# Genetic mapping of complex traits: the case of Type 1 diabetes

*Päivi Onkamo*

*Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute*
*and*
*Division of Biometry, Rolf Nevanlinna Institute*
*and*
*Finnish Genome Center*

*Faculty of Science*
*University of Helsinki*

*Academic dissertation for the degree of Doctor of Philosophy*

*To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium IV, Porthania, on January 25th, 2002, at 12 o'clock noon.*

*Helsinki, 2002*

## Supervisors

Professor Elja Arjas, PhD
Rolf Nevanlinna Institute, University of Helsinki and
National Public Health Institute, Helsinki

Professor Juha Kere, MD, PhD
Department of Biosciences at Novum, Karolinska Institute and
Finnish Genome Center and Department of Medical Genetics, University of Helsinki

## Reviewers

Professor Juni Palmgren, PhD
Department of Mathematical Statistics, Stockholm University and
Department of Medical Epidemiology, Karolinska Institute

Docent Jorma Ilonen, PhD
Department of Virology, University of Turku

## Opponent

Professor Lynn Jorde, PhD
Department of Human Genetics, University of Utah, Salt Lake City

**Research reports A39**
**January 2002**

Rolf Nevanlinna Institute
P.O.Box 4 (Yliopistonkatu 5)
FIN-00014 University of Helsinki, Finland

Printed version
ISBN 952-9528-70-1
ISSN 0787-8338
YLIOPISTOPAINO
HELSINKI 2002

PDF version
ISBN 952-9528-71-X
http://ethesis.helsinki.fi/
HELSINGIN YLIOPISTON VERKKOJULKAISUT
HELSINKI 2002

# Contents

## List of original publications

This thesis is based on the following original papers, referred to in the text by their Roman numerals (*I*-*V*). In addition, some unpublished data are presented.

I Onkamo P, Väänänen S, Karvonen M, and Tuomilehto J (1999). Worldwide increase in incidence of Type I diabetes – the analysis of the data on published incidence trends. Diabetologia 42:1395-1403

II Onkamo P, Pitkäniemi J, Tuomilehto J, and Arjas E. Increasing incidence of type I diabetes – a role for genes? Submitted to Am J Epid.

III Pitkäniemi J, Onkamo P, Arjas E, Tuomilehto-Wolf E, Tuomilehto J and the DiMe Study Group (2000). Estimation of transmission probabilities in families ascertained through a proband with variable age-at-onset disease: application to the HLA A, B and DR loci in Finnish families with type 1 diabetes. Hum Hered 50:308-317

IV Toivonen HTT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, and Kere J (2000). Data mining applied to linkage disequilibrium mapping. Am J Hum Genet 67:133-145

V Onkamo P, Ollikainen V, Sevon P, Toivonen HTT, Mannila H, and Kere J. Linkage disequilibrium mapping by data mining: utilizing covariates and phenotype measurements in search of genes for quantitative and complex traits.

Publications II and III are also found in thesis by Pitkäniemi J.

## Abbreviations

| | |
|---|---|
| ASP | affected sib pair |
| *ApoE4* | Apolipoprotein gene, allele *E4* |
| bp | base pair |
| cM | centiMorgan |
| CF | cystic fibrosis |
| DERI | The Diabetes Epidemiology Research International Group |
| Df | degrees of freedom |
| DiMe | the Childhood Diabetes in Finland Study |
| DNA | deoxyribonucleic acid |
| DS | disease susceptibility |
| DZ | dizygotic (twins) |
| EURODIAB ACE | the European collaborative research project for diabetes |
| HGS | human genome sequence |
| HLA | Human Leukocyte Antigen; HLA-B, HLA-Cw, HLA-DQB1 and HLA-DRB1 are HLA genes. |
| HPM | Haplotype Pattern Mining |
| IBD | identical by descent |
| IDDM | insulin-dependent Diabetes Mellitus |
| *INS* | Insulin gene |
| kb | kilo base pair (1 000 bp) |
| KDD | knowledge discovery in databases |
| LD | linkage disequilibrium, non-random association of alleles in consecutive loci |
| LOD (score) | logarithmic odds of the ratio of likelihoods |
| Mb | mega base pair (1 000 000 bp) |
| MCMC | Markov Chain Monte Carlo; computational method |
| MZ | monozygotic (twins) |
| NPHI | National Public Health Institute |
| OMIM | Online Mendelian Inheritance in Man |
| PIC | polymorphism information content |
| QHPM | Quantitative trait Haplotype Pattern Mining |
| SNP | single nucleotide polymorphism |
| TDT | transmission/disequilibrium test |
| VNTR | variable number of tandem repeats; abbreviation used for minisatellite loci |
| WHO | World Health Organization |
| WHO DIAMOND | World Health Organization Project of Childhood Diabetes (DIAbetes MONDiale |

# 1. Abstract

The risk of developing Type 1 diabetes depends on the action of a number of genes in combination with environmental factors. The pathogenic process is unknown, but at the time of diagnosis the autoimmune destruction of the insulin producing beta cells of pancreas has proceeded to a level where an individual is unable to survive without insulin replacement therapy. Type 1 diabetes is the second most common chronic disease of children in Finland, with approximately 0.4% of total population being affected. The incidence is one of the highest in the world and is continuously increasing. The reasons for the upward trend in the incidence are not known. The main predisposing genetic region for Type 1 diabetes is the HLA locus on chromosome 6p21.3 (*IDDM1*). Recent genome scans have identified non-HLA loci linked to Type 1 diabetes, with much weaker effects than *IDDM1*. More accurate identification of the genes and of the DNA variants involved will lead to a better understanding of the disease.

The first part of this thesis is concerned with incidence trends: a survey of global incidence trends revealed strong evidence for a global increase. A hypothetical genetic explanation for the increase was given and studied, namely the possibility of non-Mendelian transmission of diabetes susceptibility alleles, increasing the pool of predisposing alleles in the population. The transmission probabilities of HLA A, B and DR alleles from parents to offspring were estimated from a nationwide Type 1 diabetes genetic epidemiological study carried out in Finland. Existence of strong non-Mendelian transmission could be ruled out, but minor deviations in the transmission probabilities may still be possible. A simple population model for the effects of modest transmission distortion on the incidence showed that this mechanism alone could not be the cause for the observed trend.

The second part of the thesis is concerned with the development of new methods for finding complex disease loci, with special applications to Type 1 diabetes. Data mining algorithms were used for linkage disequilibrium mapping (Haplotype Pattern Mining, HPM). The approach enables one to find loci even with strong allele and locus heterogeneity and associated low penetrances, which are expected in complex diseases, given that there are a sufficiently small number of founder disease alleles in the population. The method was applied to real data from Type 1 diabetic families from the UK, where the known susceptibility gene was accurately localized with less data than had previously been used by alternative methods. The presented studies demonstrate the advantages of utilizing the data mining approach in complex trait mapping.

# 2. From epidemiology to gene mapping

## 2.1 Epidemiology

Complex diseases, such as diabetes, asthma, or coronary heart disease, have long been a focus of epidemiological and clinical research. These diseases are common, they affect both the life quality and life expectancy of patients, and they exert great demands on health care system. The pathogeneses of these diseases are complex, presumably involving both genetic and environmental risk factors.

Epidemiology focuses on patterns of disease occurrence in populations and the factors that influence these patterns. The basic measures of disease frequency are incidence, prevalence and cumulative incidence (Table 1). The study of the occurrence of diseases helps in formulating possible roles for different etiologic factors. The factors are either known from previous studies, or they may simply be educated guesses based on clinical work. When associations are found, the next step is the assessment of the biological mechanisms behind them. Due to the very nature of the study questions, the field is closely connected with statistics.

**Table 1. The basic epidemiologic measures of disease frequency. Adapted from Khoury et al 1993.**

| | Incidence | Cumulative incidence | Prevalence |
|---|---|---|---|
| Numerator | Number of new disease onsets | Number of new cases in a period of time | Number of cases at a point of time |
| Denominator | Total time of exposure | Size of population at the start of observation | Size of the population at same point or period of time |
| Examples | Birth rates, disease incidence | Lifetime risk, penetrance | Allele and genotype frequencies |

The patterns of disease occurrence describe the disease at the population level, for instance geographical clustering, but they can also be used to direct genetic research to concentrate on specific subpopulations, e.g. population sub-isolates. Studies on prevalence and incidence may tell us, for example, the maximum number of affected individuals available for a genetic study. Geographic enrichment in the disease prevalence across the population may imply either some environmental or genetic factors increasing the susceptibility in the high prevalence areas. Furthermore, the accumulation of disease in families and pedigrees can be used, for example, to estimate whether a sufficient genetic component exists to be found by a genetic study.

While epidemiological studies will continue to be an indispensable tool for research into common complex diseases, considerable efforts in finding the actual biological causes of common complex diseases have been made during the passed decade. By utilizing the technique of positional cloning, supplemented with advanced statistical methods, geneticists are now, starting to disentangle the genetic factors behind the complex diseases. The search for genetic factors is based on studying the observed correlations between genetic markers and disease, in either population-based samples of affected and healthy individuals, or pedigrees sampled on the basis of disease occurrence in the pedigree members. The task is twofold: 1) to find genetic regions where the susceptibility loci are most probably located, 2) to assess the effects of specific susceptibility loci, alleles and genotypes on the phenotype. The discipline concerned with these questions is called genetic epidemiology.

Luckily for the geneticist, the Finnish population history makes it genetically less complex than most other populations. This gives a hope that the genetic basis for multifactorial traits might be less variable in Finnish people than in groups with more mixed ethnicities. Furthermore, registries such as the Finnish population registry, the National hospital discharge registry, the drug reimbursement registry, and the diabetes registry at the National Public Health Institute, provide reliable sources of information for scientists. Even on a global scale, they provide excellent opportunities for population-based, extensive family studies.

## 2.2 Genetic epidemiology

Genetic epidemiology is defined as ”the study of the role of genetic factors and their interactions with environmental factors in the occurrence of disease in human populations” (Khoury et al 1993). Genetic epidemiology stands at the cross-roads of genetics and epidemiology, having borrowed its basic concepts and methods from both: its statistical methodology derives from traditional epidemiology, while the main force driving the development of new methodology is the rapid advance of laboratory methods in genetics. Though still remote, the final objective of the discipline is disease control and prevention.

There are basically two kinds of approaches to study the role of genetic factors in disease. First, the study questions set by a genetic epidemiologist may be something like “what is the rate of the condition at birth in certain population, and can one estimate the mutation rates for this autosomal dominant condition?” Or “what are the geographical differences in the frequency of a genetic trait?” Second, but even more important, is the study of inheritance of traits in families. By using family studies the researcher tries to find out whether diseases cluster in families, whether the possible familial clustering is related to a common environmental exposure, biologically inherited susceptibility, or perhaps life style related risk factors. Or, a question might be “what is the genetic model for a trait?”

Genetic epidemiology brings together the methodologies of epidemiology and genetics. While the former has sought answers to the questions formulated above by large population-based studies, concentrating on finding weak effects in large masses of data, the latter has traditionally sought for strong gene effects in rare monogenic diseases with carefully ascertained pedigrees. When it comes to modern day disease genetics, the interest is now in finding the moderate to weak gene effects. Thus, the statistical approaches of modern day epidemiology, conjoined with effective ascertainment schemes relying on solid knowledge of medical genetics, is the core of the methodology of genetic epidemiology. The segregation, linkage, and heritability analyses (Table 2) all use pedigree data but, through extensive modelling of the data, look for small, inherited effects in disease. The methods will be explained in detail in the chapter 2.5.

Recently, much effort has been put into developing association and haplotype analyses. They utilize population-based association of genetic markers to traits, either because of the direct effect of the polymorphism on the phenotype, or due to linkage disequilibrium (LD, non-random association of alleles in nearby loci) of a marker locus to a close disease locus. A good understanding of effects of population history is central to understanding the nature of LD. It seems, therefore, that there is an important role for population genetics in the new gene mapping methodology.

The use of computationally intensive methods for gene mapping has become possible by the rapid progress of in computer technology. Technical development has been fast, but the need for even more efficient computers still persists, as the statistical models are also evolving to become more and more complex. In the coming years, the amount of genetic data is expected to explode, as new high-throughput marker typing machinery is introduced. This will set heavy demands on the computational methods used for analysing such data, as the prevailing statistical methodology is computationally very intensive. Furthermore, a growing need for ”integrated approaches” exists; to simultaneously utilize linkage and association, or to analyse multidimensional instead of one-dimensional responses. In short, the goal is to utilize the information in the data to its full extent.

**Table 2.** Strategies for gene mapping and assessing genetic effects

| | **Heritability analysis** | **Segregation analysis** | **Linkage analysis** | **Association analysis** | **TDT (Transmission /Disequilibrium Test)** |
|---|---|---|---|---|---|
| Outcome | Heritability $h^2=V_a/V_t$; proportions of variance due to genetic and non-genetic factors | Genetic model of the disease; e.g. dominant, recessive, or some other form of gene action | Estimate of DS gene location, effects, (usually includes segregation analysis) | Disease association; fine mapping of a DS gene | Association in the presence of linkage; fine mapping of a DS gene |
| Data | Pedigree data with especially quantitative phenotypes | Pedigree data with phenotypes | Pedigree data with genotypes and Phenotypes | Case control data (population-based or family trios) with genotypes | Core families or family trios with genotypes |
| Trait | Usually quantitative | Binary:<br>affection status | Usually binary, some also handle quantitative data | Binary, modifications to quantitative exist | Binary, quantitative |
| Statistical model / estimation methods | Variance component analysis, path analysis | Maximum likelihood methods | Maximum likelihood methods, Bayesian | Simple test statistics ($\chi^2$),<br>likelihood methods | Logistic regression |
| Software: examples | SOLAR | GAP, PANGAEA, SAGE | Genehunter, Linkage, GAS, GAP SOLAR, LOKI, SAGE | Dislamb, Transmit, GAS, DISEQ, HPM | ETDT, GASSOC QTDT |

## 2.3 Genes in populations

Genes do not exist in a vacuum – rather, they constitute a pool which has been formed by millions of years of evolution moulded by the processes of mutation, selection, recombination, and genetic drift. The genetic variation which we see today has its history. How do the processes shape the allelic diversity? In particular, how do they explain the existence of hereditary diseases? How does this all impact on gene-mapping trials?

### 2.3.1 Evolutionary processes

**Mutation**

New alleles are formed at a gene locus by mutations, such as deletions, insertions, point mutations, and gene conversion (the interchange of very short stretches of DNA between homologous chromosomes during meiosis). A mutation may be silent (in the synonymous site in the codon), it may change a single amino acid, the reading frame of the gene, or produce a stop codon, creating a non-functional, truncated protein product. Indeed, most new mutations are slightly deleterious. The overall mutation rates vary between genetic areas, but for coding genes they typically are low, in the range of $10^{-7}$-$10^{-4}$ per locus per gamete. Even though rare, the mutation process constantly creates new variation to the population.

A typical Mendelian trait is observed when a mutation in a single coding gene, either as one copy (dominant) or two copies (recessive) in an individual is sufficient to cause a change in phenotype. Today, 9,521 single gene disorders (separate phenotypes with proved mode of inheritance of the mutation characterized) are known and included in the OMIM (Online Mendelian Inheritance in Man) database www.ncbi.nlm.nih.gov/Omim/, see also McKusick (1998). When the alleles of patients affected by a particular disease have been sequenced, it has often been found that most of the disease alleles are descendants of the same ancestral mutation, the remaining being new very rare mutations. A representative example is cystic fibrosis, a common single-gene disorder with recessive inheritance: two thirds of disease alleles worldwide derive from the same ancestral mutation dating back to hundreds of generations ago, while the remaining disease alleles consist of almost thousand independent new mutations (989 in fall 2001, CF mutation registry, www.genet.sickkids.on.ca/cftr-cgi-bin/FullTable). The mutations range from point mutations to deletions of different sizes to frame shift mutations and stop codons, each with a different effect on the protein product: some mutations change the amino acids in the binding sites, leading to proteins with altered function, some the efficiency at which the protein works thereby altering the flux in metabolic pathway. There is a continuum of phenotypes corresponding to these different mutation effects.

Unlike single gene disorders, the numbers and the types of different mutations affecting the complex traits are not well known. The few known exceptions include disease variants which are common in populations in general, and exert only a small increase in disease risk (the risk of a sibling of an affected person to also become affected is in range $\lambda_s$=1.2-5). For instance, *ApoE4* isoform (*E4* isoform is an allele class with a certain coding site substitution), which elevates the risk for Alzheimer's disease, is found in 6-37% of subjects in different populations. Similarly, serological specificity class *HLA-Cw6* is associated with psoriasis in all populations. Susceptibility to Type 1 diabetes, is conferred by a common serological specificity *HLA-DQ8*, which is associated with diabetes all over the world. The debate is, however, still not finished about the nature of disease polymorphisms: the main hypotheses are the Common Disease/ Common Variants and multiple rare variants (Zwick et al 2000). According to recent observations that indicate a prominent population bottleneck in Caucasians approximately 50,000 years ago, and also, predictions of allelic spectra in human diseases, based on a population genetic model (Reich and Lander 2001) it is more likely that the former hypothesis is true (Reich et al 2001).

**Natural selection**

The fate of a new mutation naturally depends on its effect on the phenotype. The phenotypic effects are "tested" by natural selection. In broad terms, selection works on the abilities of different genotypes to transmit genes to the next generation. This ability is often called fitness, and is defined by $w = 1 \pm s$, where $s$ is the fraction by which the individual's genotype is superior (or inferior) to the prevailing genotype with fitness 1. Selection can function at two stages: it may change the properties of individuals surviving ability, or alter their reproductive success, or do both. It should be noted that the selection actually works over whole population: for an individual the factors affecting success in reproduction depend on multitude of factors other than just one gene, but on average carriers of a certain genotype do slightly better (or worse) than carriers of a reference genotype. The effect of *directional* selection is to gradually fix allele with higher fitness.

Natural selection has relevance even to modern human genetics: modern humans live in a grossly different environment to our ancestors, and we have developed diseases related to the lifestyle. Nutrition, in particular, has been revolutionized in the course of just a few generations. Genes that helped early humans to survive through periods of famine now cause obesity and associated diseases, with the modern diet and increased lifespan (for instance the Type II diabetes and 'thrifty' genotype hypothesis as suggested by Neel 1962). There are probably several such genes with a beneficial effect at a certain developmental stage, but which exert deleterious effects on the health later in life. There has not been any inherent selection against these genes, so that their frequencies may be very high even today. The deleterious effects can be observed as late-onset diseases only now that the average lifetime has substantially increased. Proof of natural selection acting on human populations is difficult to obtain – the well-known exception is sickle-cell anemia and malaria: the sickle-cell allele in $\beta$–globin locus produces severe anemia in the homozygous condition, while normal homozygous individuals are more susceptible to malaria. The heterozygotes are protected from both, and thus are fitter (a situation called heterozygote superiority, or *overdominance* by population geneticists). The severe anemia presented by the sickle-cell homozygote genotype prohibits the mutation from taking over the population. The geographical distribution of the sickle-cell allele follows the prevalence of the malaria parasite, *Falciparum malaria*.

Natural selection is also believed to have had a prominent effect on the HLA area on chromosome six, which is responsible for many immune system functions. The HLA gene family is exceptional compared to other genes: the mutation rates are extraordinarily high, and therefore the amount of genetic variation is also high. The HLA locus encompasses approximately 3500 kb of DNA, and is known to harbour at least 150 genes. Furthermore, approximately 80 HLA associations to chronic and infectious diseases are known (Pile 1999). It has been postulated that the fight against infectious agents, especially parasites, has directed the evolution of the HLA area, as some HLA types succeeded better than others in defending the host (de Vries and van Rood 1979, de Vries et al 1979, Markow et al 1993, Hedrick 1998).

## Genetic drift

For every new generation, gametes produced by the previous generation are randomly sampled to build up the new genotypes. If there are *N* individuals, 2*N* gametes are sampled. The random sampling process allows for changes in allele frequencies which depend only on *N* (Figure 1).

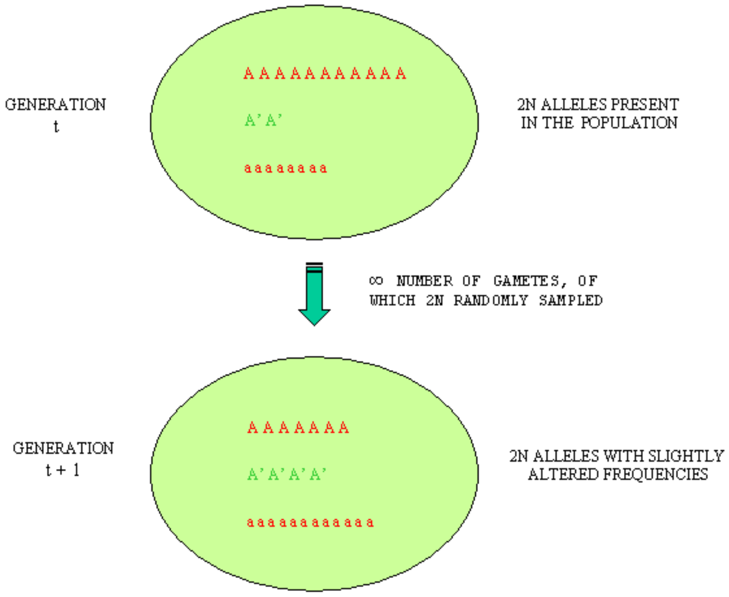


**Figure 1. The process of genetic drift.**

The process is cumulative, in the sense that the sampling for the future generation is based on the altered gene frequencies of the present generation. This sequence of cumulative changes is called genetic drift. During the process of genetic drift, some alleles are lost and some go to fixation (i.e. the fixed allele is the only one present in the population). For a rare allele variant present only in few copies in the entire population, the probability of ultimate extinction will be quite high. In fact, the rate of allele loss, or the expected time to fixation of a neutral allele, can be calculated based on statistical properties of the sampling process. The probability $F_t$ of autozygosity (i.e. the probability that an individual carries a pair of alleles that are identical by descent) in generation *t*, is given by

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t$$

(see Hartl and Clark 1989) The smaller the *N*, the more rapid is the increase in $F_t$ as a function of *t*. Obviously, $F_t$ increases with increasing numbers of generations.
The probability of not being autozygous (called allozygosity) is the same as heterozygosity (if all alleles derived from different ancestors are treated as different alleles), $H_t = 1 – F_t$, which after substitution for $F_t$ and approximation by exponential function yields

$$H_t \cong H_o \mathrm{e}^{-t/2N}$$

where $H_o$ is the heterozygosity in the start, if not 1. It is noteworthy that the variability loss due to the sampling process is quite rapid in small populations – for example, a population with effective size ($N_e$) of 40 individuals loses half of its original variation in about 25 generations, which with a generation interval 20 years yields 500 years. Based on allele frequency differences between localities in Finland, it has been suggested that drift has played an important role in the past in small and relatively isolated villages (Nevanlinna 1972).

**Polymorphic equilibria**

In a stable population, an equilibrium or *steady state* may be reached in which the rate of the formation of new disease mutations at a locus is balanced by random elimination and natural selection. For neutral polymorphisms, the equilibrium autozygosity is approximated by

$$\hat{F} = \frac{1}{4N\mu + 1}$$

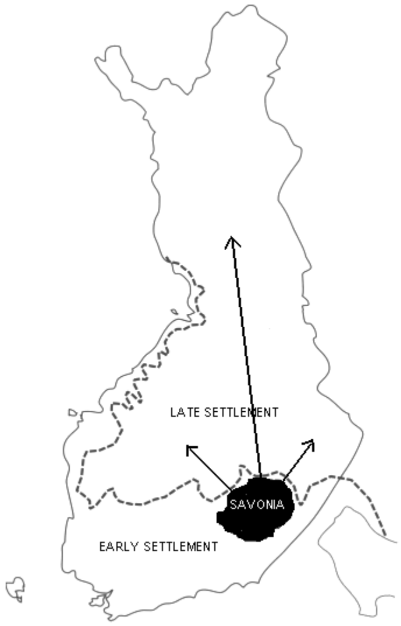
(see Hartl and Clark 1989) where *N* is the effective population size and $\mu$ is the mutation rate to neutral alleles. The quantity $4N\mu + 1$ also defines the effective number of alleles in the population, when it has reached the steady state.

If, on the other hand, the new mutations are deleterious, selection quite rapidly eliminates them from the population. If the mutation rate at the locus is $\mu$, the equilibrium frequency of deleterious alleles at the locus can be shown to be $\hat{q} = \frac{\mu}{hs}$ (where *h* is the degree of dominance, with *h* = 0 for completely recessive and *h* = 1 for completely dominant disease alleles). Thus even lethal one-gene Mendelian conditions, when there is high enough mutation rate to the disease allele, may persist in populations in low frequencies even though there is strong selection against them.

## 2.3.2 Population history

**Finnish population history**

Several, some controversial, theories of the settlement of Finland have been presented over the years. However, nowadays a plausible view is that the settlement has been a long process during which small groups of immigrants from the south, east and west have gradually mixed into the existing main population. The archaeological evidence speaks for constant habitation starting from the end of last glacial period, and thereafter a gradual development of agriculture and animal husbandry, of which the earliest evidence dates back to 3300 to 4000 years ago (Nunez 1987). Linguistic and genetic research indicates a mixed ethnic origins for Finns. Y chromosome variation in this population shows that there are two distinct male lineages in Finland (Lahermo et al 1999). The population was very small over a long period of time, only approximately 5,000 people during pre-agricultural stage, and still in 12th century no more than 50,000. The coastal areas were the first to be inhabited. The settlement of the interior started only as recently as the 1500s, when small groups from the area of Savonia migrated into the central and Eastern parts of the country (Figure 2). Thereafter, the process of emigration and founding of new communities was repeated several times, from the main villages to new uninhabited areas, producing several more or less isolated, internal subpopulations (Norio 1966, Nevanlinna 1972). The structure of these rural sub-isolates has remained amazingly stable almost to the present times, as most of the migration inside Finland has taken place from rural areas to cities. The expansion in population size started in 1700s, from 250,000, and is now levelling off at 5,180,000 (www.vaestorekisterikeskus.fi/asukasluku00-01.htm).



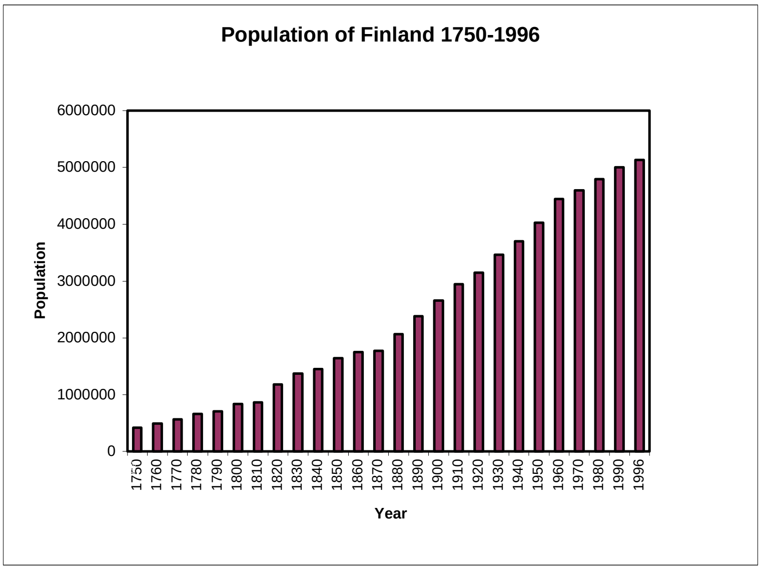


**Figure 2. Finnish population history. The southern and coastal areas have been inhabited for at least several thousands years ago, whereas the central and northern parts only in 16th century by small emigrant groups from a district in South-East of Finland, Savonia. Adapted from Norio (2000): Suomi-neidon geenit. Population size in Finland has grown from less than half a million in 1750 to 5,1 million in 1996 (adapted from Kere 2001).**

It is a well-held opinion - even outside the genetics community - that the Finnish population history makes this population especially suitable for gene mapping purposes. The small number of founders, the long period of small population size, and the expansive growth period starting in 1700s and ending after the Second World War have characterized Finnish history, and have had a major impact on its gene pool (Nevanlinna 1972). The effects of such events are discussed below. One of the purposes behind writing this thesis was to understand how the population genetic theory could be utilized to more efficiently map disease genes in a "founder" population such as Finland.

## Population history processes:
## Founder effect and population bottle-neck

As noted in the previous section, the effects of random drift are more pronounced in small populations. This is the basis for founder and bottleneck effects. Founder effects occur when a small group of individuals settles a new subpopulation: the genetic composition of the group may deviate from the source population, due to randomness in the sampling when the group is small, or because the group consists of closely related individuals (Figure 3). This shift in the genetic make-up of the new subpopulation compared to the genetic structure of the larger source population is, the founder effect. A bottle-neck may occur as consequence of the founding or population crash due to a natural disaster. The bottle-neck is the reduction in population size and thereby, the reduction in genetic diversity. After the initial settlement of a founder population, the genetic drift will be the main force acting on the subpopulation when the new population is small (and immigration negligible). Depending on the length of that time, a number of alleles may be lost, and some, even deleterious, mutations may by chance be enriched to high frequencies. An expansion in the population size will then freeze the allele frequencies to the level they were at the beginning of expansion. This freezing happens because the effect of genetic drift disappears when population is big enough. In the end, the genetic constitution of the subpopulation may be quite different from the original source population, and characterized by population specific alleles.

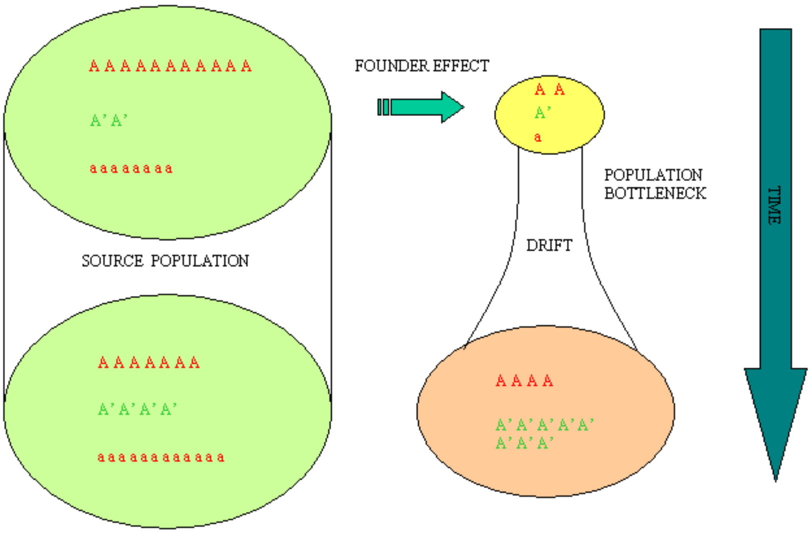


**Figure 3. Schematic diagram of founder effect and population bottleneck. A few alleles of the original source population are sampled into the new founder population (upper right corner), giving rise to founder effect (in the figure, the allele frequency of A' is elevated compared to the source population). Thereafter, subsequent genetic drift shapes the genetic constitution of the new, small population. In the example, the allele a, which is common in the source population, has disappeared from the offspring population. The source population, if large and stable, preserves its original variation in approximately the same frequencies as it had at the time of the separation of the populations.**

Many of the consequences of these processes can be observed in the Finnish population. Finns have quite a unique collection of inherited diseases – there are 36 rare single-gene disorders classified to belong to the Finnish disease heritage (Norio et al 1973, Norio 2000). On the other hand, some hereditary diseases which are relatively common in other populations are almost non-existent in Finland, including such well-known and relatively common syndromes as cystic fibrosis, Huntington's disease, and phenylketonuria. Today, all but one of the genes behind the Finnish heritage diseases have been located and cloned. The focus of research is now turning towards multifactorial diseases. Thus, the most relevant question nowadays is whether there are any special advantages – or pitfalls – in genetic mapping of complex diseases in Finnish population.

## Migration

In population genetics, migration describes the movement of individuals, and thereby their genes, from a population to another population. This exchange of genes mixes the genetic pool of the populations. Migration reduces the differentiation between subpopulations, and may even eliminate it entirely. A surprisingly low migration rate, approximately 5 % per generation, is enough to completely homogenize subpopulations (see Hartl and Clark, 1989).

**Admixture and stratification**

The term 'admixture' is used to describe a situation where individuals (chromosomes) drawn from a subpopulation have varying ethnic origins. The most obvious effect of admixture is genetic heterogeneity.

Stratification, in turn, is a situation where a population is divided into subpopulations in a non-obvious way, or the origins of people differ widely. Stratification, when present but undetected in a study sample, may easily lead to false findings. Several approaches to control and estimate the amount of stratification have been suggested recently (for example, those presented in Wijsman et al 2001).

### 2.3.3 Allele frequencies and allelic diversity

The history of evolutionary genetics has been dominated by the struggle between two competing theories describing the extent and nature of genetic variation. The neutral hypothesis assumes that most mutations are selectively neutral, with evolution being driven by mutation and drift, whereas "selectionists" claimed that most polymorphisms existing today are result of selective forces acting on the newly emerging variation. In the light of the results from the Human Genome project it seems that the amount of variation is vast (International Human Genome Sequencing Consortium 2001, Venter et al 2001), being nearer to the expectations of "neutralists" than those of selectionists.

Several studies on the amount of genetic variation in both coding and non-coding loci of humans have been carried out (Jorde et al 2000, Eaves et al 2000). The numbers of alleles per population, and the allele frequency distributions of microsatellite markers are similarly variable across all populations studied, despite differing population histories. No significant decrease in heterozygosity has been found even in isolate populations (Jorde et al 2000, Eaves et al 2000). The same seems to hold true for polymorphisms at individual SNP loci as well, although there are SNPs, which can be found in some populations only (called "population-specific" SNPs). There is, however, a potential difference between non-coding and coding SNPs: the functional relevance of amino acid altering SNPs may make them targets for natural selection (Cargill et al 1999), but this has not been proven.

The general observation of similar levels of allelic diversity of genetic markers in both isolated and large mixed populations has led to the conclusion that there would be no special advantage in using isolate populations in gene mapping studies (Eaves et al 2000). However, it should be noted that many marker alleles have been common enough to be transmitted to the Finnish population in a number of copies, even through the bottleneck. Thus, there has not been enough drift to lose these alleles, or to change their allele frequencies considerably. Conversely, the disease susceptibility alleles are often rare, and consequently many have not been present in the founders of the isolate population at all, but could be new mutations. Drift has eliminated many low frequency variants of disease alleles, which had originally been transmitted to the subpopulation.

Thus a great deal of disease allele variation present in the source population should have disappeared during the history of the isolated subpopulation (Kere 2001). This holds as well for the single-gene diseases as for complex diseases.

### 2.3.4 Linkage and linkage disequilibrium

Linkage is co-inheritance of alleles of loci that reside near to each other in a chromosome. A chromosome passed on to the next generation is a mosaic of maternal and paternal chromosomes, where the transition points are the points in which crossing over has taken place. On average, there are 3-5 crossing-overs per human chromosome per meiosis. Linkage has been extensively utilised in gene mapping: the co-segregation of a trait with particular marker alleles is evidence of the close proximity of the trait gene. The genetic distance unit, the centiMorgan (cM), is based on the observed number of recombination events between loci. Thus, linkage can be observed solely in pedigrees.

Unlike linkage, linkage disequilibrium (LD) can be detected in population samples. A pair of loci is said to be in linkage disequilibrium when, in a sample of individuals, their joint haplotype frequencies deviate from those expected under independence. As an example, consider locus 1 with alleles A and a, and locus 2 with alleles B and b, at a distance of a few centiMorgans from each other located on the same chromosome (Figure 4). At equilibrium, the frequency of the AB haplotype should equal to the product of the allele frequencies of A and B, $\pi_{AB} = \pi_A\pi_B$. If this holds, then $\pi_{Ab} = \pi_A \pi_b$, $\pi_{aB} = \pi_a\pi_B$ and $\pi_{ab} = \pi_a\pi_b$ also. Any deviation from these values imply linkage disequilibrium.

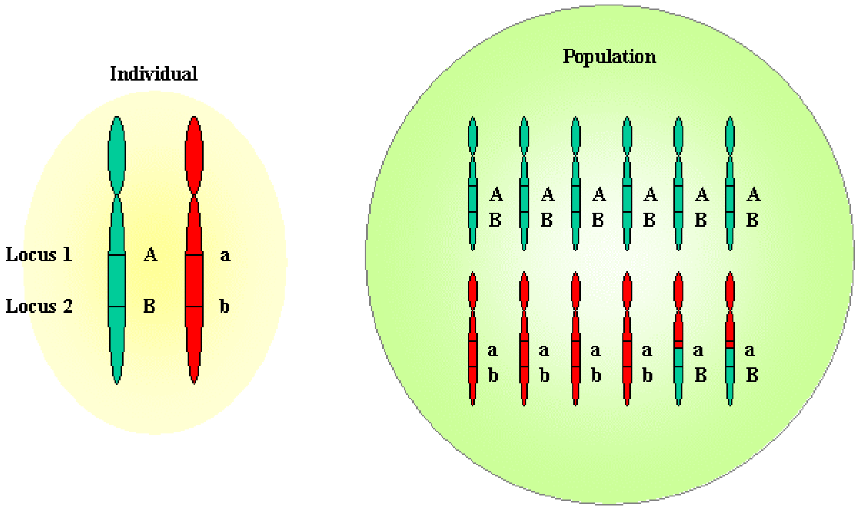


**Figure 4. Linkage disequilibrium between a pair of loci. Locus 1 and locus 2 are located a chromosome few centiMorgans apart. There are two alleles in each locus, A and a in locus 1 and B and b in locus 2. If, as in the example, certain combinations of alleles are much more frequent than others (like A and B, or a and b) there is linkage disequilibrium in the population.**

Linkage disequilibrium may arise for several reasons: a new mutation in locus 1, producing allele A´ in an AB haplotype in a population comprising of AB and Ab haplotypes will first show complete LD with allele B, as all A´ alleles in the population are associated with allele B in locus 2. Recombination will gradually mix the haplotypes so that at some point allele A´ is joined with b. Eventually the haplotype frequencies will reach equilibrium values and the LD disappears. The time required for this process depends crucially on the genetic distance of the A and B loci. The selective advantage of specific haplotypes, for instance in gene families, might lead to high frequencies of particular haplotypes and thus be seen as LD between the loci under selection and closely linked loci "hitch-hiking" withfavored alleles of the selected locus. Recent population admixture creates situations where strong LD may be found if the admixed populations are different enough in their allele and haplotype frequencies, and not too many generations have passed since the original mixing: there are long stretches of DNA in each chromosomal area with alleles originating from only one population. Undetected population stratification (if a population has an inner structure which is not known), may also lead to situation where there is LD between loci. However, these loci may even reside on different chromosomes.

Linkage disequilibrium should actually be seen as the outcome of a process, where, on one hand, new mutations, selection, drift and population admixture constantly create new LD between nearby loci, and on the other hand, recombinations and gene conversion gradually dilute it. Depending on the relative strength of these forces, the overall level and the distances to which the LD extends will vary a great deal between regions of the genome genetic areas and in some part of the genome, between populations.

**Strength of LD**

Several studies have been carried out comparing the patterns of LD in human populations with different population histories using different markers (Reich et al 2001, Eaves et al 2000, Taillon-Miller et al 2000, Abecasis et al 2001). It has been shown that in Caucasian populations there are approximately twofold amount of LD compared to Africans, probably due to a severe population bottleneck experienced by Caucasians after they left Africa (Reich et al 2001). When comparing Caucasian populations to each other, somewhat more LD is seen in small isolated populations such as Amish, Sardinians, and Finns, compared to more mixed populations. Saami people have dramatically higher levels of background LD than Finnish. This LD has probably been produced by genetic drift in a small, stable population ($N_e$=6,000) (Cavalli-Sforza and Piazza 1993). A very recent observation is that in very short distances there is less LD than there should be based on the minute amount of recombination, for which one possible explanation is gene conversion (Pritchard and Przeworski 2001, Ardlie et al 2001).

According to recent studies, it seems that the genome is in fact composed of blocks of DNA conserved as a group over long stretches of chromosome, up to 100,000 base-pairs, or 0.1 cM (Reich et al 2001). This is at least 10 times longer than was previously thought (Kruglyak 1999). In a population, there might be as few as two or four blocks in such an area (Daly et al. 2001). For mapping studies, the practical implications are that the spacing of markers need not be as dense as was previously estimated (Kruglyak et al 1999, Helmuth 2001). On the other hand, inside the blocks, increasing the number of markers does not help in refining the location of a target gene. Extensive population studies will be needed to further clarify the nature of the block structure and variation, and potential differences between populations. Currently, a haplotype map describing the block structure is under construction (Helmuth 2001).

**Background LD vs trait-associations**

An important distinction must be made between two concepts: the LD between neutral markers in a population in general, and the LD between neutral markers and a trait or disease. While the former can be observed in any population sample as *background* LD, the discovery of the latter is based on a sample of patients (and controls) in which the proportion of affected individuals in the sample has been artificially enriched. To be able to find markers in LD with a disease locus (1) the effect of the disease susceptibility locus on the phenotype must be strong enough and (2) there must exist a sufficient amount of LD between the susceptibility locus and markers neighbouring it. In addition to all evolutionary causes affecting the amount of LD, its usefulness in gene mapping also depends on the marker density, on the age of the mutation, as well as on the number and the frequencies of independent susceptibility mutations (each having a unique haplotype around the mutation).

**Measures**

A number of different statistics have been used for measuring LD. Most are defined for either two bi-allelic loci or one bi-allelic marker locus and a dichotomous trait. Those in most widespread use are $D$, $\Delta$, $\delta$, $d$ and the $\chi^2$-test statistic (Table 3). These measures are actually all derived from D. Some of these measures are dependent on population allele frequencies, so that comparisons between different populations cannot be meaningfully based on observed values of LD. Another drawback is that these point-wise LD values vary considerably from locus to locus even over very short genetic distances, due to population history, allele frequencies, and coincidence, and thus do not give a concise description of the overall LD through a genetic area. Several new methods of LD measurement have been developed, including multi-marker (haplotype) approaches allowing for multiple alleles, see chapter 2.5.2.

**Table 3. Statistics for LD. There are two alleles in the two loci, A and a in the first one, B and b in the second one. $\pi_{ij}$ is used to denote the probability of a haplotype, $\pi_A$, $\pi_a$, $\pi_B$ and $\pi_b$, the probabilities of a haplotype carrying allele *i*, as given in the small table below.**

| Measure | Definition |
|---|---|
| $D$ | $\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}$ |
| $\Delta$ | $\frac{\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}}{\sqrt{\pi_A\pi_a\pi_B\pi_b}}$ |
| $\delta$ | $\frac{\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}}{\pi_B\pi_{ab}}$ |
| $d$ | $\frac{\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}}{\pi_B\pi_b}$ |
| $\chi^2$ | $\frac{(\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB})^2 N}{\pi_{AB}\pi_{ab}}$ |

| | **B** | **b** | |
|---|---|---|---|
| **A** | $\pi_{AB}$ | $\pi_{Ab}$ | $\pi_A$ |
| **A** | $\pi_{aB}$ | $\pi_{ab}$ | $\pi_a$ |
| | $\pi_B$ | $\pi_b$ | 1 |

## 2.4 From genes to phenotypes

The etiology of the diseases, from a genetic viewpoint, can be categorized into single-gene causation, chromosomal causation, multifactorial causation with high heritability, multifactorial causation with low heritability, infectious causes and environmental causes.

The biological processes behind phenotypic characteristics differ in their complexity. In the simplest scenario, phenotypic variation can consist of two categories each of which is determined by one allele at a locus. This is the way that simple Mendelian traits arise: examples range from eye colour to traits such as some forms of idiopathic familial short stature, and diseases such as sickle cell anemia. The mode of gene action is recessive if two copies of the alternate allele are needed to produce the phenotype; this is often the case for syndromes originating from mutations which produce a non-functional protein. Dominance follows if only one copy of the mutation is sufficient for the phenotype. For instance, susceptibility to a disease is often conferred by the presence of a susceptibility allele, i.e. the allele producing protein with an altered function, regardless of the number of copies of that allele present in a cell. For both dominant and recessive characters, the heterozygote is indistinguishable from one of the homozygotes. In contrast to this, when the heterozygote is intermediate between the two homozygotes, as often is the case for quantitative traits, the joint effect of the individual alleles may be additive (the heterozygote phenotype being the sum of individual allele effects), multiplicative (product of individual allele effects), or it may not follow any easy mathematical formulation. Only rarely is the heterozygote phenotype outside the range of the homozygous phenotypes.

When there is no one-to-one correlation between phenotype and genotype, an additional set of concepts is necessary: *Penetrance* is the probability of a disease given a genotype. This can be interpreted as the lifetime risk for the disease in individuals with a certain genotype at the pre-specified locus. Low penetrance implies that the genetic predisposition conferred by the genotype is not sufficient to trigger the disease onset. A *phenocopy* is an individual expressing the trait, but who does not possess the susceptibility genotype in question (Figure 6). The oligogenic disease model is appropriate for a disease where one or a few major genes plus environmental factor(s) contribute the disease. The polygenic model points the action of several, often indistinguishable, genes, as well as environmental factors. The environmental factors may be chemical, physical, infectious, or nutritional. When there is more than one allele at a locus, or more than one locus affecting the disease, genetic heterogeneity is implied: the former situation is called allelic heterogeneity, the latter locus heterogeneity.

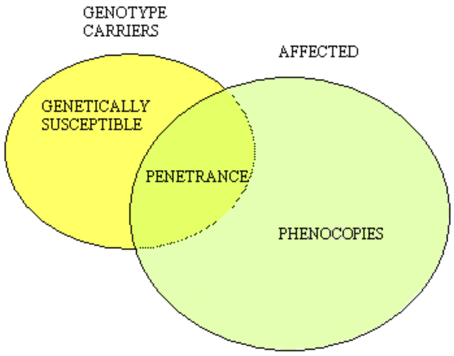


**Figure 6. Schematic diagram of relationships of concepts describing multifactorial traits. Penetrance can be interpreted as the proportion of genotype carriers who get affected during their lifetime.**

### Complex diseases

Complex diseases are inherited according to an oligogenic or polygenic model: typically, no clear inheritance pattern is seen, but the disease shows a consistent familial aggregation. The clustering in families is described by the *recurrence risk*. If we denote the probability of the disease in population by *K*, and the probability for *r*-degree relative of an affected proband to be also affected by $K_r$, then the recurrence risk is $\lambda_r = \frac{K_r}{K}$. Empirical studies on locus specific recurrence risk (Rotter and Landaw 1984), ie the genetic contribution of a specific single locus to the total recurrence risk, have shown that it is quite common that there is one locus with a large effect (major gene) and several others with weak effects (minor genes). The familial recurrence risks have also been used for estimation of the number of genes affecting a trait (Koivisto and Mannila 2001), but only models with rather low numbers of genes can yet be distinguished.

It has been hypothesized that in common diseases, the predisposing alleles would be common (Lander 1996). The rationale behind this ”common disease, common variant” hypothesis is that the high disease prevalence itself necessitates that the susceptibility alleles are common, in which case they also need to be quite old. With common alleles, it should be possible to test for association with a list of population-based SNPs and to identify disease-susceptibility mutations in that way (Risch and Merikangas 1996). Contradictory ideas have been presented by Pritchard (2001), who claims that it is more probable that there are a large number of different susceptibility alleles with quite high total frequencies. This is based on an evolutionary model for complex disease loci incorporating mutation, genetic drift, and weak purifying selection of susceptibility alleles.

### Genotype-phenotype correlation

In essence, the probability of finding a disease gene in genetic marker data is crucially dependent on the correlation between the markers and the phenotype. This, in turn, is naturally dependent on the strength of effect of the disease locus on the phenotype, and the genetic distance between the disease locus and the flanking markers, i.e. the amount of linkage disequilibrium (Figure 7). In association analysis, the frequency distribution of marker alleles in affected individuals is compared to frequency distribution in unaffected individuals. Markers with maximum difference from the distribution assuming the null hypothesis (which is complete independence between disease status and the markers in the specified genetic area) are used to infer the possible location of a susceptibility gene.

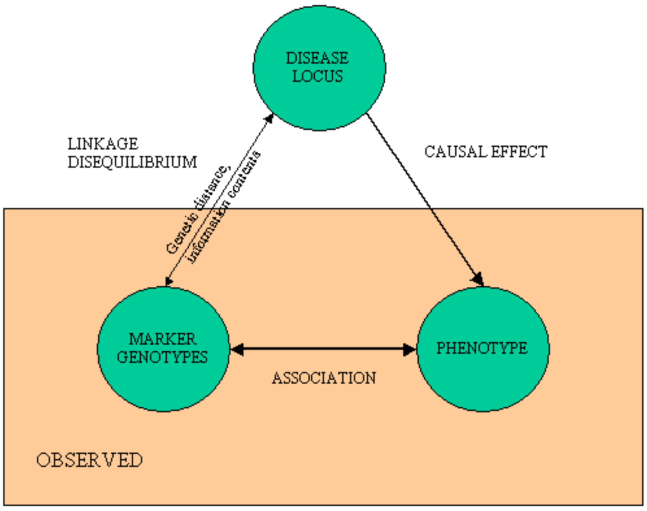


**Figure 7. Correlations between observed phenotype, observed marker data and the unknown disease locus. Only the correlation between phenotype and markers can be detected directly (modified from a lecture by J Terwilliger).**

## 2.5 Gene mapping methods

The statistical methods of disease gene mapping using genetic marker data are traditionally divided into two categories: those utilizing linkage in the pedigree data, and those which rely on population association between the phenotype and the markers. Ultimately, the two approaches just take advantage of the opposite ends of the spectrum of genetic phenomena. In linkage analysis, one searches for genetic areas which are shared between related affected individuals. The closer the relatedness, the longer are these areas and the easier they are to pinpoint. Association analysis works on the assumption that the affected individuals are related but only distantly; they share the same disease mutation as well as a short haplotype around it. The longer the time which has passed since the original ancestral mutation, the shorter is the shared segment. This gives an opportunity to very accurate localization. Essentially, both methods search for an area shared IBD as a consequence of inheritance in pedigree, the difference being in the size of the pedigree (Figure 8).

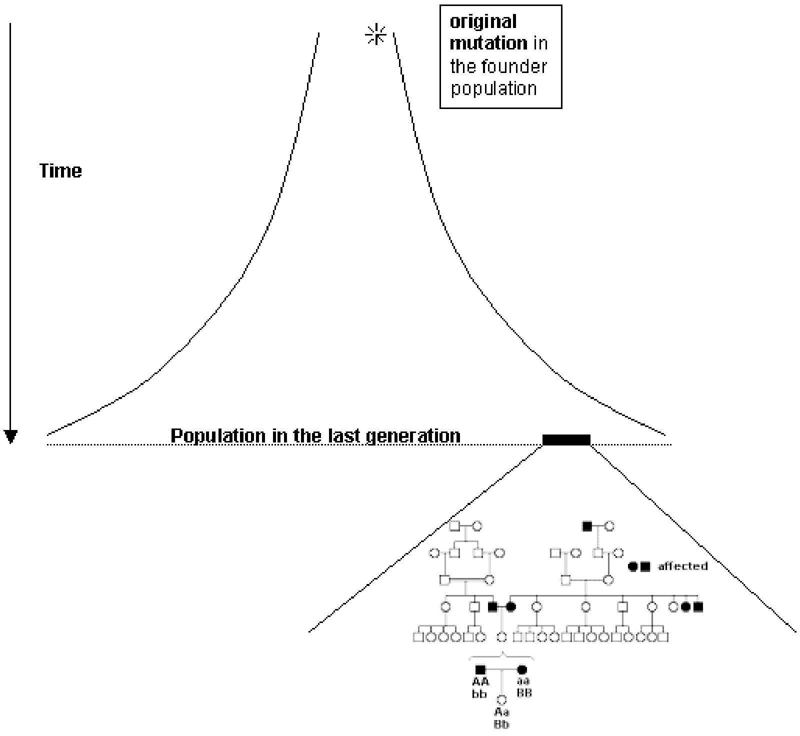

original mutation in the founder population
Time
Population in the last generation
●■ affected
AA bb
aa BB
Aa Bb

**Figure 8**. **Association vs. linkage. The disease mutation has originated in a common ancestor, and has, through random drift, spread in the population. An association of conserved haplotype around the disease mutation may still be observed in the present generation, given a suitable map density and high enough penetrance, as association analysis seeks for alleles shared in common between the affected individuals. The gene could also be mapped through linkage, in which the co-segregation of the phenotype with genetic markers is tracked in pedigrees, without paying attention to the identity of the allele.**

### 2.5.1 Linkage methods

In linkage analysis, co-segregation of known marker loci with disease status is searched for in pedigrees which include several affected and healthy individuals. Co-segregation implies that the susceptibility allele of the unknown disease locus is inherited concurrently with certain marker allele(s) more often than could be seen if the two were unlinked. In practice, pedigrees with several affected family members are obtained and genotyped for a set of markers, either covering the whole genome, or a pre-specified chromosomal area. There are several approaches to the actual statistical analysis, where the division to parametric and non-parametric methods is the most fundamental.

**Parametric linkage analysis**
In the simplest case, a single major locus is assumed to affect the trait in question. The probability of the trait value in individual *i* is given by the penetrance Pen($X_i$ | $G_i$), where $X_i$ is the phenotype of *i* and $G_i$ the disease locus genotype of individual *i*. As an example, for a dichotomous trait, denoted by *D*, which is conferred by a single locus with two alleles, the penetrances could be defined as $\text{Pen}_{\text{AA}}$=P(*D* | *G*=AA), $\text{Pen}_{\text{Aa}}$=P(*D* | *G*=Aa), and $\text{Pen}_{\text{aa}}$=P(*D* | *G*=aa). Let the *i*th person have phenotype $X_i$ and possible genotype $G_i$. Conditioning on the genotypes of each of the *n* people in a pedigree yields Ott's representation of the likelihood

$$L = \sum_{G_1} \cdots \sum_{G_n} \Pr(X_1,..., X_n \mid G_1,..., G_n) \Pr(G_1,..., G_n)$$
$$= \sum_{G_1} \cdots \sum_{G_n} \prod_{i} \text{Pen}(X_i \mid G_i) \prod_{j} \text{Prior}(G_j) \prod_{k,l,m} \text{Tran}(G_m \mid G_k, G_l)$$

in the first representation the product of the phenotype in an individual given the genotype, and the probabilities of the corresponding genotype, are summed over all genotypes. This expands into the latter form, where the product of all penetrance probabilities over all individuals, the product of prior probabilities of the *j* founder genotypes in the pedigree (those individuals who do not have known parents), and product of all transmission probabilities in triplets consisting of an offspring *m* and the parents *k*,*l*, are summed over all genotypes.

The overall likelihood of a data set consisting of several (independent) pedigrees is the product of pedigree-wise likelihoods. The inheritance model, i.e. the penetrances of the disease, are given (or guessed if not known) in advance. In reality, the *disease locus genotypes* are not known but a hypothetical disease locus is put in the vicinity of a *known marker locus* and, given the co-segregation pattern of the trait and the marker genotypes in the pedigree, the *recombination fraction* between the hypothetical disease locus and marker locus is estimated. The procedure is repeated for each marker location through the genome, the calculation of the value of the likelihood along with estimation of the most probable value of the recombination fraction. The likelihood is then compared to a likelihood assuming recombination fraction of 50% that is to the null hypothesis of "no linkage". The most commonly used way of assessing the strength of linkage is by logarithmic odds of the ratio of the two likelihoods (LOD score): $Z(\theta) = \log_{10}\left[\dfrac{L(\theta)}{L(0.5)}\right]$
The term parametric linkage analysis is used to signify that the penetrances and allele frequencies of the disease locus, as well as allele frequencies of marker loci, are required in advance.

**Non-parametric linkage methods (Affected relative pair methods)**
The class of so called non-parametric linkage methods includes methods in which the sharing of genetic areas between affected individuals in a pedigree is compared to the expected sharing given the degree of relatedness between the individuals. This type of analysis is also called affecteds-only method of linkage analysis. As a special case of the more general affected relative pair method, the affected sib-pair (ASP) method relies on the intuitive idea that in diseases with a distinct genetic component, siblings affected by

the same condition probably share susceptibility alleles transmitted through their parents. The expected sharing in a locus with no effect on a trait is ¼ for 2 alleles, ½ for one, and ¼ for 0 alleles in common between full siblings. The sharing, *Z*, is estimated in each markers locus, and compared to the probabilities under the null hypothesis. The number of alleles shared IBD at a *marker* locus provides an estimator for the number of IBD alleles at the *trait* locus. The observed and the expected sharing can be compared with a simple $\chi^2$-test, but several more refined approaches have been developed (see Davis and Weeks 1997). New methods test the sharing in the proposed disease locus directly by utilizing multipoint haplotypes. Sharing in the disease locus can be shown to be $Z_0$=0.25/$\lambda_s$, $Z_1$=0.5*$\lambda_0$/$\lambda_s$, $Z_2$=1-$Z_0$-$Z_1$, where $\lambda_s$ is the sib recurrence risk and $\lambda_0$ the risk in population. When the assumed locus has no effect on disease susceptibility, $\lambda_0 = \lambda_s = 1$ and *Z*=(0.25,0.5,0.25). The LOD score is maximized over *Z* to provide a test of linkage. These methods have been implemented in ASPEX (by D Hinds and N Risch, http://bioweb.pasteur.fr/docs/aspex/usage.html), MAPMAKER/SIBS (Kruglyak and Lander 1995), and GENEHUNTER (Kruglyak et al 1996).

The affected sib pair methods are non-parametric in the sense that nothing explicit is assumed about the inheritance model of the disease. The power of the methods is best for recessive, fully penetrant traits. In extensions of the method, quantitative phenotypes, as well as relative pairs (other than sibs), have been incorporated.

In addition to the methods mentioned, several other approaches have been applied for linkage analyses, such as Bayesian inference (Heath 1997, Uimari and Sillanpää 2001) and variance component methods (Almasy and Blangero 1998).

**Statistical power of linkage methods**

The statistical power of the parametric linkage analysis, ie the probability of rejecting the null hypothesis of “no linkage” along locations in the chromosome when the null hypothesis is false, depends on the marker map density, marker information content, number and size of the pedigrees included, the strength of the effect of the disease locus on the trait being studied, and the validity of the given disease model. To be able to observe co-segregation of a disease and a marker locus in a pedigree, it is necessary that the individuals are heterozygous for the marker loci; the probability of heterozygosity for a marker depends on the marker information content. If a marker locus and disease locus are located far apart, the probability of recombination is high (near 50%) and no linkage can be detected. In practice, a map density of 10-20 cM is sufficient to avoid too high recombination rates. Low penetrance complicates the linkage analysis: even though a person carries a susceptibility allele, he or she may be unaffected, and thus gives ambiguous information of linkage of the marker locus to the disease status. Despite the shortcomings of the parametric methods, they have proven to be very effective in finding disease gene loci in single-gene disorders. Also, they are clearly more powerful than the non-parametric methods when there is a good preconception of the actual disease model.

Compared to parametric linkage, non-parametric analysis is less powerful *if* a good approximation of the disease model for a trait is known. If a grossly incorrect disease model is used, the probability of finding the genes can be very low. For complex diseases, the underlying mechanisms are not known, and thus use of non-parametric methods may be warranted. For example, genome scans carried out for the major complex diseases, such as Type 1 diabetes, have been done on sib-pairs.

The results obtained in the study of complex traits thus far show great variability. Whereas several susceptibility genes have already been identified, for instance, for obesity, there are still only a few confirmed susceptibility loci for some other diseases, despite great efforts put to the mapping of the complex traits. The most probable reason is that there are several genes affecting disease susceptibility with very low locus specific risks, of the order of $\lambda_s \approx 1.5$. These are very difficult to identify: power analyses show that even thousands of sib-pairs or core families would be needed to identify genes with these effect sizes (Hauser et al 1996). In real studies, the typical numbers of sib-pairs collected have been only a few hundred, resulting in the situation where the actual genes found have varied widely. It is rare that the same loci are found in replicate studies. For a polygenic disease with several weak gene effects, it is much more difficult to report a specific finding than just to find *some* locus (Suarez et al 1994). One way of increasing the statistical power is to pool the data over several populations, or to do a meta-analysis and combine the results across several independent studies (Gu et al 2001).

### 2.5.2 Association analysis and LD mapping

Using linkage information is preferred when no priors exist for the potential location of the disease susceptibility genes, and genome-wide scan is warranted. On the other hand, when (1) prior candidate genes exist, or (2) initial linkage to a genetic region has been observed, association (or LD) methods are utilized. In a candidate gene approach (1), known polymorphisms in the candidate area are genotyped, and their association to the trait measured with standard association analysis methods. If, on the other hand, there is initial linkage to the trait (2), the goal is to fine-map the region, and a dense set of genetic markers ranging through the area is chosen, genotyped, and evaluated for linkage disequilibrium to a true disease gene. Both approaches rely on the assumption that at least a small part of the disease mutations originate from the same ancestral mutation, so that there will be a strong enough association (or LD) to the trait in the patients. Thus, they are considered more suitable for an isolated population setting than large, old and mixed populations. The basic difference of the approaches is that (1) association analysis is direct testing of association between the phenotype and markers, irrespectively of the exact location of the disease mutation, whereas (2) in LD mapping a statistical model of decay of LD is exploited to actually estimate the location of the disease mutation between the marker locations, based on strength of correlation of phenotype and markers.

The original simple statistics used for measuring the allelic association, such as $D$, $\Delta$, and the standard $\chi^2$-test, were developed for considering simple marker-marker or marker-disease -association, where the marker is bi-allelic. All these statistics are interrelated through simple mathematical derivations. In practice, there are differences in their sensitivity to, for instance, the population allele frequencies. They have been compared by, for example, Devlin and Risch (1995), and Guo (1997). The major drawback in using these marker-wise measures of association is that the values of the statistics vary greatly, even when pairs of loci are in the same genetic area and at very similar distances from each other. Lately, association and LD methods have been in the focus of methodological interest and improvement, as they are believed to offer new, unforeseen possibilities for genetic analyses, for example, in high-density SNP studies

(Risch & Merikangas 1996). It is expected that this kind of high-density data will enable researchers to directly measure the association between candidate genes and the phenotypes being studied.

Newer approaches are based on likelihood-based modelling of LD around a disease susceptibility gene (Terwilliger 1995, Devlin et al 1996, Lazzeroni 1998, McPeek and Strahs 1999, Service et al 1999). Genomic regions, rather than alleles, that are shared among affected individuals, are searched for. The statistical models are based on more or less simplified assumptions of the decay of LD. Regardless of these simplifications these methods have been shown to be more powerful than the traditional methods. Some of the models are robust to high levels of etiologic heterogeneity (McPeek and Strahs 1999, Service et al 1999). Some new approaches use Bayesian modelling with MCMC, for instance, Rannala and Reeve (2001). Their method uses a prior probability of the location of the disease mutation that is based on information given by human genome sequence (HGS) information. The prior is updated by estimates of LD at a set of linked markers in the region, to produce the posterior density of the location of the mutation. Most of the approaches have been developed for analysing small areas of the genome with rather low numbers of markers, and are thus best suited for fine mapping purposes. Furthermore, many are highly computationally intensive, which prevents their use with larger data sets. Some likelihood-based approaches are based on assumptions which do not hold in real data sets, and the effects of the violations of assumptions are not known. Missing and erroneous data, inherent in real data sets, may also have unforeseeable effects on the power of association methods and on the false positive rates.

It is quite clear that for complex diseases the association methods still have to be developed so as to be able to detect realistic weak gene effects, and possible interactions between genes or genes and environment. For instance, if a disease gene contains several susceptibility alleles at many different sites with low frequencies, the power of current statistical tests of association will be greatly reduced (Slager et al 2000). One of the major complications for association methods is thought to be possible population stratification, which may produce false positive findings when not correctly taken care of in the analysis. One solution that has been proposed is to use family-based controls, ie so-called pseudo-controls, but also more systematic methods that directly account for stratification (and estimate its magnitude in the data) have been presented (Pritchard and Rosenberg 1999, Satten et al 2001).

Lately, there has been increasing interest in integrating the association methods with the linkage approach; in a way this has been contemplated in the TDT test (Spielman et al 1993) and its derivatives. These approaches are based on simultaneous consideration of population association, which is seen in the sample of independent trios (father, mother and the affected offspring), and the observed transmission of the associating haplotypes from parents to affected offspring. Such an approach has been presented by for example Göring and Terwilliger (2000).

In addition, quantitative trait mapping utilising association (or LD) has been a focus of intensive study: for example, TDT tests have been extended to analyse quantitative traits (Allison 1997). TDT by variance component approaches in quantitative trait association analysis has been presented by Abecasis et al 2000), whose QTDT method also includes test statistics by Fulker et al (1999), Allison (1997), and Rabinowitz (1997). In contrast to family-based tests, an approach using population samples consisting of unrelated individuals has recently been published by Zhang and Zhao (2001). Their quantitative similarity-based association test measures the association between a candidate marker and a quantitative trait, using unlinked markers for simultaneous estimation of population stratification. As most of the populations are more or less mixed, several methodological approaches to study admixed populations have been presented recently. Strategies for disease gene mapping specially exploiting the admixture effect have also been developed (admixture mapping, McKeigue 1997, 2000).

# 3. Type 1 diabetes

Diabetes has, as a disease, been known for at least 2,000 years; there are notes of patients with symptoms like sweet odour of urine, made by Roman doctors a few centuries BC. There was no cure to the disease until 1920’s, when insulin replacement therapy was invented. Before this, the insulin-dependent type of this disease was invariably lethal, the patient dying in a matter of months or at the most couple of years after disease onset. Today, the distinction between types 1 and 2 is made on the basis of disease process: type 1 is conferred by immune-mediated destruction of β cells of pancreas and consecutive lack of insulin production. Type 1 diabetes accounts for about 10% of all diabetes. In addition to types 1 and 2, there are some other types of diabetes (Table 4).

The detailed disease pathogenesis is unknown, though the clinical symptoms in Type 1 are the result of nearly total autoimmune destruction of beta cells. The autoimmune process itself is presumed to start years before the clinical symptoms appear, even in the foetal period, which has made it difficult to find the factors that trigger the process. The disease also occurs, both naturally and induced, in other animal species, such as dogs, rats and mice. These animal models of the disease have been used intensively in the study of aetiological process, especially for finding the actual molecular triggers, which begin the autoimmune process.

**Table 4. Types of diabetes. In addition to those listed, there are rare forms with varying aetiologies (genetic syndromes, surgery, drugs, malnutrition, infections, other illnesses). Sources: http://www.niddk.nih.gov/health/diabetes/pubs/dmstats/dmstats.htm#four, http://www.childrenwithdiabetes.com/dictionary/**

| Type | Insulin production | Therapy | Prevalence | Age at onset | Genetics |
|---|---|---|---|---|---|
| **Type 1 diabetes** | None or very little | Insulin replacement therapy, diet, exercise | 0.4% of total population | Most frequently in children and adolescents, but also in adults | *HLA*, *INS*, several minor genes |
| **Type 2 diabetes** | Insensitivity to the action of insulin, but also insulin Deficiency | Diet, oral medication, exercise | Depends on population; usually 90-95% of all diabetes is Type 2 | Primarily in adults, also in young | Several loci known |
| **Gestational diabetes** | Insufficient | Appr. 95% of cases back to normal when pregnancy ends | 2-5% of all pregnancies | Pregnant women | Not known; same risk factors as in Type 2? |
| **MODY** (Maturity-Onset Diabetes of the Young) | Insufficient | Weight reduction, oral hypoglycaemic medications | Very rare | Children and adolescents | Mendelian one-gene form of diabetes |

## 3.1 Epidemiology of Type 1 diabetes

Epidemiological studies clearly show that Finland has the highest incidence of Type 1 diabetes in the world, with approximately 40/100,000 cases per year in children below 15 years of age. For comparison, the incidence varies from 0.2–40/100,000 per year elsewhere in the world (Karvonen et al 1993, 1997). The highest incidences are typically found in Caucasian populations, particularly in northern Europe, the lowest being found in Asia and South Africa. In Finland, the prevalence of this condition is approximately 0.4% of total population, which makes it the second most common chronic disease of the childhood (after asthma). Geographical differences between the counties of Finland are low (Ranta and Penttinen 2000), but on average there seems to be higher incidence in rural areas than in cities (Rytkönen et al 2001). Age-at-onset varies considerably, from 0-30 years of age, the peak being in approximately 9 years. The age-at-onset distribution has gradually shifted downwards, so that nowadays even very young children get the disease (Karvonen et al 1999, Dahlquist and Mustonen 2000). It is thought that genetic factors play an even more important part in the disease aetiology of the youngest patients (see Hodgkinson et al 2000, Komulainen et al 1999, Valdes et al 1999). In addition to these characteristics, there are well-known seasonal patterns in diabetes incidence, with a lowered incidence during summer in the northern hemisphere (Karvonen et al 1998). An intriguing aspect is that the probability of acquiring Type 1 diabetes is different if transmitted through affected father (7%) vs. affected mother (3.5%, Warram et al 1984, The Eurodiab ACE Study Group 1998). The reason for this difference is not known.

Though the diagnosis of Type 1 diabetes is inevitably much clearer than that of e.g. mental disorders, the overall criteria of diagnosis had not been agreed upon until the past decade. For that reason, many national registries including information from the decades preceding 1980s cannot be interpreted so straightforwardly. Misclassifications of types of diabetes are likely to have happened relatively frequently. Even today, in developing countries, where the burden of infectious diseases is the major issue of concern of the health care system, it is very likely that diabetes is underreported and thus no reliable statistics are available for a truly global evaluation of the disease epidemiology. For these reasons, the wide geographical variation in the incidence of type 1 diabetes was first shown in the late 1970s. During the 1960s to the early 1980s, data on incidence of Type 1 diabetes were only available for a few populations, mostly from regions with a high or intermediate risk of Type 1 diabetes. The lack of standardized data made it difficult to determine the true magnitude of the worldwide variation in incidence or time trends (LaPorte et al 1985). Therefore, international research collaborations were started in 1980s to collect aggregate data on the incidence of Type 1 diabetes, The Diabetes Epidemiology Research International Group, DERI (Rewers et al 1988), World Health Organization Project of Childhood Diabetes (DIAbetes MONDiale) in 1990 (WHO DIAMOND Project Group on Epidemics (1992), and the collaborative research project EURODIAB ACE (Green et al 1992a). The standardized procedures agreed upon for the collection of the incidence data now permit a comparative assessment of temporal trends among several populations.

As a part of this thesis (Publication I), the trends in the incidence of Type 1 diabetes all over the world, based on a systematic literature review, were analyzed. Statistical analysis of the data was performed in order to find out whether the trend is truly global. Another objective was to quantitatively evaluate the extent to which the change in incidence of Type 1 diabetes differs among populations.

## 3.2 Genetics of Type 1 diabetes

Susceptibility to Type 1 diabetes is mediated by a combination of a major effect from HLA region and probably at least ten other susceptibility loci elsewhere in genome (Risch et al 1993, European consortium for IDDM genome studies 2001). The hypothesis for a genetic basis is strongly supported by the increased concordance in MZ twins (30-50%) compared to that in DZ twins (5-27%), and in siblings (4-12%). The recurrence risk of the disease in siblings, $\lambda_s$, is 15 (Risch 1987). The HLA region is the best known genetic factor affecting the risk of type 1 diabetes (*IDDM1*, Singal and Blajchman 1973, Nerup et al 1974), with approximately 40-50% of sibling recurrence risk mediated by this region (Risch 1987). The exact role of the HLA in the pathogenesis is still controversial, but the polymorphisms in HLA class II sub-region containing the tightly linked loci *HLA-DRB1* and *HLA-DQB1* seems to be the best predictor of the disease. The class II molecules present extra-cellular antigens to helper T-cells and are essential in generation of immune response against foreign molecules. The DR specificities DR4 and DR3 (when associated with *HLA-DQB1* alleles *DQB1*0302*, also known as DQ8, and *DQB1*02*, respectively) have the highest effect on susceptibility. Even as high a proportion as 95% of diabetic patients carry these alleles (in Caucasians generally, though in Finland this figure is somewhat lower), whereas in the population the proportion of carriers is approximately 50%. The strong effect mediated by the HLA can also be observed in diabetic families with yet unaffected siblings: there is increased prevalence of diabetes-associated autoantibodies in siblings who are HLA identical to an index case already suffering from diabetes (Kulmala et al 2000). Other HLA gene loci, besides *HLA-DRB1* and *HLA-DQB1*, are also known to affect the disease susceptibility: *HLA-B* locus has been shown to have an effect (Nejentsev et al 1997, 2000).

The second most important susceptibility locus has been mapped near the insulin gene (*INS*, which has been designated *IDDM2*) on chromosome 11p15 (Bell et al 1984, Thomson et al 1989), where a VNTR polymorphism on the regulative region of the gene is associated with differential risk to Type 1 diabetes (Pugliese et al 1997). There are approximately 20 additional putative loci, showing weak positive signs in candidate gene analyses and genome screens (Davies et al 1994, Field et al 1994, European Consortium for IDDM Genome Studies 2001). The estimated contribution of these loci to the overall genetic risk is small indeed, $\lambda_s$=1.1-2.0. Several genome scans have shown the difficulty of mapping these small effect loci, for example two simultaneous publications of genome scans in Nature Genetics in 1998 showed that loci found by one study could not be identified by the other (Mein et al 1998, Concannon et al 1998). A recent meta-analysis on linkage data gathered in several populations, consisting of 800 sib pairs altogether, has finally been able to replicate some of the earlier findings with six of the candidate genes being confirmed (Cox et al 2001). An independent investigation of Scandinavian patient material (408 multiplex families) confirmed *HLA*, *INS* and *IDDM15* (European Consortium for IDDM Genome Studies 2001). The overall conclusion has been that the sizes of data sets used for mapping the additional loci have been deficient for the purpose.

Despite the recent advances in disentangling the complex genetic background of Type 1 diabetes, still some questions remain unanswered. What is the aetiological process leading to the overt disease? What are the triggers that start the production of auto-antibodies against the pancreatic Beta cells? What is the precise role of the major susceptibility gene complex, HLA, in this process? Are the numerous minor susceptibility loci that affect the disease susceptibility same in different populations, and what are their effects in the disease pathogenesis? Why is the incidence of the condition increasing, and why is the inheritance of the disease characterized by such peculiar patterns, as described in the previous section?

From the perspective of statistical methodology, it is evident that there are not many options left for improvement of linkage methods in order to find small effect loci in type 1 diabetes, or that linkage methods alone can be used to narrow down the genetic areas, as the number of recombinations gets too low even in the largest data sets. Furthermore, it is anticipated that in future genetic studies will mainly be done by searching for associations with candidate loci or candidate areas. Thus, there is clear interest in trying new approaches for association analysis. In this thesis, one such approach, an application of association rules from data mining to find interesting patterns of haplotypes in genetic marker data, is presented. This work led to a still ongoing project designing new methodology with approaches borrowed from computer science. Two of the papers have been included in this dissertation, the original publication describing the method, results with simulated data and with real data, and a consecutive paper in which the method has been extended to quantitative traits.

During 1980s and 1990s some reports had raised the issue of possible preferential (non-Mendelian) transmission of diabetes-associated alleles from parents to offspring (Jin et al 1994, Klitz et al 1986, Vadheim et al 1986, Martin-Villa et al 1990, Kockum et al 1994, Eaves et al 1999). This means that some alleles might not be segregating in the normal, Mendelian 50% probability. The interest in such a phenomenon comes from observations of patterns of inheritance of Type 1 diabetes: the difference in maternal and paternal transmission of the disease itself, where the probability of child to have diabetes is two-fold when the father is affected compared to the mother being affected (Warram et al 1984, and The Eurodiab ACE Study Group 1998) is yet unexplained. Secondly, the haplotypes DR3,DQ2 and DR4,DQ8 are common in most populations, despite them rendering susceptibility to diabetes (and some other autoimmune disorders), which might be at least partly explained by the proposed unequal probability of transmission. Thirdly, some ideas have been presented suggesting that the increasing incidence of Type 1 diabetes could be related to a change in the genetic pool of populations. Thus, we wanted to empirically assess the transmission probabilities of the diabetic alleles in the large genetic epidemiological data set we had at hand in National Public Health Institute.

We also wanted to find out to what extent possible non-Mendelian transmission could affect population allele frequencies of the distorting allele. We present a simple single gene model in order to evaluate the magnitude of the allele frequency change in time, and by assuming reasonable penetrance probabilities evaluate the effect on time trend of the incidence of Type 1 diabetes. We fit this model by applying the method of maximum likelihood, using data of newly diagnosed Finnish Type 1 diabetes cases under the age of 15, registered between 1965 and 1996.

# 4. Aims of the study

The main aims of the present investigation were to develop genetic epidemiological methodology and to apply it to Type 1 diabetes. The specific aims were

1. To evaluate global incidence trends in Type 1 diabetes
2. To study possible reasons for the increase in incidence, from a genetic point of view, with a simple population genetic model assuming elevated transmission probabilities for susceptibility alleles
3. To develop methods for estimating transmission probabilities of susceptibility alleles in variable age-at-onset disease data, and to apply it to Type 1 diabetes data
4. To develop new, more sensitive methods for finding complex disease loci by making use of association: the goal was to develop methods that could take into account all possible associations, and not to rely on restricting assumptions. With this aim in mind, data mining algorithms were applied to the problem of mapping a binary phenotype (affection status) locus using relatively dense microsatellite or SNP haplotype data. This work resulted a new method for association analysis, the Haplotype Pattern Mining. Later on, the HPM was further developed in order to be able to analyze quantitative phenotypes.

# 5. Materials and methods

## 5.1 Worldwide incidence trend analyses

**Literature search**

The incidence data for the worldwide trend analysis were collected with a literature search using MEDLINE, direct examination of reference lists of the articles, and hand searches of selected journals, and published conference abstracts. The final date considered was February 28th 1999. More than 160 original publications reporting time series of the incidence of Type 1 diabetes were found.

**Inclusion criteria**

The publications were evaluated with strict inclusion criteria to include only those studies with adequate reliability for meta-analysis. The inclusion criteria were 1) the study period was 8 years or more 2) the incidence rates were presented for each year separately 3) the number of cases per year was 5 or more 4) in the papers where the age standardization had been reported the incidences had been estimated with age-standardization according to the world population, and, 5) the type 1 diabetes was diagnosed according to the WHO definition.

Incidence data were obtained either from tables or from figures in the published articles. Altogether 37 studies from 27 countries met the inclusion criteria and were included in the analysis (Table 1, Publication I). The period of the studies ranged from 8 to 32 years, with the average length being 14.9 years. The estimates of the degree of case-ascertainment were high, ranging from 85 to 100. The studies included in the analysis were covered the period 1960 to 1996.

**Linear regression for trend estimation**

Simple linear regression under the assumption of normally distributed errors was used to fit a temporal trend for each population. Logarithmic transformation of the age-standardized incidence was explained by the calendar year as independent variable: $ln\lambda_i(t)=\alpha_i+\beta_i t$, where $\lambda_i(t)$ denote the age standardized incidence predicted at year $t$ for population $i$, the intercept $\alpha_i$ is different for each population, and $\beta_i$ is the population specific regression coefficient (the trend), respectively. In this form, the regression coefficient (×100%) can be interpreted as percentage, which is approximately the average relative increase in incidence per year. This multiplicative regression model was used, because it fits the data well, and is commonly used in estimating time trends in incidence.

The overall (global) estimate of the relative annual increase was obtained by using a pooled, centralized data set: first, for each population, the logarithms of the age standardized incidence rates and the time points were centered in order to make different length of studies and incidence levels more comparable. Then, using the method of least squares, a straight line constrained to cross the origin of the centered coordinate system was fitted to the pooled data set. The regression coefficient has the same interpretation as in the population-wise analysis. The analysis was subsequently repeated as weighted regression, where residual sum of squares was weighted by the number of cases in individual studies to give more weight to observations with a higher number of cases.

We also predicted incidences for the year 2010, with both the multiplicative and additive regression models. The additive regression model was fitted by using the age-standardized incidence was used as dependent variable without logarithmic transformation. The curves were simply extrapolated to the year 2010. In essence, the multiplicative model fits an exponential curve to the incidence, whereas additive model fits a straight line.

## 5.2 A population genetic model for incidence

**Model**

To evaluate the magnitude of effect of a possible non-Mendelian tranmission of susceptibility allele(s) in a population, given that such exists, a simple population genetic model was constructed. The model is based on following population genetic principles: there is random mating in a population (with respect to the susceptibility locus). A single diabetes-associated allele in one locus is assumed to show transmission distortion. The allele showing the transmission distortion and conferring increased susceptibility to Type 1 diabetes is denoted by 'A'; other alleles in that locus are simply collapsed to 'a'. $\tau$ is used to denote the probability of inheriting A from a heterozygous Aa parent. If inheritance is Mendelian, $\tau = 0.5$. For simplicity, penetrances of susceptibility alleles were assumed to be constant through time. All individuals with a certain genotype and in some specified age class have the same probability of Type 1 diabetes through the whole time period.

Let $k$ denote the genotype $k = 1,2,3$ for genotypes AA, Aa and aa, respectively, and $b$ the birth cohort. The genetic model was constructed as follows. The genotype frequencies are in Hardy-Weinberg equilibrium in the first generation. Let $q_k(t)$ denote the genotype frequencies of genotype $k$ ($k$=1,2,3) in generation $t$, and let $r_A^t$ and $r_a^t$ denote the allele frequency of A and a in generation $t$, respectively. All mating types, their frequencies (based on the assumption of random mating), and the genotype frequencies in the offspring are given in Table 1, Publication II.

Summing over all mating types, the expected new genotype frequencies in the generation $t$+1 are then

$$
\begin{aligned}
q_1^{(t+1)} &= (q_1^{(t)})^2 + 2q_1^{(t)}q_2^{(t)}\tau + (q_2^{(t)})^2\tau^2 \\
q_2^{(t+1)} &= 2q_1^{(t)}q_2^{(t)}(1-\tau) + 2q_1^{(t)}q_3^{(t)} + (q_2^{(t)})^2 2\tau(1-\tau) + 2q_2^{(t)}q_3^{(t)}\tau \\
q_3^{(t+1)} &= (q_2^{(t)})^2(1-\tau)^2 + 2q_2^{(t)}q_3^{(t)}(1-\tau) + (q_3^{(t)})^2
\end{aligned}
$$

Thus, the genotype probabilities in the offspring depend only on the transmission probabilities of the alleles and genotype frequencies of the previous generation. In order to obtain the genotype frequencies for annual birth cohorts between these generations, a linear approximation of the equations was used, as we were mainly interested in finding out the approximate magnitude of the effect, not its exact value.

The model was used to simulate the allele (and genotype) frequency change in a population where there is increased transmission of certain alleles. By adding age-group and genotype specific incidences, the expected population incidence can be obtained.

**Parameter estimation**

Speculating further with the population model, we wanted to estimate the transmission probability from the real annual incidence data when assuming that this would be the only factor changing the incidence through the study period. The transmission was estimated by fitting the genetic model for the increasing incidence to the data of new Type 1 diabetes cases in Finland from 1965 to 1996. The penetrances, $\lambda_{jk}$, were assumed to be constant in the age groups 0-4.99, 5-9.99, and 10-14.99 years, in order to reduce the number of parameters to be estimated. Given that Type 1 diabetes is a rare disease and that the numbers of new cases in each year, age group and genotype class are mutually independent, the number of incident cases should be Poisson-distributed. The Poisson-likelihood was expressed as

$$P\left(d_{ij};\ i\ =\ 65,...,96\ , j=1,2,3 \mid \theta\right)=\sum_{i}\sum_{j}\left[d_{ij}\log(\mu_{ij})-\mu_{ij}\right]$$

where $\theta=\left(\lambda_{jk}(j,k=1,2,3),\tau,q_0\right)$, $d_{ij}$ is the number of new cases in year $i$ in the $j$-years-old, and $\mu_{ij}$ is the expected number of new cases in a year based on genotype frequencies and penetrances. The frequency of susceptibility allele was chosen to be 0.2 at the start of the time period considered (in the 1930s). The initial value of the transmission distortion $\tau$ was set to 0.5, corresponding to no transmission distortion.

**The Finnish annual age-specific incidence data**

Data on the new cases of Type 1 diabetes in Finland were obtained from two nationwide sources: new cases between 1965 and 1986 were obtained from the Central Drug Registry of the Social Insurance Institution, and between 1987 and 1996 from the prospective childhood Type 1 diabetes registry. In Finland, all children with Type 1 diabetes are treated in hospital at the time of diagnosis and therefore case ascertainment is virtually 100% complete. Details of the data collection are described elsewhere (Tuomilehto et al 1991, 1992).

## 5.3 Estimation of transmission probabilities in HLA

In order to evaluate the hypothesis of non-Mendelian transmission in HLA loci in the Finnish population, the transmission probabilities were estimated from the data that has been collected for the diabetes study by the Genetic Epidemiological Unit in NPHI, Finland. As the data has been collected for a study of Type 1 diabetes in the families, it was not straightforward to estimate transmission probabilities in diabetes-associated loci, as transmission of alleles in these loci is not independent of the ascertainment event. Thus, an approach to take this *ascertainment effect* into consideration had to be made.

**The DiMe data**

Estimation of the transmission probabilities of the diabetes susceptibility loci (HLA A, B and DR) alleles was made using the Childhood diabetes in Finland (DiMe) study families. The DiMe study is the largest population-based genetic-epidemiological family study of Type 1 diabetes (Tuomilehto et al 1992). Nationwide, all cases under the age of 15 were identified during the recruitment period from September 1986 to April 1989. The whole data set consists of 801 participating families with at least one Type 1 diabetic child. Affected children as well as their parents and siblings were HLA genotyped at A, B and DR loci using conventional serology. HLA genotyping was done on 757 families. The details of the study procedure are described in Tuomilehto et al (1992).

### Ascertainment correction

In the case of a random sample, the estimation and testing of transmission probabilities can be done using routine methods (Jin et al 1994). However, our family data was obtained from a biased sample, i.e. there was at least one affected child per family. This means that transmission of diabetic allele(s) has almost surely occurred at least once per family, as 90% of diabetics carry known HLA susceptibility alleles. Correction by simply leaving out the index child is not a very good option in order to get unbiased estimates of the transmission rates, as the probability of a family having been included in the data set depends not only on the index child but, on the other offspring at risk at the time period of recruitment as well (as these offspring also contribute to the probability the family is ascertained to the study). Therefore it is crucial to take the method of ascertainment into account in the estimation of transmission probabilities.

Using these basic assumptions, 1) dates of birth of offspring and genotypes of the parents are known 2) for sibships in the general population, the transmitted parental alleles and affection statuses are conditionally mutually independent given the dates of birth and the genotypes of the parents, 3) independence of different sibships, it can be deduced that the genotypes of all siblings younger than the index child are always sampled independently of the events that led to the ascertainment of the family. As a consequence, there are three possible subsets of children in each sibship: 1) children who were older than 15 years at the beginning of the recruitment period 2) children who were eligible to become a proband and older than the index child, or the index child himself/herself, and 3) children younger than the index child. Genotypes, i.e. the transmission of alleles from parents to offspring, are independent of the ascertainment in sets 1 and 3, but not in set 2 through which the ascertainment of the family has taken place. Children in set 2 who were not diagnosed with Type 1 diabetes during the recruitment period are less likely to have inherited diabetic genes than a random child of the same parents. Excluding set 2 from analysis allows one to make unbiased inference without need to model the probability of the ascertainment event. The mathematical proof of the above formulation is given in Publication III.

### Estimation of transmission probabilities and the statistical tests

First, the single allele transmission probabilities and their variances (as in Jin et al 1994) were estimated from the data. Tests for allele specific deviation from the expected 50% transmission were carried out on these estimates, as described in Publication III. A locus specific (global) test for an overall deviation was also carried out. All tests were done separately for males and females, to allow for sex-specific effects. The global test for

Mendelian transmission is based on the number of observed transmissions, $n_{rs}$, of all possible alleles $r$ from all parental genotypes $r,s$, where $r \neq s$, and $r<s$. The sum of squared standardized normally distributed random variables is approximately $\chi^2$–distributed with $A(A\text{-}1)/2\text{-}A_0$ degrees of freedom, where $A$ is the number of alleles and $A_0$ the number of genotypes not seen in the data. Finally, we carried out a goodness-of-fit test, where single allele transmissions are estimated simultaneously and there are less parameters to be estimated than in the global test. The details of the testing are given in Publication III.

Only parent-child sets with complete HLA genotypes were eligible for use in the transmission estimation. In the DiMe data there were altogether 718 siblings in 471 families who were born after the oldest child in a family that was diagnosed with Type 1 diabetes during the enrollment period. The numbers of families, individuals, and transmissions eligible for the analysis are given in Table 1, Publication III.

## 5.4 Haplotype Pattern Mining

**Extension to traditional association analysis**

In association analysis, the strength of association of a trait is measured either a marker or a haplotype at a time, with either simple association measures or with statistical modelling of LD designed for the purpose. The methods reviewed in chapter 2.5.2 and in Publications IV and V offer possibilities for a refined analysis for fine scale mapping, but often are not suited for larger stretches of the genome or for genetically very heterogeneous data (or genes with very small effects). Thus, the goal for developing a new approach for association analysis was to 1) exclude restrictive assumptions typical in most current methods and 2) to account for all association observed through the whole area studied in a single analysis.

Haplotype Pattern Mining (HPM) is based on algorithms developed to find frequent patterns efficiently from large databases (Agrawal et al 1993, Agrawal et al 1996). The method uses haplotypes as input; they can be obtained with Genehunter (Kruglyak et al 1996) for example. In diseases with a reasonable genetic contribution, affected individuals are likely to have higher frequencies of associated marker alleles near the DS gene than control individuals. Combinations of marker alleles, which are more frequent in disease-associated chromosomes than in control chromosomes, are searched for in the data, without making any assumptions about the mode of inheritance of the disease. These combinations, haplotype patterns, are sorted by the strength of their association to the disease, and the resulting list of haplotype patterns is used in localizing the DS gene. The method is an algorithm-based extension of traditional association analysis.

**Data mining**

Data mining, or "knowledge discovery in databases" (KDD) can be defined as methods for discovery of useful information from large collections of data. The new information being searched for may be the detection of unexpected, new associations between variables in the data, clustering of the objects in the database, etc. These objectives are achieved by use of intelligent, computationally efficient algorithms, such as association rules. Data mining algorithms have been successfully applied to a wide set of fields: market basket analysis, direct marketing, DNA sequence search, telecommunications network alarm analysis, to mention a few (http://www.cs.helsinki.fi/~mannila/).

### Finding haplotype patterns

We look for haplotype patterns that consist of a set of nearby markers, which are not necessarily consecutive to each other. Given a marker map $M$ with $k$ markers $m_1,\ldots,m_k$, a *haplotype pattern* $P$ on $M$ is defined as a vector $(p_1,\ldots,p_k)$, where each $p_i$ is either an allele of $m_i$ or the "don't care" symbol ($*$). The haplotype pattern $P$ occurs in a given haplotype vector (chromosome) H = $(h_1,\ldots,h_k)$if $p_i = h_i$ or $p_i = *$ for all $i$, $1 \le i \le k$. For example, consider a marker map of 10 markers. The vector $P_1 = (*, 2, 5, *,3, *, *, *, *, *)$, where 1, 2, 3,… are marker alleles, is an example of a haplotype pattern. This pattern occurs, for instance, in a chromosome with haplotype (4, 2, 5, 1, 3, 2, 6, 4, 5, 3).

Our goal is to search for haplotype patterns that roughly correspond to haplotypes that are identical by descent in the disease-associated chromosomes. We allow for gaps in the haplotype patterns since mutations, errors, missing data, and recombinations can corrupt continuous haplotypes. Marker mutations and errors typically will cause only very short gaps.

The maximum length of the patterns as well as the maximum number of gaps are parameters defined by the user. It hardly would make sense to look for patterns of unrestricted length and complexity, as it is expected that in most cases the data consists of relatively sparse marker map, with even the longest conserved ancestral sequences perhaps 10 cM long.

In the basic version, the HPM method is presented in terms of the (signed) $\chi^2$ -measure of marker-disease association. A signed version of the measure is used in order to discriminate disease association from control association, the positive measure being obtained when the pattern $P$ is more frequent in cases than in controls, and negative otherwise. Given a "positive association threshold" $x$, we say that $P$ is strongly associated with the disease if $\chi^2 \ge x$. The first part of the HPM method is to output all haplotype patterns that are strongly associated with the disease status for a given value of the association threshold $x$. If pattern parameters are specified - a maximum genetic length, a maximum number of gaps, or a maximum length for gaps - the task is refined by requiring that these additional restrictions are also fulfilled. On another hand, given such a frequency threshold, all patterns exceeding the threshold can be enumerated efficiently with data mining algorithms (Agrawal et al 1993, 1996).

### Gene localization

Haplotype patterns close to the DS locus are likely to have a stronger association to it than haplotypes further away; consequently the locus is likely to be where the strongest associations are. We compute the marker frequency $f(m_i)$ of marker $m_i$ with respect to $M,H,Y,x$, where $Y$ denotes the phenotypes, as the number of patterns that contain marker $m_i$, possibly in a gap. The idea is that each haplotype pattern roughly corresponds to a continuous chromosomal region, potentially identical by descent, where gaps allow for corruption of marker data. While markers within the gaps are not used in measuring the disease association of the pattern, the whole chromosomal region of the pattern is thought to be relevant.

The marker frequency gives a score for each marker. On the condition that we assume a DS gene to be present, e.g., based on linkage analysis, we would predict the gene to be somewhere close to the markers with largest frequencies. As a point prediction we could simply give the locus of the most frequent marker. This does not, of course, imply that we assume the DS locus to really overlap with the marker; we simply predict at the granularity of marker density. Consequently, the optimal point predictions of our method are within one half of the inter-marker distance from the true loci.

**Permutation tests**

The results obtained by considering marker frequencies can be contrasted against the null hypothesis that "all the chromosomes are drawn from the same distribution", i.e., that there is no gene effect in the disease status. Permutation tests can be used for this purpose. The permutations are carried out by randomly shuffling the status fields of the chromosomes in a data set, keeping the proportions of affected and control chromosomes constant, in a fashion similar to Churchill and Doerge (1994). The *p*-values are obtained marker-wise, and the DS gene is predicted to be in the vicinity of the marker with the smallest empirical p-value. *Consecutive* markers are dependent, and thus a large number of mutually dependent *p*-values are produced. This is not a problem, since we do not use the *p*-values for hypothesis testing, but only for ranking markers.

**Quantitative traits and utilization of covariates**

HPM was extended to utilize information from quantitative traits, either as a response variable or covariates. This has been accomplished simply by measuring the strength of association with a linear model. The quantitative trait HPM, or QHPM, is carried out as follows. First, all haplotype patterns that occur at least once were searched. For each pattern we fit a linear model predicting the chosen phenotype,

$$Y_i = \alpha + \beta * P_j + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \ldots + \beta_n * X_{ni}$$

where $P_j$ is the indicator variable for the occurrence of the pattern $j$ in the chromosome being studied, $X_{ni}$ is the value for $n^{th}$ explanatory variable for individual $i$, and $Y_i$ is the trait value for that individual. The explanatory variables, or covariates, might be environmental factors, sex, and age at examination. The model was fitted using functions in publicly available statistical programming language **R** (Ihaka and Gentleman 1996). The significance of a pattern as a covariate was obtained from a t-test comparing the model to the best fitting model in which the corresponding coefficient is zero. These nominal significances (*p*-values) form the basis for the scoring function for markers used in this approach.

**Scoring functions**

For each marker in turn, all haplotype patterns that overlap with the marker are considered. The markers, in which the overlapping haplotype patterns show strong association to the phenotype are those of the most interest. The distribution of observed nominal *p*-values of all haplotype patterns overlapping a marker were compared to *p*-values distributed uniformly between 0 and 1. This is because uniform distribution is expected for mutually independent patterns not associated with the trait. We acknowledge that the patterns we observe are not mutually independent, but use the uniform distribution as an approximation of the expected distribution, under the null hypothesis of no trait association. As a measure of difference to the observed distribution of *p*-values the following heuristic scoring function was used. Let $p_i$ be the $i^{\text{th}}$ smallest *p*-value of the *n* observed *p*-values for a given marker, and $q_i$ the corresponding expectation ($i/(n+1)$) if *n* *p*-values were randomly picked from the uniform distribution. The score was defined as the mean of the distances $(p_i - q_i) \log (p_i/q_i)$ (partial Kullback-Leibler distances). This measure yields larger distances when the observed distribution is skewed towards lower *p*-values. The disease gene is predicted to be in the neighbourhood of the marker with the largest distance measure.

In addition to the modified K-L distance, two other measures for scoring function were tested, namely the Kolmogorov-Smirnov goodness-of-fit test (comparing the observed and expected distributions of the *p*-values) and adjusted $R^2$ of the linear regression model (the rationale behind being that $R^2$ should reach higher value when a new explanatory variable is added to the model).

### Simulated data with binary affection status

The performance of HPM and QHPM were evaluated with simulated data sets. The data were produced by the Populus simulator package (V. Ollikainen 2002, dissertation), to correspond to a small, isolated founder subpopulation. The population grows from the initial size of 300 to approximately 100,000 individuals over 500 years. Each individual was assigned a pair of homologous chromosomes. The length of chromosome was 100 cM for both sexes. Crossing-over is allowed to take place with equal probability along the chromosome, and without any chiasmatic interference. The density of microsatellites was 1 per cM, and that of SNPs 3 per cM of chromosome. The PIC (polymorphism information content) of the simulated microsatellites was set to 0.7, and that of SNPs 0.375 (the two alleles being equifrequent).

The disease allele was assumed to be dominant, so that each individual with at least one susceptibility allele has the same probability of disease. A high phenocopy rate was used (individuals with no susceptibility alleles are assigned as being affected): the proportion of chromosomes actually carrying the susceptibility mutation was 2.5-10%, which correspond to relative risks of first-degree relatives of $\lambda$=1.2-4.1. The sample sizes, which were ascertained from the simulated population replicates, were small, 100 affected individuals with 100 pseudocontrols, in order to keep the testing situation as realistic as possible. The effects of erroneous data were tested by introducing random errors in 0-10% of the data, and the effect of missing data by removing 0-20% of alleles randomly. For simulated SNPs, missing data were simulated by randomly removing 12.5% of the alleles. This was done in order to mimic the effect of haplotyping ambiguities with SNP markers, expected to occur whenever a family trio, both parents and the only offspring, are heterozygous in a given locus.

### Simulated quantitative trait data

The quantitative trait data was simulated in a corresponding manner than the dichotomous trait data, with an isolated population growing from 100 founder individuals to 100,000 in 20 generations. The genetic length of the simulated chromosomes was 100 cM for both males and females. Within this region, the disease locus was randomly selected, and 6 founder mutations were randomly assigned to the initial population, all of which were then associated to a different founder haplotype. No chiasmatic interference was modelled. The simulated microsatellite markers had 4 alleles with frequencies of 0.4, 0.2, 0.2, and 0.2 in the founder population. The markers were spaced 1 cM apart. We computed the liabilities for each individual using two alternative models:

$M_1 = 2x_g + x_{e1} + x_{e2} + x_r + C_1$, (difficult model) and

$M_2 = 5x_g + x_{e1} + x_{e2} + x_r + C_2$, (easy model) where

$x_g$ is an indicator variable for the presence of at least one of the disease-predisposing mutations, so the disease alleles are dominant with reduced penetrances. Variables $x_{e1}$ and $x_{e2}$ are environmental components and $x_r$ an unobserved random component, all of which follow a standard normal distribution. Constants $C_1$ and $C_2$ represent the baseline liability, and they are adjusted in an extra pre-sampling phase to make the prevalence of the disease as close to the target value of 5 % as possible. When the liability of an individual has been computed, the disease statuses are defined in the simulation in the following manner: an individual's probability of being affected is obtained from formula

$$\log \frac{p}{1-p} = M_i,$$

where $i$=1 for the difficult and $i$=2 for the easy model.

Five quantitative traits, $Q_1$ to $Q_5$, were simulated per each model. The value for each trait $Q_j$ was computed from formula

$$Q_j = jx_g + x_{e1} + x_{e2} + r,$$

where $x_g$, $x_{e1}$, and $x_{e2}$ are the genetic and environmental liability components described above and are the same for all traits, and $r$ is a random value between zero and unity.

To make the simulations more realistic, a population model with substructure was used. The substructuring might have a major impact on the amount and patterns of LD observed. Thus, we chose to divide the total population into 4 smaller subsections. A moderate amount of migration between neighboring subpopulations was assumed: the probability that an individual born in a subpopulation migrates to another is 4%. Within each subpopulation, spouses are selected randomly.

The sampling from the simulated population was done on the basis the affection status: 200 independent trios with an affected offspring were randomly sampled. For the analysis of quantitative traits no further sampling based on values of quantitative traits was done. This ascertainment scheme closely resembles real studies in the sense that data is often collected through an affected proband, and there are correlated quantitative traits which could also be genetically mapped.

### HLA data

We applied our method to a real data set, consisting of sib-pair families with type 1 diabetes from the UK (Bain et al 1990), which were genotyped for 25 polymorphic microsatellite markers. These markers covered a 14 Mb region including the entire HLA complex. The *HLA-DQB1* and *DRB1* loci are the primary constituents for Type 1 diabetes susceptibility mapped to this region. This data set was originally obtained in order to investigate the accuracy by which this locus could be mapped with currently available statistical methods. To test HPM in a setting similar in sample size to the simulated cases, only 200 out of the original 385 affected sib-pair families were used, with only one of the affected offspring selected randomly in each family. Control chromosomes were generated by including only the non-transmitted alleles or haplotypes. HPM was applied to this data set using the same parameters as described for the analysis of the simulated microsatellite data.

### *INS* data

Following Publication IV, several experiments were carried out with both different kinds of simulated data sets, as well as with real data. These included an as yet unpublished analysis of *INS* gene data. This data set contains 435 families from the UK, US and Denmark, each with two offspring diagnosed with Type 1 diabetes, and their parents (Bain et al 1992). The genetic region studied was around the *IDDM2* locus, i.e. the *INS* gene on human chromosome 11p15.5. Eight markers had been typed, 6 SNPs, one microsatellite, and VNTR locus. The microsatellite locus contained 6 alleles, whereas the VNTR locus contained 81 alleles, which can be divided into two major classes on the basis of the allele lengths. The allele classes are called I and III, where only class I alleles are associated with increased risk of Type 1 diabetes. As the markers are densely spaced, the shortest distances being only 300 bp, it was expected that there would be very strong LD between them. The SNP alleles were denoted with 1 and 2, 1 for previously known positively associated alleles and 2 for non-associated alleles, respectively, in the analyses reported here.

First, only one affected child per a family was chosen, so as to avoid problems of having familially correlated cases and controls in the data set for HPM. The data set was divided into two parts: in the first one the first child was always chosen, together with non-transmitted parental chromosomes as controls, and in the second set the younger child was used. The data was haplotyped using Genehunter2. For technical reasons, families in which there was missing information, or the offpring chromosomes had undergone recombinations, were left out of the final datasets. This resulted in 382 family trios in the first data set and 372 in the second.VNTR alleles were either used as such in the haplotype search, or the VNTR class was used.

### HPM pattern search parameters

For evaluation of the basic HPM (in Publication IV), the following parameter values were chosen. The maximum length of haplotype patterns was restricted to seven consecutive markers, which corresponds to segments of 6-8 cM. This is close to the average length of shared haplotypes in a population of approximately 500 years of age. At most two gaps were allowed per haplotype, and their lengths were limited to one marker. With these parameter values, localization time for one simulated data set on a 400 MHz Pentium PC was around one minute. After some experimenting, the association threshold for the signed $\chi^2$ measure was set to $x = 9$. To ensure that the selection of these particular values is not critical for the method and to assess the robustness of HPM in this respect, we also experimented using patterns of unlimited length, with longer gaps, and without gaps. For simulated SNP analysis, the pattern search parameters were modified slightly, to account for the higher density of markers: the maximum length of a haplotype pattern was 21 markers (approximately 7 cM). The maximum number of gaps was two and the maximum length of a gap was one marker.

### QHPM parameters

The QHPM analyses were mostly made using one set of parameter values: the maximum length of the haplotype patterns to search for was set to 7 markers, and the maximum number of gaps per pattern to 1, with the maximum gap length being 1 marker. The minimum number of occurrences of a pattern was 10 (frequency limit), to exclude patterns for which significant association could not be obtained. Experiments with other parameter settings have been described in Publication IV, in which it was shown that the method is extremely robust against different choices of parameters.

# 6. Results and discussion

## 6.1 Trends in incidence of Type 1 diabetes

**European populations**

In general the incidence of Type 1 diabetes was higher in European populations than among other non-Europeans (Figures 1 and 2, Publication I). This probably reflects the fact that Caucasoids tend to have higher level of incidence in general than Mongoloids and Africans do, though there are geographical differences in incidence depending on the admixture between racial groups and possible environmental exposures (Karvonen et al 1993). The level of increase seemed to be similar in some geographically adjacent populations: For example in the Northern European countries; Finland, Sweden and Norway where the incidence of Type 1 diabetes has been high for a long time, the increase was 1-3% per year. Adjacent countries around the Baltic Sea, Estonia, Latvia, Lithuania and Poland with an intermediate or low incidence (4-10/100,000/year) showed an upward course but not a statistically significant trend in incidence. The increase in incidence in Eastern Europe varied from 2.1% per year in East Bulgaria to 8.5% per year in Hungary. The four populations from UK had the mean incidences from 14.3 to 21.6 and increase in incidence ranged between 1.9 % and 3.7 % except for Leicestershire, where mean incidence was 7.8 with increase of 9.5%. The Leicestershire data, however, were considerably older (from 1965 to 1981) than from other UK study populations. Speculating on the observed changes in incidence, there seems to exist a tendency for incidence to be increasing in the countries with “western” life-style and high living standard, and the former socialist nations, which are in a process of change, with the possible exception of Baltic countries.

**Non-European populations**

The data for other than European populations (and especially non-Caucasians) was sparse: from Asia, data included Japan and China. From Africa we had Algeria and Libya, from South America only Mestizos in Peru, and we also had Polynesians in Hawaii. There was marked increase in incidence for all these countries. The data from the US and New Zealand were contradictory in the sense that there was marked, statistically significant increase in incidence in two of the populations included, while in the others the increase seems only modest, and do not reach statistical significance. Naturally, differences in sample sizes from different populations might have an effect on these results.

**Global trends**

The global annual increase was 3.0% (95% CI 2.59;3.33, p=0.0001) during 1960 to 1996, demonstrating a highly significant increasing trend. When the annual incidence rates were weighted with the number of cases in each individual study, the increase in incidence was 2.5% (95% CI 2.32;2.66; p=0.0001). The estimated population-wise regression lines illustrate well the increasing trends (Figures 1 and 2, Publication I).

**Conclusions**

Generally, a lower base-line level of Type 1 diabetes incidence was associated with an on average higher annual increase in incidence (Figure 3, Publication I). The association between the level of incidence and increase in incidence was assessed by calculating the correlation coefficient between the logarithms of incidences in 1983 predicted by the model and the incidence increases estimated by the multiplicative model, where the year 1983 was chosen because almost all studies covered it.

Predictions based on the linear model showed that Finland still would have the highest incidence in the world (50/100,000/year) in year 2010. However, when this prediction is compared to the actual incidence estimates from years 1997-2000, which have been 44.6/100,000/year, 50.0, 49.1, and 45.9, respectively (A Reunanen, personal communication), the prediction seems to be inadequate, as the actual increase rate is even more striking. According to the predictions, the next highest incidences will be in Norway, Prince Edward Island (Canada), Western Australia, Scotland (UK), Oxford (UK), and Sweden. Despite the large relative increases in the incidence observed in China and Peru, the absolute incidence rates in these countries would still remain low, less than 2/100,000/year. Based on these predictions, the incidence in Japan will be lower than 5/100,000/year and in Poland, Latvia and Lithuania the incidence will be under 10/100,000/year.

In addition to the findings that were made in this study, the literature review revealed other populations where an increase in the incidence of Type 1 diabetes had been reported (these studies were not adequate for the meta-analysis according to the criteria we used, but otherwise valid). These were Croatia, Denmark, Kuwait, the Netherlands, Russia, and Switzerland (Schoenle et al 1994, Ruwaard et al 1996, Bingley and Gale 1989, Green et al 1992b, Jaksic et al 1996, Choubnikova et al 1996, Shaltout et al 1995).

## 6.2 Effects of segregation distortion of susceptibility alleles on the incidence

**Expectations based on the model**

The simple population genetic model can be used to assess the extent to which the allele and genotype frequencies may increase in the course of a few generations with varying values of $\tau$. Three different values were illustrated in Figure 1 a-b, Publication II. Obviously, in the situation where the susceptibility allele is dominant with high penetrance and high initial allele frequency, the incidence is high. The relative change in incidence is most prominent in case there is a large difference in the relative genotype specific penetrances, even if the change in allele frequency is small. It is obvious from these considerations that extreme values of transmission distortion are to be needed in order for the incidence to increase as rapidly as it actually has done.

**Maximum likelihood estimate for transmission parameter based on real data**

Two models were fitted to the data: one with transmission probability fixed to 0.5 (M1) and another where transmission probability was estimated (M2). When the two models were compared using likelihood ratio test, model M2 fitted better ($\chi^2$=131.12, 1 df, p<0.001). The point estimate of transmission distortion was 0.86 and the estimated genotype frequencies were (0.06, 0.37, 0.57). The observed and fitted incidence for both models M1 and M2 of Type 1 diabetes are plotted in Figure 4, Publication II.

**Conclusions**

Fitting a population genetic model with non-Mendelian transmission as the sole factor affecting changes in the incidence of Type 1 diabetes in Finland led to an estimate of the transmission probability of 0.86. Such an extreme form of transmission distortion is unlikely biologically. On the other hand, it is evident that a biologically reasonable transmission distortion alone, even with the highest biologically reliable penetrances (eg defined with respect to a major susceptibility allele, DR4 (DQ8) carrier, and non-carrier genotypes) can only explain a small part of the rapid increase in the incidence of Type 1 diabetes observed in Finland. The observed increase could only be explained by realistic non-Mendelian transmission rates if the relative penetrance differences of the susceptibility genotypes were much greater than those known for DR4 (DQ8) today.

Based on these results, the role of other factors, probably environmental, in modifying the disease incidence should be emphasized. Environmental factors could either modify the penetrance of susceptibility gene(s), or act as triggering factors contributing directly to the incidence. Factors which have changed rapidly during the last few decades should be important in this respect, but none which have a well established association with Type 1 diabetes are known. It has been hypothesized that changes in penetrance might be linked to patterns of childhood immunization, but this has yet to be confirmed (Blom et al 1991).

## 6.3 Transmission probability estimates based on DiMe data

**Overall locus effects: HLA A, B, DR loci**

The global tests for transmission patterns in the HLA A, B, and DR loci showed some evidence for non-Mendelian transmission in the A locus for both maternal ($p = 0.04$) and paternal ($p = 0.04$) alleles, as well as for the B locus on maternal side ($p < 0.01$) and paternal side ($p = 0.05$). However, the transmission in the DR locus as a whole did not deviate from Mendelian expectations.

The variation between the loci is somewhat surprising given that A, B and DR loci are tightly linked to each other and are almost always transmitted together (the recombination rate inside the HLA area is approximately 1% per generation). However, it might imply that there is non-Mendelian transmission of some specific HLA haplotypes or segregation in some specific genotypes.

**Allele and sex specific effects**

The paternal A26 allele and maternal A32 allele were transmitted less often than expected. Also, segregation of alleles from paternal genotypes A28,A32 and A2,A3 may not happen according to Mendel’s law. The paternal B38 allele and maternal B62 allele were both transmitted at a lower frequency than expected. In the DR locus, the maternal DR2 allele was inherited at a reduced frequency.

**Conclusions**

The data do not provide direct support of the hypothesis of non-Mendelian inheritance of alleles at the HLA A, B, and DR loci. Though some single allele transmission probabilities differed significantly from 50%, these cannot be interpreted as conclusive because corrections for multiple testing were not made. Generally, the existence of strong non-Mendelian transmission can be ruled out. A comparison with estimates of transmission given by other studies does not reveal any consistent pattern of non-Mendelian transmission of any particular alleles.

## 6.4 Performance of HPM and QHPM

**Simulated data with a dichotomous trait**

The localization accuracy, which we call the fraction of data sets for which the localization was successful as a function of the allowed localization error, was shown to be good with the search parameters defined above (Figure 2a, Publication IV). For the "easiest" data set, with A=10% of affected individuals carrying the susceptibility mutation, the error made in prediction was less than 4 cM in 90% of all data sets. A clear decrease in the power was seen in the data sets with 5% carrying the susceptibility allele, and the method did not succeed much better than random guessing when only 2.5% of affected carrying the mutation.

By doubling the sample size, but keeping otherwise all search parameters constant, the localization accuracy improved significantly for low values of A (5%, 2.5%); for larger values of A there was not much difference (Figure 2b, Publication IV). Thus, the localization accuracy did not depend that crucially on the fraction of disease mutant carriers in the data, but more on the *number* of disease mutation haplotypes in the data.

The tests with simulated errors in the data, as well as those with missing data, showed that moderate proportions do not affect the localization accuracy at all (Figures 2c,d, Publication IV). Only the highest proportions tried, 10-20%, decreased the localization accuracy by 10-15%. A comparison to two simpler association methods (simple haplotype association, and haplotype patterns without gaps) showed that errors and missing data were more detrimental to those approaches.

The robustness of the method with respect to the selection of pattern search parameters, simulated data with A=10%, 1% corrupted and 20% missing, was re-analyzed (Figure 2f, Publication IV). The effect of gaps in the patterns was evaluated by either prohibiting gaps or by allowing the gaps to be up to three markers long instead of just one. In addition, a test was run where the length of the haplotype patterns was not limited. Differences started to appear at error bounds of at least 2 to 4 cM: allowing longer gaps improved the performance somewhat, whereas prohibiting gaps altogether resulted in a decreased performance.

**Localization accuracy with permutation tests**

Permutation tests were used to obtain more information about the significance of observed marker scores. The experimental results obtained with 1,000 random permutations showed that the peaks observed in marker frequencies in the vicinity of DS locus clearly surpassed those produced by background LD. The permutation surface for a simulated data set with A=7.5% is shown in Figure 3a,b, Publication IV. The prediction accuracy can be improved by permutation tests: We predicted the location of the DS gene to be at the marker with the smallest *p*-value instead of the most frequent marker. Optionally, given a threshold for the *p*-value, we made a prediction only if the best *p*-value was below the threshold (and otherwise replied “don't know”). The localization accuracy was somewhat improved by employing permutation tests (Figure 3d, Publication IV, A=5%). The improvement was less evident with A=7.5%, and with A=10% this modification had practically no effect. For A=2.5%, again, there was no improvement with the sample size of 100 affected individuals.

**Localization accuracy in SNP data**

The results (Figure 4, Publication IV) show that the HPM method performs well with the simulated biallelic data. For A=10% the accuracy is close to that of complete microsatellite data, despite the 12.5% of missing data; with smaller values of A the accuracy drops somewhat faster than with complete microsatellite data. Overall, the localization accuracy with 3 SNPs per 1 cM in these data sets is close to that of a map with one microsatellite per 1 cM.

**Localization accuracy with HLA data**

The results (Figure 5, Publication IV) demonstrate that the method was capable of mapping the disease locus to the marker located closest to *HLA-DQB1* and *DRB1*, that is marker D6S2444, even though background LD in the HLA and the telomeric end of the map was very strong. A comparison to the results of the $T_{sp}$ analysis (Herr et al 2000) shows that the mapping accuracy was similar with both approaches even though we used less information for the HPM method.

**Predicting the disease mutation site in the *INS* data set**

Haplotype patterns with highest values of the $\chi^2$-measure are given in Table 5a. All VNTR alleles were included in the Table 5a, whereas the VNTR class was used in 5b, enabling us to differentiate between patterns of VNTR class association vs. VNTR allele association. The algorithm includes sub-strings of long haplotypes thus yielding redundant haplotype patterns. Still, the abundance of “1” alleles in the associated patterns, especially in the middle of the region studied, was clear. All associated VNTR-alleles belonged to class I. For some reason, strong associations were not found for any alleles other than 655 and 714. The results were very similar for both parts of data we had.

Marker-wise association scores were high but not very informative. The association was strongest in markers 4 and 5, which are the VNTR and –23/*Hph*I, but this might be because scores are expected to be higher near the middle of region in any case. Thus, the significance of the score was evaluated by a permutation procedure. The permutation procedure was run with 10,000 iterations, which showed that the association is strong through the whole region (Figure 9, below).

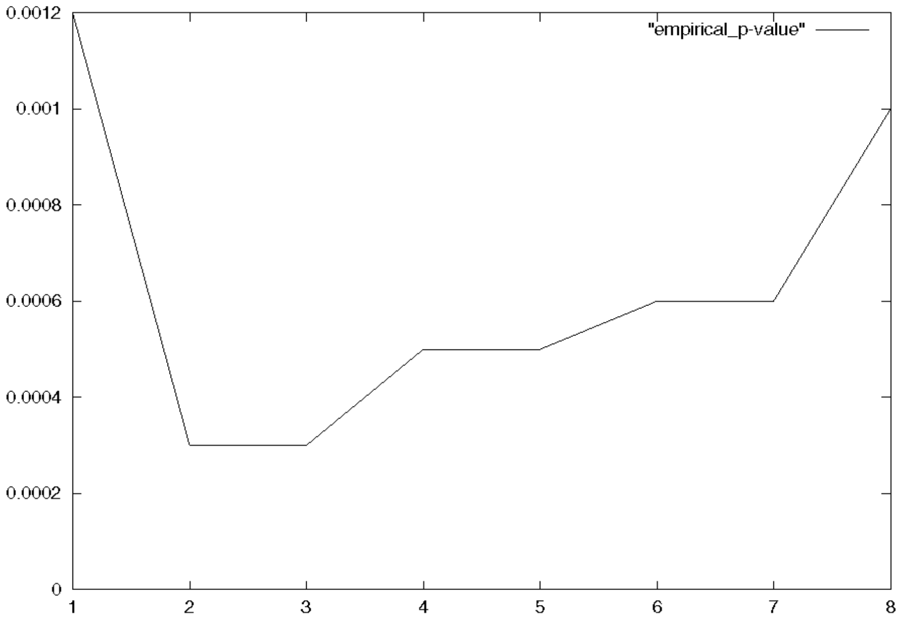


**Figure 9.** Empirical *p*-values based on 10,000 permutations. Markers 1-8 are *TH*, -2733A/C, -2221*Msp*I, VNTR class, -23/*Hph*I, +805/*Dra*III, +1127/*Pst*I, +1428*Fok*I, respectively. All family trios were included in the analysis.

**Protective haplotypes**

Protective haplotypes were searched with the same approach, by simply swapping the affected and control statuses in the data. In short, the haplotypes found were very similar to those originally published by Bennett et al (1995): the patterns corresponding to protective haplotype and very protective haplotype were strongly associated with being unaffected. The very protective haplotypes were especially pronounced when the *VNTR* class was used instead of alleles. The curves of markerwise scores were very similar to those shown in Figure 9.

**Table 5a**
Haplotype patterns in 382 family trios (VNTR alleles included). The patterns have been sorted according to their $\chi^2$ values. *N* is the total number of chromosomes in the sample; Freq is the number of chromosomes where the particular haplotype pattern has been observed, and Conf is the probability that a disease-associated chromosome carries the haplotype in question. All these values are output of HPM algorithm. Analyzed markers were TH, -2733A/C, -2221MspI, VNTR, -23/HphI, +805/DraIII, -1127/PstI, +1428FokI.

| $\chi^2$ | *N* | Freq | Conf | TH | -2733A/C | -2221MspI | VNTR | -23/HphI | +805/DraIII | -1127/PstI | +1428FokI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.333 | 1528 | 665 | 0.534 | - | - | - | - | 1 | - | - | - |
| 31.313 | 1528 | 661 | 0.535 | - | - | 1 | * | 1 | * | 1 | - |
| 31.313 | 1528 | 661 | 0.535 | - | - | 1 | * | 1 | - | - | - |
| 30.517 | 1528 | 664 | 0.534 | - | - | - | - | 1 | * | 1 | - |
| 30.509 | 1528 | 660 | 0.534 | - | - | 1 | * | 1 | * | 1 | 1 |
| 30.509 | 1528 | 660 | 0.534 | - | - | 1 | * | 1 | * | * | 1 |
| 30.352 | 1528 | 663 | 0.534 | - | - | - | - | 1 | * | 1 | 1 |
| 30.352 | 1528 | 663 | 0.534 | - | - | - | - | 1 | * | * | 1 |
| 23.942 | 1528 | 681 | 0.527 | - | - | 1 | * | * | * | 1 | - |
| 23.776 | 1528 | 680 | 0.527 | - | - | 1 | * | * | * | 1 | 1 |
| 23.667 | 1528 | 683 | 0.526 | - | - | 1 | - | - | - | - | - |
| 23.667 | 1528 | 683 | 0.526 | - | - | - | - | - | - | 1 | 1 |
| 23.396 | 1528 | 685 | 0.526 | - | - | - | - | - | - | - | 1 |
| 22.620 | 1528 | 684 | 0.525 | - | - | - | - | - | - | 1 | - |
| 19.819 | 1528 | 606 | 0.533 | - | - | 1 | * | 1 | 1 | 1 | - |
| 19.819 | 1528 | 606 | 0.533 | - | - | 1 | * | 1 | 1 | - | - |
| 19.269 | 1528 | 605 | 0.533 | - | - | 1 | * | 1 | 1 | 1 | 1 |
| 19.269 | 1528 | 605 | 0.533 | - | - | 1 | * | 1 | 1 | * | 1 |
| 19.047 | 1528 | 609 | 0.532 | - | - | - | - | 1 | 1 | 1 | - |
| 19.047 | 1528 | 609 | 0.532 | - | - | - | - | 1 | 1 | - | - |
| 18.982 | 1528 | 608 | 0.532 | - | - | - | - | 1 | 1 | 1 | 1 |
| 18.982 | 1528 | 608 | 0.532 | - | - | - | - | 1 | 1 | * | 1 |
| 16.547 | 1528 | 580 | 0.533 | - | 1 | * | * | 1 | 1 | 1 | 1 |
| 16.547 | 1528 | 580 | 0.533 | - | 1 | * | * | 1 | 1 | * | 1 |
| 16.458 | 1528 | 578 | 0.533 | - | 1 | 1 | * | 1 | 1 | 1 | 1 |
| 16.458 | 1528 | 578 | 0.533 | - | 1 | 1 | * | 1 | 1 | 1 | - |
| 16.458 | 1528 | 578 | 0.533 | - | 1 | 1 | * | 1 | 1 | * | 1 |
| 16.458 | 1528 | 578 | 0.533 | - | 1 | 1 | * | 1 | 1 | - | - |
| 16.156 | 1528 | 581 | 0.533 | - | 1 | * | * | 1 | * | 1 | 1 |
| 16.112 | 1528 | 580 | 0.533 | - | 1 | * | * | 1 | 1 | 1 | - |
| 16.112 | 1528 | 580 | 0.533 | - | 1 | * | * | 1 | 1 | - | - |
| 16.068 | 1528 | 579 | 0.533 | - | 1 | 1 | * | 1 | * | 1 | 1 |
| 16.068 | 1528 | 579 | 0.533 | - | 1 | 1 | * | 1 | * | 1 | - |
| 16.068 | 1528 | 579 | 0.533 | - | 1 | 1 | * | 1 | * | * | 1 |
| 16.068 | 1528 | 579 | 0.533 | - | 1 | 1 | * | 1 | - | - | - |

**Table 5b**

Haplotype patterns in 382 family trios (only VNTR class included). The patterns have been sorted according to their $\chi^2$ values. For VNTR class, number 1 was used for class I alleles. The statistics and the markers are as in the Table 5a.

| $\chi^2$ | *N* | Freq | Conf | TH | -2733A/C | -2221MspI | VNTR | -23/HphI | +805/DraIII | -1127/PstI | +1428FokI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.333 | 1528 | 665 | 0.534 | - | - | - | - | 1 | - | - | - |
| 31.313 | 1528 | 661 | 0.535 | - | - | 1 | * | 1 | * | 1 | - |
| 31.313 | 1528 | 661 | 0.535 | - | - | 1 | * | 1 | - | - | - |
| 30.832 | 1528 | 658 | 0.535 | - | - | 1 | 1 | * | * | 1 | - |
| 30.832 | 1528 | 658 | 0.535 | - | - | 1 | 1 | - | - | - | - |
| 30.517 | 1528 | 664 | 0.534 | - | - | - | - | 1 | * | 1 | - |
| 30.509 | 1528 | 660 | 0.534 | - | - | 1 | * | 1 | * | 1 | 1 |
| 30.509 | 1528 | 660 | 0.534 | - | - | 1 | * | 1 | * | * | 1 |
| 30.352 | 1528 | 663 | 0.534 | - | - | - | - | 1 | * | 1 | 1 |
| 30.352 | 1528 | 663 | 0.534 | - | - | - | - | 1 | * | * | 1 |
| 30.190 | 1528 | 662 | 0.534 | - | - | - | 1 | - | - | - | - |
| 30.043 | 1528 | 657 | 0.535 | - | - | 1 | 1 | * | * | 1 | 1 |
| 30.043 | 1528 | 657 | 0.535 | - | - | 1 | 1 | * | * | * | 1 |
| 30.031 | 1528 | 661 | 0.534 | - | - | - | 1 | * | * | 1 | - |
| 29.873 | 1528 | 660 | 0.534 | - | - | - | 1 | * | * | 1 | 1 |
| 29.873 | 1528 | 660 | 0.534 | - | - | - | 1 | * | * | * | 1 |
| 29.743 | 1528 | 655 | 0.535 | - | - | 1 | 1 | 1 | * | 1 | - |
| 29.743 | 1528 | 655 | 0.535 | - | - | 1 | 1 | 1 | - | - | - |
| 29.718 | 1528 | 659 | 0.534 | - | - | - | 1 | 1 | - | - | - |
| 28.975 | 1528 | 654 | 0.534 | - | - | 1 | 1 | 1 | * | 1 | 1 |
| 28.975 | 1528 | 654 | 0.534 | - | - | 1 | 1 | 1 | * | * | 1 |
| 28.940 | 1528 | 658 | 0.534 | - | - | - | 1 | 1 | * | 1 | - |
| 28.792 | 1528 | 657 | 0.534 | - | - | - | 1 | 1 | * | 1 | 1 |
| 28.792 | 1528 | 657 | 0.534 | - | - | - | 1 | 1 | * | * | 1 |
| 23.942 | 1528 | 681 | 0.527 | - | - | 1 | * | * | * | 1 | - |
| 23.776 | 1528 | 680 | 0.527 | - | - | 1 | * | * | * | 1 | 1 |
| 23.667 | 1528 | 683 | 0.526 | - | - | 1 | - | - | - | - | - |
| 23.667 | 1528 | 683 | 0.526 | - | - | - | - | - | - | 1 | 1 |
| 23.396 | 1528 | 685 | 0.526 | - | - | - | - | - | - | - | 1 |
| 22.620 | 1528 | 684 | 0.525 | - | - | - | - | - | - | 1 | - |
| 19.819 | 1528 | 606 | 0.533 | - | - | 1 | * | 1 | 1 | 1 | - |
| 19.819 | 1528 | 606 | 0.533 | - | - | 1 | * | 1 | 1 | - | - |
| 19.269 | 1528 | 605 | 0.533 | - | - | 1 | * | 1 | 1 | 1 | 1 |
| 19.269 | 1528 | 605 | 0.533 | - | - | 1 | * | 1 | 1 | * | 1 |

### Results in simulated quantitative trait data

The localization accuracy of the QHPM was analyzed with 100 replications of data in all simulation settings: both the easy and the difficult models, with all five quantitative traits Q1-Q5. The localization accuracy is illustrated by the cumulative percentage of data sets in which localization error was the same or less than the given level (on X axis in Figure 3a,c, Manuscript V). Clearly, the simulated data varied from practically impossible to very easy for this method.

Next, the QHPM approach was compared with basic-HPM. The quantitative traits were dichotomized and basic-HPM analysis carried out for these new binary variables, with the model M1 (difficult model), and all five quantitative traits. The dichotomization was made by rearranging the data with respect to the values of quantitative trait to be dichotomized, and then dividing the data into two equally sized parts. The half with lower values was labeled as controls, and that with higher values as cases. The basic-HPM was run with parameter settings $\chi^2$ -limit 6, maximum pattern length 7, and maximum number and maximum length of gaps was 1.

The analysis of dichotomized variables, compared with quantitative analysis by QHPM, led us to conclude that the probability of correct prediction is very similar with both methods when the genetic effect on the trait is sufficiently high (Figure 4, Manuscript V). However, when the genetic control of the trait decreases, the advantage of the genuinely quantitative analysis becomes clear, the probability of correct prediction is higher using QHPM than basic-HPM.

# 7. Summary and conclusions

Research on complex diseases only seems to be approaching the final goal, the prevention and cure of the diseases, very slowly. A wide spectrum of methodological approaches will evidently be needed, including epidemiological, clinical, classical and molecular genetical, and genetic epidemiological approaches. This thesis consists of the application of a range of approaches, from a meta-analysis of incidence trends to development and testing of new fine-mapping methods.

In the dissection of the genetic background of complex diseases it has become evident that understanding of not only biological mechanisms but also the population history, or population genetics, is crucial: the amount of allelic heterogeneity, the proportion of "phenocopies", and differences between populations are, in many respects, a result of population processes. Mutation, drift, selection, and isolation shape the genetic constitution of a population and thus the disease genetics. By understanding the behavior of allele frequencies and haplotypes as a function of generations, it is easier to formulate the expectations of what one can find in genetic marker data. What are the boundaries of statistical methods for finding the genes with smaller and smaller effects?

Given the current stage of human genome mapping project, the almost complete sequence of human genome, and the projects starting the evaluation of variability in human populations, it is important to predict the needs of genetic researchers in a few years. It can be anticipated that genotyping and analysis of candidate polymorphisms, probably with SNP markers, will prevail. Mass genotyping with high-throughput machinery, combined with efficient computational methods, is a direction in which research seems to be going. Currently no one knows the population genetic structure of inter-individual variability at the sequence level. The finding of a block structure in Caucasian populations gives interesting insights, but is still far from being conclusive for all genes – and non-Caucasian populations.

**HPM**

In the publication IV we were able to show that the algorithmic solution to the fine mapping problem works well even with very small proportions of a mutant allele in the affected sample. Thus, we believe that new hope for finding small effect loci can be found in applying wholly new ways of thinking to the problem. In manuscript V, an extension to quantitative traits was presented, which was shown to give good results with simulated data. For the applicability of the methods themselves, it is very important to note that in spite of them having been tested with simulated isolated population data, they seem to work as well with data from large, even mixed ethnicities (UK, publication IV). As fine mapping will take place over a smaller scale in future, there will be less need for only using data from isolated populations; the structure of haplotypes at a very small scale is dependent on ancient events. Thus, methods looking for shared haplotype structure should function at that scale independently of the population history.

Regardless of the yet unsolved statistical problems, association studies in general have recently been proposed as a powerful approach for detecting the several weak genetic effects which underlie susceptibility to complex diseases (Risch and Merikangas

1996, Lander 1996). Improved techniques for high-throughput identification and genotyping of polymorphisms, such as SNPs, offer the possibility of using high numbers of markers in genome screening and candidate gene scanning in the near future. The density of such maps that is sufficient, given the population history of modern human populations, is again strongly dependent on the population history (Reich et al 2001, Kruglyak et al 1999, Long and Langley 1999).

We believe that the approach adopted here may be extendable to the analysis of some of these complex characteristics. As a non-parametric approach the method has unique properties compared to other LD methods. For example, it is conceptually rather straightforward to extend the algorithm to find several genes simultaneously.

The method was modified to find two genes simultaneously and tested using simulated data with two interacting genes and phenocopies. Both DS loci were reliably localized (data not shown). The problem of multiple founder haplotypes (allelic heterogeneity at a DS locus) is largely bypassed by simply counting separate patterns together in the marker scores. The handling of marker inconsistencies, such as genotyping errors and mutations, might be further improved by allowing approximate pattern matching in the haplotype pattern discovery step. The proposed method also scales well to large data sets of biallelic and multiallelic markers.

Finally, data mining methods might be used as a pre-processing step for more detailed explicit statistical analysis. For example, the haplotype patterns might be used as a sample space for the reconstruction of ancestral haplotypes in DS chromosomes. The location of the DS gene, age of the mutation, and share haplotypes therein could be estimated as the model parameters.

**Increasing incidence**

Even today, the factors which have been responsible for the increasing the incidence of Type 1 diabetes, have not been identified. The Publications I and III comprise an effort to understand the development of the incidence and to browse some of the hypotheses used to explain the phenomenon.

Genetic factors have been shown to be important in the liability to Type 1 diabetes by epidemiological methods by Kaprio et al (1992), Kyvik et al (1995), and Cordell and Todd (1997). Although it is possible that the part of the population genetically predisposed for Type 1 diabetes is in fact increasing, the evidence at hand shows that this increase is modest and alone not a sufficient cause for the observed increase in incidence. Changes in the genetic composition of the human populations are usually slow. In the analysis of incidence trends in this Thesis, even the longest study period only covered 30 years, which is approximately the time period equal to one generation. It is very unlikely that a 3 to 10-fold increase in incidence during such a short time could be result of corresponding increase in the proportion of individuals with genetic susceptibility to Type 1 diabetes. Instead, the penetrances of the susceptibility genes might be changing. The penetrance is likely to be determined by an interaction between several susceptibility genes and unknown environmental factors (Todd 1997).

During recent years much attention has been played to the identification and possible control of environmental factors which may initiate or trigger the process leading to Type 1 diabetes. Although some studies suggest associations between environmental factors such as diet and viral infections with the risk of Type 1 diabetes

(Kostraba et al 1993, Dahlquist et al 1990, Virtanen et al 1994, Hyöty et al 1988, 1995, Hiltunen et al 1995), their causative role in the aetiology of Type 1 diabetes has not been shown. It is also difficult to show that any of these environmental factors has changed in such a way that a continuous global increase in the incidence of Type 1 diabetes would be easily explained.

In addition to Type 1 diabetes, the incidence and prevalence of several other autoimmune diseases are known to be increasing. Already over ten years ago, a “hygiene” hypothesis (Strachan 1989) was put forward: the microbial environment that children in modern developed countries encounter does not include many of those pathogenic organisms which have co-existed with human species for very long times. It is possible that our immune system has adapted to encountering these antigens in a certain developmental stage, and as that does not take place anymore the immune system gets “misled” into attacking the body’s own molecules and cells. However, it is still an open question as to why the attack is directed against specific cell types or not in all individuals.

# Acknowledgements

This study was carried out at the Division of Biometry, Rolf Nevanlinna Institute, during 1997-2001, of which part time in Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute, and the final stages in Finnish Genome Center in 2000-2001. I want to thank the Directors of these institutions, Jukka Sarvas at RNI, Jussi Huttunen at NPHI, and Juha Kere in FGC for providing excellent research facilities for carrying out the work.

Special thanks naturally go to my supervisors; for Elja Arjas, who has kept an eye over the work through the whole time, and given realism and stability to the plans, which experienced changes during the years. The biometrics group with several genetics-oriented young scientists, gathered up by Elja, was a right “home” for a beginning researcher in the new field of genetic epidemiology. Thank you also for very careful checking of all the manuscripts! No mistake could avoid your eyes, and you always taught accuracy for our impatient minds. Special thanks for Juha Kere, who has constantly been inspiring, looked on the bright side of all things, and believed in new ideas presented by students! Also, thank you for the argumentations we have had – they have given me new points of view. Hope to you, too! And, for passing on the great interest in complex diseases, and also for interest towards epidemiology, which was a new field to me, I warmly thank Jaakko Tuomilehto, who supervised the first three publications included in this Thesis in collaboration with Elja.

Professor Juni Palmgren and docent Jorma Ilonen are thanked for reviewing the manuscript of this thesis and for their constructive criticism. Juni, whom I have had pleasure to know from years before, has also been a real role model for a young woman in science, for both her career and attitude!

I wish to express my sincere gratitude to all people with whom I have in practice worked during these years; in chronological order; Saara Väänänen for the first years in NPHI, for her positive way of thinking, wonderful personality, and friendship; Marjatta Karvonen for her guidance to the world of epidemiology. Thanks for Valma Hyttinen and Elena Moltschanova for your company and scientific (as well as many other) opinions! Special thanks for Janne Pitkäniemi, you were practically a tutor of my work during the year and half when you were also working in RNI. Thanks for your cheerfulness and outspokenness! Also, great thanks for Mikko Sillanpää and Pekka Uimari, the “senior” computational geneticists in RNI; the “genetics” gang we formed had many intriguing conversations over coffee table! Thanks for the whole biometrics group in RNI; the atmosphere there was just great!!! That I’ll remember for years to come. I have also had the privilege and pleasure to work with the computer science group in RNI: Hannu “Haba” Toivonen, assigned Professor in the beginning of year 2002 (making a fantastic positive exception to the traditional imago of a professor!), thanks for all support, punctuality, and excellent partying company(!) … and his students Petteri Sevon, always so relaxed and open-minded himself, and Kari Vasko, who went on to work at CSC. Thanks for Vesa Ollikainen, with whom I have worked for a couple of years in FGC. Thank you for your friendship. I also want to thank “the new generation” of statistical geneticists Mikko K, Hanni, Riika, and Sampo (who considers himself as a botanist, but we know what he really is…). Keep the flag up! Thanks for the whole staff of RNI, practicalities have always been very well taken care of: Pirjo, Tarja, Matti Taskinen. Thanks for people in FGC, Elina, Mika, Päivi T, Anna, Päivi L, Michael, Timo, Riitta, and many others for nice company for the last year and a half. Thanks for Tarja Laitinen and Heikki Mannila, senior scientists whom I had pleasure to work with sometimes. Thanks also for Cambridge people, Mathias, who helped with Publication IV, Frank for the "inside information", Sarah and others: you showed me that there is (really interesting) life outside Finland :-). I would have loved to stay a while!

My warm thanks to my mother Arja, Saku, and the rest of the family. I also want to remember my grandmother, who has already passed away, but gave a model of courageous attitude of going through any difficulties life may present, whatever it takes. I also wish to thank all my friends, for their support during these years: Tuija, for your warm-hearted support from the early years on and the adventures we had, that all gave me a lot of strength and inspiration to go on; Olli, always high-spirited and inventive; Vesa, Valma,Vipe & Kepa, Mikko, and many others.

Finally, the financial support provided by ComBi graduate school, Academy of Finland, Emil Aaltonen Foundation, Yliopiston kansleri, and Suomalainen Konkordia-liitto are gratefully acknowledged.

Helsinki, January 2002

Päivi Onkamo

# References

Abecasis GR, Cardon LR, Cookson WO (2000) A General Test of Association for Quantitative Traits in Nuclear Families. Am J Hum Genet 66:279-292

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WOC (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191-197

Agrawal R, Imielinski T, Swami A (1993) Mining Association Rules between Sets of Items in Large Databases. In: Buneman P, Jajodia S (eds) Proceedings of 1993 ACM SIGMOD conference on management of data. Association for Computing Machinery, Washington, DC, pp 207-216

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. AAAI Press, Menlo Park, CA, pp. 307-328

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211

Allison DB (1997) Transmission disequilibrium tests for quantitative traits. Am J Hum Genet 60**:**676-690

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582-589

Bain SC, Prins JB, Hearne CM, Rodrigues NR, Rowe BR, Pritchard LE, Rithchie RJ, Hall JR, Undlien DE, Ronningen KS (1992) Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. Nat Genet 2:212-215

Bain SC, Todd J, Barnett A (1990) The British Diabetic Association – Warren repository. Autoimmunity 7:83-85

Bell GI, Horita S, Karam JH (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes 33:176-183

Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, Nerup J, Bouzekri N, Cambon-Thomsen A, Rønningen KS, Barnett AH, Bain SC, Todd JA (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284-292

Bingley PJ, Gale EA (1989) Rising incidence of IDDM in Europe. Diabetes Care 12: 289-295

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalynaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding region of human genes. Nat Genet 22:231-238

Cavalli-Sforza LL, Piazza A (1993) Human genomic diversity in Europe: a summary of recent research and prospects for the future. European Journal of Human Genetics. 1:3-18

Choubnikova J, Shubnikof E, Kalashnikova L (1996) The increase of type 1 diabetes incidence among children in Novosibirsk city (Abstract). 32nd Annual Meeting of the European Association for the Study of Diabetes, Vienna (Abstract)

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, Farmer J, Williams SR, Cox NJ, Bell GI, Risch N, Spielman RS (1998) A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. Nature Genetics 19:292-296

Cordell HJ, Todd JA (1995) Multifactorial inheritance in type 1 diabetes. Trends Genet 11: 499-504

Cox N, Wapelhorst B, Morrison VA, Johnson L, Pinchuk L, Spielman RS, Todd JA, Concannon P (2001) Seven regions of the genome show evidence of linkage to Type 1 diabetes in a consensus analysis of 767 multiplex families. Am J Hum Genet 69:820-830

Dahlquist GG, Blom LG, Persson LA, Sandström AI, Wall SG (1990) Dietary factors and the risk of developing insulin dependent diabetes in childhood. Br Med J 300:1302-1306

Dahlquist G, Mustonen L (2000) Analysis of 20 years of prospective registration of childhood onset diabetes time trends and birth cohort effects. Swedish Childhood Diabetes Study Group. Acta Paediatrica 89:1231-1237

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001). High-resolution haplotype structure in the human genome. Nat Genet 29:229-232

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC, Todd JA (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130-135

Davis S, Weeks DE (1997) Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. Am J Hum Genet 61:1431-1444

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1-16

De Vries RRP, Meera Khan P, Bernini LF, van Loghem E, van Rood JJ (1979) Genetic control of survival in epidemics. Journal of Immunogenetics 6:271-287

De Vries RRP, van Rood JJ (1979) HLA and infectious diseases. Arch Dermatol Res 264:89-95

Eaves IA, Bennett ST, Forster P, Ferber KM, Ehrmann D, Wilson AJ, Bhattacharyya S, Ziegler AG, Brinkmann B, Todd JA (1999) Transmission ratio distortion at the INS-IGF2 VNTR. Nat Genet 22:324-325

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nature Genetics 25:320-323

European Consortium for IDDM genome studies (2001) A genomewide scan for Type 1 –diabetes susceptibility in Scandinavian families: identification of new loci with evidence of interactions. Am J Hum Genet 69:1301-1313

Field LL, Tobias R, Magnus, T (1994) A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin-dependent diabetes mellitus. Nature Genet 8:189-194

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259-267

Green A, Andersen PK, Svendsen AJ, Mortensen K (1992b) Increasing incidence of early onset type 1 (insulin-dependent) diabetes mellitus: a study of Danish male birth cohorts. Diabetologia 35: 178-182

Green A, Gale EA, Patterson CC for the EURODIAB ACE study (1992a) Incidence of childhood-onset insulin-dependent diabetes mellitus: the EURODIAB ACE study. Lancet 339: 905-909

Gu C, Province MA, Rao DC (2001) Meta-analysis for model-free methods. Advances in Genetics 42:255-272

Guo S-W (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered 227:1-14

Göring HHH, Terwilliger JD (2000) Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet 66:1310-1327

Hartl DL, Clark AG (1989) Principles of population genetics. Sinauer Associates, Sunderland, Massachusetts.

Hauser ER, Boehnke M, Guo WW, Risch N (1996) Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. Genet Epidemiol 13:117-137

Heath SC (1997) Markov chain segregation and linkage analysis for oligogenic models. Am J Human Genet 61:748-760

Hedrick PW (1998) Balancing selection and MHC. Genetica 104:207-214

Helmuth, L (2001) Genome research: Map of the Human Genome 3.0. Science 293:583-585

Herr M, Dudbridge F, Zavattari P, Cucca F, Guja C, March R, Campbell R, Barnett A, Bain S, Todd JA. Koeleman BP (2000) Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21. Human Molecular Genetics. 9:1291-1301

Hiltunen M, Hyöty H, Karjalainen J, Leinikki P, Knip M, Lounamaa R, Åkerblom HK (1995) Serological evaluation of the role of cytomegalovirus in the pathogenesis of IDDM: a prospective study. Diabetologia 38: 705-710

Hodgkinson AD, Millward BA, Demaine AG (2000). The HLA-E locus is associated with age at onset and susceptibility to type 1 diabetes mellitus. Human Immunology 61:290-295

Hyöty H, Leinikki P, Reunanen A, Ilonen J, Surcel HM, Rilva A, Kaar ML, Huupponen T, Hakulinen A, Makela AL (1988) Mumps infections in the etiology of type 1 (insulin-dependent) diabetes. Diabetes Res 9: 111-116

Hyöty H, Hiltunen M, Knip M, Laakkonen M, Vahasalo P, Karjalainen J, Koskela P, Roivainen M, Leinikki P, Hovi T (1995) A prospective study of the role of coxsackie B and other enterovirus infections in the pathogenesis of IDDM. Diabetes 44: 652-657

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299-314

International Human Genome Sequencing Consortium (2001) Nature 409:860-921

Jaksic J, Matic I, Stojnic E, Juros A, Pelajic A (1996) Incidence of insulin dependent diabetes mellitus in children aged 0-19 in the Sibenic area. Diabetol Croat: 29-33

Jin K, Speed TP, Klitz W, Thompson G (1994) Testing for Segregation Distortion in the HLA Complex. Biometrics 50:1189-1198

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979-988

Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, Stengard J, Kesaniemi YA (1992) Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35: 1060-1067

Karvonen M, Jantti V, Muntoni S, Stabilini M, Stabilini L, Muntoni S, Tuomilehto J (1998) Comparison of the seasonal pattern in the clinical onset of IDDM in Finland and Sardinia. Diabetes Care 21:1101-1109

Karvonen M, Pitkäniemi M, Pitkäniemi J, Kohtamäki K, Tajima N, Tuomilehto J for the World Health Organization DIAMOND Project Group (1997) Sex difference in the incidence of insulin-dependent diabetes mellitus: an analysis of the recent epidemiological data. Diabetes Metab Rev 13: 275-291

Karvonen M, Pitkäniemi J, Tuomilehto J (1999) The onset age of type 1 diabetes in Finnish children has become younger. The Finnish Childhood Diabetes Registry Group. Diabetes Care 22:1066-1070

Karvonen M, Tuomilehto J, Libman I, LaPorte R for the World Health Organization DIAMOND Project Group (1993) A Review of the recent epidemiological data on the worldwide incidence of type 1 (insulin-dependent) diabetes mellitus. Diabetologia 36: 883-892

Kere J (2001) Human population genetics: Lessons from Finland. Annu Rev Genomics Hum Genet 2:103-128

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of Genetic Epidemiology. New York, Oxford; Oxford University Press.

Klitz W, Sing KL, Neugebauer M, Baur M, Ekkehard DA, Thomson G (1986) A Comprehensive Search for Segregation Distortion in HLA. Hum Immunol 18:163-180

Kockum I, Wassmuth R, Holmberg E, Michelsen B, Lernmark Å (1994) Inheritance of MHC class II genes in IDDM studied in population-based affected and control families Diabetologia 37:1105-1112

Koivisto M, Mannila H (2001) Offspring risk and sibling risk for multilocus traits. Human Heredity. 51:209-216

Komulainen J, Kulmala P, Savola K, Lounamaa R, Ilonen J, Reijonen H, Knip M, Åkerblom HK, and the Childhood Diabetes in Finland (DiMe) Study Group (1999) Clinical, autoimmune, and genetic characteristics of very young children with Type 1 diabetes. Diabetes Care 22:1950-1955

Kostraba JN, Cruickshanks KJ, Lawler-Heavner J, Jobim LF, Rewers MJ, Gay EC, Chase HP, Klingensmith G, Hamman RF (1993) Early exposure to cow's milk and solid foods in infancy, genetic predisposition, and risk of IDDM. Diabetes 42: 288-295

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and non-parametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-144

Kulmala P, Savola K, Reijonen H, Veijola R, Vähäsalo P, Karjalainen J, Tuomilehto-Wolf E, Ilonen J, Tuomilehto J, Åkerblom HK, Knip M, and the Childhood Diabetes in Finland Study Group (2000). Genetic markers, humoral autoimmunity, and prediction of Type 1 diabetes in siblings of affected children. Diabetes 49:48-58

Kyvik KO, Green A, Beck-Nielsen H (1995) Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. Br Med J 311: 913-917

Lahermo P, Savontaus ML, Sistonen P, Beres J, de Kniff P, Aula P, Sajantila A (1999) Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. Eur J Hum Genet 7:447-458

Lander ES (1996) The new genomics: Global views of biology. Science 274:536-539

LaPorte RE, Tajima N, Åkerblom HK, Berlin N, Brosseau J, Christy M, Drash AL, Fishbein H, Green A, Hamman R (1985) Geographic differences in the risk of insulin-dependent diabetes mellitus: the importance of registries. Diabetes Care 8 [Suppl 1]: 101-107

Lazzeroni LC (1998) Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am J Hum Genet 62:159-170

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research. 9:720-31

Markow T, Hedrick PW, Zuerlein K, Danilovs J, Martin J, Vyvial T, Armstrong C (1993) HLA polymorphims in Havasupai: evidence for balancing selection. Am J Hum Genet 53:943-952

Martin-Villa JM, Vicario J, Martinez -Lazo J, Serrano-Rios M, Lledo G, Damiano A, Hawkins F, Regueiro JR, Arnaiz-Villena A (1990) Lack of preferential transmission of diabetic HLA alleles by healthy parents to offspring in Spanish diabetic families. J Clin Endocrinol Metab 70:346-349

McKeigue (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 60:188-196

McKeigue (2000) Multipoint admixture mapping. Genetic Epidemiology 19:464-467

McKusick VA (1998) Mendelian inheritance in Man. Baltimore: Johns Hopkins Univ. Press.12th ed. http://www.ncbi.nlm.nih.gov/omim.

McPeek MS, Strahs A (1998) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858-875

Mein CA, Esposito L, Dunn MG, Johnson GCL, Timms AE, Goy JV, Smith AN, Sebag-Montefiore L, Merriman ME, Wilson AJ, Pritchard LE, Cucca F, Barnett AH, Bain SC, Todd JA (1998) A search for Type 1 diabetes susceptibility genes in families from the United Kingdom. Nature Genetics 19:297-300

Neel JV (1962) Diabetes mellitus: “a thrifty” genotype rendered detrimental by “progress”. Am J Hum Genet 14:353-362

Nejentsev S, Gombos Z, Laine A-P, Veijola R, Knip M, Simell O, Vaarala O, Åkerblom HK, Ilonen J (2000) Non-class II HLA gene associated with type 1 diabetes maps to the 240-kb region near HLA-B. Diabetes 49:2217-2221

Nejentsev S, Reijonen H, Adojaan B, Kovalchuk L, Sochnevs A, Schwartz EI, Åkerblom HK, Ilonen J (1997) The effect of HLA-B allele on the IDDM risk defined by DRB1*04 subtypes and DQB1*0302. Diabetes 46:1888-1892

Nerup J, Platz P, Andersen OO, Christy M, Lyngsøe J, Poulsen JE, Ryder LP, Thomse M, Nielsen LS, Svejgaard A (1974) HL-A antigens and diabetes mellitus. Lancet 2:864-866

Nevanlinna HR (1972) The Finnish population structure. A genetic and genealogical study. Hereditas 71:195-236

Norio R (1966) Heredity in the congenital nephrotic syndrome. Academic dissertation, the University of Helsinki.

Norio R (2000) Suomi-neidon geenit. Helsinki: Kustannusosakeyhtiö Otava.

Norio R, Nevanlinna HR, Perheentupa J (1973) Hereditary diseases in Finland. Annals of Clinical Research 5:109-141

Nunez MG (1987) A model for the early settlement of Finland. Fennosc Archael 4:3-18

Ollikainen V (2002) Simulation techniques for disease gene localization in isolated populations. Academic dissertation.

Pile KD (1999) Broadsheet number 51: HLA and disease associations. Pathology 31:202-212

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1-14

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220-228

Pugliese A, Zeller M, Fernandez A Jr, Zalcberg LJ, Bartlett RJ, Ricordi C, Pietropaolo M, Eisenbarth GS, Bennett ST, Patel DD (1997) The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the *INS* VNTR-*IDDM2* susceptibility locus for type 1 diabetes. Nat Genet 15:293-297

Rannala B, Reeve JP (2001) High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet 69:159-178

Ranta J, Penttinen A (2000) Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. Statistics in Medicine 19:2345-2359

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342-350

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends in genetics 17:502-510

Rewers M, LaPorte RE, King H, Tuomilehto J (1988) Trends in the prevalence and incidence of diabetes: insulin-dependent diabetes mellitus in childhood. World Health Stat Q 41: 179-189

Risch N (1987) Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 40:1-14

Risch N, Ghosh S, Todd J (1993) Statistical evaluation of multiple locus linkage data in experimental species and relevance to human studies: application to murine and human IDDM. Am J Hum Genet 53:702-714

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Rotter JI, Landaw EM (1984) Measuring the genetic contribution of a single locus to a multilocus disease. Clin Genet 26:529-542

Ruwaard D, Gijsen R, Bartelds IM, Hirasing RA, Verkleij H, Kromhaut D (1996) Is the incidence of diabetes increasing in all age-groups in The Netherlands? Results of the second study in the Dutch Sentinel Practice Network. Diabetes Care 19: 214-218

Rytkönen M, Ranta J, Tuomilehto J, Karvonen M (2001) Bayesian analysis of geographical variation in the incidence of type 1 diabetes in Finland. Diabetologia 44:Suppl1, in press.

Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466-477

Schoenle EJ, Molinari L, Bagot M, Semadeni S, Wiesendanger M (1994) Epidemiology of IDDM in Switzerland. Increasing incidence rate and rural-urban differences in Swiss men born 1948-1972. Diabetes Care 17: 955-960

Service SK, Lang DW, Freimer NB, Sandkuijl LA (1999) Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728-1738

Shaltout AA, Qeabazard MA, Abdella NA, LaPorte RE, al Arouj M, Ben Nekhi A, Moussa MA, al Khawari MA (1995) High incidence of childhood-onset IDDM in Kuwait. Diabetes Care 18: 923-927

Singal DP, Blajchman MA (1973) Histocompatibility (HLA) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. Diabetes 22:429-432

Slager SL, Huang J, Vieland VJ (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genetic Epidemiol 18:143-156

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Strachan DP (1989) Hay fever, hygiene and household size. BMJ 299:1259-1260

Suarez BK, Hampe CL, Van Eerdewegh P (1994) Problems of replicating linkage claims in psychiatry. In: Gershon ES, Cloninger CR (eds) Genetic approaches to mental disorders. American Psychiatric Press, Washington, DC, pp 23-46

Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nature Genetics 25:324-328

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777-787

The Eurodiab ACE Study Group and The Eurodiab ACE Substudy 2 Study Group (1998) Familial risk of type 1 diabetes in European children. Diabetologia 41:1151-1156

Thomson G, Robinson WP, Kuhner MK, Joe S, Klitz W (1989) HLA, insulin gene, and Gm associations with IDDM. Genet Epidemiol 6:155-160

Todd JA (1997) Genetics of type 1 diabetes. Pathol Biol (Paris) 45: 219-227

Tuomilehto J, Lounamaa R, Tuomilehto-Wolf E, Reunanen A, Virtala E, Kaprio EA, Åkerblom HK and the Childhood Diabetes in Finland (DiMe) Study Group (1992) Epidemiology of childhood diabetes mellitus in Finland - background of a nationwide study of Type 1 (insulin-dependent) diabetes mellitus. Diabetologia 35:70-76

Tuomilehto J, Rewers M, Reunanen A, Lounamaa P, Lounamaa R, Tuomilehto-Wolf E, Akerblom HK (1991) Increasing trend in Type 1 (insulin-dependent) diabetes mellitus in childhood in Finland. Analysis of age, calendar time and birth cohort effects during 1965 to 1984. Diabetologia 34:282-287

Uimari P, Sillanpää M (2001) Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genetic Epidemiology 21:224-242

Vadheim CM, Rotter JI, MacLaren NK, Riley WJ, Anderson E (1986) Preferential Transmission of Diabetic Alleles Within the HLA Gene Complex. N Engl J Med 315:1314-1318

Valdes AM, Thomson G, Erlich HA, Noble JA (1999) Association between type 1 diabetes age of onset and HLA among sibling pairs. Diabetes 48:1658-1661

Venter JC, Adams MD, Myers EW et al (2001) The Human Genome. Science 291:1304-1351

Virtanen SM, Jaakkola L, Räsänen L, Ylonen K, Aro A, Lounamaa R, Akerblom HK, Tuomilehto J (1994) Nitrate and nitrite intake and the risk for type 1 diabetes in Finnish children. Diabet Med 11: 656-662

Warram JH, Krolewski AS, Gottlieb MS, Kahn CR (1984) Differencies in risk of insulin-dependent diabetes in offspring of diabetic mothers and diabetic fathers. N Engl J Med 311:149-152

WHO DIAMOND Project Group on Epidemics (1992) Childhood diabetes, epidemics, and epidemiology: an approach for controlling diabetes. Am J Epidemiol 135: 803-816

Wijsman EM, Almasy L, Amos CI, Borecki I, Falk CT, King TM, Martinez MM, Meyers D, Neuman R, Olson JM, Rich S, Spence MA, Thomas DC, Vieland VJ, Witte JS, MacCluer JW (2001) Analysis of complex genetic traits: Applications to asthma and simulated data. In Genetic Epidemiology, Volume 21 (Suppl 1), pgs. S1-S853

Zhang S, Zhao H (2001) Quantitative similarity-based association tests using population samples. Am J Hum Genet 69:601-614

Zwick ME, Cutler DJ, Chakravarti A (2000) Patterns of genetic variation in Mendelian and complex traits. Annu Rev Genomics Hum Genet 1:387-407

# Original publications