prostate cancer. Some cases of prostate cancer are due to an autosomal susceptibility locus with an allele or alleles that collectively show a dominant and age dependent way (Carter et al. 1992; Grönberg et

al.1997a; Schaid et al. 1998). Other researches have suggested either for recessive or X-linked mode of inheritance (Monroe et al. 1995; Pakkanen et al.2007). One other reason for fruitless of linkage studies is the high prevalence of phenocopies. When the sporadic cases are analyzed as affected individuals, but they do not share the same disease locus with the hereditary family cases, linkage results are substantially questioned. At the same time, the evidence also suggests a much more complex genetic basis of prostate cancer than expected. A segregation study in 263 prostate cancer families reported that the disease is more likely caused the contributions of two to four prostate cancer susceptibility genes than one gene (Colon et al. 2003). A new analysis method to twin study data provided by Lichtenstein was analyzed (Risch 2001; Lichtenstein et al. 2000). Similarly, the results of Page of reanalyzed (Schaid 2004; Page et al. 1997). The new result suggests that the genetic basis of prostate cancer can not be fully explained by independent, rare, autosomal dominant mutations but rather by recessive and / or multiple interacting loci (Schaid 2004). Furthermore, the modifier genes and environmental factors can influence the phenotype of both high and low penetrance genes (de la Chapelle 2004).

## 2.3.3 Previous linkage findings

Since 1996, research groups worldwide have collected data from families with multiple prostate cancer cases and have preformed linkage analyses to find the susceptibility genes. Numbers of regions have been suggested to harbor hereditary prostate cancer genes. Table 2-6 shows the most significant initial linkage reports.

**Table 2-6. Putative hereditary prostate cancer susceptibility loci**

| Locus/Gene | Location | Reference |
|---|---|---|
| HPC1/RNASEL | 1q24-25 | (Smith et.al. 1996) |
| PCAP | 1q42.2-43 | (Berthon et al.1998) |
| HPCX | Xq27-28 | (Xu et al. 1998) |
| CAPB | 1p36 | (Gibbs et al.1996b) |
| HPC20 | 20q13 | (Berry et al. 2000) |
| MSR1 | 8p22-23 | (Xu et al. 2001c) |
| HPC2/ELAC2 | 17p11 | (Tavtigian et al. 2001) |

Chromosome 1q24-25 was reported in the first genomic scan as prostate cancer loci with a maximum HLOD of 5.43 (Smith et al. 1996). The following subsequent analysis of the same set of families reported strongest evidence for linkage to *HPC1* among men with an early age of diagnosis (age<65years), and the evidence increased if there were at least five men affected (Grönberg et al. 1997b). Surprisingly, subsequent reports attempting to replicate the linkage for *HPC1* were not so successful. Hence, the International Consortium for Prostate Cancer hereditary Genetics (ICPCG) performed a meta-analysis of 772 families affected by hereditary prostate cancer from North America, Australia, Finland, Norway, Sweden, and the United Kingdom (Xu 2000). Suggestive evidence of linkage to *HPC1* locus with a HLOD of 1.4 was reported. Other findings on chromosome one were reported in a genome-wide scan of 47 French and German families, in which *PCAP* located in 1q42.2-43 was detected with a maximum two-point LOD score of 2.7 (Berthon et al. 1998). In 2001, linkage analysis with 50 microsatellite markers spanning chromosome 1 in 159 hereditary cancer families was carried out, the highest LOD score was located at 1q24-25, with HLOD of 2.54 (Xu et al. 2001b).

A maximum two-point LOD score of 3.22 on chromosome 1p36 was observed in 12 families with a history of both prostate and primary brain cancers, and the locus was termed *CAPB* for cancer of the Prostate and Brain (Gibbs et al. 1996). An excess of brain and central nervous system cancers had been previously reported in high-risk prostate cancer families (Goldgar et al. 1994, Isaacs et al. 1995). In addition, 1p36 is a region of frequent loss of heterozygosity (LOH) in brain tumors (Takayama et al. 1992, Bello et al. 2000). Later studies have not been able to replicate this result.

Evidence for prostate cancer linkage at 8p22-23 was found with a peak HLOD of 1.84 in 159 pedigrees affected by HPC (Xu et al. 2001c). In prostate cancer LOH on 8p was reported to be one of the most frequent somatic alterations, occurring in >60% of cancers (Cunningham et al. 1996). In 2003, linkage to 8q22-23 was replicated with the linkage analysis among 57 families from Sweden (Wiklund et al. 2003). In their study, evidence of linkage was seen in families with early-onset prostate cancer with a peak multipoint non parametric linkage score of 2.01 (P=0.03) and in families with a small number of

affected individuals with a linkage score of 2.25 (P=0.01). Furthermore, a recent genome wide linkage analysis from Germany observed linkage at 8p22 in the family collection of 139 prostate cancer families (Maier et al. 2005b).

A genome wide scan for prostate cancer predisposition loci using a small set of high risk prostate cancer pedigrees from Utah was performed (Tavtigian et al. 2001). The first 8 pedigrees analyzed reports suggestive evidence of linkage on chromosome 17p11 and finally, the analysis was expanded to 33 pedigrees which gave a maximum multipoint LOD score of 4.3 (Tavtigian et al. 2001). In contrast, no evidence was found for linkage of *HPC2* locus in a total sample of 159 families, nor in any subset of pedigrees based on characteristics that included age at onset, number of affected members, male to male disease transmission, or rare (Xu et al. 2001a).

Linkage to a locus on 20q13 was reported with two point LOD score of 2.69 for the dominant model and 3.11 for the recessive model (Berry et al. 2000). The strongest evidence of linkage was found with the pedigree having <5 family member affected with prostate cancer, a later average age at diagnosis, and no male-to-male transmission. Two studies have confirmed this finding (Bock et al. 2001, Zhang et al. 2001). In 2003, eight genome wide scans for prostate cancer susceptibility (Cunningham et al. 2003b; Edwards et al. 2003a; Janer et al. 2003; Lange et al. 2003; Schleutker et al. 2003; Wiklund et al. 2003; Witte et al. 2003; Xu et al. 2003) were published in *Prostate* together. They reported the only LOD score >3 when they identified a linkage to chromosome 20 with HLOD scores of 4.77. Thus they confirmed their initial finding on chromosome 20 (Berry et al. 2000). Furthermore, a large study performed by the ICPCG among 1234 pedigrees failed to replicate linkage to *HPC20* (Shaid, Chang & ICPCG. 2005).

Xq27-28 was identified in a combined study of four groups representing North America, Finland and Sweden (Xu et al. 1998). There was a maximum two-point LOD score of 4.60 and *HPCX* was estimated to account for 16% of HPC overall. In 2000, a subgroup analysis among 57 Finnish HPC families was performed, which indicates that families with no male-to-male transmission and a late age at diagnosis (>65 years) accounted for

most of the *HPCX*-linked cases (Schleutker et al. 2000). Not so many studies have provided some supporting evidence for linkage to *HPCX*, only a few found the evidence in families with male-to-male transmission (Lange et al. 1999; Bochum et al 2002).

Ten other linkage regions with LOD scores >2 were reported, on chromosomes 2, 3, 4, 5, 6, 7, 9, 16, 17, and 19. A Finnish genome wide linkage analysis indicated two chromosomal regions, 3p25-26 with two-point LOD score of 2.57 and 11q14 with two-point LOD score if 2.97 (Schleutker et al. 2003). Fine-mapping with 39 microsatelite markers in 16 familes validated 3p25 as a prostate cancer susceptibility locus in Finland (Rökman et al. 2005). The maximum multipoint HLOD was 3.39 at 3p26 and 1.42 at 11q14.

# 3 Objectives of the study

In the present thesis the genetic susceptibility to prostate cancer was studied in Finnish prostate cancer families, with following specific aims:

1. To scan the whole genome for novel susceptibility loci for prostate cancer.
2. To compare the result with the previous genome wide scans of Finnish and other populations.

# 4 Materials and Methods

## 4.1 Study Objects

HPC families

The 56 multiplex Finnish prostate cancer families were selected based on informativeness for linkage analysis, and were evaluated by the number of affected relatives and the number of relatives from whom a blood sample was available for genotyping. The families selected had at least three first- or second- degree relatives affected with prostate cancer and a blood sample available for genotyping from at least one affected person per pedigree. Characteristics of the 56 families are show in Table 4-1.

**Table 4-1 Characteristic of 56 families, each family has MMTrans column (stands for male to male transmission), TotalInd column (total number of individuals in each family), DnaSample column (total number of Dna samples in each family have), AffectedNumber column (total number of affected persons in the family), AverageAge column (average age of prostate cancer patient at diagnosed date).**

| FamID | MMTrans | Bilineal | TotalInd | DnaSample | AffectedNumber | AverageAge |
|---|---|---|---|---|---|---|
| 2001 | | | 26 | 16 | 4 | 59.63 |
| 2004 | Y | N | 25 | 9 | 3 | 72 |
| 2011 | Y | N | 10 | 5 | 3 | 61 |
| 2013 | Y | N | 19 | 10 | 4 | 70.25 |
| 2015 | Y | N | 40 | 9 | 6 | 57.04 |
| 2019 | N | N | 22 | 11 | 3 | 69.66 |
| 2031 | Y | N | 19 | 3 | 3 | 71.67 |
| 2032 | N | N | 10 | 3 | 3 | 73.33 |
| 2038 | N | N | 21 | 1 | 4 | 65.25 |
| 2045 | | | 19 | 8 | 4 | 63.27 |
| 2051 | N | N | 27 | 8 | 5 | 72 |
| 2062 | N | N | 60 | 13 | 6 | 57.53 |
| 2066 | N | N | 35 | 25 | 4 | 62.25 |
| 2069 | N | N | 13 | 8 | 3 | 65 |
| 2082 | Y | N | 15 | 11 | 3 | 68.33 |
| 2083 | N | N | 48 | 5 | 4 | 71.25 |
| 2091 | Y | N | 21 | 6 | 3 | 67.67 |
| 2097 | Y | N | 37 | 1 | 6 | 75.83 |
| 2101 | | | 43 | 2 | 2 | 60.95 |
| 2102 | N | N | 35 | 7 | 4 | 58.93 |
| 2106 | N | N | 28 | 14 | 3 | 70.33 |
| 2109 | Y | N | 39 | 7 | 3 | 53.45 |
| 2111 | Y | N | 35 | 1 | 4 | 65.25 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2113 | | | 42 | 4 | 2 | 57.78 |
| 2119 | Y | N | 9 | 6 | 3 | 74 |
| 2139 | Y | N | 11 | 5 | 3 | 66.67 |
| 2143 | Y | N | 5 | 3 | 3 | 67 |
| 2176 | | | 31 | 9 | 7 | 69 |
| 2215 | Y | N | 25 | 5 | 4 | 67.68 |
| 2232 | N | N | 32 | 7 | 3 | 54.54 |
| 2236 | N | N | 36 | 7 | 4 | 60.38 |
| 2238 | Y | N | 46 | 9 | 5 | 56.65 |
| 2248 | N | N | 20 | 11 | 4 | 60.25 |
| 2264 | | | 17 | 12 | 4 | 66.5 |
| 2275 | N | N | 25 | 16 | 4 | 70.75 |
| 2279 | N | N | 20 | 5 | 4 | 47.86 |
| 2283 | Y | N | 19 | 6 | 4 | 70.5 |
| 2291 | Y | N | 55 | 0 | 5 | 51 |
| 2292 | Y | N | 20 | 8 | 4 | 49.44 |
| 2308 | | | 13 | 7 | 4 | 64 |
| 2351 | Y | N | 14 | 6 | 3 | 61.33 |
| 2359 | N | N | 12 | 7 | 3 | 73.67 |
| 2362 | N | Y | 23 | 10 | 4 | 60.25 |
| 2364 | N | N | 14 | 5 | 3 | 58.67 |
| 2374 | N | N | 23 | 17 | 3 | 62.67 |
| 2375 | N | N | 26 | 18 | 3 | 52.33 |
| 2391 | Y | N | 52 | 5 | 5 | 61.4 |
| 2393 | N | N | 17 | 11 | 3 | 69 |
| 2394 | N | N | 45 | 35 | 5 | 57.6 |
| 2400 | N | N | 14 | 10 | 3 | 57.33 |
| 2402 | N | N | 20 | 13 | 4 | 67 |
| 2403 | N | N | 25 | 14 | 3 | 67 |
| 2404 | N | Y | 19 | 8 | 5 | 64.6 |
| 2408 | N | N | 26 | 3 | 3 | 70 |
| 2413 | N | N | 18 | 12 | 4 | 61.5 |
| 2427 | N | N | 32 | 24 | 6 | 68.83 |

20 families showed male to male (M-M) transmission, 501 DNA samples were gathered, 490 microsatellites markers spanning whole genome were genotyped, and in these families, there were 214 affected persons in total. The diagnoses of the patients and family histories were initially obtained from questionnaires and subsequently confirmed from medical records, parish records and the Finnish Cancer Registry. Written informed consent was obtained from all living patients and their relatives who had donated blood sample for the study.

## 4.2 Methods

### 4.2.1 Flow chart

General process is shown in figure 4-1. Allele frequencies were estimated by using SIB-PAIR (Duffy, 2006) program, then Mendelian errors were checked by PEDCHECK (O'Connel and Weeks, 1998) program, after that Genotype data were ready to be analyzed. Two point parametric analyses were carried out through FASTLINK (Cottingham et al. 1993; Schäffer et al. 1994) program, and multipoint parametric and non parametric analyses were carried out through GENEHUNTER (Kruglyak et al. 1996; Kruglyak and Lander 1998) program.

Detailed process which contains perl scripts were shown in figure 4-2, pedigree file (Exp1.pre, containing family information included in appendices) and map file (Exp1.map, containing the distance of the marker included in appendices) are processed by *SibToPre.pl (*Perl script use to combine Exp1.pre and Exp1.map for next step, included in appendices) to generate input file (Exp1.in, included in appendices) for SIB-PAIR (Duffy, 2006) program. SIB-PAIR reads in input file and generate frequency file (Exp1_fre.txt, describing alleles' frequencies of each marker, included in appendices), the frequency file is then processed by *freToDat.pl* (Perl script use to change the format of frequency file, included in appendices) to generate DAT file (containing locus number, penetrance parameters, and other importance factors), which is compatible for FASTLINK (Cottingham et al. 1993; Schäffer et al. 1994) and GENEHUNTER (Kruglyak et al. 1996; Kruglyak and Lander 1998).

Pedigree file (Exp1.pre) has to be reformatted by MAKEPED, and a PED file (Exp1.ped) must be generated. Now, PEDCHECK can be used to check DAT file (exp1.dat included in appendices) and PED file (exp1.ped), until there are no error reports for both Level 1 and Level 2 checks. If there are Level 1 errors, one must simply locate the line from PED file (which would be indicated by PEDCHECK program), and make the corresponding changes, after Level 1 checks have been done (No error report). Then, it is possible to

continue to Level 2 checks, locate the error, and run Level 3 and Level 4 checks, which would give you suggestions concerning how to clean the errors.

After passing PEDCHECK (O'Connel and Weeks, 1998) program's level 2 checking, ped file (exp1.ped) and dat file (exp1.dat) are ready for linkage analysis. FASTLINK (Cottingham et al. 1993; Schäffer et al. 1994) should be used for two point analysis, and GENEHUNTER (Kruglyak et al. 1996; Kruglyak and Lander 1998) for single point nonparametric, and multipoint parametric and nonparametric analysis.

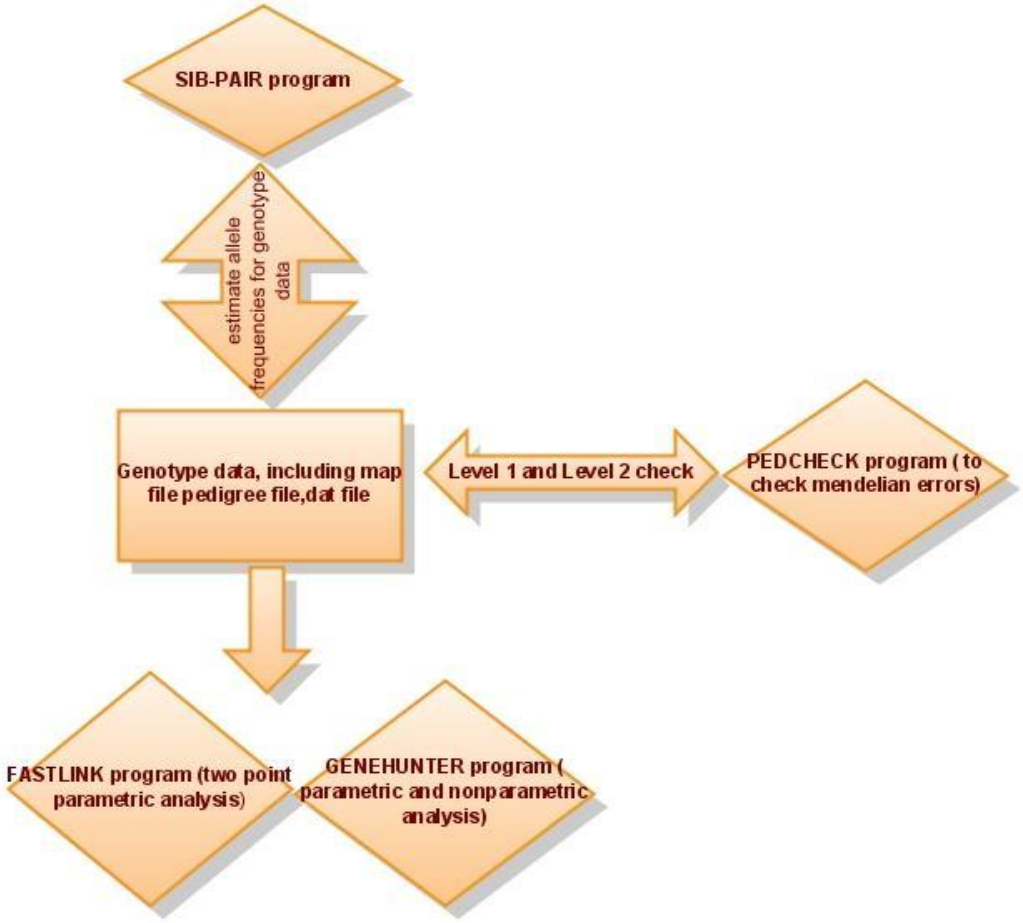


**Figure 4-1 General flow chart**

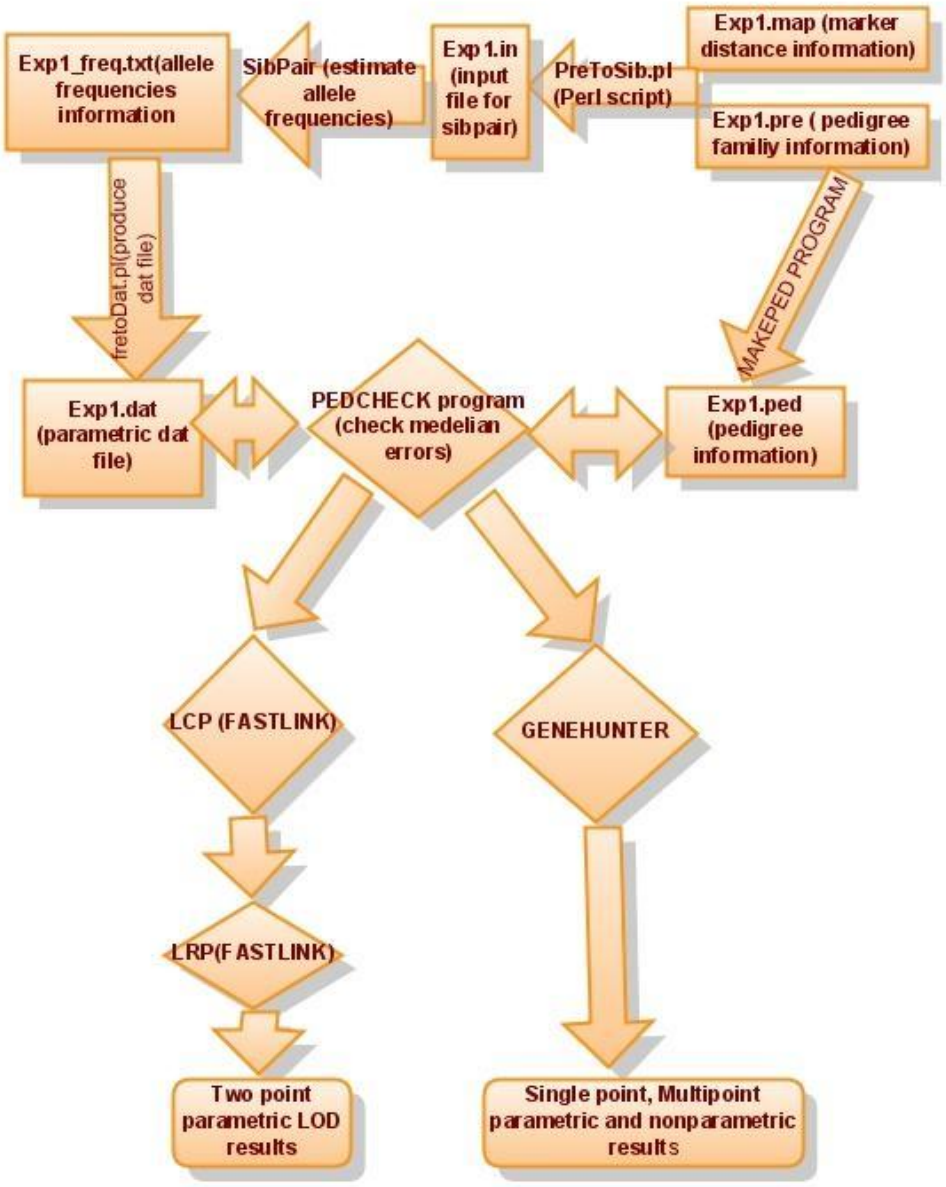


**Figure 4-2 Detailed flow chart of whole process**

## 4.2.2 SIB-PAIR

Program SIB-PAIR (Duffy, 2006) performs a number of simple nonparametric and robust analyses of family data. It is modeled on the Genetic Analysis System (Young, 1995) in terms of the command language and types of analysis. In this study, SIB-PAIR is used to estimate allele frequencies in codominant genetic systems. Two approaches are available, First approach is a straight allele count; the second approach uses contribution of each pedigree weighted by the number of founders it contains. Alternatively, the imputed and observed genotypes in the founders can be counted, or the MLE of the founder allele frequencies calculated by MCMC (Monte-Carlo Markov Chain).

SIB-PAIR is command line oriented, and prints output to the standard output (command line screen). Output can be diverted it to a file using the "out" command. The data set contains one record (newline character delimited) per individual. Records must be sorted into pedigrees. Records take the format (shown in figure 4-3) used by GAS (linkage software)

S*et loc <marker name> marker <distance to previous marker> cm*

*Pedigree-id person-id father-id mother-id sex locus-value 1…locus value n*

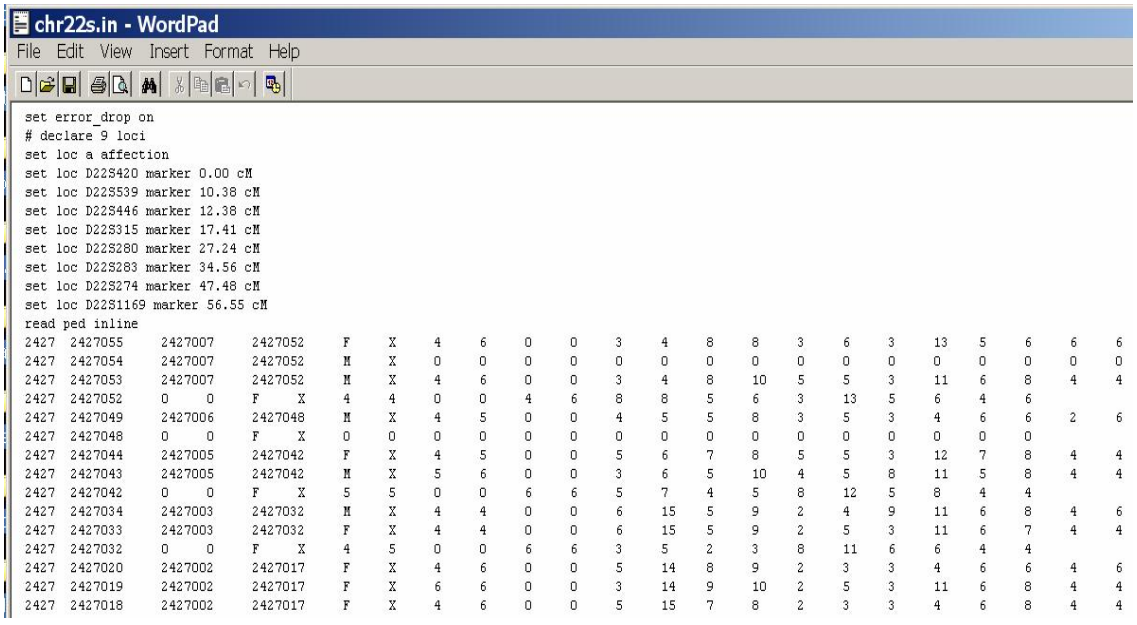


**Figure 4-3 Input file format for SIBPAIR program**

Start the program, and read in the input file by command “include chr22s.in”. The “run” command actually starts the initial processing of the pedigree. By typing command “descrie”, the program print allele frequencies (shown in figure 4-4) for marker loci, segregation ratios for binary trait, or means, variances, familial correlations and a sibship variance test for a quantitative trait.

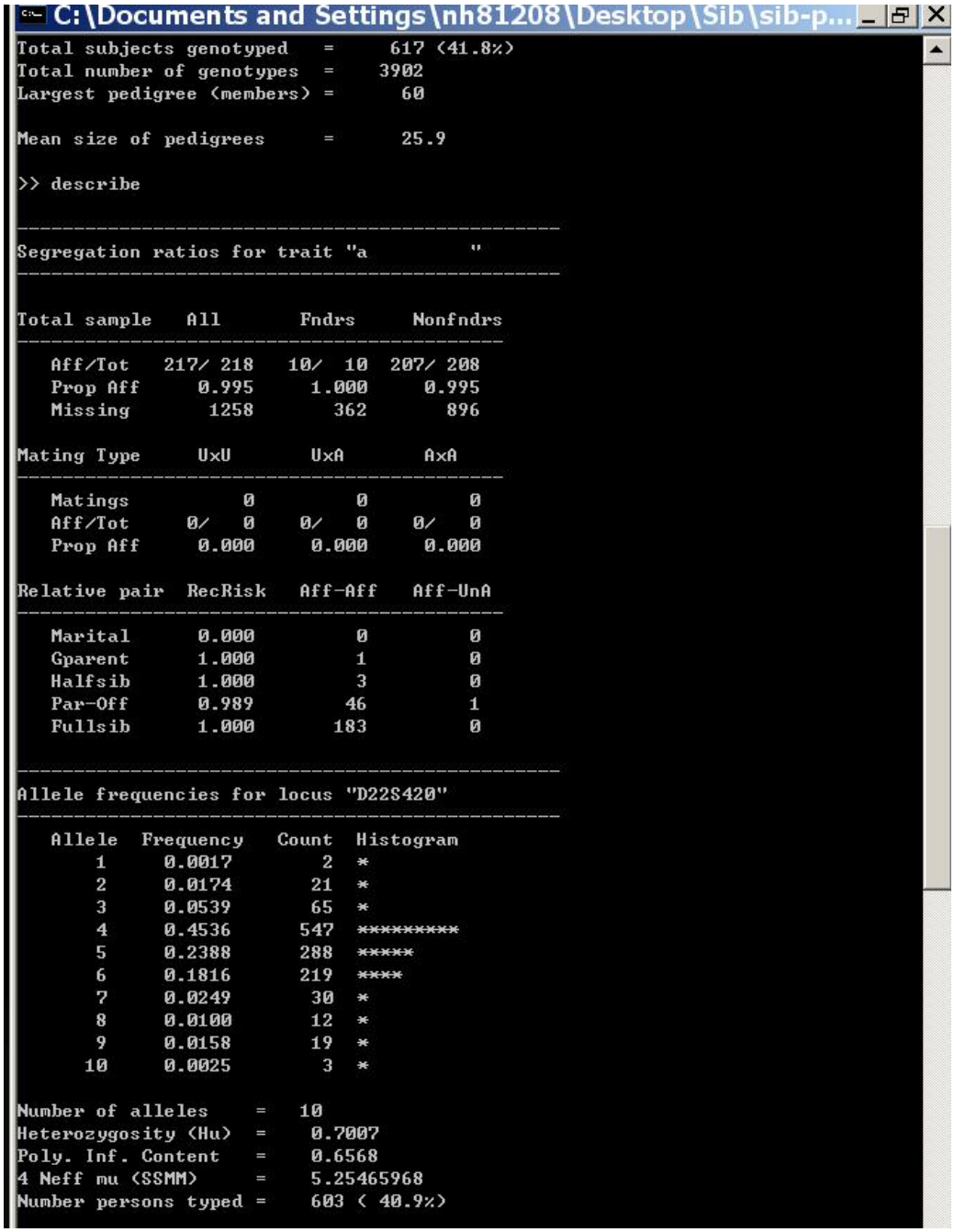


**Figure 4-4 allele frequencies calculated by SIBPAIR**

## 4.2.2 PedCheck

Prior to performance of linkage analysis, all Mendelian inconsistencies in the pedigree data should be eliminated. Identification of erroneous genotypes manually is very difficult and time consuming. In fact, sometimes the errors are not found until the running of linkage-analysis software. In the case of very large pedigrees, the effort required for finding the erroneous genotypes and to cross-reference pedigree and marker data that need down coding can be very hard and difficult.

**Nuclear – Family Algorithm**

The nuclear-family algorithm was implemented in level 1 check of PEDCHECK program. The nuclear-family algorithm uses the known genotypes to check for inconsistencies between parents and offspring for each marker. An error is reported if one or more of the following condition is presented: the alleles of a child and a parent are inconsistent; the child is consistent with each parent separately but with both parents; more than four alleles are presented in a sibship; more than three alleles are presented in a sibship with a homozygous child; more then two alleles are presented in a sibship with two different homozygotes among the sibs; an allele is out of bounds; at an X-linked locus, a male is not coded as "homozygous" as required by the LINKAGE programs; or an individual has only one allele available in an autosomal system (O'Connel and Weeks, 1998).

If the errors have been corrected after level 1 check, then there still maybe errors due to the revised information, which will not be found out by this nuclear-Family algorithm. Moreover, the nuclear-family algorithm ignores loop information in the pedigree file system (O'Connel and Weeks, 1995). Thus, to identify errors that can not be detected by this algorithm; a genotype-elimination algorithm will be used, which would be implemented in level 2 check.

**Genotype – Elimination Algorithm**

Genotype Elimination is implemented via the extended version of Lange-Goradia algorithm (Lange and Goradia 1987; Lange and Weeks 1989) for set-recoded genotypes (O'Connel and Weeks 1995). The Lange-Goradia algorithm uses the nuclear-family relationships recursively to eliminated invalid genotypes in the pedigree; the recursion is continued until no more genotypes can be eliminated (O'Connel and Weeks 1998). The power of the genotypes-elimination algorithm is stronger than the nuclear-family algorithm because it can find inconsistencies resulting from the elimination of certain genotypes, on the basis of more complex pedigree relations. For each pedigree and locus, the genotype-elimination algorithm detect the first component nuclear family that is found with an error that has not been found already by the nuclear family algorithm, and it outputs the inferred genotypes lists for each member of that nuclear family.

If a complete genotype elimination algorithm finds no errors, then the genotypes are consistent with the inheritance laws of Mendelian, and linkage analysis is ready to be performed. Although genotype elimination algorithm will always identify the errors, and the diagnostic output of the inferred-genotypes lists does not always allow you to for easy identification of the source of the problem, because the genotype lists for untyped individuals may be long. Thus, two more error checking algorithms were used to locate the source of the subtle errors

**Critical-Genotype algorithm**

In a complicated case, one might want to invoke the critical-genotype algorithm, which attempts to identify the critical genotypes, this would output the list of the genotype, if any, in the pedigree, untyping one typed individual in the family, by scoring him/her as having an unknown genotype, would make solve the inconsistencies. It would be better to apply the genotype-elimination algorithm again after untyping, to determine if the inconsistency has been eliminated.

**Odds – Ratio algorithm**

If the critical genotype algorithm identifies several critical genotypes at a locus, then one must decide a priori which critical genotype is most likely to be erroneous. To distinguish

between alternative critical genotypes, an odds-ratio statistic based on single locus likelihoods of the pedigrees is implemented. First, for each individual with a critical genotype, one must identify the valid typings that eliminate the inconsistency. Second, for the particular locus, the likelihoods of the pedigree data for each alternative valid typing at each critical genotype need to be computed and stored, holding all other critical genotypes at their original value; that is , for each alternative genotype, one must compute the likelihood *L* (pedigree and that alternative genotype). Let *Lmax* be the largest likelihood obtained. Then, for each alternative genotype, the quantity *Lmax/L* should be formed, which gives an odds ratio against that alternative genotype versus any genotype with likelihood *Lmax*. Any genotype with a value of *Lmax* will have an odds ratio of 1, and, in general, the best-supported genotypes will have an odds ratio close to or equal to 1 (O'Connel and Weeks 1998).

In a routine genotyping, a large majority of genotyping errors will be identified and diagnosed by simple level 1 error checking, which is computationally extremely fast. After clearing the errors reported by level 1 checking. The genotype-elimination algorithm which was implemented in level 2 check, would produced the output of the error genotypes, when this output does not indicate the source of the error, level 3 and level 4 checking can be used to identify the individual most likely to be involved.

### 4.2.3 FastLink

Genetic linkage analysis is a statistical technique used to map genes and find the approximate location of disease genes. For this purpose there is a standard software package called LINKAGE. FASTLINK (Cottingham et al. 1993; Schäffer et al. 1994) is a significantly modified and improved version of the main programs of LINKAGE that runs much faster sequentially, can run in parallel, and allows the user to recover gracefully from a computer crash.

The LINKAGE package contains a series of programs for maximum likelihood estimation of recombination rates, calculation of lod score tables, and analysis of genetic

risks. The input files to the LINKAGE are pedigree and genotypic data, and model definition file which describe locus description, recombination rates, and gene order. The pedigree and genotypic data must be processed prior to analysis by a series of preparatory programs such as MAKEPED that accompany the analytic programs in the LINKAGE package. Estimation of recombination rates and calculation of the maximum lod score: ILINK for general pedigrees; CILINK for three-generation reference pedigrees; MLINK for lod score tables and risk analysis; LINKMAP for location score for general pedigrees; CMAP for three generation reference pedigrees.

Running the LINKAGE programs requires following steps:

1 Input pedigree and genotypic data

MAKEPED(program) transfer original pedigrees data to PEDFILE used for LINKAGE analysis.

Type command:

Makeped <input file original pedigree data file name> <PEDFILE name>

2 Description of loci

Use PREPLINK(program) to construct DATAFILE, to describe general information on loci and locus order, information on recombination, and program specific information.

Type commands “preplink”, and set the parameter as required for the analysis.

3 Analysis

The data file and pedigree file constructed in step 1 and step 2 serves as input (shown in Figure 4-5),

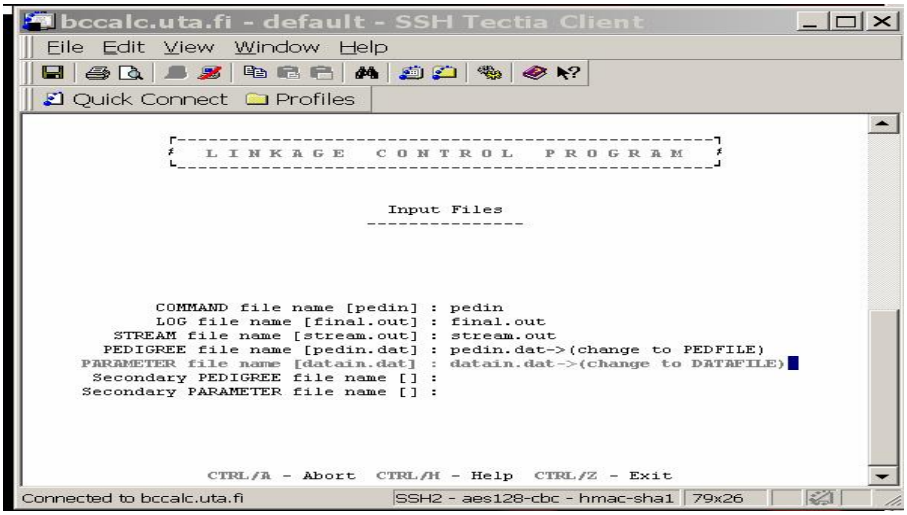


**Figure 4-5 input file for analysis**

use LCP (program) to set other parameter, choose which program to use(ILINK or MLINK, shown in Figure 4-6),

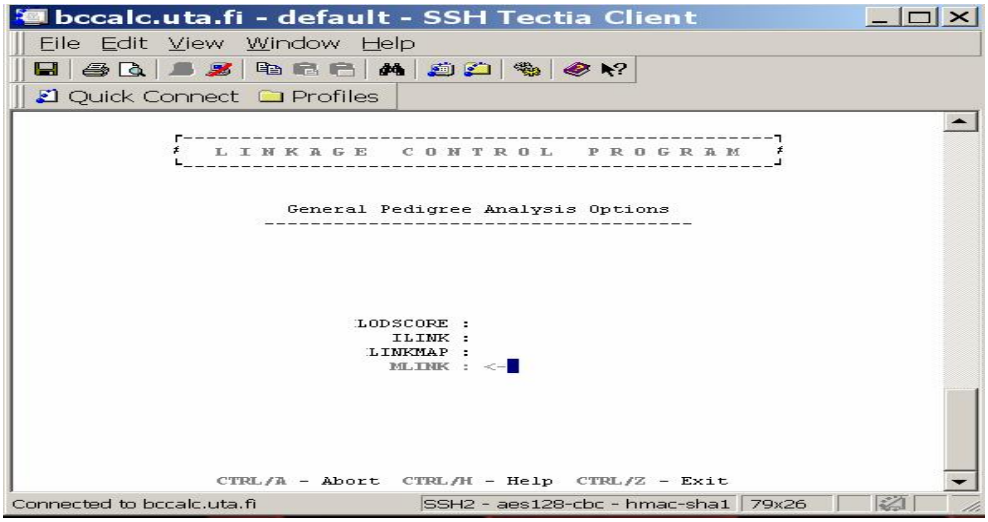


**Figure 4-6 choose analysis program**

And specify the locus order (shown in figure 4-7),

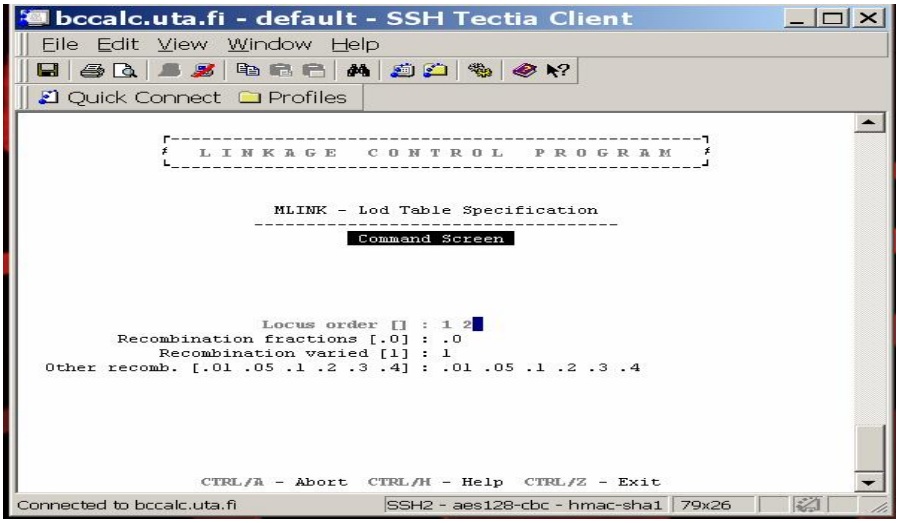


**Figure 4-7, specify the locus order**

the executable program PEDIN would be generated, simply type “pedin”, it would start analysis.

4 Result

Use “steam.out” file (which was generated automatically after the execution of “pedin”) as input for LRP (program), choose what kind of format you want the result to look like.

## 4.2.4 GENEHUNTER

Utilizing a second speedup of the HMM, the GENEHUNTER package expanded on the method of Kruglyak (Kruglyak et al. 1996) of extracting complete multipoint data from sibpairs (MAPMAKER/SIBS) by extending it to general pedigrees of modest size (twice the number of nonfounders, 2N – the number of founders, F ^20). Besides traditional lod score computation, GENEHUNTER included a new (model-free) nonparametric linkage (NPL) statistic, information content mapping and haplotype reconstruction. The NPL analysis is robust to uncertainty about the mode of inheritance, is more powerful than other general pedigree model-free methods and loses little power relative to traditional lod score analysis (i.e., when there are no errors in the description of the inheritance model).

Running GENEHUNTER program, GENEHUNTER is a command line oriented program, after starting GENEHUNTER, it requires following commands (interpretation of the command were included in Appendices)

```
npl:1> photo chr7.out
'photo' is on: file is 'chr7.out'

npl:2> ps on
Postscript output is now 'on'

npl:3> skip large off
Large pedigrees are now used but trimmed.

npl:4> map function kosambi
The Kosambi map function is now in use.

npl:5> haplotype off
Haplotype output is now 'off'

npl:6> off end 5
Scanning will now be done 5.0 cM beyond the ends of the map

npl:7> increment step 5
Scanning will be done in 5 steps per map interval

npl:8> count recs on
Count recs is now 'on'

npl:9> analysis BOTH
The current analysis type is 'BOTH'

npl:10> load marker chr7.dat

npl:11> scan chr7.pre
// analyze pedigree data

npl:12> total

.....................................................
…
…
file to store postscript plot [npl_plot.ps]: npl.ps
file to store postscript plot [lod_plot.ps]: lod.ps
file to store postscript plot [info_content.ps]: info.ps

npl:13> quit
```

# 5 Results

### *Two point LOD results*

The graphical presentation of the two-point LOD results is presented in Figure 1. For four chromosomes, 3, 13, 17 and X, the maximum LOD scores reached values over two (2.03 - 2.67), which can be considered as statistically significant evidence of linkage. For chromosome 2, 3, 6, 8, 12, 14, and 17, the maximum LOD scores reached values over 1 (1.07 - 1.54), which can be considered as suggestive evidence of linkage.

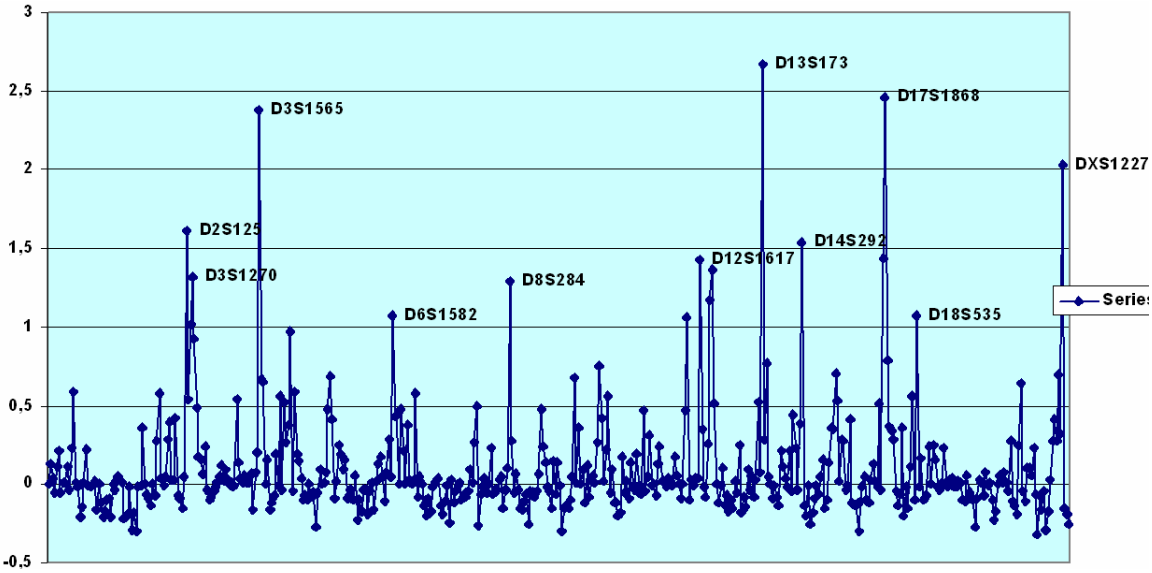


**Figure 1 Two Point FASTLINK results of 490 markers on 56 families are plotted in a chromosomal order starting from chromosome 1 on the left and ending with chromosome X on the right. For each marker the maximum two point LOD result is presented. On four chromosomes, logarithm of odds (LOD) scores over 2 were detected.**

### *Combined analysis results for markers which gave two point LOD scores over 2*

Followed by the two-point linkage analysis, an initial combined analysis of prostate cancer families was performed, which yielded the two-points LOD, multipoint parametric HLOD, single point and multipoint non-parametric NPL scores. In Table 5.1, the combined analyses results are shown for chromosomes 3, 13, 17, X.

| Table 5.1 Summary of genome wide linkage scan results from chromosome 3, 13, 17, and X. | | | | |
|---|---|---|---|---|
| | 3q26 | 13q33 | 17q21 | Xq27 |
| Marker | D3S1565 | D13S173 | D17S1868 | DXS1227 |
| Two-points LOD | 2.38 | 2.67 | 2.46 | 2.03 |
| θ | 0.1 | 0.2 | 0.1 | 0.2 |
| Multipoint HLOD | 0.64 | 0.92 | 3.40 | 1.87 |
| α | 0.21 | 0.28 | 0.5 | 0.33 |
| Location(cM) | 184.7 | 92.79 | 70.03 | 192.6 |
| Multipoint NPL | 0.98 | 1.84 | 2.80 | 2.49 |
| P value | 0.16 | 0.035 | 0.003 | 0.007 |
| Single Point NPL | 1.34 | 2.04 | 2.51 | 2.03 |
| P value | 0.09 | 0.022 | 0.007 | 0.022 |

The most significant results were obtained from chromosome 13 with marker D13S173 at 13q33, which gave the best two point LOD score of 2.67 at recombination fraction (θ) of 0.2, LOD score 2.6 at θ=0.1, LOD score 2.02 at θ=0.3. Flanking marker D13S1241 proximal to D13S173 gave peak two-point LOD score of 0.52 at θ =0.3, and marker D13S285 distal to D13S173 gave peak two-point LOD score of 0.77 at θ=0.2.The maximum multipoint HLOD score in this region was 0.90 at position 88.5 cM (location of D13S173). The multipoint NPL score of 1.81 (Pvalue=0.03) was given at this same location, and the single-point NPL score of 2.04 (Pvalue=0.022) was given at marker D13S173. The important families which contribute the most to the observed LOD scores are 2248, 2394, 2400, and 2427.

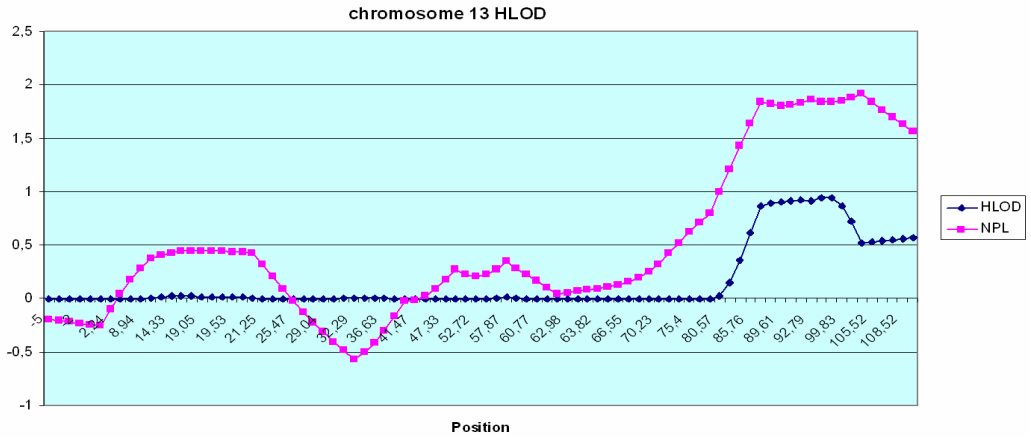


**Figure 2, Multipoint HLOD (blue) and NPL (red) scores of chromosome 13**

Same magnitude of linkage was also observed on chromosomes 3, 17, and 23. The second most significant parametric LOD score for 56 multiplex families was obtained from chromosome 17 with marker D17S1868 at 17q21, which gave the best two point LOD score was 2.46 (θ=0.1). Flanking marker D17S1872 proximal to D17S1868 gave peak two-point LOD score of 1.44 at θ=0.2. Marker D17S787 distal to D17S1868 gave peak two-point LOD score of 0.79 at θ=0.3. The maximum multipoint HLOD score in this region was 3.40 at position 70.0 cM. The multipoint NPL score of 2.80 (Pvalue=0.003) was given at this same location, and the single-point NPL score of 2.51 (Pvalue=0.007) was given at marker D17S1868. The families which contribute the most to the observed LOD scores are, 2062, 2069, 2079, 2372, and 2427.

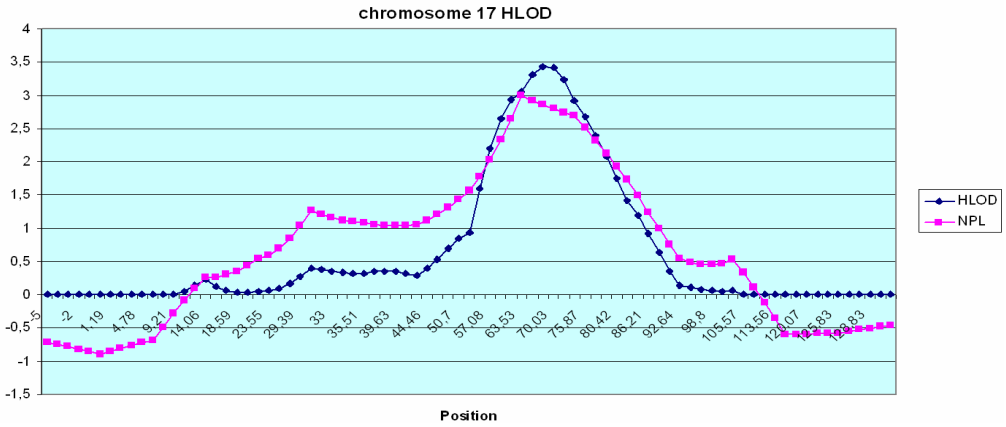


**Figure3, Multipoint HLOD (blue) and NPL (red) scores of chromosome 17**

The third significant LOD score come from chromosome 3 with marker D3S1565 at 3q26, which gave the best two point LOD score was 2.38 (θ=0.1). Flanking marker D3S1614 proximal to D3S1565 gave the peak two-point LOD score of 0.2 (θ=0.3). Marker D3S1262 distal to D3S1565 gave the peak two-point LOD of score 0.67 (θ=0.3). The maximum multipoint HLOD score in this region was 0.64 at position 184.7cM. The multipoint NPL score of 0.98 (Pvalue=0.16) was given at this same position, and the single-point NPL score of 1.34 (Pvalue=0.09) was given at marker D3S1565. The families which contribute the most to the LOD scores are, 2248, 2275, 2292, 2375, 2400, and 2427.

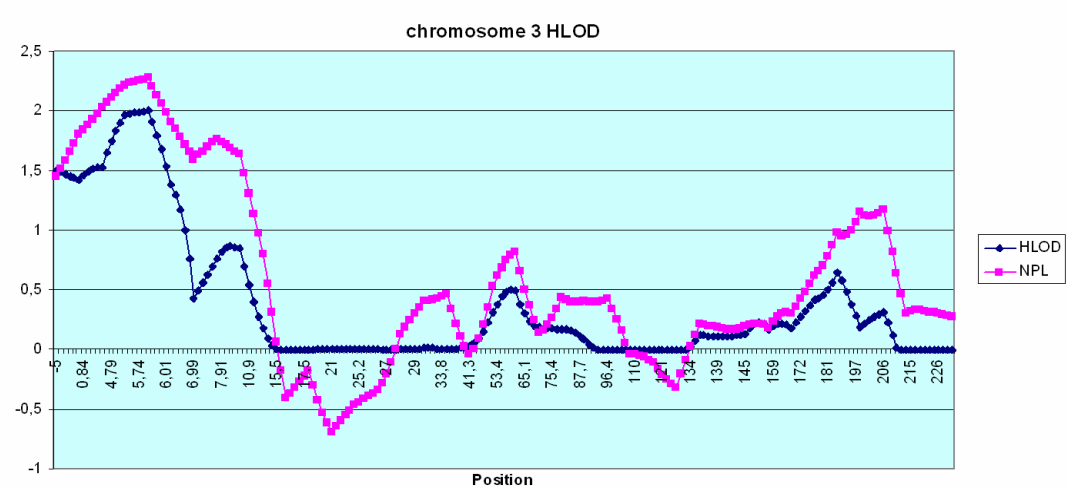


**Figure 4, Multipoint HLOD (blue) and NPL (red) scores of chromosome 3.**

The fourth and the last significant LOD score over 2 comes from chromosome X with marker DX1227 at Xq27, which gave the peak two point LOD score of 2.03 (θ=0.2). Flanking marker DXS1047 proximal to DXS1227 gave the peak two-point LOD score of 0.7 at θ=0.3. The maximum multipoint HLOD score in this region was 1.87 at 192.6 cM. The multipoint NPL score of 2.49 (Pvalue=0.007) was given at this same position, and the single-point NPL score of 2.05 (Pvalue=0.02) was given at marker DXS1227. The families which contribute the most to this LOD score are, 2069, 2102, 2232, 2236,2359,2400,2403.

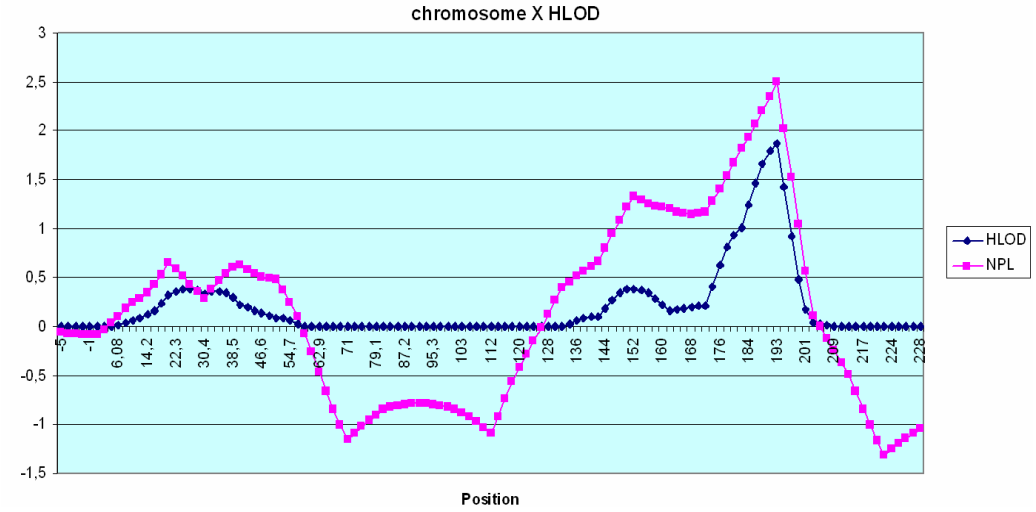


**Figure 5, Multipoint HLOD (blue) and NPL (red) scores of chromosome X**

***Combined analysis results which gave two point LOD score over 1 (part 1)***

The combined analyses results of markers from chromosome 3, 12, 14 and 17, which gave two point LOD score over 1, are shown in table 5.2.

**Table 5.2 Summary of genome wide linkage scan results from chromosome 3, 12, 14, and 17.**

| | 3p26 | 12q21 | 14q32 | 17q12 |
|---|---|---|---|---|
| Marker | D3S1270 | D12S351 | D14S292 | D17S1872 |
| Two-points LOD | 1.32 | 1.36 | 1.54 | 1.44 |
| θ | 0.2 | 0.2 | 0.1 | 0.2 |
| Multipoint HLOD | 2.00 | 2.07 | 1.82 | 1.74 |
| α | 0.35 | 0.37 | 0.43 | 0.37 |
| Location(cM) | 5.80 | 95 | 127.8 | 80.42 |
| Multipoint NPL | 2.27 | 2.15 | 2.26 | 1.92 |
| P value | 0.013 | 0.017 | 0.014 | 0.029 |
| Single Point NPL | 1.53 | 0.95 | 1.797 | 1.65 |
| P value | 0.06 | 0.17 | 0.038 | 0.05 |

In previous table 5.1, chromosome 3 with marker D3S1565 at 3q26 gave the peak two point LOD scores of 2.38; the maximum multipoint parametric HLOD and multipoint nonparametric NPL scores were all below 1. However, the peak two point LOD score of 1.32 (θ=0.2) was given with marker D3S1270 at 3p26, corresponding maximum multipoint parametric HLOD score of 2.00 was given at position 5.80 cM, the multipoint NPL score of 2.27 (Pvalue=0.013) was given at this same location, and the single point NPL score of this marker was 1.53 (Pvalue=0.06).

Chromosome 12 with marker D12S351 at 12q21 gave the peak two point LOD score of 1.36 (θ=0.2) where the two point LOD score is 1.36 (θ=0.2) with marker D12S1617. Flanking marker D12S326 proximal to D12S351 gave the peak two point LOD score of 1.17(θ=0.2), marker D12S346 distal to D12S351 gave the peak two point LOD score of 0.51 (θ=0.3).The maximum multipoint HLOD score in this region was 2.07 at position 95 cM. The multipoint NPL score of 2.15 (Pvalue=0.017) was given at this same location, and the single point NPL score of 1.53 (Pvalue=0.06) was given at the marker D12S351.

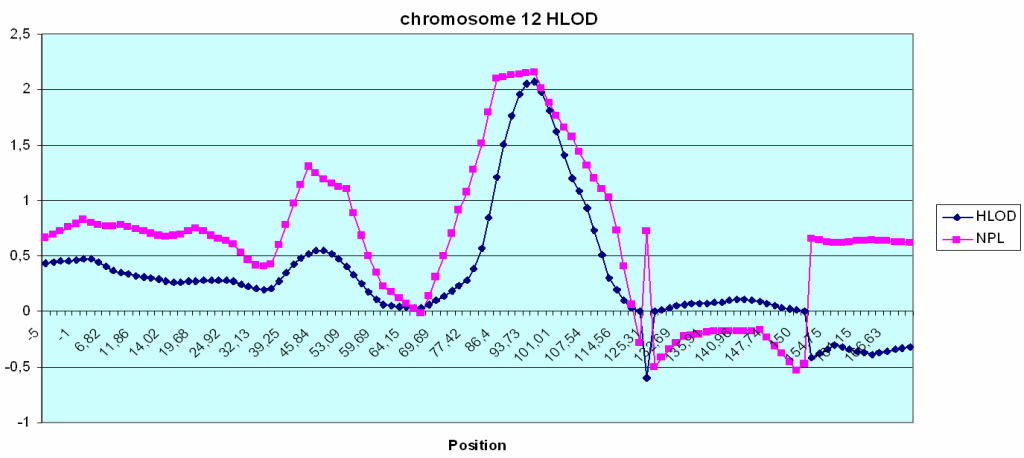


**Figure 6, Multipoint HLOD (blue) and NPL (red) scores of chromosome 12**

Chromosome 14 with marker D14S292 at 14q32 gave the peak two point LOD score of 1.54 (θ=0.1). Flanking marker D14S985 proximal to D14S292 gave the peak two point LOD score of 0.39 (θ=0.3). The maximum multipoint HLOD score in this region was 1.82 (Pvalue=0.83) at position 127.8 cM. The multipoint NPL score of 2.26 (Pvalue=0.014) was given at this same position, and the single point NPL score of 1.797 (Pvalue=0.038) was given at marker D14S292.

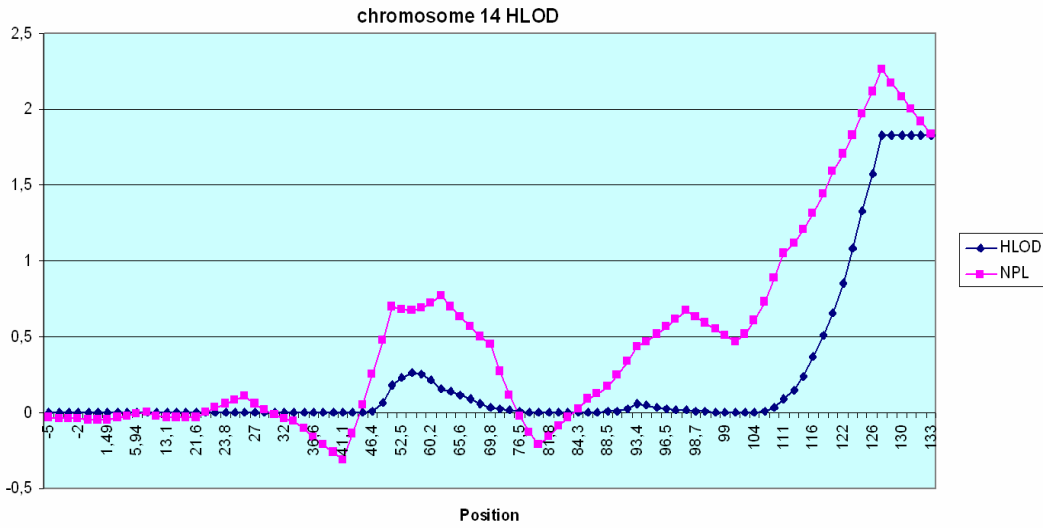


**Figure 7, Multipoint HLOD (blue) and NPL (red) score of chromosome 14**

***Combined analysis result which gave two point LOD score over 1 (part 2)***

The combined analysis results of markers from chromosome 2, 6, and 8, which gave two point LOD score over 1, are shown in table 5.3.

**Table 5.3 Summary of genome wide linkage scan result from chromosome 2, 6 and 8.**

| | 2q37 | 6p21 | 8q24 |
|---|---|---|---|
| Marker | D2S125 | D6S1582 | D8S284 |
| FastLink LOD | 1.61 | 1.07 | 1.29 |
| θ | 0.2 | 0.2 | 0.2 |
| Multipoint | 1.01 | 1.054 | 1.15 |
| HLOD | 0.35 | 0.25 | 0.27 |
| α | 253.03 | 56.14 | 142.9 |
| Location(cM) | | | |
| Multipoint NPL | 1.37 | 1.216 | 1.45 |
| P value | 0.08 | 0.11 | 0.07 |
| Single Point NPL | 1.86 | 1.33 | 1.07 |
| P value | 0.034 | 0.093 | 0.14 |

Chromosome 2 with marker D2S125 at 2q37 gave the peak two point LOD score of 1.61 (θ=0.2). Flanking marker D2S338 proximal to D2S125 gave the peak two point LOD score of 0.05 (θ=0.4). The maximum multipoint HLOD score in this region was 1.01 at position 253.03 cM, the multipoint NPL score of 1.37 (Pvalue=0.08) was given at this same location, and the single point NPL score of 1.86 (Pvalue=0.034) was given at marker D2S125.

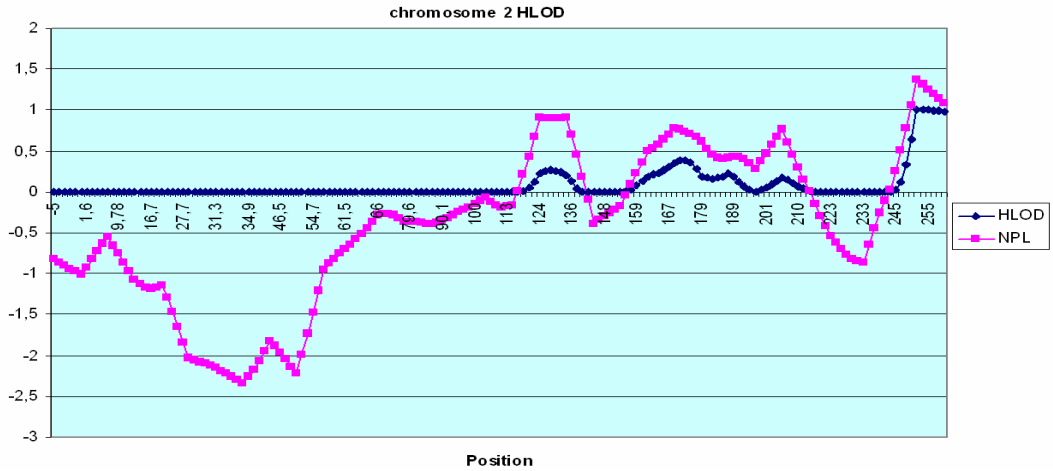


**Figure 8, Multipoint HLOD (blue) and NPL (red) scores of chromosome 2**

Chromosome 6 with marker D6S1582 at 6p21 gave the peak two point LOD score of 1.07 (θ=0.2). Flanking marker D6S1610 proximal to D6S1582 gave the peak two point LOD score of 0.05 (θ=0.4), marker D6S257 distal to D6S1582 gave the peak two point LOD score of 0.43 (θ=0.3). The maximum multipoint parametric HLOD score of in this region was 1.054 at position 56.14 cM, the multipoint NPL score of 1.216 (P=0.11) was given at this same location, and the single point NPL score of 1.33 (Pvalue=0.093) was given at marker D6S1582.

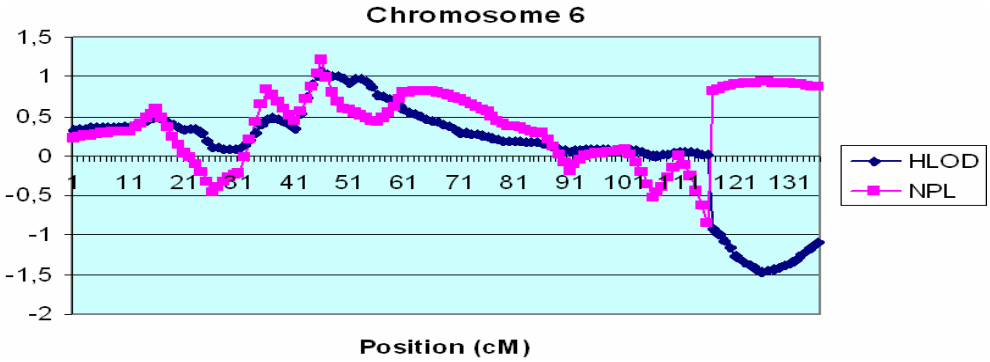


**Figure 9, Multipoint HLOD (blue) and NPL (red) scores of chromosome 6**

Chromosome 8 with marker D8S284 at 8q24 gave the peak two point LOD score of 1.29 (θ=.0.2). Flanking marker D8S1774 proximal to D8S284 gave the peak two point LOD score of 0.1 (θ=0.3), marker D8S272 distal to D8S284 gave the peak two point LOD score of 0.28 (θ=0.3). The maximum parametric HLOD score in this region was 1.15 at position 142.9 cM. The multipoint NPL score of 1.45 (Pvalue=0.07) was given at this same location, and the single point NPL score of 1.07 (Pvalue=0.04) was given at the marker D8S284.

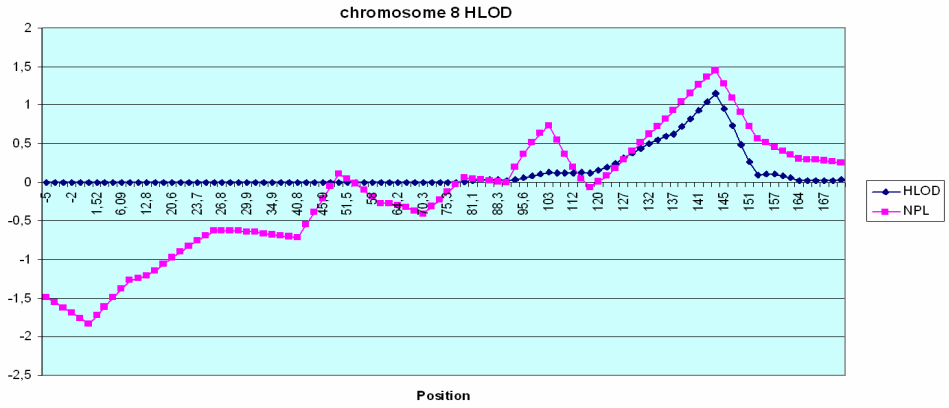


**Figure 10, Multipoint HLOD (blue) and NPL (red) scores of chromosome 8**

# 6 Discussion

Many complex human diseases like prostate cancer, which involve multiple genetic and environmental determinants, have increased in incidence during the past two decades. Considerable effort and expense have been expended on researches aimed at detection of genetic loci contributing to the susceptibility to complex human diseases. Linkage analysis is a valuable tool for identification of genetic loci predisposing to prostate cancer. The presence of genetic heterogeneity in population necessitates large-scale studies to provide significant statistical power to identify major loci, and genetic linkage studies are likely to be most successful in ethnically homogeneous populations, where the effects of genetic heterogeneity are minimized. Finland is one of the countries with a relatively homogeneous population, as demonstrated by frequent founder effects in both monogenic and complex diseases (Peltonen et al. 2000).

The present genome wide scan is carried out similarly as the previous genome wide scan (Schleutker et al, 2003). To test for Mendelian errors in the data set, PEDCHECK(O'Connel and Weeks, 1995) was used. Finnish marker allele frequencies are estimated from all genotyped individuals by SIBPAIR (Duffy 2006) program. Standard two point analyses are performed with FASTLINK(Cottingham et al. 1993; Schäffer et al. 1994). Multipoint parametric and non parametric analyses were performed by GENEHUNTER-PLUS (Kruglyak et al. 1996; Kruglyak and Lander 1998), and for chromosome X GENEHUNTER (versions 1.3) was used. For the parametric analyses, a dominant affected-only model is used; all unaffected man and all women were treated as unknown; and the frequency of the disease allele is set to 0.003.

The present genome wide scan, using allele sharing at 490 markers, spanning all 22 autosomes and X chromosome, among 56 families from Finland affected with HPC, was evaluated through linkage analyses. The combined analyses detected several possible prostate susceptibility loci from chromosome 2, 3, 6, 8, 12, 13, 14, 17, and X.

***Markers with significant evidence of linkage from chromosomes 3, 13, 17, X***

At chromosome 3, two possible loci were detected, one locus was marker D3S1565 at 3q26, which gave the peak two point LOD score of 2.38, corresponding multipoint HLOD and NPL are below 1, and the single point NPL score is 1.34. Another locus was marker D3S1270 at 3p26, which gave the peak two point LOD score of 1.32, corresponding multipoint HLOD and NPL score are above 2, and the single point NPL is 1. 53. In the previous genome wide scan, there was significant evidence of linkage at 3p25-26 (Schleutker et al. 2003), also a small peak was shown at 3q26. In 2005, fine mapping with 39 markers in 16 families, including muliplex families, were carried out at 3p25-26 and 11q14 region (Rökman et al. 2005), the results provide strong evidence for the existence of a prostate cancer susceptibility gene at 3p26 in Finland prostate cancer families, and two candidate genes *CHL1* and *CNTN6* were choose for mutation screening, no obvious deletion from these two genes were found (Rökman et al. 2005). The present genome wide scan, supports the previous finding of 3p26 region, also suggests that new locus at 3q26 might be a susceptibility locus site for further research. Most recently, it was reported DNA copy number gains at chromosome 3q25-q26 in 50% of prostate carcinomas using CGH (Sattler et al. 2000). Similar frequencies of 3q gains were reported by others research groups, although various sites on 3q are to be involved (Cher et al, 1996; Strohmeyer et al, 2004). The fine mapping of the 3q25-q26 amplification unit in prostate cancer and the identification of *TLOC1/SEC62* as the gene with the highest frequency of gene copy number gains compared with other positional candidates, over expression at the mRNA and protein levels indicates a biological role *of TLOC1/SEC62* in prostate cancer development (Jung et al.2006).

At chromosome 13, one locus with marker D13S173 was detected, the results provide strong evidence of linkage at 13q33. The maximum multipoint HLOD score in this region was 0.90 at position 88.5 cM (location of D13S173). The multipoint NPL score of 1.81 (Pvalue=0.03) was given at this same location, and the single-point NPL score of 2.04 (Pvalue=0.022) was given at marker D13S173. In the previous genome wide scan of 10 HPC families, there was no report of linkage to this region (Schleutker et al. 2003). Chromosome 13 is one of the most frequently altered chromosomes in cancer, including carcinoma of the prostate. Two known tumor suppressor genes, *RB1* and *BRCA2*, map to

chromosome 13 (Hyytinen et al, 1999a). Three distinct regions at q14, q21-22, and q33, showed the most frequent loss-of-heterotsygosity (LOH), suggesting their involvement in the development of prostate cancer (Hyytinen et al, 1999a). The *XPG/ERCC5* gene, a DNA repair gene that when mutated in the germline leads to xeroderma pigmentosum, has been also mapped to 13q33, LOH at *XPG* in prostate cancer supports that the 13q33 region contains a gene important in the development of prostate cancer, however not enough mutations were detected, *XPG* may be not the target gene involved (Hyytinen et al, 1999b).

At chromosome 17, two loci were detected. One locus with marker D17S1872 at 17q12 gave the peak two point LOD score of 1.44 ($\theta$=0.2), another locus with maker D17S1868 at 17q21 gave the peak two point LOD score 2.67 ($\theta$=0.2). The multipoint HLOD score in this region was 3.40 at 70 cM. The multipoint NPL score of 2.80 (Pvalue=0.03) was given at this same position. In this genome wide scan, the marker D17S1868 was defined in the MAP file at position 93.89 cM, and D17S1872 was defined at position of 83.14 cM, the multipoint analysis has the peak HLOD score at 70 cM, which was far before the region of D17S1868 (93.89 cM), and even before D17S1872 ( 83.14 cM). It should be around 17q12, which need further research to decide. In the genome-wide scan (GWS) based on 175 families from the University of Michigan Prostate Cancer Genetics Project (PCGP), the strongest result for prostate cancer linkage was on chromosome 17q21 LOD scores of 2.36 in all families and 3.28 in the subset of families with four or more affected men. At the sametime, in a subsequent combined genome-wide linkage analyses of 1,233 families (Xu et al, 2005) from 10 independent studies of hereditary prostate cancer (including the 175 families from our PCGP study), chromosome 17q21 was one of the top five regions with linkage evidence of prostate cancer. Furthermore, the strongest linkage signal for hereditary prostate cancer in the PCGP study is within 5 cM of the BRCA1 gene (on chromosome 17q21), suggesting the presence of a susceptibility locus close to the BRCA1 gene region. Germ-line, loss-of-function mutations in the BRCA1 gene on chromosome 17q21 substantially increase the lifetime risk of developing breast and ovarian cancer. BRCA1 mutations may also lead to an increased risk of other cancers; some studies have reported an increased risk of prostate cancer among male

carriers of deleterious BRCA1 mutations in breast and ovarian cancer families (Risch et al, 2001, Brose et al, 2002). Thompson and Easton (2002) reported an ~2-fold increased relative risk of prostate cancer in BRCA1 mutation carriers compared with noncarriers, although the effect was restricted to men who were young at their time of diagnosis (<65 years of age).

At chromosome X, marker with DXS1227 at Xq27 gave the peak two point LOD score of 2.03 (θ=0.2). . The maximum multipoint HLOD score in this region was 1.87 at 192.6 cM. The multipoint NPL score of 2.49 (Pvalue=0.007) was given at this same position, and the single-point NPL score of 2.05 (Pvalue=0.02) was given at marker DXS1227. This combined analyses results provide strong evidence of linkage to Xq27 region. HPCX locus at Xq27-28 was previously observed in a combined population of 360 prostate cancer families collected at four independent sites in North America, Finland and Sweden. The X chromosome locus appears to have a prominent role in prostate cancer predisposition in the Finnish study population, in which large fraction of families (over 40%) are estimated to be X-linked ( Xu et al. 1998). Later linkage analysis was carried out with 22 markers for the HPCX region of HPC families in Finland. Result indicate that the most of HPCX positively came from the families having (no male-to-male) NMM transmission and a late age of diagnosis (Baffoe et al, 2005). Confirmation of linkage to HPCX has been tried in several other studies (Lange et al. 1999; Goode et al. 2001, etc), and the study which support the linkage to HPCX the most, was based on 143 Utah extended high risk PRCA families; and a multipoint TLOD of 2.74 (P=0.0002) was obtained (Farnham et al. 2004). However, no trait-causing mutation or variants have been reported in this region.

***Markers with suggestive evidence of linkage from chromosome 2, 6, 8, 12, 14***

At chromosome 6, marker D6S1582 at 6p21 gave the peak two point LOD score of 1.07 (θ=0.2). The maximum multipoint parametric HLOD score of in this region was 1.054 at position 56.14 cM, and the multipoint NPL score of 1.216 (P=0.11) was given at this same location. The result is interesting because it is very close to the region identified in two previous aggressive prostate cancer studies (Shaid 2005; Lange et al. 2006). The

ICPCG analysis (Shaid 2005) reported a non parametric LOD=3.00 at a position of 42 cM, and a recessive HLOD=2.20 at 43 cM. In a study of 71 families with elevated risk of prostate cancer, University of Michigan researchers reported a nonparametric LOD of 2.09 at 30 cM, and a parametric HLOD=1.52 at that position in the recessive model (Lange et al. 2006). The international ACTANE Consortium also reported an HLOD of 1.41 under a rare dominant model near D6S2439 (42 cM) in a study of 64 families from 5 countries (Edwards et al. 2003)

At chromosome 8, marker D8S284 at 8q24 gave the peak two point LOD score of 1.29 ($\theta$=0.2). The maximum parametric HLOD score in this region was 1.15 at position 142.9 cM. The multipoint NPL score of 1.45 (Pvalue=0.07) was given at this same location, and the single point NPL score of 1.07 (Pvalue=0.04) was given at the marker D8S284. This site was also reported in the previous genome wide scan (Schleutker et al. 2003). Compare to the analyses results; the increasing number of HPC families has not brought any big changes to the analysis result. A recent genome wide linkage scan of 323 Icelandic high risk prostate cancer families produced a suggestive linkage singal on chromosome 8q24, with a LOD score of 2.11 at microsatellite marker D8S529 (located at 148.25 cM; Amundadottir et al. 2006). Analysis of several markers in the region identified a single allele, which was described as allele -8, of microsatellite marker DG8S737 that was associated with prostate cancer risk in case-control studies involving Caucasian men from Iceland, Sweden, and the Chicago area of the United State as well as African-American men from Michigan area of the United States, overall, the estimated odds ratio (OR) for carriers of the variant -8 allele of microsatellite DG8S737 was 1.62 (P=2.7 x 10-11), 13% of the controls carried at least one copy of the variant allele (Amundadottir et al. 2006).

At chromosome 12, marker D12S351 at 12q21 gave the peak two point LOD score of 1.36 ($\theta$=0.2).The maximum multipoint HLOD score in this region was 2.07 at position 95 cM. The multipoint NPL score of 2.15 (Pvalue=0.017) was given at this same location, and the single point NPL score of 1.53 (Pvalue=0.06) was given at the marker D12S351.

It was reported in the previous genome wide scan (Schleutker et al. 2003), that some evidence of linkage was observed on chromosome 12 at 12q22-23 region. The present analyses confirmed the evidence of linkage, and the evidence of linkage is even stronger in 12q21.33, this region is very close to 12q22. It has been reported, that a novel gene TU12B1-TY in the region at 12q22-q23.1 frequently was deleted in pancreatic cancer (Yatsuoka et al. 2004). However, there is no report of 12q21 being susceptibility locus for prostate cancer.

The present genome wide scan, also reveals areas which are not covered in the previous genome wide scan (Schleutker et al, 2003). At chromosome 2, marker D2S125 at 2q37 gave the peak two point LOD score of 1.61 ($\theta$=0.2). The maximum multipoint HLOD score in this region was 1.01 at position 253.03 cM, the multipoint NPL score of 1.37 (Pvalue=0.08) was given at this same location, and the single point NPL score of 1.86 (Pvalue=0.034) was given at marker D2S125. Chromosome 14 with marker D14S292 at 14q32 gave the peak two point LOD score of 1.54 ($\theta$=0.1). The maximum multipoint HLOD score in this region was 1.82 (Pvalue=0.83) at position 127.8 cM. The multipoint NPL score of 2.26 (Pvalue=0.014) was given at this same position, and the single point NPL score of 1.797 (Pvalue=0.038) was given at marker D14S292. Further research is needed for these new areas.

***Next step***

Prostate cancer represents a significant world wide public health burden. Epidemiological and genetic epidemiological studies have consistently provided data supporting the existence of inherited prostate cancer susceptibility genes. Segregation analyses of prostate cancer suggest that a multi-gene interaction mode may best explain familial clustering of this disease. Multiple major genes are likely in contribution to prostate cancer susceptibility, and there are advantages in modeling interactions in linkage analyses, current studies to identify these major genes contribute to primarily rely on single gene approaches. This is particularly true when exploring for novel regions of linkage using genome wide scans. Among a dozen genome wide scans for prostate cancer susceptibility genes published to data, all most all of these haven't modeled gene-gene

interactions. The interactions of genes that lead to a disease may be described as consistent with either a heterogeneity model in which alterations in any of several genes is sufficient or an epistemic model in which several simultaneous genetic alterations are required. Linkage analysis methods that model interaction may increase the statistical power to detect linkage when interaction among genes exists.

Thanks to the advent of genotyping platforms that allow genome wide searching for association between hundreds of thousands of random polymorphisms and disease phenotypes in large samples of unrelated individuals, Linkage disequilibrium (LD) studies are becoming increasingly popular. LD analysis has a wide range of applications. Population geneticists utilize LD analyses to assess the population structure and population history. Another focus of LD analysis is in mapping complex trait loci. The international HapMap project (www.hapmap.org) generates genome wide and densely spaced sequence variation data in several human populations from Asia, Africa and Europe. This type of data will certainly promote multi-locus LD measures in order to assess the variability of background correlation across genomic regions.

# 7 References

Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Le Roux L, Olafsdottir A, Blondal T, Andresdottir M, Gretarsdottir OS, Bergthorsson JT, Gudbjartsson D, Gylfason A, Thorleifsson G, Manolescu A, Kristjansson K, Geirsson G, Isaksson H, Douglas J, Johansson JE, B 0Š1lter K, Wiklund F, Montie JE, Yu X, Suarez BK, Ober C, Cooney KA, Gronberg H, Catalona WJ, Einarsson GV, Barkardottir RB, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K.A common variant associated with prostate cancer in European and African populations. Nat Genet. 2006. 38: 652-8.

Baffoe-Bonnie AB, Smith JR, Stephan DA, Schleutker J, Carpten JD, Kainu T, Gillanders EM, Matikainen M, Teslovich TM, Tammela T, Sood R, Balshem AM, Scarborough SD, Xu J, Isaacs WB, Trent JM, Kallioniemi OP, Bailey-Wilson JE.A major locus for hereditary prostate cancer in Finland: localization by linkage disequilibrium of a haplotype in the HPCX region. Hum Genet. 2005. 117:307-16.

Bello MJ, de Campos JM, Vaquero J, Kusak ME, Sarasa JL, and Rey JA. High resolution analysis of chromosome arm 1p alterations in meningioma Cancer Genet. Cytogenet. 2000. 120:30-36

Berthon P, Valeri A, Cohen-Akenine A, Drelon E, Paiss T, Wohr G, Latil A, Millasseau P, Mellah I, Cohen N,Blacnche H, Bellane Chantelot C, Demenais F, Teillac P, LeDuc A, de Petriconi R, Hautmann R, Chumakov I, Bachner L, Mailand NJ, Lidereau R, Vogel W, Fournier G, Mangin P, Cussenot O. Predisposing gene for early-onset prostate cancer, localized on chromosome 1q42-43. Am. J. Hum. Genet. 1998. 62:1416-124

Berry R, Schroder JJ, French AJ, McDonnell SK, Peterson BJ, Cunningham JM, Thibodeau SN and Schaid DJ. Evidence for a prostate cancer susceptibility locus on chromosome 20. Am. J. Hum, Genet. 2000.67:82-91

Bock CH, Cunningham JM, McDonnell SK, Schaid DJ, Peterson BJ, Pavlic RJ, Schroder JJ, Klein J, French AJ, Marks A, Thibodeau SN, Lange EM, and Cooney KA Analysis of the prostate cancer susceptibility locus HPC20 in 172 families affected by prostate cancer. Am. J. Hum. Genet. 2001. 68:795-801

Botstein D, White RL, Skolnick M,Davis RW. Construction of a genetic linkage map in man using restriction length polymorphisms. Am. J. Hum. 1980. 32:314-331

Brose MS, Rebbeck TR, Calzone KA, Stopfer JE, Nathanson KL, Weber BL. Cancer risk estimates for BRCA1 mutation carriers identified in a risk evaluation program. J Natl Cancer Inst. 2002. 94:1365-72.

Carter BS, Beaty TH, Speinberg GD, Childs B and Walsh PC. Mendelian inheritance of familial prostate cancer. Proc. Natl. Acad Sci, 1992. 89:3367-3371

Carter, B.S., Bova, G.S., Beaty, T.H., Steinberg, G.D., Childs, B., Isaacs, W.B and Walsh, P.C Hereditary prostate cancer: Epidemiologic and clinical features. J. Urol. 1993. 150:197-802.

Canning C, Thompson EA, Sjolnick MH. Probability functions on complex pedigrees. Adv. Appl. Prob. 1978. 10:26-91

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. Nature. 2004. 429:446-52

Cher ML, Bova GS, Moore DH, et al. Genetic alterations in untreated metastases and androgen-independent prostate cancer detected by comparative genomic hybridization and allelotyping. Cancer Res 1996. 56:3091 – 102.

Conlon EM, Goode EL, Gibbs M, Stanfold JL, Badzioch M, Janer M, Kolb S,Hood L, Ostrander EA, Jarvik GP, and Wijsman EM. Oligoneic segregation analysis of hereditary prostate cancer pedigrees: Evidence for multiple loci affection age at onset. Int J. Cancer . 2003. 105:630-635

Conneally PM, Rivas ML. Linkage analysis in man. Adv. Hum. Genet. 1980. 10:209-206

Cottingham Jr. RW, Idury RM, Schäaffer AA: Faster Sequential Genetic Linkage Computations. Am. J. Hum. Genet . 1993. 53:252-263

Cunningham JM, Shan A, Wick MJ, McDonnel SK, Schaid DJ, Tester DJ, Qian J, Takahashi S, Jenkins RB, Bostwick DG and Thibodeau SN. Allelic imbalance and microsatelite instability in prostatic adenocarcinoma. Cancer Res. 1996. 56: 4475-4482

Cunningham JM, McDonnel SK, Marks A, Hebbing S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Shaid DJ, Thibodeau SN and Mayo Clinic, Rochester, Minnesota. Genome linakge screen for prostate cancer susceptibility loci: Results from the mayo clinic familial prostate cancer study. Prostate. 2003b. 57:335-346

de la Chapelle. Genetic predisposition to colorectal cancer. Nat Rev Cancer. 2004. 4:769-780

de la Chapelle A. Disease gene mapping in isolated human populations: The example of Finland. J Med Genet 1993. 30::857-65

Dong C and Hemminki K. Modification of cancer risks in offspring by sibling and parental cancer from 2,112,616 nuclear families. Int.J. Cancer. 2000. 92:144-150

Dorman JS, Trucco M, LaPorte RE, Kuller LH. Family studies: The key to understanding the genetic and environmental etiology of chronic disease? Genet. Epidemiol. 1988.5:305-310

Duffy. 2006. SibPair program. http://www.qimr.edu.au/davidD/#sib-pair

Edwards S, Meitz J, Eles R, Evans C, Easton D, Hopper J, Giles G, Foulkes WD, Narod S, Simard J, Badzioch M, Mahle L and international ACTANE Consortium. Results of a genome wide linkage analysis in prostate cancer families ascertained through the ACTANE consortium. Prostate. 2003a. 57:270-279

Elston RC, Stewart J. A general model for the genetic analysis of pedigrees. Hum. Hered. 1971. 21:523-542

Elston RC. Segregation analysis. Adv. Hum. Genet. 1981. 11:63-120

Edwards S, Meitz J, Eles R, Evans C, Easton D, Hopper J, Giles G, Foulkes WD, Narod S, Simard J, Badzioch M, Mahle L. Results of a genome-wide linkage analysis in prostate cancer families ascertained through the ACTANE consortium. Prostate 2003. 57:270–279

Goldgar DE, Easton DF, Cannon-Albright LA, and Skolnick MH. Systematic population-based assessment of cancer risk in first degree relatives of cancer probands. J.Natl. Cancer Inst. 1994. 86:1600-168

Farnham JM, Camp NJ, Swensen J, Tavtigian SV, Albright LA. Confirmation of the HPCX prostate cancer predisposition locus in large Utah prostate cancer pedigrees. Hum. Genet. 2005. 116:179-85

Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE. Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Lancet 1994. 343:692 – 5.

Gann, PH and Giovannucci. Prostate cancer and Nutrition.
(www.prostatecancerfoundation.org/atf/cf/%7B705B3273-F2EF-4EF6-A653-E15C5D8BB6B1%7D/Nutrition_Guide.pdf)

Goode El, Standford JL, Peters MA, Janer M, Gibbs M, Kolb S, Badzioch MD, Hood L, Ostander EA, Jarvik GP. Clinical characteristics of prostate cancer in an analysis of linkage of four putative susceptibility loci. Clin. Cancer. Res. 2001. 7:2739-49

Gronberg H, Damber L, Damber JE, and Iselius L. Segregation analysis of prostate cancer in Sweden: Supprot for dominant inheritance. Am.J.Epidemiol. 1997a. 146:552-557

Gronberg H, Xu J, Smith JR, Carpten JD, Isaacs SD, Freije D, Bova GS, Danber JE, Bergh A, Walsh PC, Collins FS, Trent JM, Meyers DA and Isaacs WB. Early age at diagnosis in families providing evidence of linkage to the hereditary prosate cancer locus (HPC1) on chromosome1. Cancer res. 1997b. 57: 4707-4709

Haldane JBS, Smith CAB. A new estimate of the linkage between the genes for colour-blindness and haemorphilia in man.. Annals of Eugenics. 1947.14:10-31

Hasstedt SJ. A mixed model likelihood approximation on large pedigrees. Computers and Biomedical Research. 1982. 15:295-307

Isaacs SD, Kiemeneu LA, Baffoe-Bonnie A, Beaty TH, and Walsh PC. Risk of cancer in relatives of prostate cancers in relatives of prostate cancer probands. J. Natl. Cancer Inst. 1995. 87: 991-996

Janer M, Friedrichsen DM, stanford JL, Badzioch MD, Kolb S, Deutsch K, Peters MA; Goode EL, Welti R, DeFrance HB, Iwasaki L, Li S, Hood L, Ostrander EA, and jarvik GP. Genomic scan of 254 hereditary prostate cancer families. Prostate. 2003. 57:309-319

Jorde LB. Linkage Disequilibrium and the Search for Complex Disease Genes. Genome Res. 2000. 10: 1435-1444

Jung V, Kindich R, Kamradt J, Jung M, Muller M, Schulz WA, Engers R, Unteregger G, Stokle M, Zimmermann R, Wullich B. Genomic and expression analysis of the 3q25-q26 amplification unit reveals TLOC1/SEC62 as a probable target gene in prostate cancer. Mol. Cancer. Res. 2006. 4:169-76

Keetch, D.W., Rice, J.P., Suarez, B.K and Catalona, W.J. Familial aspects of prostate cancer: A case control study. J.Urol.1995. 154:2100-2102

King MC, Lee GM, Spinner NB, Thomson G, Wrensch MR. American Review of Public Health. Genet. epidemiol. 1984. 5:-52

Kruglyak L, Daly MJ, Lander ES. Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am. J. Hum. Genet. 1995. 56:519-27

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. 1996.58:1347-1363.

Kruglyak L, Lander ES. Faster multipoint linkage analysis using Fourier transforms. J. Comput. Bio. 1998.5:1-7.

Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994. 265:2037-48

Lander ES, Botstein D. Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. Sciences. 1987. 236:1567-1570

Lange EM, Gillanders EM, Davis CC, Brown WM, Campbell JK, Jones M, Gildea D, Riedesel E, Albertus J, Freas-lutz D, Markey C, Giri V; Dimmer JB, Montie JE, Trent JM and Cooney KA. Genome wide scan for prostate susceptibility genes using families from the University of Michigan Prostate Cancer Genetics Project finds evidence for linakge on chromosome 17 near BRCA1. Prostate. 2003. 57:326-334

Lange K, Elston RC. Extensions to pedigrees analysis I. Likelihood calculations for simple and complex pedigrees. Hum. Hered. 1975. 25:95-105

Lange K, Weeks DE. Efficient computation of lod scores: genotype elimination, genotype redefinition, and hydrid maximum likelihood algorithms. 1989. Ann. Hum. Genet. 53:67-83

Lange K, Goradia TM. An algorithm for automatic genotype elimination. Am. J. Hum. Genet. 1987 40:250-6

Lathrop GM, Lalouel JM, Julier C, Ott J. Multilocus linkage analysis in humans: Detection of linkage and estimation of recombination. Am. J. Hum. Genet. 1985. 37:482-498

Lathrop GM, Lalouel JM, White NW. Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis. Genetic Epidemiology. 1986. 3:39-52

Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Karprio J,Koskenvuo M, Pukkala E, Skytthe A and Hemminki K. Environmental and heritable factors in the causation of cancer--analyzes of cohorts of twins from Sweden, Denmark, and Finland. N.Engl. J.Med. 2000. 343:78-85

Lilienfeld AM. A methodological problem in testing a recessive genetic hypothesis in human disease. Am. J. Public Health . 1959. 49:199-204

MacMahon B, Pugh TF. Epidemiology: Principles and methods.1970. Boston, Little, Brown.

Majumder PP, Chakraborty R, Weiss KM. Relative risk of diseases in the presence of incomplete penetrance and sporadics. Stat. Med. 1983. 2:13-24

Maier C, Vesovic Z, Bachman N, Herkommer K, Braun AK, Surowy HM, Assum G, Paiss T and Vogel W. Germline mutations of the MSR1 gene in prostate cancer families from Germany. Hum. Mutat. 2006. 27:98-102

Mausner JS, Kramer S. Epidemiology: An introductory Text, 2nd Ed. 1985. Philadelphia, PA, WB Saunders.

McGuffin P, Huckle P. Simulation of Mendelism revisited: The recessive gene for attending medical school. 1980. Am. J. Hum. Genet. 1990. 46:994-999

Monroe KR, u MC, Kolonel LN, Coetzee GA, WIlkens LR, Ross RK, and Henderson BE. Evidence of an X-linked or recessive genetic component to prostate cancer risk. Nat. Med. 1995. 1:827-829

Morton NE. Sequential tests for the detection of linkage. Am. J. Hum. Genet. 1955. 7:277-318.

Morton NE,Chung cs. Genetic Epidemiology. New York: Academic Press, 1978. pp 3-11

Morton NE. Linkage and association. In Rao DC, Elston RC, Kuller LH, Feinleib M, Carter C, Havlik R (eds), Genetic Epidemiology of Coronary Heart Disease: Past, Present, and Future. 1984. New York, Alan R Liss, PP 245-265

Morton NE. Linkage disequilibrium maps and association mapping.. JCI. 2005. 115:1425-1431

Nupponen NN and Carpten JD. Prostate cancer susceptibility genes: Many studies,many results, no answers. Cancer Metastatics Rev. 2001. 20:155-164

O'Connell JR, Weeks DE. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. Am. J. Hum. Genet. 1998. 63:259-66

Page, W.F.,Braun, M.M.,Partin, A.W., Caporaso, N. and Walsh, P. Hereditary and prostate cancer: A study of World War II veteran twins. Prostate.1997. 33:240-245

Pakkanen S, Baffoe-Bonnie A, Matikainen MP, Koivisto PA, Tammela TLJ, Deshmukh SOuL,Baily-Wilson J and Schleutker J. Segregation analysis of 1546 prostate cancer families in Finland show recessive inheritance. Hum. Genet. 2007. 121:257-67

Parkin, D.M., Whelan, S.L., Ferlay, J., Raymond, L., and Young, J. Cancer Incidence in Five Continents, 1997 Vol. VII (IARC Scientific Publications No. 143) Lyon, IARC

Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet. 2000. 1:182-90.

Ponder, B.A. Cancer genetics. Nature. 2001. 411:336-341

Rao DC, Keats BJB, Llouel JM, Morton NE, Yee S. A maximum likelihood map of chromosome 1. Am. J. Hum. Genet. 1979. 30:516-529

Renwick JH. The mapping of human chromosomes. Annual Review of Genetics. 1971.5:81-120

Risch N Genetic linkage and complex diseasea, with special reference to psychiatric disorders. Genet epidemiol. 1990a. 7:3-16.

Risch N. The genetic epidemiology of cancer: Interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol. Biomarkers Prev. 2001. 10: 733-741

Risch HA, McLaughlin JR, Cole DE, et al. Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. Am J Hum Genet. 2001. 68:700-10.

Roberts DF. A definition of genetic epidemiology. In Chakraborty R, Szathmary EJE (eds), Diseases of Comples Etiology in Small populations: Ethnic Differences and Research Approaches. 1985. New York, Alan R Liss, pp9-20.

Rökman A, Ikonen T, Mononen N, Autio V, Matikainen MP, Koivisto PA. ELAC2/HPC2 involvement in hereditary and sporadic prostate cancer. Cancer Res . 2001. 61:6038-41;

Rökman A, Ikonen T, Seppälä EH, Nupponen N, Autio V, Mononen N, et al. Germline alterations of the RNASEL gene, a candidate HPC1 gene at 1q25, in patients and families with prostate cancer. Am J Hum Genet 2002. 70:1299-304

Rökman A, Baffoe-Bonnie AB, Gillanders E, Fredriksson, H, Autio V, Ikonen T,Gibbs KD, jr Jones M,Gildea D, Freas-lutz D, Markey C, matikainen MP,Koivisto pa, Tammela TL, Kallioniemi OP, Trent J, Bailey-wilson JE and Schleutker J. Hereditary prostate cancer in Finland: Fine mapping validates 3p26 as a maor predisposition locus. Hum. Genet. 2005. 116:43-50

Sattler HP, Lensch R, Rohde V, et al. Novel amplification unit at chromosome 3q25-27 in human prostate cancer. Prostate 2000. 45:207 – 15.

Schaid DJ, McDonnel SK, Blute ML and Thibodeau SN. Evidence for autosomal dominant inheritance of prostate cancer. Am. J. Hum. Genet. 1998. 62:1425-1438

Schaid DJ. The complex gentics epidemiology of prostate cancer. Hum. Mol. Genet. 2004. 13:103-21

Schaid DJ, Chang BL, and International Consortium For Prostate Cancer genetics. Descritption of the international consortium for prostate cancer genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. Prostate.2005. 63:276-290

Schäffer AA, Gupta SK, Shriram K, Cottingham Jr. RW: Avoiding Recomputation in Linkage Analysis. Hum. Heredity . 1994. 44:225-237.

Schleutker J, Matikainen M, Smith J, Kovisto P, Baffoe-Bonnie A, Kainu T, Gillanders E, Sandila R, Pukkala E, Carpten J, Stephan D, Tammela T, Brownstein M, Baily-wilson J, trent J, and Kallioniemi OP .A genetic epidemiological study of hereditary prsotate cancer (HPC) in Finland: Frequent HPCX linkage in families with late onset disease. Clin, Cancer Res. 2000. 6: 4810-485

Schleutker J, Baffoe-Bonnie AB, Gillanders E, Kainu T, Jones MP, Freas-lutz D, Markey C, Gilda D, Riedesel E, Albertus J, Gibbs KD . Jr Matikainen M, Koivisto PA, Tammela T, Bailey Wilson JE, Trent JM and Kallioniemi OP. Genome wide scan for linkage in Finnish Hereditary prostate cancer (HPC) families identifies noval susceptibility loci at 11q14 and 3p25-26. Prostate. 2003. 57:280-289

Seppälä EH, Ikonen T, Autio V, Rökman A, Mononen N, Matikainen MP. Germ-line alterations in MSR1 gene and prostate cancer risk. Clin Cancer Res 2003. 9:5252-6

Smith CAB. The detection of linkage in human genetics. Journal of the Royal Statistical Society, 1953. Series B 14:153-192.

Smith CAB . The development of human linkage analysis. Annals of Hum. Genet. 1986. 50:293-311

Smith JR, Freije D, Carpten JD, Gronberg H,Xu J,Isaacs SD, Brownstein MJ,Bova GS, Guo H,Bujnovszky P, Nusskern DR, Damber JE,Bergh A, Emanuelsson M, Kallioniemi OP, Walker Daniels J, Baily-Wilson JE, Beaty TH, Meyers DA,Walsh PC, Collins FS, Trent JM and Isaacs WB. Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome wide search. Science 1996. 274:1371-134

Steinberg, GD; Carter BS; Beaty TH; Childs B; Walsh PC. Family history and the risk of prostate cancer. Prostate 1990. 17:337-47.

Strohmeyer DM, Berger AP, Moore DH, et al. Genetic aberrations in prostate carcinoma detected by comparative genomic hybridization and microasatellite analysis: association with progression and angiogenesis. Prostate 2004. 59:43 – 58.

Sturt E. Amappin function for human chromosomes. Ann Hum. Genet. 1976. 40: 147-63

Susser M . Separating heredity and environment. Am. J. of Preventive Medicine 1985. 1:5-23

Susser M, Susser E. Separating heredity and environment. I. Genetic and environment indices. In Susser M (eds). Epidemiology, Health and Society: Selected Papers. 1987a. New York, Oxford University Press, pp 103-114

Susser M, Susser E. Seprarating heredity and environment. II. Research designs and strategies. In Susser M (eds). Epidemiology, Health and Society: Selected Papers. 1987b. New York, Oxford University Press, pp 114-128

Takayama H, Suzuki T, Mugishima H, Fujisawa T, Ookuni M, Schwab M, Gehring M, Nakamura Y, Sugimura T and Terada M. Deletion mapping of chromosome s 14 q and 1p in human Neuroblastoma . Oncogene. 1992. 7:1185-1189

Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffarri S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc, G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A,

Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommmens J and Cannon Albright LA. A candidate prostate cancer susceptibility gene at chrmosome 17p. Nat Genet. 2001. 27:172-80

Thompson D, Easton DF. Cancer incidence in BRCA1 mutation carriers. J Natl Cancer Inst. 2002. 94:1358-65.

Weiss KM, Chakraborty R, Majumder PP, Smouse P. Probelms in the assessment of relative tisks of chronic diseases among biological relatives of affected individuals. Journel of Chronic Diseases. 1982. 35:539-551

Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends. Genet. 2002. 18:19-24

Whittemore AS, Halpem J. A class of test for linkage using affected pedigree members. Biometrics. 1994a. 50:118-27

Whittemore AS, Halpem J. Probability of gene identify by desent computation and applications. Biometrics. 1994b. 50:109-17

Wiklund F, Gillanders EM, Albertus JA, Bergh A, Damber JE, Emanuelsson M, FreasLutz DL, GIldea DE, Goransson I, Jones MS, Josson BA, Lindmark F, Markey CJ, Riedesel EL, Stenman E, Trent JM and Gronberg H. Genome wide scan of swedish families with hereditary prostate cancer: suggestive evidence of linkage at 5q11.2 and 19q13.3. Prostate. 2003. 57:290-297

Witte JS, Suarez BK, Thiel B, Lin J, Yu A, Banerjee TK, Burmester JK, Casey G, and Catalona WJ. Genome wide scan of brothers: Replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. Am. J.Hum. Genet. 2003. 67:92-97

Xu J, Combined analysis of hereditary prostate cancer linkage to 1q24-25; Results from 772 hereditary prostate cancer families from the international Consortium for prostate cancer genetics, Am. J.Hum. Genet. 2000. 66: 945-57

Xu J, Meyers DA, Freijie D, Isaacs S, Wiley K, Nusskern D, Ewing C, Wilkens E, Bujnovszky P, Bova GS, Walsh P, Isaacs W, Schleutker J, Matikainene M, Tammela T, Visakorpi T, Kallioniem OP, Berry R, Shaid D, French A, McDonnell S, Schroder J, Blute M, Thibodeau S, Trent J. Evidence for a prostate cancer susceptibility locus on the X chromosome. Nat genet. 1998 20: 175-19

Xu J, Zheng Sl, Hawkins GA, Faith DA, Kelly B, Isaacs SC, Wiley KE, Change B, EWing CM, Bujnovszky P, Carpten JD, Bleecker ER, Walsh PC, Trent JM,Mayeres DA,and Isaacs WB. Linakge and association studies of prostate cancer susceptibility: Evidence for linkage at 8q22-23. Am. J. Hum. Genet. 2001. 69:341-50

Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiely KE; Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, Freas-lutz D,Markey C, Meyers DA, Walsh PC, Trent JM and Isaacs WB. Genome wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. Prosate. 2003. 57:320-325

Xu J, Dimitrov L, Chang BL, et al. A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. Am J Hum Genet. 2005. 77:219 – 29.

# 8 Appendices

Appendix 1: Exp1.pre (pedigree file contain family and genotype information)

(each column (left to right), stands for Family ID, Individual ID, Father ID, Mother ID, sex (1=male, 2=female), Affection_Status (0=unknown, 1=healthy, 2=Affected), Liability_class, and Genotypes (two column for each marker, stands for allele1 and allele2))

| | | | | | | | gen01 | | gen02 | | gen03 | | gen04 | | gen05 | | gen06 | | gen07 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2427 | 55 | 7 | 52 | 2 | 0 | 3 | 5 | 9 | 8 | 9 | 6 | 6 | 8 | 8 | 1 | 3 | 3 | 4 | 4 | 10 |
| 2427 | 54 | 7 | 52 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 53 | 7 | 52 | 1 | 0 | 3 | 5 | 10 | 6 | 9 | 6 | 10 | 8 | 13 | 3 | 9 | 3 | 4 | 4 | 4 |
| 2427 | 52 | 0 | 0 | 2 | 0 | 3 | 9 | 10 | 6 | 8 | 6 | 10 | 8 | 13 | 1 | 9 | 3 | 3 | 4 | 10 |
| 2427 | 49 | 6 | 48 | 1 | 0 | 3 | 5 | 5 | 7 | 9 | 9 | 9 | 7 | 12 | 24 | 7 | 3 | 4 | 4 | 10 |
| 2427 | 48 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 44 | 5 | 42 | 2 | 0 | 3 | 9 | 9 | 9 | 11 | 9 | 10 | 7 | 8 | 3 | 7 | 3 | 4 | 4 | 10 |
| 2427 | 43 | 5 | 42 | 1 | 0 | 3 | 5 | 9 | 8 | 9 | 6 | 7 | 4 | 8 | 3 | 4 | 4 | 7 | 7 | 10 |
| 2427 | 42 | 0 | 0 | 2 | 0 | 3 | 5 | 9 | 8 | 9 | 7 | 9 | 4 | 7 | 4 | 7 | 3 | 7 | 10 | 10 |
| 2427 | 34 | 3 | 32 | 1 | 0 | 3 | 5 | 9 | 5 | 6 | 9 | 9 | 4 | 7 | 24 | 7 | 3 | 3 | 4 | 4 |
| 2427 | 33 | 3 | 32 | 2 | 0 | 3 | 5 | 9 | 5 | 6 | 9 | 9 | 4 | 7 | 24 | 7 | 3 | 3 | 4 | 4 |
| 2427 | 32 | 0 | 0 | 2 | 0 | 3 | 5 | 9 | 6 | 9 | 9 | 9 | 4 | 7 | 24 | 7 | 3 | 3 | 4 | 5 |
| 2427 | 20 | 2 | 17 | 2 | 0 | 3 | 9 | 10 | 5 | 10 | 9 | 9 | 7 | 12 | 3 | 8 | 3 | 6 | 4 | 7 |
| 2427 | 19 | 2 | 17 | 2 | 0 | 3 | 9 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 8 | 2 | 2 | 7 | 7 |
| 2427 | 18 | 2 | 17 | 2 | 0 | 3 | 9 | 10 | 10 | 11 | 9 | 10 | 7 | 12 | 24 | 3 | 2 | 4 | 4 | 7 |
| 2427 | 17 | 0 | 0 | 2 | 0 | 3 | 5 | 10 | 10 | 11 | 9 | 9 | 7 | 12 | 3 | 8 | 2 | 6 | 4 | 7 |
| 2427 | 16 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 15 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 14 | 15 | 16 | 2 | 0 | 3 | 9 | 9 | 5 | 11 | 8 | 9 | 8 | 8 | 3 | 3 | 3 | 4 | 4 | 4 |
| 2427 | 13 | 15 | 16 | 2 | 0 | 3 | 5 | 5 | 5 | 9 | 8 | 9 | 7 | 8 | 24 | 3 | 3 | 3 | 4 | 7 |
| 2427 | 12 | 15 | 16 | 2 | 0 | 3 | 9 | 9 | 5 | 11 | 9 | 10 | 8 | 12 | 3 | 3 | 3 | 4 | 4 | 4 |
| 2427 | 11 | 15 | 16 | 2 | 0 | 3 | 9 | 9 | 9 | 11 | 6 | 10 | 8 | 12 | 3 | 3 | 3 | 4 | 4 | 4 |
| 2427 | 10 | 15 | 16 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 9 | 15 | 16 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 8 | 15 | 16 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 7 | 15 | 16 | 1 | 2 | 1 | 5 | 5 | 9 | 9 | 6 | 8 | 8 | 12 | 3 | 3 | 3 | 4 | 4 | 7 |
| 2427 | 6 | 15 | 16 | 1 | 0 | 3 | 5 | 5 | 9 | 9 | 8 | 9 | 7 | 8 | 24 | 3 | 3 | 3 | 4 | 7 |
| 2427 | 5 | 15 | 16 | 1 | 2 | 1 | 5 | 9 | 9 | 11 | 6 | 10 | 8 | 12 | 3 | 3 | 3 | 4 | 4 | 7 |
| 2427 | 4 | 15 | 16 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2427 | 3 | 15 | 16 | 1 | 2 | 1 | 5 | 5 | 5 | 9 | 9 | 10 | 7 | 12 | 24 | 3 | 3 | 3 | 4 | 7 |
| 2427 | 2 | 15 | 16 | 1 | 2 | 1 | 9 | 9 | 5 | 11 | 9 | 10 | 7 | 12 | 24 | 3 | 3 | 4 | 4 | 7 |
| 2427 | 1 | 15 | 16 | 1 | 2 | 1 | 5 | 5 | 9 | 9 | 6 | 8 | 8 | 8 | 3 | 3 | 3 | 4 | 4 | 7 |

....

Appendix 2: Exp1.map (contains distance information between markers)

| Chromosome | Location | DistanceFromAbove | MarkerName | Heterozygosity |
|---|---|---|---|---|
| 21 | 1.000 | 0.00 | D21S1256 | 0.65 |
| 21 | 2.000 | 0.00 | D21S1914 | 0.86 |
| 21 | 2.100 | 9.09 | D21S1909 | 0.85 |
| 21 | 3.000 | 9.97 | D21S1252 | 0.80 |
| 21 | 3.100 | 5.04 | D21S2055 | 0.88 |
| 21 | 4.000 | 5.38 | D21S266 | 0.59 |
| 21 | 4.100 | 11.9 | D21S1446 | 0.79 |

Appendix 3: PreToSib.pl (perl script used to combine pedigree file and map file to generate input file for SIBPAIR program)

```
##change *.pre file into *.in for sib analysis
##how to use
##type following in the command line in the directory where the file
##were presented
##    perl PreToSib.pl exp1.map exp1.pre exp1.in

open MAP,"<".$ARGV[0] or die "can't open $ARGV[0]";
open PRE,"<".$ARGV[1] or die "cant open $ARGV[1]";
open SIB,">".$ARGV[2] or die "cant open $ARGV[2]";

print SIB "set error_drop on\n";
print SIB "set loc a affection\n";
my $location;
while(<MAP>){
      if($_=~/^\d+/){
            my @array=split /\s+/, $_;
            $location=$location+$array[2];
            print SIB "set loc $array[3] marker ",$location," cM\n"
      }
}
print SIB "read ped inline\n";
while(<PRE>){
      my @array=split /\s+/, $_;
      for my $index(0..3){
            print SIB "$array[$index]\t";
            }
      if($array[4]==2){
            print SIB "F\t";
      }elsif($array[4]==1){
            print SIB "M\t";
      }
      if( $array[5]==0){
            print SIB "X\t";
      }elsif($array[5]==2){
            print SIB "Y\t";
      }elsif($array[5]==1){
            print SIB "N\t";
      }
      for my $i(7..$#array){
            print SIB "$array[$i]\t";
      }
      print SIB "\n";
}
```

Appendix 4: Exp1.in (file generated by perl script PreToSib.pl, which is compatible for SIBPAIR program)

```
set error_drop on
set loc a affection
set loc D21S1256 marker 0 cM
set loc D21S1914 marker 0 cM
set loc D21S1909 marker 9.09 cM
set loc D21S1252 marker 19.06 cM
set loc D21S2055 marker 24.1 cM
set loc D21S266 marker 29.48 cM
set loc D21S1446 marker 41.38 cM
read ped inline
2427  55    7     52    F     X     3     7     5     6     3     3
      4     4     1     3     3     4     1     7
2427  54    7     52    M     X     0     0     0     0     0     0
      0     0     0     0     0     0     0     0
2427  53    7     52    M     X     3     8     3     6     3     7
      4     9     3     9     3     4     1     1
2427  52    0     0     F     X     7     8     3     5     3     7
      4     9     1     9     3     3     1     7
2427  49    6     48    M     X     3     3     4     6     6     6
      3     8     19    7     3     4     1     7
2427  48    0     0     F     X     0     0     0     0     0     0
      0     0     0     0     0     0     0     0
2427  44    5     42    F     X     7     7     6     8     6     7
      3     4     3     7     3     4     1     7
2427  43    5     42    M     X     3     7     5     6     3     4
      1     4     3     4     4     6     4     7
2427  42    0     0     F     X     3     7     5     6     4     6
      1     3     4     7     3     6     7     7
2427  34    3     32    M     X     3     7     2     3     6     6
      1     3     19    7     3     3     1     1
2427  33    3     32    F     X     3     7     2     3     6     6
      1     3     19    7     3     3     1     1
2427  32    0     0     F     X     3     7     3     6     6     6
      1     3     19    7     3     3     1     2
2427  20    2     17    F     X     7     8     2     7     6     6
      3     8     3     8     3     5     1     4
2427  19    2     17    F     X     7     8     0     0     0     0
      0     0     3     8     2     2     4     4
2427  18    2     17    F     X     7     8     7     8     6     7
      3     8     19    3     2     4     1     4
2427  17    0     0     F     X     3     8     7     8     6     6
      3     8     3     8     2     5     1     4
2427  16    0     0     F     X     0     0     0     0     0     0
      0     0     0     0     0     0     0     0
2427  15    0     0     M     X     0     0     0     0     0     0
      0     0     0     0     0     0     0     0
2427  14    15    16    F     X     7     7     2     8     5     6
      4     4     3     3     3     4     1     1
2427  13    15    16    F     X     3     3     2     6     5     6
      3     4     19    3     3     3     1     4
......
```

Appendix 5: Exp1_fre.txt (output file from SIBPAIR program, which is converted to DAT file by perl script fretoDat.pl)

```
-----------------------------------------------
Segregation ratios for trait "a         "
-----------------------------------------------

Total sample   All       Fndrs     Nonfndrs
-------------------------------------------
   Aff/Tot   217/ 224   10/  10  207/ 214
   Prop Aff     0.969     1.000     0.967
   Missing       1254       364       890

Mating Type     UxU       UxA       AxA
-------------------------------------------
   Matings          0         0         0
   Aff/Tot     0/   0    0/   0    0/   0
   Prop Aff     0.000     0.000     0.000

Relative pair  RecRisk   Aff-Aff   Aff-UnA
-------------------------------------------
   Marital      0.000         0         0
   Gparent      1.000         1         0
   Halfsib      1.000         3         0
   Par-Off      0.979        46         2
   Fullsib      0.966       183        13

-----------------------------------------------
Allele frequencies for locus "D21S1256"
-----------------------------------------------
   Allele  Frequency   Count  Histogram
       1     0.0017        2  *
       2     0.0051        6  *
       3     0.3774      440  ********
       4     0.2264      264  *****
       5     0.0369       43  *
       6     0.0129       15  *
       7     0.2196      256  ****
       8     0.0918      107  **
       9     0.0240       28  *
      10     0.0043        5  *

Number of alleles    =   10
Heterozygosity (Hu)  =    0.7482
Poly. Inf. Content   =    0.7091
4 Neff mu (SSMM)     =    7.41622752
Number persons typed =    583 ( 39.4%)

-----------------------------------------------
Allele frequencies for locus "D21S1914"
-----------------------------------------------
   Allele  Frequency   Count  Histogram
       1     0.0049        6  *
       2     0.0786       96  **
       3     0.1751      214  ****
......
```

Appendix 6: fretoDat.pl (perl script used to convert Exp1_fre.txt into exp1.dat)

```
#turn exp1_fre.txt to exp1.dat
#how to use
#perl fretoDatafile.pl exp1_fre.txt exp1.dat

open FRE,"<".$ARGV[0] or die "can't open $ARGV[0]";
open PAR,">".$ARGV[1] or die "can't open $ARGV[1]";

my %hash_chr1,$str,$number,@array;
while(<FRE>){
            if($_=~/Allele frequencies for locus "(D\d+S\d+\w?)"/){
                  $str=$1;
                  $number++;
                  push (@array,$str);
            }
            if($_=~/\d+\s+\d+\s+.*/){
                  my @array=split /\s+/, $_;
                  push(@{$hash{$str}},$array[2]);
            }
}

print PAR " ",$number+1," 0 0 5  \n";
print PAR " 0 0.0 0.0 0\n";
print PAR "  ";
for my $i(1..$number+1){
      print PAR "$i  ";
}
print PAR "\n";
print PAR "1   2  \n";
print PAR "  0.99700  0.003  \n";
print PAR " 2  \n";
print PAR "  0.00100000  1.0000  1.0000 \n";
print PAR "  0.5000  0.5000  0.5000\n";
foreach my $key(@array) {
      print PAR "3  ",$#{$hash{$key}}+1,"  \n";
      print PAR "  ";
      foreach my $array(@{$hash{$key}}){
            print PAR "$array  ";
      }
      print PAR "\n";
}
print PAR " 0 0  \n";
print PAR "  ";
for my $in(1..$number){
      print PAR "0.1000  ";
}
print PAR "\n";
print PAR " 1 0.10000 0.450000  \n";
```

Appendix 7: Exp1.dat (DAT file converted by fretoDat.pl)

```
8 0 0 5
 0 0.0 0.0 0
  1  2  3  4  5  6  7  8
1   2
  0.99700  0.003
 2
  0.00100000  1.0000  1.0000
  0.5000  0.5000  0.5000
3  10
  0.0017   0.0051   0.3774   0.2264   0.0369   0.0129   0.2196   0.0918
0.0240  0.0043
3  12
  0.0049   0.0786   0.1751   0.0606   0.1080   0.1841   0.1538   0.1718
0.0336  0.0229  0.0049  0.0016
3  10
  0.0414   0.2206   0.0617   0.0462   0.0848   0.2977   0.1821   0.0453
0.0087  0.0116
3  11
  0.2013   0.0025   0.3077   0.1088   0.0172   0.0106   0.0859   0.1440
0.1146  0.0065  0.0008
3  22
  0.0417   0.0408   0.2456   0.0515   0.0029   0.0107   0.0718   0.1563
0.0311  0.0291  0.0379  0.0262  0.0583  0.0165  0.0427  0.0466  0.0427
0.0010  0.0301  0.0019  0.0087  0.0058
3  11
  0.0828   0.0270   0.6590   0.0352   0.0008   0.0189   0.1025   0.0623
0.0090  0.0008  0.0016
3  11
  0.4295   0.1210   0.0019   0.1029   0.0590   0.0038   0.1867   0.0629
0.0067  0.0038  0.0219
 0 0
  0.1000  0.1000  0.1000  0.1000  0.1000  0.1000  0.1000
 1 0.10000 0.450000
```

Appendix 8: interpretation of genehunter's command

```
npl:1> photo chr7.out
'photo' is on: file is 'chr7.out'
//output was directed to the chr7.out file, so everything printed on
//the screen would be saved into chr7.out file.

npl:2> ps on
Postscript output is now 'on'
//the "total stat" command will prompt the user for filenames in which
//to store postscript graphs for total LOD score, total NPL statistic,
//and total information content.

npl:3> skip large off
Large pedigrees are now used but trimmed.
//pedigree individuals will be trimmed off until the pedigree is small
//enough to be analyzed within the current setting of 'max bits'. This
//trimming is done such that the maximum amount of linkage information
//is retained – the first individuals to be eliminated will be
//unaffected individuals to at the bottom of the pedigree as these
//individuals add very little to the NPL statistic and will affect the
//LOD score somewhat depending on the proposed penetrance of the
//disease allele.

npl:4> map function kosambi
The Kosambi map function is now in use.
//choose a cM <-> rec-frac conversion function
//this command controls which mapping function is used to convert
//centiMorgans to recombination-fractions and back again both in the
//input and output of the program and in the internal calculations. The
//default 'map function' is Kosambi

npl:5> haplotype off
Haplotype output is now 'off'
//determine likely haplotypes for individuals

npl:6> off end 5
Scanning will now be done 5.0 cM beyond the ends of the map
//select how far to comput score beyond ends of the map
//this command controls how far before the first marker and after the
//last marker in a map scores will be calculated. Here subsequent scan
//commands will begin calculating scores 5.0 cM before the first marker
//and continue stepping through until 5 cM after the last marker. The
//default value of 'off end' is 0.0 cM.

npl:7> increment step 5
Scanning will be done in 5 steps per map interval
//choose the scan step size
//scan command will calculate scores at 5 equally spaced positions
//between each marker.

npl:8> count recs on
Count recs is now 'on'
//turn recombination counting on
//Turning this option on activate the recombination-counting mechanism
//in the "scan" command. After each pedigree is scanned, the observed
//recombinations (and resulting distances) are shown for each map
```

```
//interval alongside the actual distance of the interval. When there
//are significantly more recombinants than expected in an interval or
//set of intervals, this can often indicate an error or errors in the
//genotype data.

npl:9> analysis BOTH
The current analysis type is 'BOTH'
//select what type of linkage analysis to perform?
//BOTH: both NPL and LOD scores will be produced.

npl:10> load marker chr7.dat
// load marker locus data
//This command reads in the marker locus data (allele frequencies for
//each genetic marker, frequency and penetrance information for the
//disease). The format of this file must be identical to the linkage
//parameter file (output from the PREPLINK program).

npl:11> scan chr7.pre
// analyze pedigree data
//The main analysis command in GENEHUNTER is the ‘scan’ command. For
//each pedigree found in the file indicated, the ‘scan’ command will
//compute LOD scores and NPL sharing statistics at many position in the
//genetic map.
…
…


npl:12> total
//show total scores from a scan of muliple pedigrees
//The ‘total’ command can only used after a successful ‘scan’ command
//of multiple pedigrees. It will display the same 5 columns of output
//as the ‘scan’ command produced for each pedigree.
Totalling
pedigrees: ..................................................
…
…
file to store postscript plot [npl_plot.ps]: npl.ps
file to store postscript plot [lod_plot.ps]: lod.ps
file to store postscript plot [info_content.ps]: info.ps

npl:13> quit
//quit the program.
```