Department of Public Health
University of Helsinki, Finland

and

Department of Mathematics and Statistics
University of Helsinki, Finland

Statistical analysis of the associations of hereditary factors with the risk of Type 1 diabetes and Diabetic Nephropathy based on the familial data collected through ascertaiment from population-based registers

Janne Pitkäniemi

A C A D E M I C D I S S E R T A T I O N

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in the Arppeanum, Snellmaninkatu 3, on 16th of May 2008 at 12 noon.

Helsinki, 2008

Supervisors	Professor Elja Arjas Department of Mathematics and Statistics University of Helsinki
	Professor Timo Hakulinen Finnish Cancer Registry
	Professor Jaakko Tuomilehto Department of Public Health University of Helsinki
Reviewers	Professor Esa Läärä Department of Mathematical Sciences University of Oulu
	Professor Suvi Virtanen Department of Health Science University of Tampere
Official opponent	Associate professor Soumitra Ghosh Wisconsin Medical College

ISSN 0355-7979
ISBN 978-952-10-1379-9
ISBN 978-952-10-1380-5 (pdf)

Helsinki 2008
Helsinki University Print

List of original publications

This thesis consists of the following five papers referred in the text by Roman numerals (I-V):

Paper I : 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. The DiMe Study Group. Pitkäniemi J, Onkamo P, Arjas E, Tuomilehto-Wolf E, Tuomilehto J. Hum Hered. 2000; 50(5):308-17.

Paper II: Increasing incidence of Type 1 diabetes--role for genes? Pitkäniemi J, Onkamo P, Tuomilehto J, Arjas E. BMC Genet. 2004 2; 5:5.

Paper III: Class I and II HLA genes are associated with susceptibility and age at onset in Finnish families with Type 1 diabetes. Pitkäniemi J, Hakulinen T, Nasanen J, Tuomilehto-Wolf E, Tuomilehto J; DiMe Study Group. Hum Hered. 2004;57(2):69-79.

Paper IV: Full likelihood based genetic risk analysis for sibships ascertained from a population-wide registry of incident cases. Pitkäniemi J, Varvio S-L, Corander J, Lehti N, Partanen J, Tuomilehto-Wolf E, Tuomilehto J, Thomas A, Arjas E and the DiMe Study Group (submitted).

Paper V: Genetic random effects model for family data with long-term survivors: analysis of diabetic nephropathy in Type 1 diabetes. Pitkäniemi J, Moltchanova E, Haapala L, Tuomilehto J. Hakulinen T. Genet Epid. 2007; 31(7):697-708.

Publications I and II are also found in Dr. Päivi Onkamo's thesis.

Abbreviations

ACE-LTS	Additive genetic, common environment and different environment LTS model
CI	Confidence interval (classical), credible interval (Bayesian)
DiMe	The Childhood Diabetes in Finland study
DERI	The Diabetes Epidemiology Research International study
DN	Diabetic Nephropathy
DNA	Deoxyribonucleic acid
$f(\,)$	Probability density function
$F(\,)$, c.d.f	Cumulative distribution function
HapMap	A project that constructs a SNP based haplotype map of human chromosomes
HLA	Human Leukocyte Antigen; HLA-A, HLA-B and HLA-DRB1 genes
IDDM1-IDDM18	Names of the genes linked/associated with T1D
LD	Linkage disequilibrium
LTS	Long-term survivor, non-susceptible, immune
MCMC	Markov Chain Monte Carlo
MHC	Major Histocompatibility Complex
NPHI	National Public Health Institute
$p(\,)$	Probability function
$S(\,)$	Survival function
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes

Abstract

In genetic epidemiology, population-based disease registries are commonly used to collect genotype or other risk factor information concerning affected subjects and their relatives. This work presents two new approaches for the statistical inference of ascertained data: a conditional and full likelihood approaches for the disease with variable age at onset phenotype using familial data obtained from population-based registry of incident cases. The aim is to obtain statistically reliable estimates of the general population parameters. The statistical analysis of familial data with variable age at onset becomes more complicated when some of the study subjects are non-susceptible, that is to say these subjects never get the disease. A statistical model for a variable age at onset with long-term survivors is proposed for studies of familial aggregation, using latent variable approach, as well as for prospective studies of genetic association studies with candidate genes. In addition, we explore the possibility of a genetic explanation of the observed increase in the incidence of Type 1 diabetes (T1D) in Finland in recent decades and the hypothesis of non-Mendelian transmission of T1D associated genes. Both classical and Bayesian statistical inference were used in the modelling and estimation. Despite the fact that this work contains five studies with different statistical models, they all concern data obtained from nationwide registries of T1D and genetics of T1D. In the analyses of T1D data, non-Mendelian transmission of T1D susceptibility alleles was not observed. In addition, non-Mendelian transmission of T1D susceptibility genes did not make a plausible explanation for the increase in T1D incidence in Finland. Instead, the Human Leucocyte Antigen associations with T1D were confirmed in the population-based analysis, which combines T1D registry information, reference sample of healthy subjects and birth cohort information of the Finnish population. Finally, a substantial familial variation in the susceptibility of T1D nephropathy was observed. The presented studies show the benefits of sophisticated statistical modelling to explore risk factors for complex diseases.

1. Introduction

All the five papers presented in this thesis introduce new ideas of statistical inference in genetic-epidemiological family studies of complex diseases with variable age at onset. In papers I-III classical statistical inference is applied and in papers IV-V Bayesian inference is preferred to obtain estimates of the model parameters and their variation. A common theme in papers I-IV is the estimation of population parameters of interest when the observed familial data have been selected – ascertained – from a population-based disease registry of incident cases of T1D. Phenotype-based selection of families, ascertainment, is a common way of collecting data in genetic studies of rare human diseases. Although different statistical models are used in each of the five papers: multinomial model (paper I), Poisson model (paper II), fixed effect long-term survivor model (paper III), discrete time hazard model (paper IV) and random effects long-term survivor model (paper V) all of these models are applied to the genetics of T1D or Diabetic Nephropathy.

All the data used in the analysis comes from two nationwide population-based registries of T1D in Finland. None of the currently available statistical methods for genetic analyses of variable age at onset phenotype is directly applicable to the problems of population-based family data: how to model the variable age at onset phenotype (trait) and account for population-based ascertainment of variable age at onset disease in the statistical analysis. Statistical analyses of non-randomly collected family data in genetics dates back to the founders of the classical statistical inference (1) and has gained increasing interest in recent years, because of the rapid improvement in computational capabilities needed by complex statistical models (2,3).

Survival analysis of times to event of interest has been utilized many decades in epidemiological studies and has been applied increasingly to the genetic studies of complex diseases with variable age at onset (4). Many common complex diseases, such as coronary heart disease and different forms of diabetes, share two important characteristics: disease can be diagnosed at any age and not all subjects become affected during their lifespan. It thus seems natural to assume that subjects can be classified into those who are susceptible to the

disease and those who are not – called non-susceptible (5). In papers III and V the framework of long-term survivors (6) is modified for genetic studies of familial aggregation and association.

Problems of statistical inference in all of the five papers in this dissertation originated from the familial data collected in studies of the genetics of T1D or its complications. The detailed pathogenesis of T1D remains unsolved, although the primary role of Human Leucocyte Antigen region (HLA), located in chromosome 6q, in T1D risk has been well established since the early 1970's (7) and later confirmed by genome-wide studies (8,9). Yet there are still many unresolved issues related to the more detailed characterization of the gene(s) in the HLA region (10). In this work, articles I, III and IV concern hypothesis of the genetic influence of HLA-A, HLA-B and HLA-DRB1 loci on T1D risk. Paper II is not directly related to HLA but explores possible genetic effects to explain the observed increasing trend in the incidence of T1D in Finland, assuming a causal variant with effects identical to those observed for the HLA-DR3,4 genotype. In paper IV, familial aggregation of the partially latent susceptibility and age at onset of nephropathy among Type 1 diabetics is modelled and estimated.

The aims of this dissertation were to develop statistical methods to analyze registry-based data and on the other hand characterize genetic effects of HLA A, B and DR loci both to age at onset and to susceptibility of T1D. Next, an overview of the statistical techniques in genetic epidemiology with focus on variable age at onset phenotype, ascertainment and genetic association is presented. A short introduction to the epidemiology and especially genetic epidemiology of T1D is given next. The study designs and data collection in the DiMe and the DERI studies are then described. The proposed statistical models and inference are described and results concerning T1D are presented, with discussion of the models and results at the end of the thesis.

2. Genetic epidemiology

Genetic epidemiology may be viewed as the study of the joint action of genes and environmental factors in causing disease in human populations and their patterns of inheritance in families (11). It is then evident that genetic epidemiology focuses on the familial, especially genetic, determinants of disease. The process of establishing genetic basis of a phenotype in genetic epidemiology, depending on the questions asked and data at hand, can be classified as follows (articles in this thesis are given by Roman numerals):

- Analysis of familial aggregation: Does the potentially genetically determined phenotype aggregate in families? (paper V)
- Segregation analysis: Find a genetic model that adequately explains patterns of phenotypes in families (paper V)
- Linkage analysis: Determine the approximate chromosomal location of disease gene(s) using information on the known locations (markers) in the chromosome(s)
- Association analysis: Localize the gene in more detail, look for association with candidate genes, characterise possible causal genes (papers I-IV)
- Gene effects – gene expression, microarrays etc. Does the polymorphism affect mRNA?

2.1 Summary of the basics of Human Genetics

Human DNA is organized in 46 chromosomes – 22 homologous pairs and two sex-specific chromosomes (called X and Y). Each chromosomal strand is made of sequence of nucleotide bases of four types: adenine (A), cytosine (C), guanine (G) and thymine (T), which are bound by covalent bonds. Double-stranded DNA is replicated by breaking the two strands and then constructing a new complementary strand for each. A single strand of DNA may also act as a template for a complementary strand of RNA. In this transcription RNA is identical to DNA, but T is replaced by U (Uracil). Genes are regions of DNA from which the transcribed DNA encodes to tell the cell how to construct aminoacids to make proteins. Genes contain regions of variable length called exons and introns. Mature RNA is processed by cutting out introns.

In translation, the splices the exonic sequencies produce mRNA, which codes for proteins. The ability of genes to alterate protein function is the basis of genetic influence on phenotype(s).

According to Mendel's first law, each parent donates one of each pair of the chromosomal strand with equal probability. At a given locus, many different forms of gene representing individual mutations may exist, called alleles. Alleles are differentiated by the repeats of nucleotide bases. An unordered pair of alleles at the same locus is called a genotype. Genotype can be treated as unobserved variable in the statistical analysis. When the parental source of alleles is known genotype can be called ordered genotype. A combination of alleles at multiple loci along same chromosomal strand is known as haplotype. A heterozygous individual carries two different alleles, while an homozygous individual has two copies of the same allele. Observable, already known locations in the chromosomal strands are called markers, and two types of markers are currently widely used in genetic epidemiological studies: microsatellite and single nucleotide polymorphism markers (SNPs). Markers can be highly polymorphic, that is genes usually have many possible alleles or markers that contain only some simple aminoacid sequences. Some regions in human genome, like HLA region in chromosome 6, are highly polymorphic. Markers at the HLA region have until recently been typed with serological methods, based on the antibody production, rather than modern PCA-based genotyping or more recent high throughput methods (9).

Our ability to locate disease genes in linkage and association studies in humans is based on the cross-over (meiotic recombination) events during meiosis. Meiosis is a special case of cell division where sperm and egg cells are created. In meiosis the DNA material can be changed in a cross-over event when DNA from one parental strand switches to another parental strand. There are several phases in meiosis as described in Figure 1, where cross-over occurs in Prophase 1. There homologous chromatids pair up and form physical connections (chiasmata). DNA strands break up and swap material from one chromosome to material from another. This results gametes having material from both homologues of a chromosomal pair.

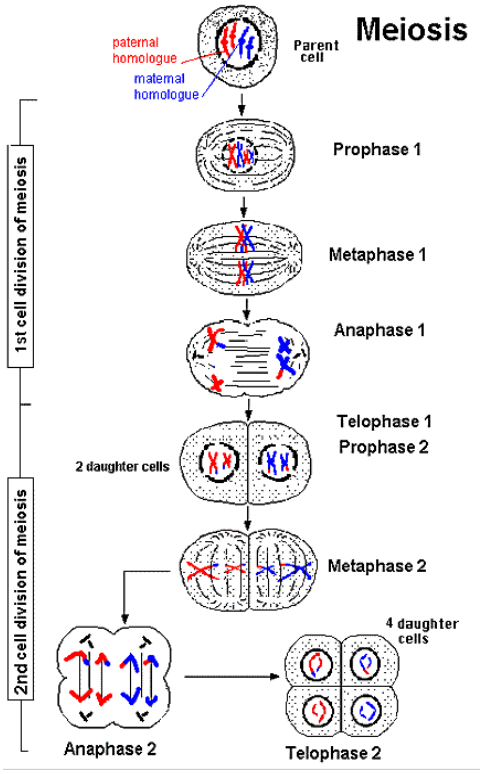

Figure 1. Phases of meiosis (www.accessexcellence.org/AB/GG/meiosis.html)

In Figure 2, a cross-over event has occurred somewhere between A and B loci. During prophase I the four available chromatids are in tight formation with one another. While in this formation, homologous sites on two chromatids can mesh with one another, and may exchange genetic information. In Figure 2, a part of the maternal chromatid containing A7, are joined with paternal chromatid containing B2 to form a recombinant chromatid and similarly another recombinant chromatid containing A5 and B3 is formed.

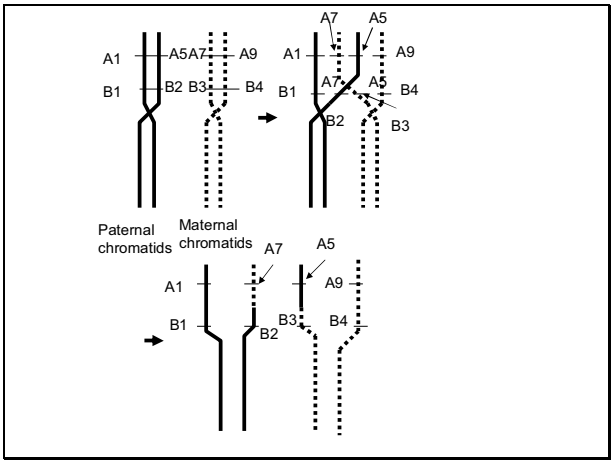

Figure 2. Crossing-over event, producing two recombinant and two non-recombinant chromatids.

The frequency of recombination between two locations depends on their distance, because recombination can occur with small probability at any location along chromosome. Under the assumption that crossover event occur randomly the expected number of crossover events can be modeled using the Poisson model, resulting a Haldane's map function. This is the most simply function that relates recombination fractions to map distances. Genetic Linkage analysis relies on the fact that if a marker is passed down through a family consistently with the disease then it can be said that gene with a functional effect is located close to that marker. Linkage analysis has been less successful to locate genes influencing complex diseases than monogenic diseases, mainly due to lower statistical power (12). Therefore, genetic association studies have been used to dissect genetics of complex diseases and later in this thesis principles of genetic association are described.

2.1.1 Human Leucocyte Antigen

The Major Histocompatibility Complex (MHC) is a large gene-dense region in the genome of most mammalians. It plays an important role in the immune system and reproduction vertebrates. In humans MHC is known as the Human Leukocyte Antigen system (HLA). HLA consists of series of closely linked and highly polymorphic genes, spanning about 4 Mb on the short arm of the chromosome (6p21.1-6p21.3). HLA genes can be classified into three major subgroups: class I, class II and class III. This grouping is based on the encoding functions and

expression of the molecules (13). HLA-A and HLA-B loci belong to class I and HLA-DR to class II. Locations of the HLA-A, HLA-B and HLA-DR loci, studied in the papers I, III and IV, are shown in Figure 3.

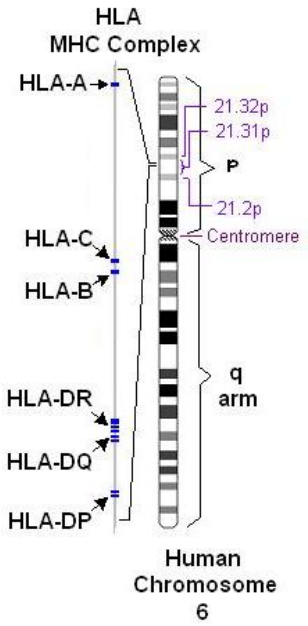

Figure 3. Locations of HLA-A, HLA-B and HLA-DR loci on chromosome 6.

At the moment, all the known loci at the HLA have currently nearly 3,000 alleles that have been detected in the human population (http://www.ebi.ac.uk/imgt/hla/stats.html, last updated October 2007), of these the most variable are HLA-B and HLA-DR(B1). Known variant alleles according to the international ImmunoGenetics –project (IMGT-HLA) database for loci studied in here are HLA-A 617, HLA-B 960 and HLA-DRB1 626. The first complete DNA sequence and gene map of HLA was published in 1999, and estimated 224 genes in the region (14). To keep track on the alleles of various loci at the HLA region, there is a nomenclature concerning HLA and it is evaluated at every two years. Two systems of the nomenclature can be applied to the naming of HLA alleles:

1) The older system is based on serological (antibody-based) recognition. In this system antigens were assigned letters and numbers (e.g. HLA-DR4 or, shortened, DR). Because all the HLA genotype data of the Childhood Diabetes in Finland study

(DiMe) in this dissertation is obtained with serological methods we use this notation of the HLA-A, HLA-B and HLA-DR alleles.

2) The “newer” system is based on the nucleotide sequencing based recognition of the alleles and thus allows for more refined definition of alleles. A "HLA" is used in conjunction with a letter * and four or more digit number (e.g. HLA-B*0801, A*68011, A*240201N N=Null) in order to designate a specific allele at a given HLA locus.

The MHC genes and the cell surface protein molecules encoded by the MHC play a critical role in T-cell recognition and thus in human immune system. The HLA system, by virtue of its extreme polymorphism, ensures that few individuals are identical and thus the population at large is well equipped to deal with attack. Because some HLA antigens are recognised on all of the tissues of the body (rather than just blood cells), the identification of HLA antigens is described as "Tissue Typing". Routine tissue typing identifies the alleles at three HLA Class I loci (HLA-A, HLA-B, HLA-C) and three Class II loci (HLA-DR, HLA-DP) and is based on the structure of the antigens produced and their function. Class I genes encode glycoproteins expressed on the surface of almost all nucleated cells of the body and class I molecules are encoded by the A, B and C region in humans. Class II genes encode glycoproteins expressed primarily on antigen-presenting cells (macrophages, dendritic cells, B cells) where they present the processed antigen to the T helper cells. These molecules are encoded by the class II region. HLA Class I genes and the HLA Class II genes each spread over approximately one third of the length of HLA region. The remaining section sometimes called class III, contains loci responsible for complement, hormones, intracellular peptide processing and other developmental characteristics, including tumour necrosis factor (TNF). These loci are not actually a part of the HLA complex, but since they are located within the HLA region are included in the nomenclature, because its components are either related to the functions of HLA antigens or are under similar control mechanisms to the HLA genes. The HLA region has been shown to link and/or associate to many diseases, including T1D (9).

2.2 Fundamental concepts in genetic epidemiology

The first step in studies of genetic epidemiology uses descriptive methods to document familial aggregation for trait that cannot be explained solely by environmental factors. In the next step, heritability is studied. When the phenotype is continuous, heritability means the proportion of variation in trait due to genetic factors. Variance component or path models are commonly used statistical methods in heritability studies. When the phenotype is dichotomous (e.g. healthy, affected) there is a threshold in the latent liability (continuous) describing the cutoff point which divides subjects into the two groups. Segregation analysis tests models of inheritance on family data and does not require observed genetic markers, which are polymorphic genes whose physical locations are known. Any possible phenotype can be analysed in segregation analysis e.g. dichotomous, continuous or censored continuous. Modern genetic association studies use information on the historical recombinant events that could cause linkage disequilibrium, and there are sophisticated statistical methods for the analyses of genetic data to localise and characterize in more detail gene responsible of variation in phenotype. Genetic association studies have become more popular when dense marker maps have become available using microsatellite markers or SNP's.

Segregation analysis compares patterns of trait in families to some pre-specified inheritance model. In order to relate genes or markers to the phenotype, we need a mathematical function describing the relationship. This function is called penetrance and it gives a probabilistic model that relates genotypes to phenotype using a conditional probability function: P(phenotype | gene(s) and/or environmental factors). In all of the studies presented here, phenotype is a variable age at onset phenotype. The simplest directly observably phenotype is binary (affected/healthy), but it can be continuous (cholesterol, blood pressure). This dissertation focuses on phenotype with variable age at onset, because T1D can be diagnosed at any age in contrast to phenotypes expressed immediately at birth.

2.3 Statistical inference in genetic epidemiology

Statistical inference is based on assuming a probabilistic model which could have produced the data. Because both classical (frequentist) and Bayesian statistical inference is used in this thesis, a short overview of both is given. The classical inference is based on the mathematical

probability density function of the data $Y = (Y_1,...,Y_n)$ given the unknown but fixed model parameters θ: $f(Y \mid \theta)$. If the observations can be assumed to be independent and identically distributed (i.i.d.), the likelihood function of the observations is simply a product of the density functions of each observation, i.e. $f(Y \mid \theta) = \prod_{i=1}^{n} f(Y_i \mid \theta)$. Maximum likelihood estimator is the value of θ, which maximizes the above likelihood function. The classical statistical inference relies on the assumption that the mechanism which generated the data remains unchanged, and the data generation process can be repeated. This classical framework has been applied to the data analysis in papers I-III. Note that in the context of classical inference the density function describes the generation of the data. In the genetic epidemiology observations are rarely independent because commonly we observed data from relatives (pedigrees). In order to analyze data of dependent observations one can use multivariate distributions or introduce unobserved latent random variables. Familial dependency in the age at onset and susceptibility were accounted for using latent variable approach in paper IV in this thesis.

The research questions in papers IV and V were modelled using Bayesian inference. The core of the Bayesian inference is the joint probability density of the observed data *Y* and θ, which is treated as random variable in the analysis unless value is known. The statistical inference is based on the posterior distribution of

$$f(\theta \mid Y) = \frac{f(Y,\theta)}{f(Y)} = \frac{f(Y \mid \theta)g(\theta)}{f(Y)} \propto f(Y \mid \theta)g(\theta),$$

where $f(Y)$ is the marginal probability density function of the data obtained either by summing over $f(Y \mid \theta)g(\theta)$ if θ is discrete or integrating if θ is continuous, $f(Y \mid \theta)$ is the likelihood function and $g(\theta)$ is the prior density function which incorporates researchers beliefs (in probability formulation). If the value of θ is unknown it needs to be defined. This is done by the researcher by defining a prior distribution of θ. We end up with a posterior density, which in fact is the prior multiplied by the likelihood. If the prior is vague, inference is dominated by the data and similarities with classical approach can occur. Bayesian methods provide natural tools to deal with missing/latent data by sampling from the posterior distribution of missing data than the classical approach. Because the posterior distribution rarely can be solved analytically, developments in numerical methods and computer capacity

have made possible to apply Bayesian modelling increasingly also to genetic epidemiology (15). Computationally intensive methods are needed to obtain samples of the posterior density. One of the oldest methods is Metropolis-Hastings algorithm (16), which was used in paper IV to obtain samples from the joint posterior of missing susceptibility status and the model parameters. In paper V several sampling algorithms were used, all implemented in the OpenBugs program (17).

2.4 Phenotype with variable age at onset

In the survival analysis a non-negative random variable T_i^* is modelled. This is the 'true' time of the event of interest of subject *i*, e.g. 'true' age at onset of disease. Censoring refers to a case where we observe the true time T_i^* only if it does not exceed subject *i*'s censoring time U_i, otherwise we observe U_i (right censoring). Type I censoring is assumed and it occurs when subjects are put on a test at time 0 and are followed for a fixed period of time. Then we observe failure times of those subjects that failed and while for the rest we know that they did not fail before and are said to be censored. Observed data of variable age at onset disease data consists of pairs of survival times and censoring indicators $(T_1,\delta_1),\ldots,(T_n,\delta_n)$, where $T_i = \min(T_i^*, U_i)$ and $\delta_i = 1$ if subject *i* failed during the follow-up and 0 if censored.

A phenotype with variable age at onset can be characterized by the distribution function $F(t_i;\theta) = p(T_i^* \le t_i;\theta)$, survival function $S(t_i;\theta) = 1 - F(t_i;\theta) = p(T_i^* > t_i;\theta)$ and density function $f(t_i;\theta) = \frac{dF(t_i;\theta)}{dT_i}$. Hazard function $\lambda(t_i;\theta) = \frac{f(t_i;\theta)}{S(t_i;\theta)}$, expresses limiting probability that the event of interest occurs in a given time interval, conditionally on that the event of interest has not occurred before the beginning of the interval and divided by the width of the interval. Survival models provide natural means to deal with censoring - an important feature of age at onset phenotype. Assuming independence of the observations, the likelihood function can then be written $L(T;\theta) = c \cdot \prod_{i=1}^{n} \left\{ f(T_i;\theta)^{\delta_i} \left(1 - F(T_i;\theta)\right)^{1-\delta_i} \right\}$, where *c* is a constant and *c* is not a function of the model parameters. Survival models can be written using the distribution function or the hazard function. One choice is the Weibull distribution

with a shape parameter γ and scale parameter μ having density

$f(T_i;\gamma,\mu)=\frac{\gamma}{\mu}\left(\frac{T_i}{\mu}\right)^{\gamma-1}\exp\left(-\left(\frac{T_i}{\mu}\right)^{\gamma}\right)$ (paper V) or the Gamma density function with shape parameter (a) and scale parameter (s) $f(T_i;a,s)=\frac{1}{s^a\cdot\Gamma(a)}T_i^{a-1}\exp\left(-\frac{T_i}{s}\right)$ (paper III), where $\Gamma(\cdot)$ is the gamma-function. Measured covariates z_i can be modelled using the Weibull model by defining the log of the scale parameter of subject i μ_i to be a function of the covariates and regression coefficients: $\log(\mu_i)=\beta_0+\beta' z_i$, where β is a (p×1) vector of regression coefficients and $z_i=(z_{i1},...,z_{ip})$ is vector of the observed covariates values.

Survival models have long history in epidemiology, and the variable age at onset has been analysed using survival models developed for genetic-epidemiological studies during past few decades. Proportional hazards models have been applied to both the segregation and linkage (4). This semi-parametric survival model can be written using the hazard function $\lambda_i(t)=\lambda_0(t)\exp(\beta' z_i)$. In this model formulation covariate information (observed genetic markers) is related to the hazard of disease age at onset through exponential function that multiplies the baseline hazard function. In order to include familial correlation into event time models a frailty model have been introduced by Vaupel et al. (18). It was later developed for genetic studies (19-21). Frailty is a latent random effect variable that can be added into the exponential part of the previous hazard model. In genetic epidemiology a shared frailty model indicates that all members of the same family will have the same frailty. A discrete time formulation of the survival model, in which the time scale is divided into small time periods with constant baseline hazard (piecewise constant hazard model) is applied in paper IV of this work. In the analysis of event data likelihood can be based on the Poisson density. In paper II we make a use of the Poisson probability density $f(d;\mu)=\frac{e^{-\mu}\mu^d}{d!}$, where μ is the parameter (mean number of events) and d is the observed number of events. This model is useful when the disease is rare, like T1D is.

Survival models described above can be extended to incorporate an idea that not all subjects will develop outcome even if followed forever. The long-term survivor model elegantly combines binary susceptibility phenotype and survival phenotype (6). It is assumed that

censoring distribution is proper, e.g. $G(\infty) = \lim_{T\to\infty} G(T) = 1$, and that the cumulative distribution function $F(\infty) = \lim_{T\to\infty} F(T) = p$ is strictly less than one. Then the case $p<1$ corresponds to the presence of non-susceptibles in the population. In order to allow true survival times t* to be infinite and to introduce a proper c.d.f. to describe lifetimes of non-susceptible subjects, a latent Bernoulli random variable D_i for each subject *i*, which is one (susceptible) with probability p and 0 (non-susceptible) with probability 1-*p* is introduced. Because we do not know who is non-susceptible, of all subjects, D_i is a partially latent random variable. The c.d.f. for true survival times is then

$$F(T_i;\theta) = p(T_i^* \le t_i;\theta) = p\{T_i^* \le t_i \mid D_i = 1;\theta\}\, p\{D_i = 1;\theta\} + p\{T_i^* \le t_i \mid D_i = 0;\theta\}\, p\{D_i = 0;\theta\}$$
$$= pF_0(T_i;\theta) + 0 = pF_0(T_i;\theta),$$

which essentially rescales the possibly improper *F* to a new distribution F_0 with total mass 1. The likelihood function for the observed survival times with non-susceptibles can then be written as

$L(T;\theta) = c \cdot \prod_{i=1}^{n} \left\{ f(T_i;\theta)^{\delta_i} \left(1 - p \cdot F_0(T_i;\theta)\right)^{1-\delta_i} \right\}$. This likelihood was applied in paper III, when ascertainment was taken into account. Maller et al. (6) worked on the problems of right extremes, sufficient follow-up and testing of the presence of immunes and provided many statistical properties of the LTS model. In the Bayesian formulation of this model (22), susceptibility status D_i is treated as partially latent variable and sampled as a part of the joint posterior distribution as it has been done in the paper V. There are also some very recent non-Bayesian applications of this model in the genetic studies (23,24).

2.5 Ascertainment

Phenotype-based collection of families, ascertainment, is a common type of data collection for rare diseases in genetic epidemiology. This leads to families in which affected individuals are over-represented compared to the population as a whole, and thus possibly falsely increasing the statistical power when estimating genetic parameters such as recurrence risk ratios (25).When the aim of the study is to make inference of the model parameters in the general population from which the observed data has been ascertained, it must also be remembered that statistical inference has to take into account the process of data collection. In

naive statistical analysis ignoring ascertainment process, the statistical inference of the model parameter estimates are biased – this is called ascertainment bias. Crucial for the proper statistical analysis of such data is that the ascertainment process should be well documented and clearly defined. Problems related to ascertainment have been studied extensively from the early days of modern statistics (1). In the classical ascertainment model a binary phenotype (healthy, affected) is assumed and a person is affected with a constant probability η independently of others. An affected subject becomes a proband with constant probability π. In other words, a single individual causes a family to enter the sample (proband) independently of all other persons in a particular sibship. If we observe all sibships with at least one affected (π=1, complete ascertainment), we can compute the conditional probability of observing $R=r$ affected subjects given the size of sibship $S=s$ and at least one affected $p(R=r \mid S=s, R\geq 1; \eta)=\frac{p(R=r \mid S=s;\eta)}{p(R\geq 1 \mid S=s;\eta)}=\frac{\binom{s}{r}\eta^r(1-\eta)^{s-r}}{1-(1-\eta)^s}$. The likelihood function is the product of the previous probabilities over the possible sizes of sibships and the numbers of affected subjects.

The past studies of ascertainment focused either on the definition of discrete sampling units or on construction of more complex conditional likelihood functions based on the above definition of the proband (26-29). The problem of ascertainment can be handled by choosing the appropriate study design: a case-parent triad design in which the formulation of the conditional likelihood is based on the probability of the genotype of the case conditional on the case status and the possible parental genotypes (30). This transmission/disequilibrium test (TDT) based on conditional likelihood avoids ascertainment bias by conditioning on the parental origin and its different modifications have gained much attention as a way to avoid population stratification (28). However, one cannot estimate the absolute risk, because data are restricted to cases only and information concerning the population is not available.

Development of computationally intensive methods has produced novel statistical methods for the problems of complex ascertainment (2). Recently, Burton (3) explored the possibility of heterogeneity in the risk of disease in the general population. This expands the classical framework of ascertainment where a fixed probability of disease was assumed. Analysis of family data through probands with variable age at onset in disease registry data has not been

studied much. In their article Langholz et al. (31) concluded that in the estimation of rate ratios in addition to the rare disease assumption, it is also essential that the covariate distribution does not depend on calendar time or birth order (exchangeability assumption). In paper I of this dissertation, we construct a likelihood function conditional on the ascertainment by considering all sibs that could have become probands during the recruitment period. In paper V we incorporate the relevant parts of the Finnish population to the full likelihood function.

2.6 Genetic association

If a genotype occurs in study population more commonly among affected subjects than among healthy subjects, this is called a genetic association. This association could be noncausal because the marker could be in linkage disequilibrium with the truly causal gene. As in epidemiology, association between genetic marker and a phenotype might exist because 1) the polymorphism has truly causal role (direct association) 2) polymorphism has no causal role but is associated with nearby causal variant (indirect association) and 3) there is population stratification/admixture (confounded association) (32). Observations of genetic association can be confounded by population stratification, where unequal proportions of alleles in latent subgroups can lead to spurious associations. To avoid this it is possible to collect data from well mixed population, to match according to geographical or ethnic origin, to seek markers for subpopulation structure (33) or to search for genomic control to control false positive rate (34). The characterization and mapping of genes using indirect method is dependent on the association between causal variants and nearby markers. This association of loci is called a pairwise linkage disequilibrium (LD) in population genetics. LD can be defined as joint occurrence of pairs of alleles in two linked loci more or less than expected by random formation of haplotypes from alleles based on their frequencies. In genetic association the same allele(s) are associated with the phenotype in a similar manner across the whole study population. The amount of LD can be estimated by the square of standard correlation coefficient between alleles at pair of loci and the allele at the disease locus or by some other less common measure (35).

Genetic association studies are therefore based on the genealogical history of the sampled data. The idea is that a disease mutation arises on a particular haplotype during the genealogical history, and so individuals who inherit the mutation will also inherit the same alleles at nearby marker loci. At time of a disease mutation occurs, it is in complete linkage disequilibrium with loci nearby. LD is diluted in each meiosis due to recombination between the two loci, recurrence of the same mutation and gene conversion. It has been argued that association based methods will become more powerful than traditional linkage methods in mapping genes with small effects on complex diseases because of population genealogy and increasing number and density of marker information (microsatellite markers and SNP's) (36). It has become clear that because the initial LD and the decay varies due to assumptions of large population, mutation and selection, both within genome and between populations therefore making it difficult to adopt general rules of studying complex diseases (32).

Traditional epidemiological study designs can be employed in studies relying on the genetic association: cross-sectional, case-control, cohort and other more modern variations of these classical study designs (nested case-control, case-cohort). More genetically oriented study designs are based on observing information of the familial relationships: case-parent triads, case-parent-grandparent sets, pedigrees and many variations of these including sibpairs and twins studies. In this dissertation a cohort design is applied in papers III, IV and V. The study design in paper I can be seen as case-parent triads but rather than testing a hypothesis of recombination being 0.5 or LD being zero as in TDT-test we test the Mendel's first law. The study design of paper II can be viewed as an aggregated cohort of the Finnish population from 1965 to 1996.

3. Type 1 diabetes

3.1 Pathogenesis of Type 1 diabetes

The detailed pathogenesis of T1D is unknown, but it is believed that joint action of environmental and genetic factors is needed to either initialise or speed up the immune-mediated selective destruction of the insulin-secreting β cells, in the pancreas. In T1D patients, the pancreatic islets containing the β cells exhibit insulinitis, i.e. inflammation characterized by the presence of T lymphocytes accompanied by macrophages, B lymphocytes, without involvement of the glucagon-secreting α-cells. T1D usually presents itself with symptomatic hyperglycemia or ketoacodosis. The classic symptoms of hyperglycemia are polyuria (frequent and abundant urination) followed by polydypsia (thirst and high intake of liquids), and weight loss. Because high levels of glocose in the blood produce osmotic diuresis (excessive loss of water, Na+ and K+ with urine), the symptoms can progress to dehydration, which might include blurred vision, fatigue, and nausea. If the loss is massive, for example if precipitated by withdrawal of insulin or infection, the modulating effect of insulin in hepatic activity is suppressed and hepatic ketone body synthesis and release is increased and ketoacidosis might develop. Accumulation of ketons produces first ketonuria (increased excretion of ketons in urine), and can progress to ketoacidosis, and if not quickly corrected, to coma and death.

T1D accounts for approximately 10% of all diabetes and it has been a lethal disease until the late 1920's, when the insulin treatment was discovered. During the last decades several other forms of diabetes in children and young adults have been discovered (MODY – mature onset diabetes of the young (37); LADA - latent autoimmune diabetes of adulthood (38), and other less frequent forms), which earlier might have confounded the T1D phenotype. Although T1D commonly occurs in children, with peaks of incidence around the puberty or pre-puberty (39), it can also develop in adults. Non-genetic factors potentially associated with T1D include gender (40), birth order (41), diet (42), breast feeding (42,43), infections (44) and composition of the drinking water (45).

3.2 Epidemiology of T1D

The incidence of T1D shows variation according to many factors including: gender, age and geographical areas. A significant spatial variation worldwide has been observed from 0.2 to 40/100,000 (46). Several epidemiological studies have shown that Finland has the highest incidence of T1D in the world: during the 1990's there were approximately 40/100,000 new cases per year in children below 15 years of age (47), whereas the lowest incidence has been observed in East Asia. Geographical differences in the incidence of T1D have been detected within Finland as well.

The average increase in the incidence of T1D between 1965 and 1996 has been estimated to be 0.67/100,000 per year (47). Findings concerning the age at onset distribution with respect to calendar time are conflicting. While a study of the Finnish population indicated that age at onset has become younger (48), a later study which accounted for the registry-based ascertainment of the T1D cases, showed that the observed shift at the age at onset of T1D could be explained by random variation (49).

3.2.1 Genetic epidemiology of T1D

Although T1D can be best identified as a chronic T-cell mediated autoimmune disease attacking insulin-producing pancreatic cells, genetic susceptibility plays an important role on its onset and development. As we have also seen above, the association of specific HLA alleles and haplotypes with T1D is very strong, this genetic locus is estimated to account for <50% of genetic contributions to disease susceptibility (50, 51). Diabetogenic alleles are not fully penetrant, implying that not every individual who inherits the gene develops the disease.

Family studies have also shown the importance of genetic factors in determing T1D risk. Sibling cumulative risk of affected subjects is approximately 6% while population risk is 0.5%, varying between populations. Further evidence for genetic factors comes from the large population-based twin studies (52) and recent large population-based twin studies indicate a clear genetic component in the risk of T1D (53). HLA region has constantly shown linkage as well as association in all of the genome-wide linkage or candidate gene/genome-wide association studies. IDDM2 (region containing INS gene, chromosome 11) and IDDM12

(region containing CTL4 gene, chromosome 2) has been detected in multiple association studies. Other loci IDDM3-IDDM18 in different chromosomes have been identified but further replications are needed to avoid false-positive findings. It has been postulated that multiple genes in various chromosomes could contribute to the risk of T1D with unknown combined effects (54). At the moment there are 32 locations linked/associated with T1D (see Figure 4), but HLA seems to confer the highest risk of all of the detected locations.

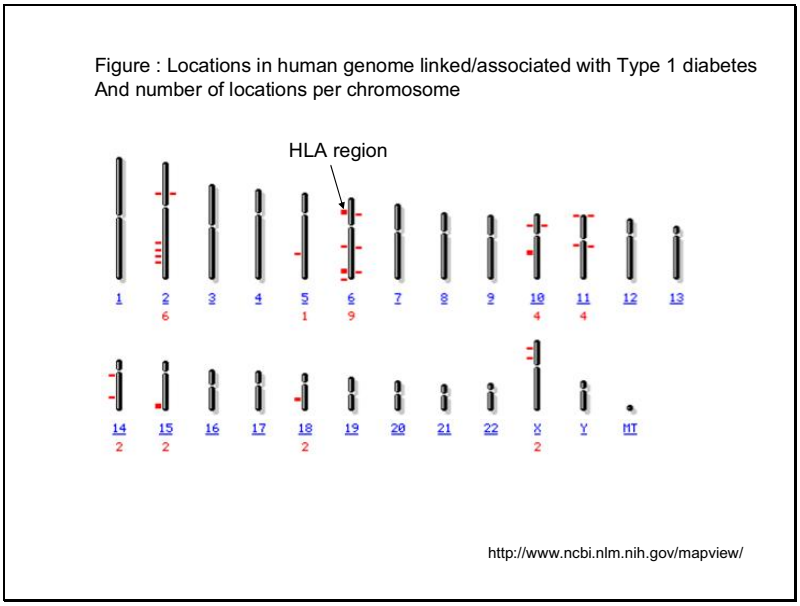

Figure 4. Locations in human genome linked/associated with T1D and the number of locations per chromosome.

One of the first genetic associations with human diseases was with HLA and T1D (55) and the association was later pinpointed to the HLA-DR3 and DR4 containing chromosomes. Interestingly HLA-DR3/DR4 heterozygotes were observed to confer higher risk of T1D than DR3/DR3 or DR4/DR4 homozygotes. Further locating the gene have been complicated by the strong long-range LD. Very recent genome-wide scans have provided strong evidence of IDDM1 (HLA region) (56) and this is consistent across different populations and studies.

3.3. Dime and DERI Studies

3.3.1 The DiMe Study and nationwide registry of T1D

The Childhood Diabetes in Finland (DiMe) study was a large population-based genetic-epidemiologic family study of T1D (57). Nationwide, all T1D cases under the age of 15 in Finland diagnosed during the calendar period from September 1987 to April 1989 were identified. The cut-off age was chosen purely for practical reasons. Newly diagnosed children under the age of 15 years with T1D were hospitalized in the pediatric wards in Finland and therefore were easier to recruit than older subjects with T1D. T1D status was checked against the data of the National Drug Registry. Of the 801 cases in the study 800 had also been registered in the Drug Registry and one person died soon after the diagnosis. The participation rate in the study was approximately 95%. Parents and siblings of the 801 probands were also asked to participate. Extensive questionnaires were filled in and blood samples were taken from participants. Probands, their parents and siblings were HLA genotyped at A, B, C and DR loci using conventional serology (57,58). The ethical committees of the National Public Health Institute and local hospitals approved the study protocol. Details of study procedures, especially data collection, are described elsewhere (57). A polymorphism information contents (PIC) was used to choose loci to the statistical analysis (11). Its value describes the polymorphism within a sample and it is a function of the number of detectable alleles and their frequency. Statistical analyses of the DiMe data were restricted to the HLA-A, HLA-B and HLA-DR loci, which had the highest polymorphic information content (0.74, 0.89 and 0.73 respectively).

3.3.2 The Diabetes Epidemiology Research International Study (DERI)

The original cohort consisted of subjects with T1D diagnosed before the age of 18 years between 1965 and 1979 (n = 5,126), who were included in the nationwide register of Finnish T1D patients and comprised the Finnish contribution to the Diabetes Epidemiology Research International (DERI) study (59,60). This register was initially based on the Social Insurance Institution's Drug Reimbursement Register, which lists patients approved to receive free-of-charge medication for certain diseases, including diabetes. Their siblings (n=10,168) were identified through the national population register of Finland. The diabetes status of the siblings was ascertained through several sources: from the nationwide Hospital Discharge Register for the years 1970–1998, from the nationwide Finnish Diabetes Register for children

and young adults for the years 1965–1998, and from the Central Drug Register through searching the records using the personal identification number assigned to all residents of Finland. Because of these multiple data sources, the case ascertainment was virtually complete. By the end of the year 1998, a total of 537 families that included 616 siblings with T1D were identified among the original DERI cases (61). Because of missing information on the time at onset of DN, nine families were excluded from the dataset used in the statistical analysis. Therefore we had in the analysis 528 families, with total of 1134 T1D patients and by the end of follow-up 321 DN cases were observed.

In order to identify patients with DN, copies of original medical records, death certificates, and autopsy data for the probands and siblings were systematically reviewed by one of the co-authors (V.H.). Overt nephropathy was defined when a patient repeatedly had either a urinary albumin excretion rate of >200 µg/min or >300 mg/24 h, a 24-h urinary protein excretion rate of>0.5 g, or a positive urinalysis for protein using a reagent strip. Microalbuminuria was defined as a urinary albumin excretion rate of 20–200 µg/min or 30–300 mg/24 h. Albumin elevations because of pregnancy, urinary tract infections, or other renal diseases alone were not considered as diagnostic for DN. The urinary albumin excretion rate had decreased in some patients because of the initiation of antihypertensive medication; in such cases, the classification of DN was based on findings before the initiation of drug treatment. Microalbuminuric patients were grouped together with normoalbuminuric subjects in all analyses.

4. Aims

The aim of this investigation is to develop novel statistical tools for genetic-epidemiological studies and to assess the associations of T1D with genetic factors, especially HLA. To achieve this aim following aspects are considered:

- Construction of conditional likelihood of variable age-at-onset phenotype in population-based ascertainment to assess Mendelian transmission of HLA-A, HLA-B and HLA-DR loci.

- Derivation of the Poisson likelihood for the incidence trend with possible genetic influence in order to consider whether non-Mendelian transmission of disease alleles lead to increase in T1D observed in Finland.

- Modelling of the genetic association of variable age at onset disease with long-term survivors in order to study the association between HLA and both susceptibility and age at onset of T1D.

- Introduction of the familial segregation model with long-term survivors and to consider the familial aggregation of susceptibility and age at onset of T1D nephropathy.

- Derivation of the full likelihood function for variable age at onset phenotype in population-based registry data in order to make population-based analysis of the HLA-A, HLA-B and HLA-DRB1 genotypes/haplotypes and the age at onset of T1D.

5. Results

5.1. Statistical models and inference

Observing a family during the recruitment period does not depend only on the index child, but rather on all siblings at risk of T1D during the recruitment period. This is accounted for in the two approaches proposed for the ascertainment based on the variable age at onset in this work:

1. In the conditional likelihood (papers I and III) approach the first born child who is diagnosed in the ascertainment "window", is considered a proband. Then the two sets of siblings are eligible for the statistical analysis: a) those born before c_0-w so that they could not become probands and b) those born after the proband (younger than proband).

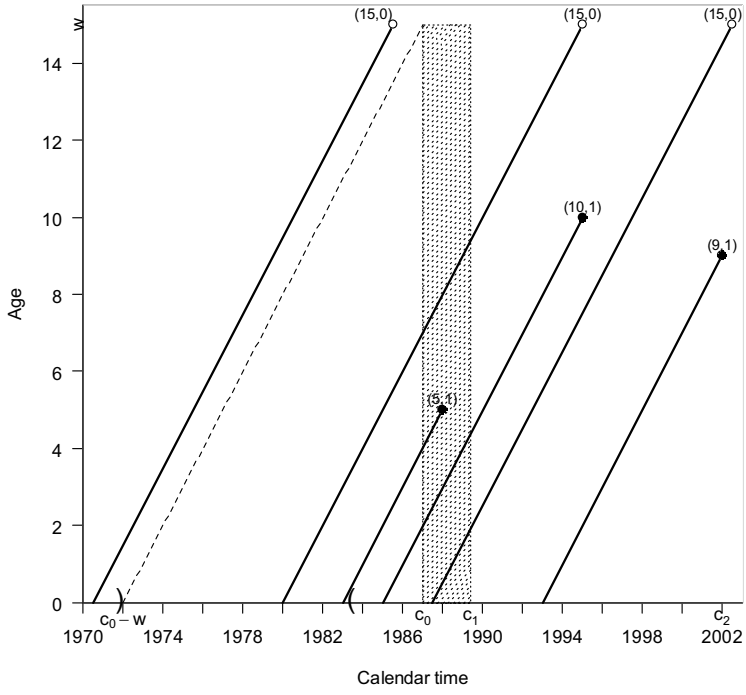

Figure 5. Lexis diagram illustrating construction of the conditional likelihood (siblings born before 1972 and after the proband included in the analysis) of the population-based ascertainment of all Finnish sibships which included an individual younger than 15 years who was diagnosed with T1D during the recruitment period 1.1.1987-30.4.1989 (dotted area).

2. In the full likelihood (paper V) the ascertained family data is complemented by the demographic information about the number of subjects at risk and external reference sample of the genetic composition of non-ascertained population in order to formulate statistically coherent likelihood function (cf Figure 6).

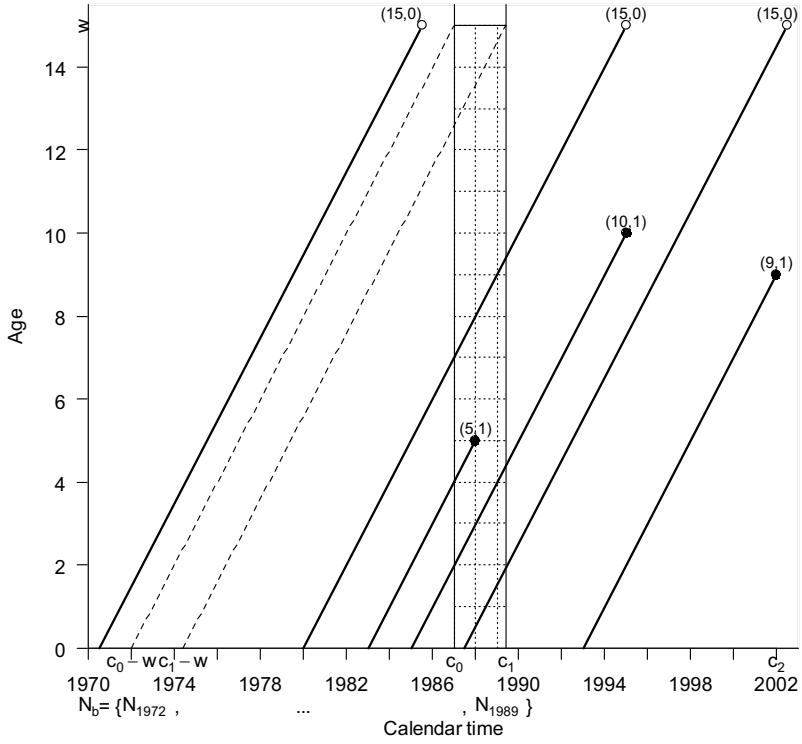

Figure 6. Lexis diagram of the population-based ascertainment of all Finnish sibships, that included an individual younger than 15 years who was diagnosed with T1D during the recruitment period 1.1.1987 - 30.4.1989. The start and end points of the recruitment window are denoted by c_0 and c_1, respectively. Further, w is the maximum age at which an individual can be ascertained (w equals 14 years here). The time interval ($c0$-w, $c1$) contains all individuals at risk, such that they could have become the probands during the recruitment. The time point c_2 denotes the end of the follow-up period for the recruited siblings. The sizes of birth cohorts $N_{1972},...,N_{1989}$ are marked in the figure.

In more detail, we consider the situation where the proband has to be diagnosed between 1.1.1987 and 30.4.1989 and has to be less than 15 years of age at the time of diagnosis. The

ascertainment is illustrated by Figures 5 and 6, were a hypothetical family is draw in the Lexis diagram, together with the number of people at risk of T1D. Notice that all families with at least one diabetic child diagnosed with T1D during the recruitment period are observed and all subjects in these families are included in the analysis based on the full likelihood.

5.1.1. Conditional likelihood

Following symbols were used in the construction of the conditional likelihood function: let T_{ij} be the calendar time of diagnosis of IDDM of the *j*th child in family *i*. We follow the convention that if there is no such diagnosis $T_{ij} = \infty$. Let $Y_{ij} = min(T_{ij}, c_1)$, and let $\delta_{ij} = 1$ if $c_0 \leq Y_{ij} < c_1$, $c_0 - w \leq b_{ij} \leq c_1$ and $Y_{ij} - b_{ij} < w$. Here Y_{ij} is the onset time of IDDM of child *j* in family *i* if $\delta_{ij} = 1$ (in which case this child is a proband). Let *r* index the alleles of a locus (the total number of different alleles being *A*), a random variable Z_{ijp} is created, which gets value *r*, $r = 1,\ldots,A$, if child *j* in family *i* got allele *r* from parent *p*, and writing the genotype of a child *j* in family *i* $Z_{ij} = (Z_{ijm}, Z_{ijf})$, where *m* and *f* stand for mother and father, respectively. Consider the observed HLA genotype of the *j*th child in the *i*th family, where g_{ijp} is the allele inherited from parent *p* (where *m* indicates mother and *f* father). Let $G_{ip} = [\, G_i^{p1}, G_i^{p2} \,]$ be the observed genotype at that locus of parent *p* of the *i*th sibship, where superscripts 1 and 2 indicate the two alleles, because grandparental origin is not known. Let K_i be the birth order of the index child (proband), that is diagnosed within the recruitment window. With this notation the observed data can be written $\{Y_{ij}, \delta_{ij}, \{g_{ijp}, G_i^{p1}, G_i^{p2}; p = m, f\}, b_{ij}; 1 = 1,\ldots,I, j = 1,\ldots,J_i\}$.

First, we assume that i) the birth dates of offspring and the genotypes of the parents are known, ii) for the sibships in the general population, transmitted parental alleles and affection statuses are conditionally mutually independent and iii) the sibships are independent. Assumption (ii) implies that the conditional independence is $\{Z_{ij}, Y_{ij}, \delta_{ij}; j \leq k\} \perp \{Z_{ij}, Y_{ij}, \delta_{ij}; j > k\} | \{b_{ij}; j = 1,\ldots,J_i\}, G_i^m, G_i^f$. Because the truth of the ascertainment event $\{K_i=k\}$ can always be deduced from $\{Z_{ij}, Y_{ij}, \delta_{ij}; j \leq k\}$ it will also be

true that $\{Z_{ij}, Y_{ij}, \delta_{ij}; j \le K_i\} \perp \{Z_{ij}, Y_{ij}, \delta_{ij}; j > K_i\} | \{b_{ij}; j = 1,...,J_i\}, G_i^m, G_i^f$. But this shows that, given $\{b_{ij}; j = 1,...,J_i\}$ *and* G_i^m, G_i^f, the genotypes Z_{ij} of all siblings younger than the index child K_i are always sampled independently of the events that led to the ascertainment of the family. In other words, given genotypes of all siblings younger than the index child are always sampled independently of the events that led to the ascertainment of the family.

In summary, inclusion of the sibship in the data does not depend only on the proband but on the other siblings who where at risk of T1D at the time period of recruitment as well. This is in contrast to the classical framework presented earlier and leaving out the proband only is not a good method of correction. Thus, 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 (obviously, the same conditional independence applies to all older siblings, in case there were any), (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. Excluding set 2 from analysis allows one to make unbiased inference without need to model the probability of the ascertainment event. A likelihood function for the transmission of HLA-A, HLA-B and HLA-DR locus alleles is then constructed using sets 1 and 3 based on the multinomial transmission probabilities. The details of the above formulation are given in the paper I and the same principles were applied in the formulation of the likelihood function in paper III.

5.1.2 Full likelihood

Because the above conditional likelihood leads to substantial loss of data and reduction in statistical power, a full likelihood was presented in paper V. The full likelihood is based on the hazard model for the risk of T1D covering the not ascertained population of Finland. Next, some notation and available data sets are presented in order to be later able to define the full likelihood:

1) In the ascertained registry families with T1D, let $i = 1, \ldots, I$ indexing all the ascertained families, $j = 1, \ldots, J_i$ indexing the individuals (siblings) in the ith family, and δ_{ij} being the indicator for right censoring of the follow-up ($\delta_{ij} = 1$ if individual j in family i was diagnosed with T1D in the recruitment window, and $\delta_{ij} = 0$ if right censored, cf. Figure 6). Let $p=m, f$ index the parents of the ith family. Further, for the siblings in the family data, let $G_{ijl} = \{G_{ijlm}, G_{ijlf}\}$ be the marker genotypes over the set of loci of interest, $l = 1,..., N_L$. The alleles received from the mother (m) and the father (f) are indexed accordingly, and an analogous indexing is later used for haplotypes as well. Parental genotypes are denoted by G^M_{ijlp}, G^F_{ijlp}. Finally, let δ_{ij} be the disease onset indicator for individual j in family i during the fixed follow-up period (calendar time), and correspondingly, let X_{ij} represent the age at onset or the age at a censoring event. The observed data of the ascertained sibships are thus collectively represented by the set $\{X_{ij}, \delta_{ij}, G_{ijlp}, G^M_{ijlp}, G^F_{ijlp}; i = 1,...,768,\ j = 1,...,J_i, l = \text{HLA} - \text{A}, - \text{B}, - \text{DRB1},\ p = m,f\}$.

2) In the HLA genotyped subjects of the Bone Marrow Donor registry, let $\{G^C_{rl} = \{G^C_{rlm}, G^C_{rlf}\}; r = 1,...,19836, l = A, B, DRB1\}$ be the set of genotypes for the unrelated individuals in the BMDR database, using a notation analogous to the familial data. These reference individuals are known not to have acquired T1D before the age w. The genotype and haplotype frequencies for this reference population are collectively denoted by $q^C_g = (q^C_{g1},...,q^C_{gn_g})$ and $q^C_h = (q^C_{h1},...,q^C_{hn_h})$, respectively.

3) Demographic data is defined by the numbers of subjects born during the time interval (c_0-w, c_1); $\{N_b; b = 1972,...,1989\}$.

The first part of the risk model specifies the hazard of acquiring T1D for individual j in family i as a function of age a. We use a discrete time hazard model and index age $a = 0,...,14$ corresponding age intervals $[0,1),[1,2),\ldots[14,15)$. For the genotype effect model, the hazard is assumed to be of form

$$\lambda_{ija} = \lambda_a \exp(\beta[G_{ijlm}, G_{ijlf}])$$

Here λ_a is a baseline hazard in the population and $\beta[G_{ijlm},G_{ijlf}]$ are the genotype effects representing the molecular marker information at l = HLA-A, HLA-B, HLA-DRB1, each locus being considered and analyzed separately.

An assumption of the conditional independence of individual disease onset times given the above hazard model parameters leads to the likelihood expression for all data:

$$\prod_{i=1}^{I}\prod_{j=1}^{J_i}\left[\prod_{a=0}^{X_{ij}-1}\left(1-\lambda_{ija}\right)\right]\lambda_{ijX_{ij}}^{\delta_{ij}}\left(1-\lambda_{ijX_{ij}}\right)^{1-\delta_{ij}}$$

$$\times \prod_{k_1:b_{k_1}\in(c_0-w,c_1-w)}\left(1-\left[S_{k_1}\left(c_0-b_{k_1};\beta\right)-S_{k_1}\left(15;\beta\right)\right]\right)$$

$$\times \prod_{k_2:b_{k_2}\in(c_1-w,c_0)}\left(1-\left[S_{k_2}\left(c_0-b_{k_2};\beta\right)-S_{k_2}\left(c_1-b_{k_2};\beta\right)\right]\right)$$

$$\times \prod_{k_3:b_{k_3}\in(c_0,c_1)}S_{k_3}\left(c_1-b_{k_3};\beta\right).$$

The next three factors in (2) are the contributions of the individuals (indexed here with k_1, k_2 and k_3) in the background population, not belonging to the DiMe families and considered individually, who were born between $c_0 - w$ and c_1 and who therefore were at risk of being diagnosed with T1D in the "ascertainment window" (cf. Figure 6.). Technical details of the calculations of the Bayesian modeling and calculations of the marginal survival function in likelihood above are given in the paper IV.

5.1.3 A transmission distortion model for incidence of Type 1 diabetes

In the following, k is used to denote the genotype ($\{k = 1,2,3\}$ for genotypes *AA*, *Aa* and *aa*, respectively), b the birth cohort and τ the transmission probability. Let $q_k^{(t)}$ denote the genotype frequencies of genotype k in generation t, and let $r_A^{(t)}$ and $r_a^{(t)}$ denote the allele frequency of '*A*' and '*a*' in generation t. The expected new genotype frequencies in generation $t+1$ can be derived from the basic theorems of the population genetics. As the incidence of

T1D in our data depends on the genetic susceptibility in children aged 14 years or under, the genetic change should be calculated individually for every birth cohort. They are now treated as genotype frequencies of the distinct annual birth cohorts, with the interval between two consecutive generations chosen to be 25 years. In order to obtain the genotype frequencies for annual birth cohorts between these generations, a linear approximation of the generation specific genotypes was used, giving then $q_k(b+1) = q_k(b) + \left(\frac{q_k^{(t+1)} - q_k^{(t)}}{25} \right)$ where q_{bk} is the genotype frequency in the birth cohort born in year b. The rate of change of the allele and genotype frequencies depends on the deviation of τ from the Mendelian expectation, 0.5. The incidence is now a function of the birth cohort genotype frequencies and the penetrance parameters.

The following notation was used: i= calendar year, j= age, $b=i-j$= year of birth (of a birth cohort), N_b= size of the birth cohort obtained from the national population registry (constant), d_{ij}= number of new cases of T1D in year i in the j-years-old, N_{ijk}=number of genotype k carriers in year i in age class j, $q_{bk} = q_{(i-j)k}$=the frequency of genotype k in the cohort born in year b, λ_{ijk}=penetrance for genotype k in year i at age j, q_0=frequency of allele A in the year in which insulin treatment was introduced (1930).

The observed data consist of $\{d_{ij},\ N_{ij}\ ,\ i = 65,...,96,\ j = 0,...,14\}$. In order to reduce the number of parameters to be estimated, we suppose that λ_{ijk} does not depend on i, i.e. $\lambda_{ijk}=\lambda_{jk}$ and further, that λ_{jk} is constant in each of the age groups 0-4.99, 5-9.99, 10-14.99. We index these three age groups by j=1,2,3. Since T1D is a rare disease and we assume that the numbers of new cases in each (i,j,k) cell are mutually independent, it is natural to have $d_{ij} \sim \text{Poisson}(\mu_{ij})$, where $\mu_{ij} = q_{1b}\lambda_{j1} + q_{2b}\lambda_{j2} + q_{3b}\lambda_{j3}$. The log-likelihood is then $\sum_i \sum_j \left[d_{ij}\, log(\mu_{ij}) - \mu_{ij} - log(d_{ij}!) \right]$, where μ is a function of λ_{jk} ($j,k=1,2,3$), τ and q_0.

5.1.4 Long-term survival model for genetic association

To develop a model for the genetic association between susceptibility and age at onset of disease and observed markers, some notation is first introduced. Let $a_l = (1,...,A_l)$ index alleles at locus $l=1,...L$, corresponding loci *A,B,DR* and let $\alpha_l = (\alpha_{l1},...,\alpha_{lA_l})$ be a $A_l \times 1$ vector of regression coefficients for the allele effects at locus *l* on the susceptibility. Further, let $\alpha_h = (\alpha_1,...,\alpha_H)$ be a $H \times 1$ vector of regression coefficients for the three locus haplotype effects on the susceptibility, where H indexes all possible three locus haplotypes: 1,1,1, …., A_A,A_B,A_{DR}. Regression coefficients for the age at onset are denoted respectively: $\beta_l = (\beta_{l1},...,\beta_{lA_l})$ for the allele effects at locus *l* and $\beta_h = (\beta_1,...,\beta_H)$ haplotype effects. The vectors of the observed marker genotypes at locus *l* $g_{ij}^l = (g_{ijm}^l, g_{ijf}^l)$ are coded as a $n \times A_l$ matrix Z_l^α, with each row Z_{ijl}^α being a vector of the genotype scores of locus *l* for the susceptibility. Here, the upper index α indicates that this matrix codes for the susceptibility effects. Respectively, a matrix Z_l^β of the locus specific genotype scores that are used to model the age at onset effects, where each row Z_{ijl}^β is a vector of the genotype scores for subject *j* in family *i*. The two matrixes Z_l^α and Z_l^β allow different dominance effects of the loci *l* on the susceptibility and age at onset. Respectively $n \times H$ matrixes Z_h^α and Z_h^β are defined using haplotype scores to model dominance of the haplotypes. When susceptibility is modelled with multiplicative model, genotype scores are the number of alleles/haplotypes carried by a subject *j* in family *i* (0 not a carrier, 1 one copy of the allele and 2 is subject carries two copies of the allele of interest). Alternative dominance models are presented by Thomas et al. and Cordell et al. (56, 57).

In papers III and V we choose to model the probability density of dichotomous partially latent susceptibility D_{ij} for subject *j* in family *i* using logit-function: $\text{logit}\left(p\left(D_{ij} \mid g_{ij};\theta\right)\right) = \ln\left(\frac{p\left(D_{ij} \mid g_{ij};\theta\right)}{1-p\left(D_{ij} \mid g_{ij};\theta\right)}\right) = \xi_{ij}$ and let ξ be a (n×1) vector of linear predictors of subject *j* in family *i*. In paper III we model the genetic association of T1D susceptibility with HLA marker data using logistic regression model for the allele effects: $\xi = \alpha_0 + \sum_l Z_l^\alpha \alpha_l$, here the model parameters of interest are $\theta = (\alpha_0, \alpha_1,...,\alpha_l)$. In the

haplotype model the linear predictor for subject j in family i is $\xi = \alpha_0 + Z_h^{\alpha}\alpha_h$, where $\theta = (\alpha_0, \alpha_h)$ and linear predictor for the joint allele and haplotype effects model is $\xi = \alpha_0 + \sum_l Z_l^{\alpha}\alpha_l + Z_h^{\alpha}\alpha_h$, where $\theta = (\alpha_0, \alpha_e, \alpha_1, ..., \alpha_l, \alpha_h)$. Age at onset is modelled with 2-parameter gamma distribution: $f(Y_{ij} \mid g_{ij}; a_{ij}, s)$, where the shape parameter is in case of allele effects model $log(a) = \beta_0 + \sum_l Z_l^{\beta}\beta_l$, haplotype $log(a) = \beta_0 + Z_h^{\beta}\beta_h$ and joint model $log(a) = \beta_0 + \sum_l Z_l^{\beta}\beta_l + Z_h^{\beta}\beta_h$. For the age at onset we used dominant model in all three fitted models in paper III. Genotype/haplotype scores for the allele effects/haplotype were assigned to the elements of the matrix Z_{ijl}^{β} and Z_{ijh}^{β} according to dominant model, because multiplicative assumption of allele/haplotype effects on the mean age at onset was considered too restrictive. Details of the likelihood function for the ascertained sibships and construction of the joint probability density of siblings family *i* are given in paper III, but in principle was constructed using conditional likelihood described earlier in this work. The simulation study in paper III showed that the existence of non-susceptible subjects in the population may lead to biased estimates of genetic association, when the disease and age-at-onset parameters are analysed separately.

5.1.5 Modelling familial aggregation and segregation in long-term survivors

In order to study clustering of nephropathy of T1D in Finnish families statistical models for the aggregation and segregation were developed. The aim was to estimate the proportion of subjects susceptible to diabetic nephropathy while accounting for unmeasured familial factors (61). This was achieved by using latent variables to account for familial dependencies both in susceptibility and age at onset.

5.1.5.1 Shared LTS random effects model

In order to introduce the LTS shared random effects model, some notation is first laid out. Let $\alpha = (\alpha_1, ..., \alpha_p)$ be a (1×p) vector of regression coefficients for the effects of the covariates p-

observed covariates $z_{ij} = (z_{ij1},...,z_{ijp})$ on the susceptibility and $\beta = (\beta_1,...,\beta_p)$ corresponding vector for covariates influencing the age at onset. The a_i and b_i are the unobserved random effects of family i that describe the variation between families (sibships), quantified by variances σ_α^2 and σ_β^2, for the susceptibility and age at onset that is not explained by other covariates in the shared LTS model. The probability density of the age at onset is modelled with a 2-parameter Weibull density: $f(T_{ij};\gamma_0,\mu_{ij}) = (\gamma_0/\mu_{ij})(T_{ij}/\mu_{ij})^{\gamma_0-1} exp(-(T_{ij}/\mu_{ij})^{\gamma_0})$, where γ_0 is the shape and μ_{ij} is the scale parameter. In the shared random effects model we simply assume

$$D_{ij} \sim \text{Bernoulli}(p_{ij}),\ \ \text{logit}(p_{ij}) = \alpha_0 + \alpha' z_{ij} + a_i$$

$$T_{ij} \mid D_{ij} = 1 \sim \text{Weibull}(\gamma_0, \mu_{ij}),\ \ \mu_{ij} = \exp(\beta_0 + \beta' z_{ij} + b_i)$$

The Weibull function was chosen on the basis of statistical simplicity in the long-term survivor model. In the shared random effects model, the parameters of interest are $\theta = (\alpha_0, \sigma_\alpha^2, \beta_0, \sigma_\beta^2, \gamma_0)$, where α_0 is the risk of susceptibility in the base population and the parameters of the age at onset are: β_0 is the "baseline" age at onset scale parameter of the Weibull distribution, $a_i \sim N(0, \sigma_\alpha^2)$ is the random effect of the susceptibility, $b_i \sim N(0, \sigma_\beta^2)$ is the random effect of the scale parameter of the Weibull age at onset distribution and γ is the inverse of the shape parameter. Simulation study of sibships with long-term survivors showed that all coverage rates of 95% credible intervals from the shared LTS model were at least at the nominal level. As expected the model is sensitive to the choice of informative priors for the variances of the random effects. However, with larger families the estimates had better coverage even with informative priors.

5.1.5.2 ACE-LTS model for nuclear families

In the ACE-LTS variance component model (cf. Figure 7) we define variance components of the additive genetic, common family environment and common sibling environment separately for susceptibility $(\sigma_\alpha^2, \sigma_{vc}^2, \sigma_{vcs}^2)$ and age at onset $(\sigma_{at}^2, \sigma_{vct}^2, \sigma_{vcst}^2)$. We use the reparametrization introduced by Burton et al. (62) in the generalized mixed linear model

framework. The random effect b_i in the model on previous page is replaced by independent additive random latent variables F_i, G_i, H_i, $\mathrm{R}^{\mathrm{P}}_{\mathrm{ij}}$ and $\mathrm{R}^{\mathrm{C}}_{\mathrm{ij}}$ for susceptibility and FT_i, GT_i, HT_i, $\mathrm{RT}^{\mathrm{P}}_{\mathrm{ij}}$ and $\mathrm{RT}^{\mathrm{C}}_{\mathrm{ij}}$. These describe the effects of additive polygenic, common family environment and shared sibling environment. Trait (susceptibility and age-at-onset) variation can be factored to variance components by replacing b_i for fathers with following: $FT_i + GT_i + \mathrm{RT}^{\mathrm{P}}_{\mathrm{ij}}$, for mothers $FT_i - GT_i + \mathrm{RT}^{\mathrm{P}}_{\mathrm{ij}}$ and for children $FT_i + HT_i + \mathrm{RT}^{\mathrm{C}}_{\mathrm{ij}}$. Corresponding random effects of a_i for susceptibility are replaced with $F_i + G_i + \mathrm{RT}^{\mathrm{P}}_{\mathrm{ij}}$, $F_i - G_i + \mathrm{RT}^{\mathrm{P}}_{\mathrm{ij}}$ and $F_i + H_i + \mathrm{RT}^{\mathrm{C}}_{\mathrm{ij}}$.

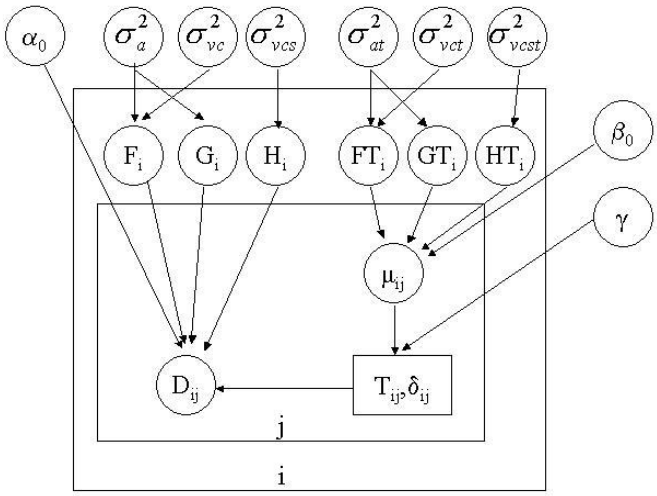

Figure 7. Directed acyclic graph of additive genetic variance component (random effects) long-term survivor model for nuclear families. Here D_{ij} is the binary susceptibility variable and the pair describes the failure and the age at onset for subject j in family i. The $\sigma^2_{\alpha}, \sigma^2_{vc}$ and σ^2_{vcs} are the variance components of the additive genetic, common family environment and common sibling environment separately for susceptibility and $\sigma^2_{at}, \sigma^2_{vct}$ and σ^2_{vcst} are the corresponding variance components for the age at onset.

The choice of priors and the formulation of the joint posterior density is given in paper IV, together with Random-walk MCMC algorithm to obtain samples from the joint posterior of ACE-LTS model. In the simulation study we observed reasonable coverage rates for all parameters with non-informative priors. However, the ACE-LTS model is even more sensitive than the shared LTS model to the choice of priors of variances of random effects since informative priors clearly resulted in even poorer coverage rates.

5.2 Results of the data analysis

5.2.1 Transmission of HLA-A, HLA-B and HLA-DRB1 alleles

To analyze transmission of alleles at the HLA-A, HLA-B and HLA-DRB1 loci we made series of the goodness-of-fit tests: maternal alleles at the A locus (χ^2=13.7, 13 df., p=0.40), paternal alleles at the A locus (χ^2=17.06, 12 df., p=0.15), maternal alleles at the B locus (χ^2=39.59, 35 df., p=0.27) and paternal alleles at the B locus (χ^2=24.84, 31 df., p=0.78), maternal at the DR locus (χ^2=17.4, 27 df., p=0.92) and paternal at the DR locus (χ^2=13.3, 17 df., p=0.72), where single allele transmissions are estimated simultaneously and there are less parameters to be estimated than in the global test. Single allele transmission probabilities were calculated to reveal individually the alleles inherited in a non-Mendelian fashion. Even though some single allele transmission frequencies were statistically significantly different from 50%, these findings cannot be considered conclusive as the significance levels have not been corrected for multiple comparisons. Therefore, we conclude that the existence of strong transmission distortion in the considered loci is excluded by our study.

5.2.2 Transmission distortion of susceptibility alleles and incidence of T1D in Finland

Two models were fitted using Poisson model described earlier: one with transmission probability fixed at 0.5 (M0) and another where transmission probability was estimated from the data (M1). According to the likelihood ratio test the model M2 resulted in a significantly better fit (χ^2=131.12, 1 df., p<0.001). The point estimate of transmission distortion τ was 0.86 with estimated gene frequencies (0.06, 0.37, 0.57). In Figure 8 the observed incidence of T1D and expected incidence based on the models M1 and M2 are plotted.

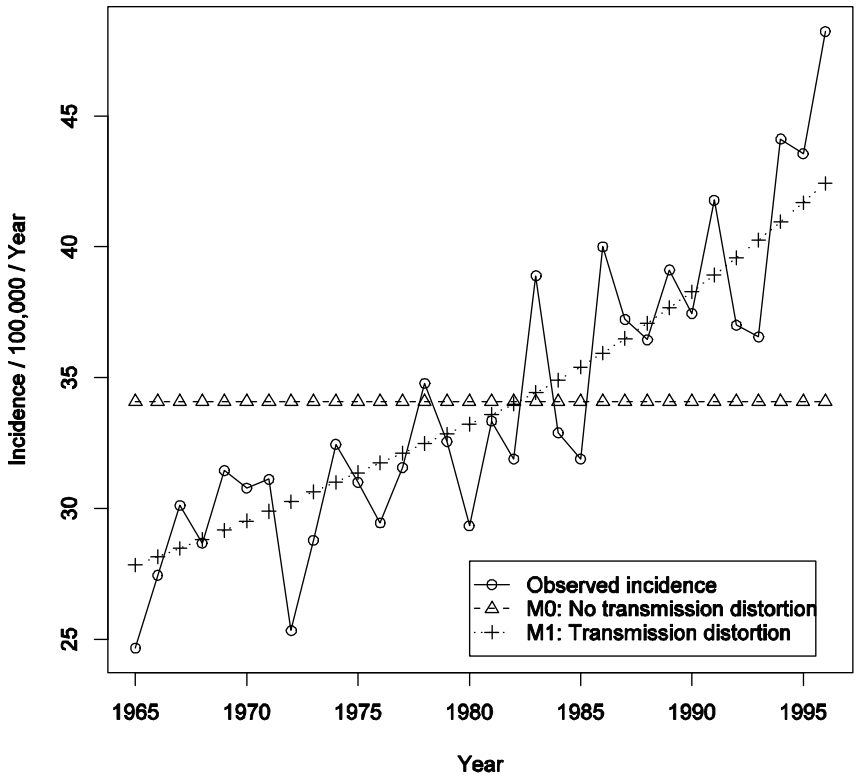

Figure 8. The observed incidence of T1D in Finland from 1965 to 1996 and expected incidence under two models: (M0) no transmission distortion (transmission probability fixed at 0.5) and (M1) allowing transmission distortion (transmission probability has been set to the estimated value of 0.998).

The fitted model allows only non-Mendelian transmission of susceptibility gene alleles as an explanation of the observed increase in the incidence of T1D. As shown by the Figures 1 a-b in paper II, the alleles frequencies may increase in the course of few generations with increasing transmission probability. It is obvious from these simple theoretical considerations that, given the current knowledge of the risk alleles of T1D, only an extreme transmission distortion of susceptibility alleles with high risk could explain the observed rapid increase in T1D from 1965 to 1988 in Finland.

5.2.3 HLA-A, HLA-B and HLA-DRB1 associations with susceptibility and age at onset of T1D

To study the association of HLA-A, HLA-B and HLA-DR locus with susceptibility and age at onset of T1D we fitted four models to the DiMe sibships data at A, B and DR locus: locus-specific allele main effects (M0), joint allele main effects (M1), three locus haplotype effects (M2) and three locus joint allele main and haplotype effects (M3). No statistically significant susceptibility or age at onset effects were detected either at the A or at the B locus. At the DR locus both DR3 (regression coefficient 1.17, p=0.012) and DR4 (regression coefficient 1.29, p=0.005) were associated with susceptibility to T1D. These correspond to increased susceptibility of 10.7% with DR3 heterozygote, and 12.5% with DR4 heterozygote with respect to the overall mean. The mean age at onset of a DR3 carrier was 18.5 years (p=0.012). The only significant allele main effect with the age at onset of T1D was observed for DR6 (mean age at onset 12.1, p=0.010).

To assess independent allele effects and take into account the linkage disequilibrium between A, B and DR loci, we performed the joint allele main effects analysis of A, B and DR loci (M1), and the joint allele main and three locus haplotype effects analyses (M3). Only B62 was significantly associated with increased susceptibility to T1D, after adjusting for haplotype effects (p=0.016). At the B locus, B8 carriers (p=0.0001) had older age at onset than all T1D cases. At the DR locus, DR3 (p=0.001) and DR6 (p=0.0233) showed significant effects with the earlier age at onset of T1D compared with the overall mean (after adjusting for haplotype effects).

When three locus haplotypes main effects were investigated we found three haplotypes associated with higher than average susceptibility (A1B8DR3; A2B62DR4; A2B8DR3). However, all these associations disappeared in the joint allele main effects and haplotype analysis (M3). Older than average age at onset was observed in A3B18DR4, and it remained significant after adjusting for allele main effects (p=4.0e-5). Due to the ascertainment correction and subsequent loss of data we have very low power to detect any significant association especially with such complex model with many parameters to be estimated.

5.2.4 Familial aggregation of susceptibility and age at onset of T1D nephropathy

In paper V, we analyzed the susceptibility and age at onset of DN by simulating the joint posterior distribution of the model parameters of three different models: (i) the measured covariates only (M0), (ii) latent random effects only (M1) and (iii) both measured covariates and variance component model (M2). Gender was assumed to influence both the susceptibility and age at onset. The age at onset of T1D was assumed to influence the time to the onset of DN, because if the age at onset of T1D was very young this might lead to an earlier development of DN than in T1D patients with a later age at onset.

The estimated proportion of male T1D patients susceptible to DN was 53.5% (46.5%; 61.1%, 95% credible interval) and among female T1D patients 37.5 % (31.2%; 45.0%) when no familial clustering was accounted for, based on the model M0. However, the difference increased when shared sibship effects were accounted for in model M2: males 58.9% (45.7%; 70.8%) and females 33.8% (23.1%; 46.5%). A large familial variation in the susceptibility of DN was reflected by the posterior mean of the variance 2.96 (0.72; 7.00). The posterior mean of the time to the onset of DN from the diagnosis of T1D was 19.9 years (17.4; 23.0). No significant difference between males and females was observed in the time to DN from the diagnosis of T1D, after taking into account the familial effects. The variance of time to DN related to familial clustering was small 0.08 (0.03; 0.17) and remained virtually unchanged when sex and age at onset of T1D were adjusted for.

5.2.5 HLA A, B and DRB1 associations with age at onset of T1D

The results concerning full likelihood based analysis of the association between T1D and HLA-A, HLA-B and HLA-DRB1 genotypes are presented using predictive probabilities of T1D free survivals. The heterozygous genotype DR3/DR4 at the DRB1 locus was associated with the lowest predictive probability of the T1D free survival to age 15, the estimate being 0.936 (0.926; 0.948, 95% credible interval), compared to the average population T1D free survival probability 0.995. The effect of DR4 homozygote was also strong with associated probability 0.962 (0.954; 0.969) of T1D free survival. Carriers of DR 3/3 genotype, with T1D predictive probability 0.994 (0.990; 0.998), were close to the population average probability

of T1D free survival. Carriers of DR1/DR2, a common genotype in the reference population and of DR2/DR6, had virtually no risk of T1DM before age 15, with predictive probability for T1D free survival being very close to 1 (0.999; 1.000). At the A locus A1/9 and A2/3 carriers had the lowest predictive probability of the T1D free survival to age 15, 0.983 (0.976; 0.988) and 0.993 (0.992; 0.995) respectively. At the B-locus carriers of B8/22 and 8/15 genotypes had 0.963 (0.944; 0.977) and 0.980 (0.974; 0.984) predictive probability of the T1D free survival to age 15, while carriers of B7/35 and B5/B7 had virtually no risk of T1D before the age 15. To illustrate the age dependency of the HLA-DRB1 genotypes, predictive disease free survivals for some HLA-DRB1 high risk genotypes are plotted in Figure 9.

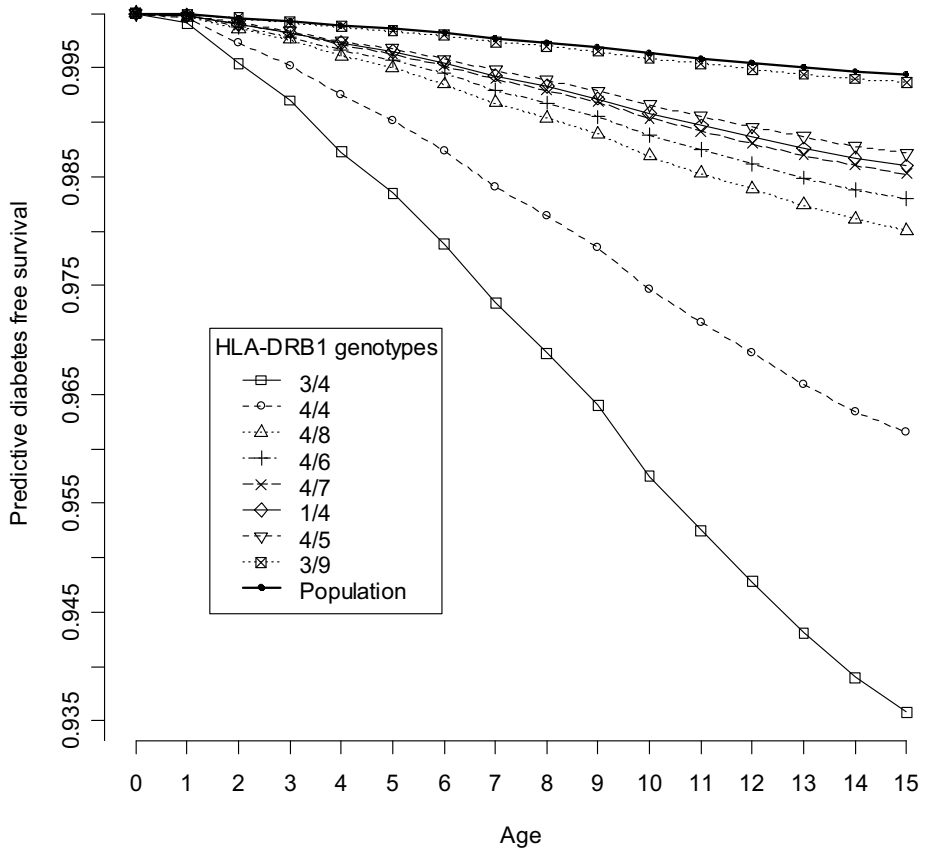

Figure 9. Predicted disease-free survival of T1D for some high risk HLA-DRB1 genotypes.

6. Discussion

It is both time and financial resources consuming effort to collect registry data which serves as data source for many future studies. Therefore, it would be atmost importance to analyze such data in a statistically coherent way. In this work, I present new statistical methods for the analysis of population-based familial data ascertained through an incident case(s) in genetic association studies of complex human diseases and for the genetic analysis of variable age at onset diseases with non-susceptible subjects. New knowledge is provided concerning the genetic effects of HLA to the T1D risk and familial aggregation of diabetic nephropathy.

The analysis of disease registry data in this thesis is based either on modelling of the disease risk in the entire population (paper IV) or restricting statistical analysis to those individuals that did not contribute to the ascertainment event (paper I). We have replicated the well known association between HLA-DR3,4 and T1D in our population-based analyses of population based registry of T1D. This is in contrast to unsuccessful replications of associations in many other complex diseases (67). Further, we do not find any support of non-Mendelian transmission of HLA-A, HLA-B and HLA-DRB1 alleles, when accounting for the ascertainment. We failed to reject the hypothesis of Mendelian transmission of alleles at the HLA-A, HLA-B and HLA-DRB1 loci in paper I, despite the results of earlier studies suggesting non-Mendelian transmission (68,69). Partly these earlier findings can be explained by lack of proper ascertainment correction and a proper analysis would use large population-based random sample. As we observed, non-Mendelian transmission of T1D susceptibility alleles does not look like the explanation of the observed increase of T1D incidence in Finland nor have there been proposed any other genetic mechanism that would explain this phenomenon (65). Although we were not able to make very strong conclusion concerning the association of HLA-A, HLA-B and HLA-DR serological genotypes with susceptibility and age at onset, recent findings with more accurate HLA-DRB1 data seem to support such an association (66).

Finding complex disease genes with modest effects has turned out to be much more difficult than expelling genes for monogenic diseases. With increasing ability to genotype large materials and combining various already existing data sources (disease registries, population based-biobanks and demographic data) a substantial gain can be obtained in the search for genes of complex diseases like T1D (63). One possibility to increase statistical power is to increase sample size by incorporating the whole population in the statistical analysis. Another advantage of our analysis using full likelihood is that population average incidence can be used as "natural" reference level when genotypes are assessed to confer increased or reduced risk. Population stratification has been named as one of the reasons for the failure to find or replicate genetic associations with diseases (67) and many approaches have been proposed (33,70) and it should be further studied with respect to ascertainment of family data.

A complex variable age at onset phenotype with non-susceptible subjects is modelled and estimated both with observed genetic marker data (paper III) and without (paper V). Long-term survival segregation model revealed a substantial familial contribution to the susceptibility of T1D related nephropathy and suggestive findings of marker alleles at the HLA region to be associated with both to susceptibility and age at onset of T1D were detected. In search for complex disease genes one alternative is to explore sophisticated statistical models for complex phenotypes, like long-term survivor model. This may be motivated by the emerging new subtypes of T1D during past decades and more is to be expected with the rapid development of our ability to assess metabolic pathways of diseases. These more refined phenotypes may be more accurate in search for disease-causing genes, requiring more complex statistical models. A long-term survivor model, in which age at onset distribution can be used to make inference concerning susceptibility, has been recently applied to studies of genetic association (23,24). However, in neither of these studies the random variable describing susceptibility status was sampled as part of the model like the Bayesian model presented in paper V. Neither was the probability of susceptibility further modelled as a function of latent or observed variables. As shown in paper V, the hierarchical LTS model is sensitive to the choice of priors (64), especially when estimating variance components of the susceptibility and this should be further explored. Although segregation analysis has been applied less often since the introduction of modern genetic technology, it remains a valuable tool when combined with modern statistical models for genetic epidemiology as describing influence of genetic and environmental factors at the population

level and possibly evaluating the importance of residual variation after some functional genes has been identified.

With the development of large disease or population-based biobanks (63) and increasing accuracy of genetic information (HapMap project) computer intensive statistical methods, such as Bayesian methods applied in this thesis, can be utilized to analyze such a considerable amount of data. As this thesis provides new ways to take into account the collection of registry data and complex variable age at onset disease, it is hoped that future registry-based studies will benefit from the statistical methods proposed here.

Acknowledgements

This study was carried out at the Division of Biometry, Rolf Nevanlinna Institute during 1997-2001 and later at the Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute and finally jointly at the Department of Public Health, School of Medicine, University of Helsinki and Department of Mathematics and Statistics, University of Helsinki. I thank the Directors of these institutes, Jukka Sarvas and Lassi Päivärinta at the RNI, Jussi Huttunen and Pekka Puska at the NPHI and Jaakko Kaprio at the Department of Public Health for providing facilities to carry out of the work.

I would like to express my sincere gratitude to my supervisors – professor Elja Arjas, who during the years taught me a lot about statistical inference and who's way of thinking about the problems facing us during the this work was impressive. His persistence and realism at the finishing stages was invaluable in the completion of this work. The group led by him at the RNI is really a fantastic place to be when studying statistical genetics with a lot of interaction between scientists working in different fields. A special thank goes to my long time supporter professor Jaakko Tuomilehto, who originated my studies of statistical methods in genetic epidemiology at the University of Southern California. His support and enthusiasm in research has been vital for me over the years. My warm thanks to professor Timo Hakulinen for meeting with me regularly at the Finnish Cancer Registry, it helped me to write two articles for this dissertation.

The reviewers of this study, professor Esa Läärä and professor Suvi Virtanen are gratefully acknowledged and I express my gratitude for the careful and well-pointed corrections that they made in addition to the constructive suggestions both in the formulas and the text that helped significantly to improve the summary.

I express my gratitude to my collaborators Dr. Päivi Onkamo and Dr. Elena Moltchanova. Päivi was the persistent person who pushed out the papers and her genetic point of view was necessary for the first two papers during my studies at the RNI. It was absolutely fantastic to

be a part of the group of people at the RNI: Pasi Korhonen, Samuli Ripatti, Mikko Sillanpää, Pekka Uimari, Kari Auranen and Jukka Ranta, with whom I had several discussions about various statistical topics. This group was definitely a great place to learn various aspects of biostatistics, including Bayesian statistical methods. Later, it has been pure fun to work with a talented young scientist such as Elena Moltchanova to pursue interesting ideas of statistical genetics. My interest in statistical genetics started at the University of Southern California and I wish to express my gratitude to professor Duncan Thomas for showing a true passion for statistical methods in genetic epidemiology and professor Bryan Langholz for all the mental and material support during my years at the University of Southern California. Genetic Analysis Workshops were absolutely fantastic mind broadening events for starting researcher and I still feel privileged to have been able to take part in the meetings. I like to thank my co-authors Sirkka-Liisa Varvio, Eva Tuomilehto-Wolf, Valma Harjutsalo, Nella Lehti, Andrew Thomas, Jukka Partanen and Jurkka Näsänen for their time and effort in various phases when preparing the papers of this dissertation. I wish to mention Dr. Marjatta Karvonen, who has always provided me excellent support in listening to any problems that I have had in my mind, starting from the days of the joint research on the epidemiology of Chlamydia Pneumonia. My co-worker Laura Haapala has patiently listened to my explanations of the statistical problems across the table. This work has required considerable amount of my time at the Department of Public Health and would have not been possible without the wonderful attitude of professor Seppo Sarna. There are many people at the Diabetes Unit of the National Public Health Institute, who have made this work possible and are not listed - my warm thanks to all of them for the support they provided me over the years.

My special thanks goes to my wife Maaria and children Joonas, Krista and Teemu. I express my deep appreciation to my wife for the wonderful support she provided at the difficult moments of this work. My warm thanks to my mother, who at early age supported my interest in fauna and flora. I also want to remember my late grandmother, who passed away quite some time ago. The discussions with her taught a young man what is meant by persistence and respect. Finally I would like to note, that I was thrilled to found out that my defence takes place exactly one hundred years after my great-grandfather Pitkäniemi FM defended his dissertation at the University of Helsinki in May 1908.

The financial support provided by the Academy of Finland, Biometry Group of the University of Southern California, ComBi graduate school and Doctoral Program of Public Health (DPPH) are acknowledged with thankfulness.

References

1. Fisher R (1934) The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugenics 6:13-25.

2. Clayton DG (2003) Conditional likelihood inference under complex ascertainment using data augmentation. Biometrika 90:976-981.

3. Burton PR (2003) Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using gibbs sampling. Genet Epidemiol 24:24-35.

4. Gauderman WJ, Thomas DC (1994) Censored survival models for genetic epidemiology: A gibbs sampling approach. Genet Epidemiol 11:171-188.

5. Farewell VT (1977) A model for binary variable with time censored observations. Biometrika 64:43-46-46.

6. Maller R, Zhou X (1996) Survival analysis with long-term survivors. John Wiley & Sons Ltd, England.

7. Tuomilehto-Wolf E, Tuomilehto J (1991) HLA antigens in insulin-dependent diabetes mellitus. Ann Med 23:481-488.

8. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM (1994) A genome-wide search for human Type 1 diabetes susceptibility genes. Nature 371:130-136.

9. Hirschhorn JN (2003) Genetic epidemiology of Type 1 diabetes. Pediatr Diabetes 4:87-100.

10. Steenkiste A, Valdes AM, Feolo M, Hoffman D, Concannon P, Noble J, Schoch G, Hansen J, Helmberg W, Dorman JS, Thomson G, Pugliese A, 13th IHWS 1 Diabetes Component participating investigators (2007) 14th international HLA and immunogenetics workshop: Report on the HLA component of Type 1 diabetes. Tissue Antigens 69 Suppl 1:214-225.

11. Thomas DC (2004) Statistical methods in genetic epidemiology. Oxford University Press, Inc., New York.

12. Dawn Teare M, Barrett JH (2005) Genetic linkage studies. Lancet 366:1036-1044.

13. Redondo MJ, Fain PR, Eisenbarth GS (2001) Genetics of type 1A diabetes. Recent Prog Horm Res 56:69-89.

14. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. (1999) Nature 401:921-923.

15. Lunn DJ, Whittaker JC, Best N (2006) A Bayesian toolkit for genetic association studies. Genet Epidemiol 30:231-247.

16. Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo methods in practice. Chapman and Hall, London.

17. Thomas A, O'Hara B, Ligges U, and Sturtz S. (2006) Making BUGS Open. *R News* 6(1): 12-17.

18. Vaupel JW, Manton KG, Stallard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16:439-454.

19. Siegmund KD, Province MA, Higgins M, Williams RR, Keller J, Todorov AA (1998) Modeling disease incidence rates in families. Epidemiology 9:557-562.

20. Siegmund KD, Todorov AA (2000) Linkage analysis for diseases with variable age of onset. Hum Hered 50:205-210.

21. Sun W, Li H (2004) Ascertainment-adjusted maximum likelihood estimation for the additive genetic gamma frailty model. Lifetime Data Anal 10:229-245.

22. Meehyang C, Schenker N, Taylor JMG, Zhuang D (2001) Survival analysis with long-term survivors and partially observed covariates. Can J Stat 29:421-436.

23. Wienke A, Lichtenstein P, Yashin AI (2003) A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. Biometrics 59:1178-83; discussion 1184-5.

24. Locatelli I, Rosina A, Lichtenstein P, Yashin AI (2007) A correlated frailty model with long-term survivors for estimating the heritability of breast cancer. Stat Med 26:3722-3734.

25. Guo SW (1998) Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet 63:252-258.

26. Morton NE (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1-16.

27. Elston RC, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31:62-69.

28. Kraft P, Thomas DC (2000) Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66:1119-1131.

29. Olson JM, Cordell HJ (2000) Ascertainment bias in the estimation of sibling genetic risk parameters. Genet Epidemiol 18:217-235.

30. Self SG, Longton G, Kopecky KJ, Liang KY (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53-61.

31. Langholz B, Ziogas A, Thomas DC, Faucett C, Huberman M, Goldstein L (1999) Ascertainment bias in rate ratio estimation from case-sibling control studies of variable age-at-onset diseases. Biometrics 55:1129-1136.

32. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121-1131.

33. Ripatti S, Pitkäniemi J, Sillanpaa MJ (2001) Joint modeling of genetic association and population stratification using latent class models. Genet Epidemiol 21:S409-14.

34. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933-1944.

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

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

37. Kadowaki T, Kadowaki H, Mori Y, Tobe K, Sakuta R, Suzuki Y, Tanabe Y, Sakura H, Awata T, Goto Y (1994) A subtype of diabetes mellitus associated with a mutation of mitochondrial DNA. N Engl J Med 330:962-968.

38. Tuomi T, Carlsson A, Li H, Isomaa B, Miettinen A, Nilsson A, Nissen M, Ehrnstrom BO, Forsen B, Snickars B, Lahti K, Forsblom C, Saloranta C, Taskinen MR, Groop LC (1999) Clinical and genetic characteristics of type 2 diabetes with and without GAD antibodies. Diabetes 48:150-157.

39. Soltesz G, Patterson CC, Dahlquist G, EURODIAB Study Group (2007) Worldwide childhood Type 1 diabetes incidence--what can we learn from epidemiology? Pediatr Diabetes 8 Suppl 6:6-14.

40. Karvonen M, Pitkäniemi M, Pitkäniemi J, Kohtamaki K, Tajima N, Tuomilehto J (1997) Sex difference in the incidence of insulin-dependent diabetes mellitus: An analysis of the recent epidemiological data. world health organization DIAMOND project group. Diabetes Metab Rev 13:275-291.

41. Guo SW, Tuomilehto J (2002) Preferential transmission of Type 1 diabetes from parents to offspring: Fact or artifact? Genet Epidemiol 23:323-334.

42. Knip M, Akerblom HK (2005) Early nutrition and later diabetes risk. Adv Exp Med Biol 569:142-150.

43. Virtanen SM, Rasanen L, Ylonen K, Aro A, Clayton D, Langholz B, Pitkäniemi J, Savilahti E, Lounamaa R, Tuomilehto J (1993) Early introduction of dairy products associated with increased risk of IDDM in Finnish children. The Childhood in Diabetes in Finland Study Group. Diabetes 42:1786-1790.

44. Filippi C, von Herrath M (2005) How viral infections affect the autoimmune process leading to Type 1 diabetes. Cell Immunol 233:125-132.

45. Parslow RC, McKinney PA, Law GR, Staines A, Williams R, Bodansky HJ (1997) Incidence of childhood diabetes mellitus in Yorkshire, northern England, is associated with nitrate in drinking water: An ecological analysis. Diabetologia 40:550-556.

46. Karvonen M, Viik-Kajander M, Moltchanova E, Libman I, LaPorte R, Tuomilehto J (2000) Incidence of childhood Type 1 diabetes worldwide. Diabetes Mondiale (DiaMond) project group. Diabetes Care 23:1516-1526.

47. Tuomilehto J, Karvonen M, Pitkäniemi J, Virtala E, Kohtamaki K, Toivanen L, Tuomilehto-Wolf E (1999) Record-high incidence of type I (insulin-dependent) diabetes mellitus in Finnish children. The Finnish childhood type I diabetes registry group.[see comment]. Diabetologia 42:655-660.

48. 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.

49. Moltchanova E, Penttinen A, Karvonen M (2005) A hierarchical bayesian birth cohort analysis from incomplete registry data: Evaluating the trends in the age of onset of insulin-dependent diabetes mellitus (T1DM). Stat Med 24:2989-3004.

50. Todd JA, Bell JI, McDevitt HO (1987) HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature 329:599-604.

51. Morel PA, Dorman JS, Todd JA, McDevitt HO, Trucco M (1988) Aspartic acid at position 57 of the HLA-DQ beta chain protects against type I diabetes: A family study. Proc Natl Acad Sci U S A 85:8111-8115.

52. 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.

53. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J (2003) Genetic liability of Type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: A nationwide follow-up study. Diabetes 52:1052-1055.

54. Paterson AD (2006) Genetic epidemiology of Type 1 diabetes. Curr Diab Rep 6:139-146.

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

56. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of Type 1 diabetes. Nat Genet 39:857-864.

57. Tuomilehto J, Lounamaa R, Tuomilehto-Wolf E, Reunanen A, Virtala E, Kaprio EA, Akerblom HK (1992) Epidemiology of childhood diabetes mellitus in finland--background of a nationwide study of type 1 (insulin-dependent) diabetes mellitus. The childhood diabetes in Finland (DiMe) study group. Diabetologia 35:70-76.

58. Tuomilehto-Wolf E, Tuomilehto J, Cepaitis Z, Lounamaa R (1989) New susceptibility haplotype for Type 1 diabetes. DIME study group. Lancet 2:299-302.

59. International evaluation of cause-specific mortality and IDDM. Diabetes epidemiology research international mortality study group. (1991) Diabetes Care 14:55-60.

60. Major cross-country differences in risk of dying for people with IDDM. diabetes epidemiology research international mortality study group. (1991) Diabetes Care 14:49-54.

61. Harjutsalo V, Katoh S, Sarti C, Tajima N, Tuomilehto J (2004) Population-based assessment of familial clustering of diabetic nephropathy in Type 1 diabetes. Diabetes 53:2449-2454.

62. Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ (1999) Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and gibbs sampling. Genet Epidemiol 17:118-140.

63. Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR (2005) Genetic epidemiology and public health: Hope, hype, and future prospects. Lancet 366:1484-1498.

64. Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1:515-534.

65. Rewers M, Zimmet P (2004) The rising tide of childhood Type 1 diabetes--what is the elusive environmental trigger? Lancet 364:1645-1647.

66. Valdes AM, Erlich HA, Noble JA (2005) Human leukocyte antigen class I B and C loci contribute to Type 1 diabetes (T1D) susceptibility and age at T1D onset. Hum Immunol 66:301-313.

67. Sillanpaa MJ, Auranen K (2004) Replication in genetic studies of complex traits. Ann Hum Genet 68:646-657.

68. Tuomilehto-Wolf E, Tuomilehto J and the DiMe Study Group: Preferential zygotic assortment of the HLA-A2,Cw1,B56,DR4,DQ8 haplotype contributes to the high incidence of insulin-dependent diabetes mellitus (IDDM) in Finland (abstract), IGES 1993.

69. 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.

70. Sillanpaa MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P (2001) Bayesian association mapping for quantitative traits in a mixture of two populations. Genet Epidemiol 21 Suppl 1:S692-9.