UMEÅ UNIVERSITY MEDICAL DISSERTATIONS
ISSN 0346-6612-946; ISBN 91-7305-836-X

MAPPING GENETIC DISEASES IN NORTHERN SWEDEN

Elísabet Einarsdóttir

Department of Medical Biosciences
Unit of Medical and Clinical Genetics
Umeå University

Umeå 2005

Department of Medical Biosciences,
Unit of Medical and Clinical Genetics
Umeå University
SE-90187 Umeå, Sweden

ISSN 0346-6612-946

ISBN 91-7305-836-X

Cover illustration by Nicolas Heluani

Layout by Tommy Sund

Printed by Solfjädern Offset AB in Umeå, Sweden

TABLE OF CONTENTS

ABBREVIATIONS

AITD	autoimmune thyroid disease
BD	Bothnia dystrophy
BMI	body-mass index
CAPN10	calpain-10
cM	centimorgan
CTLA4	cytotoxic T-lymphocyte antigen 4
DNA	deoxyribonucleic acid
GD	Graves' disease
GWS	genome-wide scan
HLA	human leukocyte antigen
HSAN	hereditary sensory and autonomic neuropathy
HT	Hashimoto's thyroiditis
IDDM	insulin-dependent diabetes mellitus (= T1DM)
IGT	impaired glucose tolerance
LD	linkage disequilibrium
LOD	logarithm of odds, lodscore
Mb	megabase (one million bases)
MHC	major histocompatibility complex (called HLA in humans)
MODY	maturity-onset diabetes of the young
NGFB	nerve-growth factor beta
NIDDM	non-insulin-dependent diabetes mellitus (=T2DM)
NOD	non-obese diabetic (mice)
SNP	single-nucleotide polymorphism
T1DM	type 1 diabetes mellitus (=IDDM)
T2DM	type 2 diabetes mellitus (=NIDDM)

ABSTRACT

The population of northern Sweden has previously been shown to be well suited for the mapping of monogenic diseases. In this thesis we have tested the hypothesis that this population could also be used for efficient identification of risk genes for common diseases. In ***Paper I*** we have hypothesised that despite the admixture of Swedish, Finnish and Sami, the northern Swedish population consists of sub-populations geographically restricted by the main river valleys running through the region. This geographic isolation, in combination with founder effects and genetic drift, could represent a unique resource for genetic studies. On the other hand, it also underlines the importance of accounting for this e.g. in genetic association studies. To test this hypothesis, we studied the patterns of marriage within and between river valley regions and compared allelic frequencies of genetic markers between these regions. The tendency to find a spouse and live in the river valley where one was born is strong, and allelic frequencies of genetic markers vary significantly between adjacent regions. These data support our hypothesis that the river valleys are home to distinct sub-populations and that this is likely to affect mapping of genetic diseases in these populations. In ***Paper II***, we tested the applicability of the population in mapping HSAN V, a monogenic disease. This disease was identified in only three consanguineous individuals suffering from a severe loss of deep pain perception and an impaired perception of heat. A genome-wide scan combined with sequencing of candidate genes resulted in the identification of a causative point mutation in the nerve growth factor beta (NGFB) gene. In ***Paper III***, a large family with multiple members affected by familial forms of type 1 diabetes mellitus (T1DM) and autoimmune thyroiditis (AITD) was studied. This syndrome was mapped to the IDDM12 region on 2q33, giving positive lodscores when conditioning on HLA haplotype. The linkage to HLA and to the IDDM12 region thus confirmed previous reports of linkage and/or association of T1DM and AITD to these loci and provided evidence that the same genetic factors may be mediating these diseases. This also supported the feasibility of mapping complex diseases in northern Sweden by the use of familial forms of these diseases. In ***Paper IV***, we applied the same approach to study type 2 diabetes mellitus (T2DM). A non-parametric genome-wide scan was carried out on a family material from northern Sweden, and linkage was found to the calpain-10 locus, a previously described T2DM-suspectibility gene on 2q37. Together, these findings demonstrate that selecting for familial forms of even complex diseases, and choosing families from the same geographical region can efficiently reduce the genetic heterogeneity of the disease and facilitate the identification of risk genes for the disease.

PUBLICATIONS

I. **Einarsdottir E**, Escher S A, Egerbladh I, Beckman L, Sandgren O, Golovleva I, Holmberg D.
The population structure of northern Sweden and its implications for mapping genetic diseases
Manuscript

II. **Einarsdottir E**, Carlsson A, Minde J, Toolanen G, Svensson O, Solders G, Holmgren G, Holmberg D, Holmberg M.
A mutation in the nerve growth factor beta gene (NGFB) causes loss of pain perception.
Hum. Mol. Genet. 2004 Apr 15; 13(8): 799-805.

III. **Einarsdottir E***, Soderstrom I*, Lofgren-Burstrom A, Haraldsson S, Nilsson-Ardnor S, Penha-Goncalves C, Lind L, Holmgren G, Holmberg M, Asplund K, Holmberg D.
The CTLA4 region as a general autoimmunity factor: an extended pedigree provides evidence for synergy with the HLA locus in the etiology of type 1 diabetes mellitus, Hashimoto's thyroiditis and Graves' disease.
Eur. J. Hum. Genet. 2003 Jan; 11(1): 81-4.

IV. **Einarsdottir E***, Mayans S*, Ruikka K, Escher S A, Lindgren P, Eliasson M, Holmberg D.
Polymorphisms in the calpain-10 gene show linkage to type 2 diabetes in a population from northern Sweden
Manuscript

* These authors contributed equally to the work

INTRODUCTION

A better understanding of the genetic factors underlying a disease may help to identify individuals at risk for this disease even before they fall ill, and a more detailed understanding of the mechanisms contributing to the disease may also lead to more effective and efficient treatment of the disease.

Finding the genetic factors contributing to a disease involves comparing patients to their healthy relatives or to control individuals, and identifying genetic factors common among patients but uncommon among healthy people. The difference between the two groups is the genetic component of the disease. This "maps the disease" to a certain part of the genome and allows for further study of the genetic factors involved.

Understanding the genetic basis of a disease is often difficult, especially if the disease is caused by a complex interaction of genetics and environment. Several strategies have been employed to simplify this task. One example is the sub-phenotype approach, where the analysis of only one of the symptoms or phenotypes of a disease is performed at a time. Working with animal models of a disease may also be a way to isolate the genetic factors underlying a disease. Using inbred mouse-strains is a particularly popular way to study human genetic diseases. Another common approach is to analyse as many patients as possible, and hope that the sheer number of samples will allow for the identification of the relevant genetic factors.

An approach that has become more and more emphasised is to strive to simplify the genetic background under investigation. The hope is that this will enable a better focus on the relevant genetic factors and a reduction in background noise. A common way to achieve this is to focus on populations that have a reduction in genetic diversity, and have been geographically or culturally isolated. The whole population is thus based on the descendants of a small number of individuals, with most of the genetic variation in the population being the variation originally present in these few individuals. Examples of these kinds of populations include the Finns, the Icelandic population, and the Sardinians.

The population of northern Sweden has a history involving a mixing of three ethnic groups followed by a dramatic expansion over the last two centuries from a founder population (Andersson-Palm 2000, Bylund 1994, Bylund 1960). While the total population is quite heterogeneous, it has been hypothesised that certain parts of the region are home to small, isolated population-groups. This thesis describes our work of characterising the sub-structures of this population and gives examples of how this may affect the mapping of genetic diseases.

BACKGROUND

GENETIC DISEASES

A disease is a visible trait or phenotype that is detrimental to the individual and is usually described by a set of diagnostic criteria defined by a consortium of specialists in the field. How common the disease is can be measured by the incidence of the disease (e.g. how many people fall ill per year) or by its prevalence (what proportion of the population has the disease at a particular time point). A disease can be largely infectious in origin (e.g. measles) or largely governed by genetic factors (e.g. albinism), but most diseases are a combination of both, and how a person responds to an infectious disease is also very often influenced by genetic factors. The term genetic disease in this thesis is used for diseases that are mainly governed by genetic factors or diseases that have a strong genetic component.

Mendelian diseases

Mendelian or monogenic diseases are genetic diseases for which the correlation between the disease and a particular mutation or mutations in one gene is very strong. These diseases are called Mendelian since they conform to the Mendelian laws of inheritance. A Mendelian or monogenic disease is usually due to one drastic mutation, causing a critical protein to be lost or the function of a protein to be severely impaired. A person carrying this mutation is likely to get the disease (the disease has a high penetrance), similarly very few individuals lacking mutations in this particular gene will succumb to the disease (the likelihood of phenocopies is low).

When looking at a family tree (or pedigree), Mendelian diseases most often show a clear dominant or recessive inheritance, one copy [dominant] or two copies [recessive] of the mutation are required for a person to be affected, and the disease can be traced through several generations. Figure 1 illustrates the inheritance of a dominant Mendelian disease within a family.

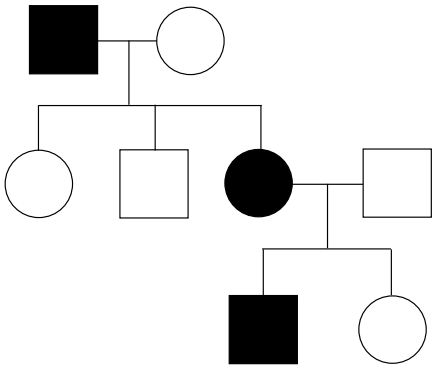

Figure 1
A simple pedigree. Squares= men, circles= women. Filled symbols= Affected individuals, open symbols= unaffected individuals.

The Mendelian diseases have traditionally been the best studied in the field of medical genetics. Finding the mutation responsible for the disease and identifying the protein that is lost or defective may be relatively easy, especially if large families with many affected individuals can be identified.

Complex diseases

Complex genetic diseases are, as the name implies, more complex than Mendelian diseases. These arise as a result of a complex interaction between genetic factors, environmental factors and lifestyle choices (smoking, age, infections, physical inactivity, etc.). The genetic factors involved in susceptibility to complex diseases are often numerous, most of them only contributing slightly to the risk of disease, and these factors may differ considerably between individuals (Figure 2). Many models have been proposed to explain the contribution of different genes to complex diseases, but the genetic factors underlying most complex diseases are still largely unknown.

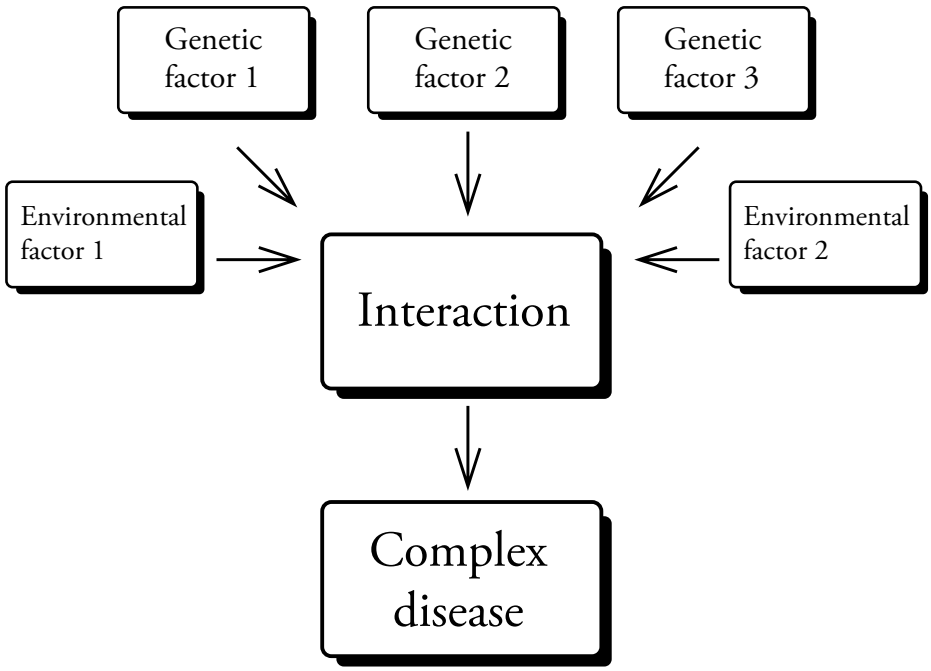

Figure 2
Combinations of genetic and environmental factors interact to influence the susceptibility to complex diseases.

Familial forms of complex diseases

A complex genetic disease is, as mentioned above, caused by an interaction of a number of genetic factors, each one in most cases contributing relatively little to a disease. Monogenic forms of these diseases can sometimes be found, these rare diseases may look very similar to the more complex forms of the disease, but are caused by only one or a few mutations, e.g. the MODY genes in type 2 diabetes mellitus (Froguel et al. 1992,

Horikawa et al. 1997, Kristinsson et al. 2001, Stoffers et al. 1997, Yamagata et al. 1996a, Yamagata et al. 1996b). These monogenic forms of complex diseases usually follow a Mendelian inheritance within a pedigree, and are inherited in a dominant or recessive manner.

Familial forms of complex diseases are usually more complex than monogenic forms of the diseases, but may still be seen as simplified versions of the more common and complicated forms of the disease as well as enriching for strong genetic effects. A familial form of a complex disease may arise when a genetic factor with a particularly strong influence on the disease arises within a family and contributes so much to disease susceptibility that other genetic factors become less important.

Theoretically, an apparent Mendelian inheritance pattern could also arise if the genetic diversity in a population is significantly decreased, so that by chance, a number of the more powerful susceptibility factors contributing to the disease have become fixed, and are shared by most of the members of a pedigree. If one more genetic risk factor is added in certain individuals, this may "tip the scales" and result in an apparent Mendelian inheritance of the disease (Varilo and Peltonen 2004). Identifying and dissecting the genetic factors involved in these familial diseases may thus be an effective strategy to begin the work of understanding the more classical complex forms.

GENETIC METHODS

Genetic markers

Mapping a genetic disease involves linking a disease to a genetic region in the genome or to a genotype or genetic trait and polymorphic genetic markers are used for this means. A polymorphic genetic marker is any genetic trait that differs between individuals. Such markers are located at fixed points in the genome, and can be visualised clearly.

The different alleles or variants of a marker allow for distinguishing between individuals. We have two copies of our chromosomes (except for X and Y), and therefore two alleles for each marker, one residing on each chromosome. Homozygosity implies the presence of two copies of the same allele, heterozygosity being when one has two different alleles for a given marker.

In the simplest form of genetic mapping, a genetic marker is analysed in a family, and alleles that are shared by the people with a disease are identified. It is likely that the mutation causing the disease is located close to this marker, and the marker is said to be linked to the disease. The position of the marker in the genome is known, so linkage to the marker tells you what part of the genome to focus on to find the gene or genetic factor involved in the disease. Genetic markers are usually not part of genes, and not the cause of a disease *per se*, but in this way they can give an indication as to the general region to study.

Microsatellite markers

Microsatellite markers are short repeats of DNA sequences, each repeat usually being 2-4 bases in length. These markers have about 5 to 15 alleles and each allele represents a particular number of times that the 2-4 base DNA sequence is repeated. Microsatellites are very common in the human genome, relatively stable over time and generations, they can be quite polymorphic and are relatively easy to analyse; all this has made them the most commonly used genetic markers when mapping genetic disease.

A drawback of using microsatellites when mapping disease is that although they are fairly well spread throughout the genome, there are regions in the genome with fewer microsatellites. The average spacing between microsatellites is also often large, making the resolution for mapping too low. The large number of possible alleles also makes automated analysis of microsatellites difficult.

Single-nucleotide polymorphisms (SNPs)

Single-nucleotide polymorphisms or SNPs are another class of genetic markers. They consist of a single DNA base that varies between individuals, and each SNP has two alleles. An SNP is formed when a mutation arises (a single base is mutated) somewhere in evolution, and two alleles now co-exist in the population. SNPs are not as informative as microsatellites since they only have two alleles, but they are found throughout the genome in greater quantities than microsatellites. This makes them an ideal tool for

investigating regions of interest in more detail, where a high resolution of markers is needed. SNPs have also become popular for large-scale automated analysis of markers, as the analysis of their two alleles is easier to automate than the more numerous alleles of a microsatellite. The majority of SNPs are, like the microsatellites, not directly involved in susceptibility to a disease *per se.* However, some SNPs have been shown to have quantitative effects on genes important in diseases, thus influencing the risk of such diseases (Gibson et al. 2001, Krempler et al. 2002, Ueda et al. 2003).

Genomic distances

The distance between two points in the genome can be defined in two distinct ways, using either the genetic or physical distance. Genetic distance (the main unit is centimorgan [cM]) relates to the likelihood of recombinations between two points (1cM equals a 1% likelihood of recombinations). Since recombinations are more likely to happen between two points the farther they are away from each other, a relative distance between the two can be obtained. Genetic distances are often used when looking at linkage between markers and diseases and the size of a region linked to a disease is often reported in cM.

Physical maps (the main unit is megabase [Mb], equalling one million bases) tell you the distance in basepairs between two points. The physical maps with the highest resolution give the exact sequence of DNA. These maps are especially useful when a particular region within the genome has been identified, e.g. to find genes or polymorphisms that could be causing the disease. The actual physical length of a cM varies between different regions of the genome and between men and women, but *on average*, one cM is approximately equal to one Mb (Lander et al. 2001).

Linkage and lodscores

Linkage is seen between a genetic marker and a disease when the alleles of a certain marker tend to be inherited differently in people with the disease when compared to their healthy relatives. This can be studied visually if the number of markers and individuals being studied is low enough, but methods to describe the statistical likelihood of linkage between a marker and a disease have also been developed. One of these, the lodscore (Morton 1955), is based on deviation from the expected recombination fraction of 0.5 under the null hypothesis of no linkage, i.e. the comparison of the likelihood of your data assuming linkage (theta < 0.5) and the likelihood of your data assuming no linkage (theta = 0.5). A lodscore over 3 to 3.6 can be seen as an indication of significant linkage between a marker and a disease or trait (Lander and Kruglyak 1995).

Parametric or model-based linkage analysis involves building a model of how a disease is inherited. This involves estimating the disease frequency in the population, the frequency of the disease-causing mutation in the population, what fraction of people have the disease for other reasons, etc., parameters that are often unknown (Morton 1955). This method can be very powerful if the model is estimated accurately, and may be feasible for

many of the Mendelian diseases. It is, however, most often quite impossible to estimate a realistic model of a complex disease governed by a number of genetic factors, incomplete penetrance and lifestyle factors. This is where non-parametric or model-free linkage methods are applied (Kruglyak et al. 1996). These methods are based on the assumption that a non-affected individual may still have the disease-causing mutations and they may fall ill at a future point. The only certainty is that the people that have the disease do indeed have the disease, so allele-sharing methods are used to see what parts of the genome the patients have in common. Taking into account how the patients are related, one can then determine if they share more genetic information than they should by chance. The lodscore obtained in non-parametric linkage is based on deviation from the expected distribution of shared genetic regions among the affected individuals under the null hypothesis of no linkage.

Association and LD

A marker and a disease are said to be associated when there is an over-representation of a certain allele of a specific marker in cases (a group of patients) when compared to a group of matched controls (individuals that are as similar to the patients as possible but do not have the disease). This could be an indication that this polymorphism or allele is the true factor causing the disease, or it could be due to linkage disequilibrium (LD). LD occurs when, many generations ago, a mutation occurred close to a marker in an individual with a particular allele. The two are so close to each other that they have a strong tendency to be inherited together, and the descendants of this individual have inherited his/her chromosome with both the allele and the mutation. The people with the disease today will have the disease due to the original mutation, and will also have a tendency to have the same allele in the marker that lies close to the mutation (Figure 3). This association will in most cases break down over time until linkage equilibrium is attained.

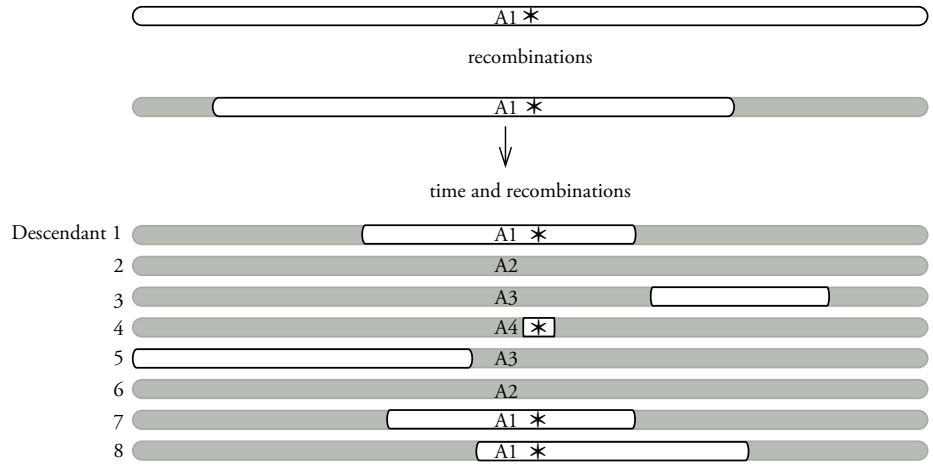

Figure 3
Linkage disequilibrium between a mutation (indicated by a star) and a genetic marker. A1 indicates allele 1, the allele in LD with the mutation. White bars represent the ancestral chromosome, grey bars any other chromosomes. Only descendant nr. 4 has the mutation but not A1.

Regions of linkage disequilibrium can be quite useful when mapping a disease and the assumption that these regions may be larger in isolated populations has been one of the stronger arguments for working with such populations (Kruglyak 1999, Laan and Paabo 1997, Peltonen 2000, Shifman and Darvasi 2001, Varilo et al. 2003, Varilo and Peltonen 2004). Identifying a genomic region that the patients share due to LD may aid in the definition of the smallest common region shared by the patients and, thus, in identifying the region contributing to the disease.

If a mutation is closely linked to a certain allele on a chromosome that has expanded its frequency in a population, this may result in an increase in the LD of the population for that particular locus. The Sami population has been closely examined as a population with a history of genetic drift, and this population has been shown to have larger areas of LD than the populations to which it has been compared, in at least some genetic regions (Johansson et al. 2005). On the basis of this, it has been suggested that initial mapping of a complex disease should preferably be performed in such a population, using another population with shorter regions of LD to further fine-map the genetic factors underlying the disease (Varilo and Peltonen 2004).

A common approach to get the most out of the strengths of linkage and association, and to minimise the weaknesses, has been to combine the two methods when mapping a disease. A group of families is first used to look for linkage to a disease. When a genomic region has been identified and fine-mapped as much as possible, an association study is performed on particularly interesting factors found in this genomic region. This may allow for a higher resolution than using only linkage, as well as giving more population-based information than if only families were studied.

Homozygosity mapping

When a disease is very rare, and the population being studied can be assumed to be relatively homogenous, it can be assumed that the same mutation is causing a disease in most of the affected individuals in the population. This is a result of a mutation arising in one individual and showing a geographical clustering in a particular region as the result of the descendants of this individual tending to live close to their parents.

An extension of this effect is when a rare recessive disease that affects multiple individuals in a family is being studied. The two copies of a mutation that each individual carries can be assumed to come from the same source or founder, especially if the parents of the patients are related to each other and/or there has been a significant amount of inbreeding in earlier generations. Identifying this mutation thus involves looking for regions in the genome where the patients are homozygous and have two copies of the same ancestral genetic region with the associated mutation (Lander and Botstein 1987). Healthy parents should each carry one copy of this region and healthy siblings one or none.

Interaction of loci

In some genetic diseases, two or more genetic loci interact to contribute to susceptibility or protection from the disease. An example of this is when locus B only increases the risk of getting a disease if a certain allele or variant of locus A is present. Conditioning, or stratifying the genetic data with respect to locus A will thus allow for better studies on the effects of locus B. The human HLA region (called MHC in other species) has been shown to play a large role in susceptibility to many autoimmune diseases (Bilbao et al. 2003, Czaja et al. 2002, Grams et al. 2002, Koo et al. 2003, Patil et al. 2001), and stratifying on the basis of HLA haplotype is commonly performed when mapping these diseases (King et al. 2000).

Genome-wide scans and candidate-gene approaches

There are two main approaches to mapping a disease, the genome-wide scan (GWS) and the candidate-gene investigation. The approach chosen will depend on the amount of prior knowledge accumulated regarding the genetics of the disease in question.

When little is known about the aetiology of the disease, or when the disease is very complex, a genome-wide scan may be required. This involves analysing markers (usually polymorphic microsatellite markers) that are evenly spaced throughout the whole genome, and looking for linkage to the disease. If linkage is found to a certain marker, more markers are added in that particular region to increase the resolution of the scan.

This method has the advantage that little prior information is needed about the genetic factors contributing to the disease and important genetic factors are less likely to be missed due to erroneous assumptions about which genes or genomic regions are involved in the disease. A drawback to using genome-wide scans is that the marker resolution is lower than if the focus is on one or a few candidate-regions and an important region may be overlooked due to this.

Genome-wide scans are very popular due to the availability of ready-made panels or groups of markers that have been optimised to work well together. This allows for analysis on a high-throughput scale, increasing the feasibility of genome-wide scans with hundreds of microsatellite markers on hundreds of individuals. This approach is now the standard when studying most complex genetic diseases.

The candidate-gene approach can be very powerful if prior knowledge exists about what genes may be involved in the disease or if a certain region of the genome has previously been implicated in the disease. A candidate region or region of interest is the region believed to contain the mutation causing a disease, and this is tested by analysing a set of polymorphic genetic markers in this region to see if linkage can be found to a particular marker.

Finemapping

When a genetic region involved in susceptibility to a disease has been identified, this region is studied closer. This finemapping usually involves adding a number of genetic markers (e.g. microsatellites or SNPs) to increase the resolution of the linkage analysis. This should result in taller, narrower linkage peaks, helping to reduce the regions of interest. Finemapping may also allow for the study of haplotypes within families that show linkage to the region. A haplotype is a set of loci or markers, located close to each other on the same chromosome. Since the markers are so close to each other, the likelihood of recombinations (exchange of genetic material between a pair of chromosomes) between them is low, and the combination of alleles found in a parent is likely to be passed on unchanged to an offspring.

Finemapping and haplotype analysis of a disease can be a very powerful tool to narrow down the region of interest further. If affected individuals that only have a part of the haplotype linked to the disease (due to recombinations) can be found, this part of the haplotype may perhaps be excluded (it is not required for the disease). Similarly, when working with a Mendelian disease, a certain region may be more interesting if an individual in the pedigree does not have this part of the haplotype and is healthy (indicating that this part of the haplotype is required for the disease). A downside of working with individuals that are too closely related is that they will share a relatively large genomic region or haplotype. This can make it hard to narrow down the region further and decide which genetic factors within the haplotype are the ones actually involved in the disease.

Having identified the smallest common genetic region shared by the patients, the next step is usually to study the genetic factors in this region and identify potential candidate genes (genes that could possibly be involved in the disease).

POPULATION GENETICS

We humans are genetically very similar (most estimates of the similarity of our genomes range from 99,5 to 99,8% (Kidd et al. 2004)). Despite this, different populations can be shown to differ genetically from each other. This has little to do with the concept of races, but rather with the history of each population. Geographically close populations tend to also be genetically similar, but factors such as the age of the population, its size, the amount of immigration, expansion, and the level of geographical and social isolation are crucial to forming the gene pool we see today.

Genetic drift and isolated populations

It has been shown that monogenic diseases can have strikingly different frequencies in certain population groups when compared to others (Pastinen et al. 2001, Peltonen et al. 1999). Examples of this can be found in many populations, and examples from northern Sweden include acute intermittent porphyria in Arjeplog (Lundin et al. 1997) and infantile genetic agranulocytosis in Överkalix (Carlsson and Fasth 2001). These diseases are almost exclusively found near these locations.

A similar pattern of deviation from countrywide prevalence as the one seen for monogenic diseases may also be seen for more complex genetic disease (Peltonen 1996). This clustering may occur in parts of the country that are sparsely populated, indicating an enrichment of individuals with a particular disease in a particular region. Different diseases can therefore cluster in different geographical regions. In some cases, the prevalence of a disease can vary considerably between two nearby countries.

These geographic differences in the prevalence of genetic diseases are likely to be the result of genetic drift, random changes in the frequencies of genetic variants or alleles in a population, occurring by chance rather than due to selection. Genetic drift has the potential to dramatically alter the frequencies of genetic variants in a population, especially if the population is small.

A founder effect is an extreme form of genetic drift, and occurs when a small number of individuals from a larger population forms the basis of a new population by moving off to found another population. Similarly, population bottlenecks may have drastic effects on the genetic composition of a population. This form of genetic drift occurs when a large part of a population is lost, e.g. due to famine or disease, resulting in survival of only a subset of the population. The survivors are subsequently the basis of a new population that then expands to form a larger population.

In both founder effects and population bottlenecks, the frequencies of genetic variants represented in these people are not likely to be completely representative of the parent population. The genetic composition of this new and expanding population may thus be drastically different from the original (Figure 4). Genetic variants may become less common or be lost, or may randomly expand their frequencies when compared to other variants.

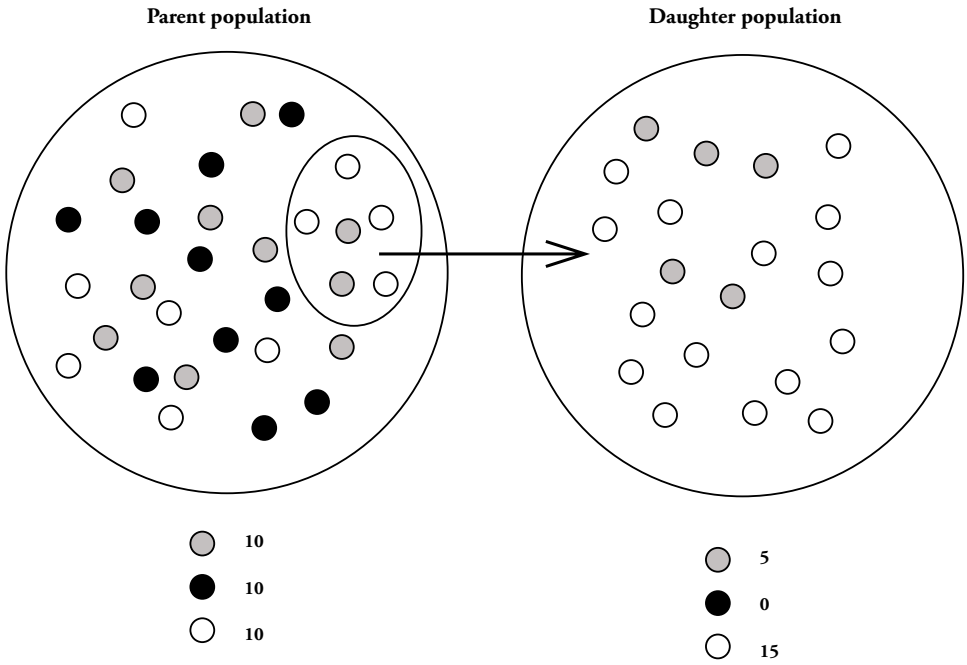

Figure 4
An example of a founder effect, resulting in a reduction in diversity and random changes in the frequencies of genetic variants due to non-random sampling from the parent population. Small circles represent three different genetic variants found in the parent population. The genetic variant represented by black dots is completely lost in the daughter population and the frequency of the genetic variant represented by grey dots is severely decreased as a result of the sampling from the parent population.

The new population, whether it is composed of survivors of a plague or famine, or settlers moving into a previously uninhabited region of land, will have a drastically decreased genetic variation when compared to the larger population from which it originated (Shifman and Darvasi 2001) since numerous genetic variants present in the parent population will not have been represented in the small number of settlers or founders. Adding to this effect, extensive inbreeding is likely to occur within the individuals in this new population, potentially fixing a number of genetic variants. The inbreeding is likely to be reduced as the population expands, but the frequencies of genetic variants in the population may remain significantly different from the original population. This effect is likely to be strongly enhanced if a population lives in relative geographic or social isolation, maintaining the genetic differences between the two populations.

Genetic drift may thus be the reason for the different prevalence of Mendelian and complex genetic disorders when comparing population-groups. It may also explain "negative founder effects", where certain disease-alleles are almost absent or completely absent in the population. Examples of this include certain common alleles causing Cystic fibrosis in Caucasians that are almost completely absent in the Finnish population (Pastinen et al. 2001).

These random events in the history of a population make isolated populations attractive for the mapping of genetic diseases. Studying populations with a high prevalence of a disease, or a disease that shows a strong geographic clustering is likely to aid in studying that disease. An unusually high prevalence of a disease can be an indication that the genetic risk factors for the disease have become enriched in this population due to founder effects, hopefully aiding in their identification. Another quality of this kind of population that potentially aids in mapping genetic diseases, particularly complex diseases, is the reduction in total genetic diversity. If a disease may be caused by numerous genetic factors in the general population but most of them were lost during a population bottleneck, the remaining genetic factors may be easier to elucidate. Also, studying LD in these populations and identifying shared ancestral haplotypes is likely to be especially useful for the mapping of complex diseases.

The population of northern Sweden

Northern Sweden is an example of a population that has been extensively investigated for the purpose of genetic research. Northern Sweden can be defined in many ways, but the focus of our work involves Norrbotten and Västerbotten, the two northernmost counties in Sweden (Figure 5).

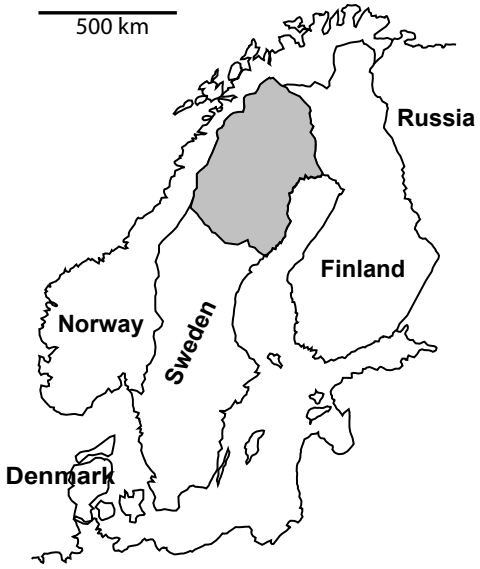

Figure 5
A map of Scandinavia with the two northernmost counties of Sweden shaded grey.

The population of northern Sweden mainly comprises people of Sami, Finnish and Swedish origins. The Sami lived in the inland areas of Sweden, especially in the northernmost regions. An early Finnish population has been suggested to have migrated and settled down along the northern part of the Gulf of Bothnia, followed by a successive Swedish colonisation in the coastal areas that was strengthened by encouragements from the Swedish State in the early 14th century.

Colonisation took place much later in both the forested areas near the coast and the inland areas, during the 19th century in particular. In the inland areas, the process was characterised by an initial colonisation by a relatively small number of pioneers from the coastal parts within the regions and from Finland, whereas the subsequent settlements were performed mainly by their descendants. Hence, the immigration was small and this clonal colonisation resulted in clusters of settlements representing several generations (Bylund 1994, Bylund 1960).

The population of northern Sweden has been shown to be well suited for the mapping of monogenic and complex diseases. This is in contrast with the known admixture of Sami, Finns and Swedes, the three large population-groups comprising the population, and despite most parts of northern Sweden not having been isolated to the same extent as other populations used for genetic research (e.g. Sardinians). A number of recessive Mendelian diseases can be found in this population, and there are indications that some complex diseases show a geographical clustering in this population. All this suggests hidden sub-structures in the population and suggests that closer study of the genetic composition of the population residing in northern Sweden might be an effective approach to further understand the factors contributing to genetic diseases in this population.

River valley genetic isolates:
The general demographic pattern in northern Sweden consists of population aggregates mainly along the coast and rivers, and scattered populations between the river valleys. Initially, the main reasons for this were access to arable land and pastures and communications by water. The major roads were likewise built along the rivers. Cross-region communication possibilities increased in the 20th century, not least via railways built through the area. Some notable exceptions to this pattern of population concentrations were due to inland industries, mining in particular.

A plausible hypothesis would be that distinct genetic sub-populations might have formed in the large river valleys that run from the mountains in the west to the coast in the east and which are largely isolated from each other by vast forests. These sub-populations are hypothesised to have arisen through founder effects during the settling of each region. Large variations in the small founder populations, due to genetic drift, followed by generations of geographical isolation, maintained the genetic differences of each river valley.

AIMS OF THE STUDIES

The aim of the work described in this thesis was to test the hypothesis that the population of northern Sweden consists of genetically distinct sub-populations, and to study how this would affect the mapping of monogenic and complex genetic diseases.

Paper I THE SUB-ISOLATES OF NORTHERN SWEDEN:

- To study the genetic structure of the population of northern Sweden by testing our hypothesis that distinct sub-isolates exist within each river valley region.

Paper II HSAN V AND NGFB:

- To study the applicability of the population of northern Sweden for the mapping of the monogenic disease HSAN V, shared by only a small number of individuals. Also, to identify the genetic region or mutation responsible for the disease.

Paper III T1DM AND AITD:

- To test the feasibility of mapping familial forms of T1DM and AITD in a family from northern Sweden. Also, if possible, to identify the genetic regions or factors responsible for T1DM and AITD in this family.

Paper IV T2DM:

- In light of the success in mapping other genetic traits, to study if this could be applied to familial forms of a classical complex genetic disease. In the process, to learn more about the genetic risk factors for T2DM in northern Sweden.

RESULTS AND DISCUSSION

PAPER I: THE SUB-ISOLATES OF NORTHERN SWEDEN

Historical data shows that people in northern Sweden did not move very far to meet a spouse and they tended to reside in their region of origin throughout their lifetime (Andersson 1897). This is likely to have maintained any existing genetic differences between the local populations, and inbreeding and genetic drift can be assumed to have added to this effect. The creation (due to founder effects) and persistence (due to geographic isolation) of distinct genetic sub-populations within the populations of the larger river valleys in northern Sweden might thus be plausible.

We tested this hypothesis by dividing Norrbotten and Västerbotten counties into regions focusing on major rivers from the mountains in the west to the eastern coast (Figure 1, ***Paper I***), and studying differences between those regions.

We studied marriage patterns in the descendants of a couple that married around 1650 and lived in Lycksele in river valley region B (Figure 1 in ***Paper I***), and saw a strong tendency of spouses to originate from the same river valley region (Figure 2 in *Paper I*). This was supported by data on marriage patterns in 1870 and 1900 in six parishes, each representing a river valley region (Figure 3 in ***Paper I***). We thus concluded that this restricted spouse selection, with a preference towards a spouse originating from the same river valley region, appeared to be a general phenomenon in the region and that until the beginning of the 20th century relatively little interaction took place between individuals from different regions.

We next tested if significant differences could be found in the genetic composition of the populations inhabiting the river valley regions. We did this by studying protein and blood group markers previously used to study Finnish and Sami admixture in northern Sweden (Nylander and Beckman 1991). In addition to confirming reports of gradients in the allelic frequencies of these markers, an additional trend was found, suggesting that these allelic frequencies are restricted by the river valley regions (Figure 4 and Table 2 in ***Paper I***). This was further supported by our analysis of the mutation causing Bothnia dystrophy (Burstedt et al. 1999), shown in Figure 5 of ***Paper I***, and these data thus support the hypothesis that the geographic regions defined in ***Paper I*** are home to genetically distinct sub-populations.

Our hypothesis was that the population of each of the larger river valleys consisted of distinct genetic sub-isolates, arisen due to a small number of founders and random events during the colonisation process, and maintained by preferential contact between individuals within the same river valley.

We have shown that genetic differentiation exists between the populations and that any genetic differentiation arising between the populations of the river valleys is likely to have been maintained due to geographical isolation. These data support our hypothesis of sub-isolates arising and being maintained within the river valleys of northern Sweden.

The existence of sub-isolates within the larger population is likely to have implications on linkage to genetic diseases. Genetic drift such as founder effects or population bottlenecks might explain the high incidence of monogenic diseases and occurrence of families with a clustering of complex diseases, and explain the geographical clustering of complex diseases seen in certain parts of the region. The genetic admixture of the total population may thus be countered by a local increase in genetic homogeneity. This has the potential of making the mapping of genetic diseases of individuals within the same genetic sub-population a powerful tool to reduce the heterogeneity of complex genetic disorders and select for common founder mutations.

On the other hand, the existence of genetic sub-populations should be taken into consideration in studies of association to genetic diseases. As these studies are dependent on careful matching of cases and controls, and are sensitive to hidden confounding factors, sub-populations with distinct differences in allele frequencies are likely to make these studies more complex. Controlling for genetic sub-isolates or matching cases to controls from the same sub-population may thus be recommended.

PAPER II: HSAN V AND NGFB

In ***Paper II***, we studied a family from northern Sweden with three individuals suffering from a severe loss of deep pain perception preventing them from feeling pain from bone fractures and joints, and with an impaired perception of heat (Minde et al. 2004). The disorder is rare, with only around 20 known cases, and this family is the only family in Sweden known to suffer from this disease (Jan Minde, *personal communication)*.

Mapping a genetic disease, albeit one with an apparently classical Mendelian inheritance, using only three patients, is quite a challenge. Our hypothesis was that although achieving this would be difficult or impossible in a more genetically heterogeneous population, we might be able to achieve this due to the genetic makeup of the individuals involved.

Genealogical analysis of the family involved showed that the origin of the disease can be traced back to a couple moving up along the Torne river from Övertorneå north to Vittangi in river valley region F (see Figure 1 in ***Paper I***) in 1674 (Figure 6) and historical records support sensory defects having affected the family for generations (Jan Minde, *personal communication*). The tendency of the couple's descendants to remain in or around Vittangi and select a spouse from the same region has resulted in extensive

inbreeding throughout several generations. This may be the reason that this recessive disease has surfaced within this family.

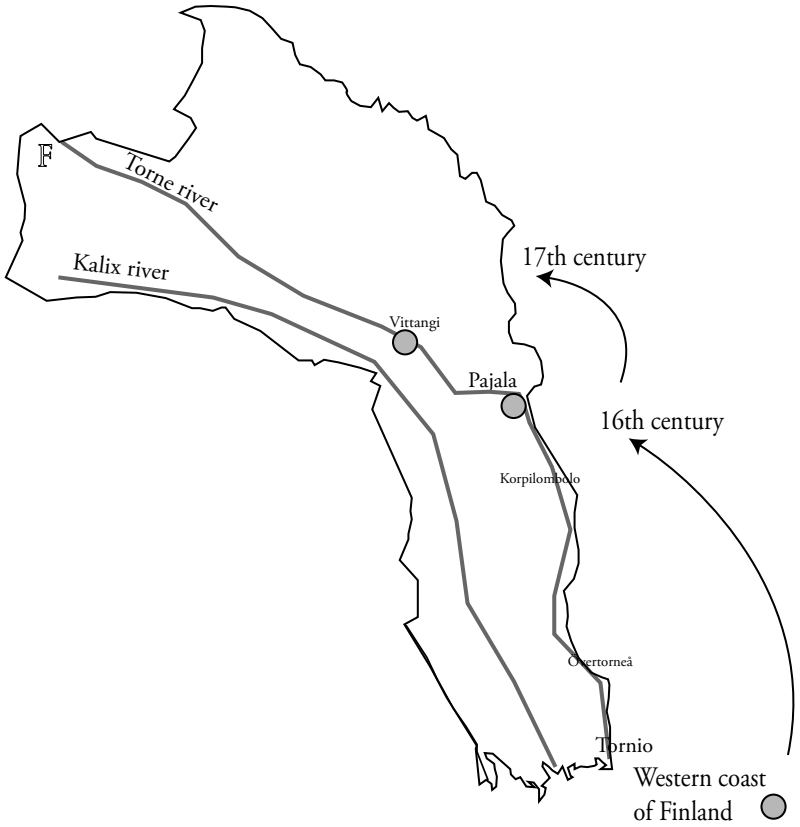

Figure 6
The presumed migration of the mutation causing HSAN V up along the Torne river in geographic region F (defined in Figure 1 of *Paper I*).

We performed a 10cM genome-wide scan to identify the genetic region or regions involved in the disease and identified a haplotype on 1p11-13 for which all three affected individuals were homozygous. Subsequent sequencing analysis of the best candidate gene in this region, nerve growth factor beta (NGFB), revealed a point mutation in exon 3 of the gene (Figure 2, ***Paper II***), replacing a basic arginine amino acid (CGG) with a non-polar tryptophan (TGG). This particular amino acid is located in a region of the protein that is highly conserved between different neurotrophins (a group of nerve growth factors of which NGFB is a member) as well as NGFs of different species. This thus supports the idea that this mutation is the cause of the disease in this family.

A number of individuals in the family were observed to have milder neurological symptoms (e.g. Charcot joints); these were found to either be heterozygous for the mutation in NGFB or lacked the mutation. A phenotype-genotype correlation study is currently underway to look at the neurological symptoms of a number of family members, even those previously reported as healthy. An overrepresentation of mild neurological symptoms in heterozygous individuals would argue for a dose-effect of the mutation and the disease being semi-penetrant even in those individuals.

The functional effects of the mutation in NGFB are now being studied further. The mutation is found in a region of the protein important for its binding to and/or activation of one of its receptors, suggesting that the mutation in NGFB is interfering with this function. Alternatively, the binding of the mutated NGFB to its other receptor might be affected, or even the stability of the NGFB mRNA or protein itself disrupted in some way.

In conclusion, we have identified a disease-causing mutation in a genetic sub-population significantly affected by founder effects, inbreeding and genetic drift. It would thus seem that, despite the known admixture of Swedish, Finnish and Sami genetic material in the region, this region of Sweden is well suited for mapping monogenic diseases.

PAPER III: T1DM AND AITD

Having seen that the population of northern Sweden can be utilised to successfully map a Mendelian disease using only three patients, we moved on to explore if this could also be applied to more complex diseases (***Paper III***).

A large 5-generation family (Figure 1 in ***Paper III***), from the region around Överkalix in river valley region F (defined in Figure 1 in ***Paper I***) was studied. This family has multiple members affected by type 1 diabetes mellitus (T1DM), Graves' disease (GD) or Hashimoto´s thyroiditis (HT), the latter two collectively known as autoimmune thyroiditis (AITD). This allowed us to study the genetic factors responsible for the susceptibility to these diseases in the population of northern Sweden.

The diseases in this family appeared to follow a Mendelian inheritance and constituted familial forms of these otherwise very complex genetic diseases. The members of this family have, except for the last two generations, almost exclusively resided in a small geographical region with a small number of inhabitants. A small number of founders in the region and genetic drift may have reduced the genetic diversity of the population and fixed a number of genetic factors contributing to the diseases. This might thus have resulted in a simplified inheritance pattern of these otherwise complex genetic diseases, potentially simplifying the study of the genetic factors contributing to these diseases.

It has been shown that AITD and T1DM have a tendency to occur together, indicating that the diseases may share a common underlying cause (Payami et al. 1989). In addition to a strong contribution to disease susceptibility by the human leukocyte antigen (HLA) locus (Barlow et al. 1996, Cudworth and Woodrow 1975, Yanagawa et al. 1993), each of these diseases has been reported to be associated and/or linked to the IDDM12 T1DM susceptibility region on chromosome 2q33 (Donner et al. 1997a, Donner et al. 1997b, Nistico et al. 1996). Since all three diseases are found within the same family, and have

been shown to be linked to common genetic factors, we hypothesised that they might share common genetic causes in this family.

A candidate gene approach was used to study regions previously implicated in susceptibility to the three diseases. Loci in the HLA region on chromosome 6p21 and in the IDDM12 region on chromosome 2q33 were deemed to be especially interesting, since all three diseases have been mapped to those regions and since the orthologous regions have been shown to confer susceptibility to diabetes in non-obese diabetic (NOD) mice, an animal model for diabetes (Aitman and Todd 1995). Treating all three diseases (T1DM, HT and GD) as one disease, we performed a two-point parametric linkage analysis using a model of dominant inheritance with 90% penetrance.

Positive lodscores were obtained in a region on 2q33 (IDDM12) indicative of the presence of a susceptibility locus in this region. All of the affected members of the family were found to carry at least one copy of the HLA haplotype DR4-DQA1*0301-DQB1*0302 previously shown to confer susceptibility to T1DM (Kockum et al. 1993). Conditioning on HLA haplotype provided a two-point lodscore of 4.20 at $\theta = 0.0$. The corresponding maximum lodscore obtained without conditioning for HLA was 1.45 at $\theta = 0.23$ (Figure 2C in ***Paper III***). This analysis thus indicates linkage of the HLA and IDDM12 regions to T1DM and AITD and indicates that the genetic factors in the IDDM12 region are dependent on genetic factors within the HLA region.

Sequence analysis of CTLA4, a candidate gene in the region, revealed a combination of three previously identified disease-associated alleles (the G allele of the A/G SNP at position 49 in exon 1, the T allele of the C/T SNP in position 1822 in the intron between exon 1 and 2, and the 106bp allele of the CTLA4 (AT)n microsatellite in the 3′ untranslated region) (Figure 2A in ***Paper III***) .

The observed linkage of susceptibility factors for T1DM/AITD to HLA and to the IDDM12 region confirms previous reports of linkage and association of T1DM, GD and HT to these loci. It also provides evidence that the same genetic factors may be mediating each of the three diseases. Two of the polymorphisms found in the CTLA4 gene, the 49A/G SNP in exon 1 and the AT microsatellite located in the 3′ untranslated region of CTLA4, have been suggested to contribute functionally to these diseases (Kouki et al. 2000, Takara et al. 2003, Wang et al. 2002). Despite this, it seems unlikely that these polymorphisms alone can explain the dominant inheritance pattern and high penetrance seen in this family. A more likely scenario involves an accumulation of risk factors each contributing only a limited increase in risk, and a number of additional polymorphisms could be present in strong linkage disequilibrium in the same genomic region but in several different genes. Susceptibility factors in familial forms of complex genetic diseases may thus be constituted by the combined effect of minor risk factors accumulating in a risk haplotype such as the one seen in this family and a relative paucity of genetic variation in other genomic regions.

Additional genetic factors located in the disease critical region cannot be excluded and an interesting facet that has surfaced in this aspect is the identification of CT60, a SNP located 6,1 kilobases downstream of the CTLA4 gene, and shown to influence the risk of T1DM and Graves' disease in a British population (Ueda et al. 2003). The mechanism for this is believed to be the effect the SNP has on the ratio of the two alternative forms of CTLA4, one membrane-bound and the other soluble.

A large genome-wide scan project has now been initiated in our research group to try to gain more information on the genetic risk factors for T1DM and AITD in the general population of northern Sweden. Also, to see if the genetic linkage we identified in "family 4" can be replicated and studied in other families.

It is interesting to note that the occurrence of inbreeding in previous generations appears to be lower in this family than in the one suffering from HSAN V, despite a high level of endogamy in both cases. The different rates of inbreeding may possibly be explained by the size of the populations in which the individuals lived. The couple that presumably introduced the NGFB mutation to the region were the founders or co-founders of Vittangi, and Vittangi is even today only home to about 2000 individuals. This can be compared to the region of Överkalix, which has twice as many inhabitants today and is closer to the coast and to the towns bordering on Finland. This may have increased the effective population size available to the members of the Överkalix family and thus not encouraged inbreeding to the same extent as in the family from Vittangi.

We conclude that families like the one studied here constitute a very useful tool when mapping genetic diseases, and in particular complex genetic diseases. When they arise, they may show a sufficient reduction in genetic heterogeneity and enrichment for strong genetic risk factors to allow studies aimed at mapping complex genetic diseases based on only one or a few families to be feasible.

PAPER IV: T2DM AND CALPAIN-10

We have shown that mapping a complex genetic disease (the autoimmune syndrome described in ***Paper III***) is feasible by the use of only one family, drawing on the strengths of the population we have at our hands and studying familial forms of complex diseases. Continuing this work, we proceeded to study type 2 diabetes mellitus (T2DM), another genetic disease that affects an increasingly large proportion of the world's population (Zimmet et al. 2001) and is known for its genetic complexity.

Extensive linkage and association studies have been undertaken to identify the genetic factors contributing to T2DM (see (Parikh and Groop 2004) for a recent review) but despite this, much remains unknown about the disease. Figure 7 shows some of the genes and genetic regions implicated in susceptibility to T2DM.

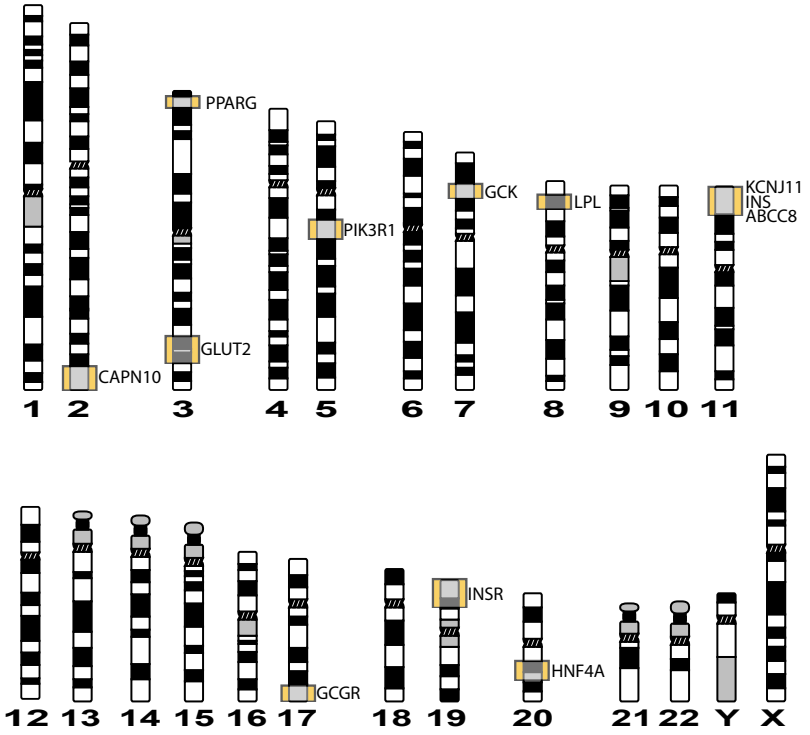

Figure 7

Some of the genetic regions linked to or associated with type 2 diabetes mellitus (T2DM) in humans.

In light of the known complexity of T2DM, the fact that an increasing number of individuals in the northern Swedish population suffer from T2DM (Eliasson et al. 2003), and the notion that this population is well suited to mapping complex diseases, we decided it would be interesting to study the genetic factors contributing to T2DM in this population. Due to the complexity of the disease, and the assumption that it is caused by a number of genetic factors even in the relatively homogeneous sub-populations of northern Sweden, we performed a non-parametric genome-wide scan to identify these genetic factors. By selecting for familial forms of the disease, and choosing families from only the northernmost part of Sweden (mostly river valley region F), we hoped to enrich for stronger, and more homogeneous genetic risk factors for the disease.

A 10cM genome-wide scan (see Figure 2 in ***Paper IV***) revealed linkage to NIDDM1, a region on 2q36.1-qter previously implicated in susceptibility to T2DM (Hanis et al. 1996), and containing calpain-10, one of only a few genes so far convincingly identified as being involved in susceptibility to T2DM (Horikawa et al. 2000). Finemapping of this

locus yielded a maximum multipoint allele-sharing lodscore of 3.20 (Figure 3 and Table 2 in ***Paper IV***).

Linkage was found to four polymorphisms in calpain-10, UCSNP-44, -43, -19 and –63 (see Figure 1 in ***Paper IV***), previously reported to be associated with T2DM (Horikawa et al. 2000). These polymorphisms are unlikely, however, to be directly involved in susceptibility to T2DM in our family material since the calpain-10 haplotypes in the families showing linkage to 2q37 do not seem to differ from the haplotypes in the families that do not show linkage. In addition, calpain-10 haplotypes previously shown to confer susceptibility to T2DM in certain combinations (Horikawa et al. 2000) do not appear to be over-represented in the families showing linkage to 2q37. Looking at what fraction of the families actually contribute to the linkage peak on chromosome 2q37, we see that more than half show positive linkage to the region. This suggests that this region of the genome harbours genetic factors important for the susceptibility to T2DM in the general population of (at least) river valley region F.

Calpain-10 is a known T2DM-susceptibility gene in this region, and polymorphisms in this gene show evidence of linkage to T2DM. Calpain-10 was originally identified as a T2DM risk gene by a linkage approach. This gene and its function were unknown at the time of the original analysis, and might have been overlooked if the linkage to the region containing it had not been so convincing, and if the original investigators had had any obvious candidate genes to work with (Horikawa et al. 2000) .

The relative lack of other genomic regions involved in linkage to T2DM in this material, in addition to the strength of the peak on 2q37, supports our hypothesis that the sub-populations within the region may exhibit distinct genetic characteristics and that the population within each region may be more genetically homogeneous than the population as a whole.

We are currently in the process of analysing and collecting further samples from families with T2DM. The results attained from these additional samples will hopefully allow for a further refinement of the linkage to 2q37 and the exclusion of a number of potential candidate genes. This should allow us to better determine if calpain-10 harbours the genetic variants contributing to T2DM, or if another gene close to calpain-10, does.

To complement these results, we are collecting and analysing a set of samples from cases with T2DM and matched controls to look at association of T2DM to the four previously studied polymorphisms in calpain-10. This analysis should help us to determine if these polymorphisms are likely to be true genetic risk factors for T2DM in northern Sweden, or if other polymorphisms in the region in and around calpain-10 need to be studied further.

CONCLUDING REMARKS

The aging of the general population and the increasing prevalence of diseases such as heart attacks, stroke, type 2 diabetes mellitus, and hypertension (to a large extent explainable by our increasingly sedimentary lifestyle and unhealthy eating habits, in combination with the increasing age of the population), are all factors that bring enormous costs to the health-care systems of the world. This drives us to study factors mediating risk of developing such diseases. Vaccination to protect from infection or autoimmunity, lifestyle changes such as losing weight and quitting a smoking habit are all likely to be effective ways of combating these diseases. Despite this, the genetic components of the diseases remain important risk factors and may be the most difficult factors to elucidate.

The results of the work discussed in this thesis clearly illustrate how identifying the genetic factors underlying a genetic disease becomes more and more difficult as the disease complexity increases. When studying Mendelian diseases, one may be able to identify a haplotype or mutation involved in the disease, but in complex diseases it is usually not that simple. There, we have to work with groups of apparently harmless polymorphisms that together have a noticeable effect, significant heterogeneity even between individuals in the same population, very common susceptibility alleles with slight but important effects on the disease prevalence, as well as figuring out how the environment or lifestyle of the patient is involved. It is for this reason that ways to limit the genetic heterogeneity in studies of complex genetic diseases have been met with enthusiasm.

The utilisation of isolated populations or populations that have undergone population bottlenecks is widely considered an effective strategy to elucidate the genetic factors contributing to complex genetic diseases. This may, however, only be valid if solid knowledge exists about the structure of the population, its history and characteristics. Any population that has been strongly influenced by genetic drift, especially if this is enhanced by geographical or social isolation, will inevitably develop its own characteristics due to random events in its population history. This makes each population unique and may make each population uniquely suitable for mapping certain genetic diseases or traits, and less adequate for mapping others (Wright et al. 1999). A good understanding of the characteristics of the particular population at hand is thus needed and will aid in determining the strengths of each population. We believe we have demonstrated that the population of northern Sweden is well suited for mapping both monogenic diseases and familial forms of several complex diseases.

Disease-mapping efforts in northern Sweden have been aided by several factors other than the genetic composition of the population itself. These include a high-quality healthcare system, contributing to a low frequency of undiagnosed cases and a positive attitude towards genetic research. Access to large biobanks and disease registers further

aid the studies of diseases in this population. Finally, a tradition of genealogical interest and the existence of comprehensive church-registers have aided the tracing of families to common founders, thus aiding in the identification of common founder mutations or haplotypes. All this, in addition to the genetic material itself, makes a strong argument for mapping of genetic diseases in northern Sweden and is likely to contribute significantly to the success of projects of this kind.

The future of mapping genetic diseases in northern Sweden seems bright. We have seen that the tendency to live within the region where one is born still prevails to a large extent, hopefully maintaining the sub-population structure that is likely to be contributing to the usefulness of the population for genetic studies. A number of monogenic and complex diseases found in northern Sweden remain to be studied, and it seems likely that numerous new mapping projects will be established in the near future.

New developments emerging to complement classical mapping of genetic diseases include studies on the extent of LD in different populations and the HapMap project. These efforts are likely to add to our current knowledge of genetic diseases, and may be especially useful for mapping complex diseases. It seems likely that we will, in the near future, use our knowledge of the sub-populations in northern Sweden to build our own river valley specific LD maps, panels of haplotype-tagging SNPs and haplotype maps for each genetic sub-population. With this as a starting point, the population of northern Sweden is likely to continue to be an excellent tool for studying genetic diseases and making complex genetic diseases somewhat less complex.

ACKNOWLEDGEMENTS

Thanks to my supervisor **Dan Holmberg** for an endless supply of crazy ideas and for giving me the chance to do real research. Thanks for all those energy- and inspiration-inducing five-minute meetings.

Thanks to my co-supervisor **Gösta Holmgren** for friendly smiles, an interest in anything connected to genetics, and for having built up genetic research in Umeå.

Thanks to Holmbergs höns, past and present members. **Mia E**, thanks for knowing so much, and being quick to make up good/plausible answers even when you don't know. Thanks **Ingela, Marie, Anna L-B, Nadia, Vinicius, Ingegerd, Ann-Sofi, Sofie, Ulrika, Johan** and **Mario** for good times in the lab and for making life so much more interesting.

Thanks **Sofia** for working with me on the diabetes project and for always being so organized... you are my hero, I have no idea how you do it! Thank you **Kurt** for the mozzie-repellent and the karate-socks, I may need those if I ever get out of building 6M. Thank you so much for proofreading this thesis and the two manuscripts.

Thanks to **Lisbeth Ä, Martin, Urban, Mia S, Irina, Linda, Carin S, Solveig L, Åsa L, Tomas, Pia, Lotta, Björn H, Anna-Karin** and everyone who's been around for various tips and tricks, conversations at lunch, and ideas on how to cope with life in the lab.

Thank you **Anna C** for suffering through the perils of mapping human disease with me when the rest of the group was doing something totally different (immunology!).

Thanks, **Lisbet Lind,** for always being there when I needed to be sarcastic about something and thanks for your help with my questions about genetics.

Thank you, **Maria L**, for sharing my appreciation for a quiet lab and its effect on productivity.

Thank you **Susann**, for bullying me and allowing me to bully you when needed. :=) Thanks for those numerous "it's us against the world" moments.

Petter. Thanks for all the p-values and lovely lodscores you helped us produce from our data and for patiently explaining the basics of statistical analysis to us over and over again. **Mikael**. Thanks for always being so enthusiastic about making life easier for the rest of us, even if it takes you hours or days of work.

Monica. Thanks for the interesting and rewarding times on the HSAN project, and for patiently answering my questions about genetics. Thank you **Kristina L** for your never-ending enthusiasm about everything and for you knowledge about all things immunological.

Stefan Escher. Thanks for the breakfast meetings on Thursdays, the help on the endless geography of northern Sweden, and all the discussions we've had where we would usually find that we agreed completely on everything as usual.

P-A. None of us would survive very long without your help. Thanks also for being such a nice person, even if you terrorise students with questions about when they will finish their thesis and the odds of getting a job after defending it.

Damerna på 3:e våningen: **Kerstin**, **Clara**, och **Birgitta**. Tack för att ni sliter för att göra livet lättare för oss andra.

Birgitta Berglund. Thank you for bringing order to our chaos.

Mia Svensson. Thanks for being my best friend when everything seemed impossible. Thanks for sharing an interest in aikido, red wine and sushi; thanks for defending your thesis a couple of months before me so I could draw on your experience.

Magnus. Thanks for talking to a shy Icelandic student in Umeå and for keeping me sane my first year in Umeå. **Anders.** Thank you for the massages, all the acupuncture needles in my body when it was hurting, and thanks for feeding me when I grew tired of my own strange grey stews. Thanks **Camilla** for being evil, but in a nice way. :=)

My "old" friends. Ásta, Helgabára and Ella. Thanks.

Louise, Greger & **Sigrid** och **Erika** & **Peter**. Tack för bastu, god mat och kul sällskap. Tack för att jag har fått "låna" Tommy av er. :=)

Pabbi, mamma og **Þorbjörg**, þið eruð best. Thank you so much **Nicolas** for being nice to my sister and thanks for the graphics for my thesis.

Tommy. Tack för allt and THANK YOU for introducing me to the wonderful (but endlessly complex) world of scaleable vector graphics. Sayonara Powerpoint! Ég elska þig.

So many people to thank and so little room in which to do it! If I missed adding your name here and you have somehow helped me make this thesis possible... sorry about that and thank you so much. :=)

This work was supported by a stipend from the Swedish Institute

REFERENCES

Aitman, TJ, Todd, JA: Molecular genetics of diabetes mellitus. *Baillieres Clin Endocrinol Metab* 9:631-656, 1995

Andersson, T: *Den inre omflyttningen. I. Norrland*. Malmö, 1897

Andersson-Palm, L: *Folkmängden i Sveriges socknar och kommuner 1571-1997 med särskild hänsyn till perioden 1571-1751*. Göteborg, Sweden, 2000

Barlow, AB, Wheatcroft, N, Watson, P, Weetman, AP: Association of HLA-DQA1*0501 with Graves' disease in English Caucasian men and women. *Clin Endocrinol (Oxf)* 44:73-77., 1996

Bilbao, JR, Martin-Pagola, A, Perez De Nanclares, G, Calvo, B, Vitoria, JC, Vazquez, F, et al.: HLA-DRB1 and MICA in autoimmunity: common associated alleles in autoimmune disorders. *Ann N Y Acad Sci* 1005:314-318, 2003

Burstedt, MS, Sandgren, O, Holmgren, G, Forsman-Semb, K: Bothnia dystrophy caused by mutations in the cellular retinaldehyde-binding protein gene (RLBP1) on chromosome 15q26. *Invest Ophthalmol Vis Sci* 40:995-1000, 1999

Bylund, E: *Koloniseringen av Botnia-regionen*. Höganäs, Bokförlaget Bra Böcker, 1994

Bylund, E: Theoretical considerations regarding the distribution of settlement in inner north Sweden. *Geografiska Annaler* 53, 1960

Carlsson, G, Fasth, A: Infantile genetic agranulocytosis, morbus Kostmann: presentation of six cases from the original "Kostmann family" and a review. *Acta Paediatr* 90:757-764, 2001

Cudworth, AG, Woodrow, JC: Evidence for HL-A-linked genes in "juvenile" diabetes mellitus. *Br Med J* 3:133-135., 1975

Czaja, AJ, Doherty, DG, Donaldson, PT: Genetic bases of autoimmune hepatitis. *Dig Dis Sci* 47:2139-2150, 2002

Donner, H, Braun, J, Seidl, C, Rau, H, Finke, R, Ventz, M, et al.: Codon 17 polymorphism of the cytotoxic T lymphocyte antigen 4 gene in Hashimoto's thyroiditis and Addison's disease. *J Clin Endocrinol Metab* 82:4130-4132., 1997a

Donner, H, Rau, H, Walfish, PG, Braun, J, Siegmund, T, Finke, R, et al.: CTLA4 alanine-17 confers genetic susceptibility to Graves' disease and to type 1 diabetes mellitus. *J Clin Endocrinol Metab* 82:143-146., 1997b

Eliasson, M, Lindahl, B, Lundberg, V, Stegmayr, B: Diabetes and obesity in Northern Sweden: occurrence and risk factors for stroke and myocardial infarction. *Scand J Public Health Suppl* 61:70-77, 2003

Froguel, P, Vaxillaire, M, Sun, F, Velho, G, Zouali, H, Butel, MO, et al.: Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* 356:162-164, 1992

Gibson, AW, Edberg, JC, Wu, J, Westendorp, RG, Huizinga, TW, Kimberly, RP: Novel single nucleotide polymorphisms in the distal il-10 promoter affect il-10 production and enhance the risk of systemic lupus erythematosus. *J Immunol* 166:3915-3922., 2001

Grams, SE, Moonsamy, PV, Mano, C, Oksenberg, JR, Begovich, AB: Two new HLA-B alleles, B*4422 and B*4704, identified in a study of families with autoimmunity. *Tissue Antigens* 59:338-340, 2002

Hanis, CL, Boerwinkle, E, Chakraborty, R, Ellsworth, DL, Concannon, P, Stirling, B, et al.: A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet* 13:161-166, 1996

Horikawa, Y, Iwasaki, N, Hara, M, Furuta, H, Hinokio, Y, Cockburn, BN, et al.: Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY. *Nat Genet* 17:384-385, 1997

Horikawa, Y, Oda, N, Cox, NJ, Li, X, Orho-Melander, M, Hara, M, et al.: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 26:163-175, 2000

Johansson, A, Vavruch-Nilsson, V, Edin-Liljegren, A, Sjolander, P, Gyllensten, U: Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations. *Hum Genet* 116:105-113, 2005

Kidd, KK, Pakstis, AJ, Speed, WC, Kidd, JR: Understanding human DNA sequence variation. *J Hered* 95:406-420, 2004

King, AL, Yiannakou, JY, Brett, PM, Curtis, D, Morris, MA, Dearlove, AM, et al.: A genome-wide family-based linkage study of coeliac disease. *Ann Hum Genet* 64:479-490, 2000

Kockum, I, Wassmuth, R, Holmberg, E, Michelsen, B, Lernmark, A: HLA-DQ primarily confers protection and HLA-DR susceptibility in type I (insulin-dependent) diabetes studied in population-based affected families and controls. *Am J Hum Genet* 53:150-167., 1993

Koo, JW, Oh, SH, Chang, SO, Park, MH, Lim, MJ, Yoo, TJ, et al.: Association of HLA-DR and type II collagen autoimmunity with Meniere's disease. *Tissue Antigens* 61:99-103, 2003

Kouki, T, Sawai, Y, Gardine, CA, Fisfalen, ME, Alegre, ML, DeGroot, LJ: CTLA-4 gene polymorphism at position 49 in exon 1 reduces the inhibitory function of CTLA-4 and contributes to the pathogenesis of Graves' disease. *J Immunol* 165:6606-6611., 2000

Krempler, F, Esterbauer, H, Weitgasser, R, Ebenbichler, C, Patsch, JR, Miller, K, et al.: A functional polymorphism in the promoter of UCP2 enhances obesity risk but reduces type 2 diabetes risk in obese middle-aged humans. *Diabetes* 51:3331-3335, 2002

Kristinsson, SY, Thorolfsdottir, ET, Talseth, B, Steingrimsson, E, Thorsson, AV, Helgason, T, et al.: MODY in Iceland is associated with mutations in HNF-1alpha and a novel mutation in NeuroD1. *Diabetologia* 44:2098-2103, 2001

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

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

Laan, M, Paabo, S: Demographic history and linkage disequilibrium in human populations. *Nat Genet* 17:435-438, 1997

Lander, E, Kruglyak, L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11:241-247, 1995

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

Lander, ES, Linton, LM, Birren, B, Nusbaum, C, Zody, MC, Baldwin, J, et al.: Initial sequencing and analysis of the human genome. *Nature* 409:860-921, 2001

Lundin, G, Lee, JS, Thunell, S, Anvret, M: Genetic investigation of the porphobilinogen deaminase gene in Swedish acute intermittent porphyria families. *Hum Genet* 100:63-66, 1997

Minde, J, Toolanen, G, Andersson, T, Nennesmo, I, Remahl, IN, Svensson, O, et al.: Familial insensitivity to pain (HSAN V) and a mutation in the NGFB gene. A neurophysiological and pathological study. *Muscle Nerve* 30:752-760, 2004

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

Nistico, L, Buzzetti, R, Pritchard, LE, Van der Auwera, B, Giovannini, C, Bosi, E, et al.: The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Hum Mol Genet* 5:1075-1080., 1996

Nylander, PO, Beckman, L: Population studies in northern Sweden. XVII. Estimates of Finnish and Saamish influence. *Hum Hered* 41:157-167, 1991

Parikh, H, Groop, L: Candidate genes for type 2 diabetes. *Rev Endocr Metab Disord* 5:151-176, 2004

Pastinen, T, Perola, M, Ignatius, J, Sabatti, C, Tainola, P, Levander, M, et al.: Dissecting a population genome for targeted screening of disease mutations. *Hum Mol Genet* 10:2961-2972, 2001

Patil, NS, Pashine, A, Belmares, MP, Liu, W, Kaneshiro, B, Rabinowitz, J, et al.: Rheumatoid arthritis (RA)-associated HLA-DR alleles form less stable complexes with class II-associated invariant chain peptide than non-RA-associated HLA-DR alleles. *J Immunol* 167:7157-7168, 2001

Payami, H, Joe, S, Thomson, G: Autoimmune thyroid disease in type I diabetic families. *Genet Epidemiol* 6:137-141, 1989

Peltonen, L: Identification of Disease Genes in Genetic Isolates. *Methods* 9:129-135, 1996

Peltonen, L: Positional cloning of disease genes: advantages of genetic isolates. *Hum Hered* 50:66-75, 2000

Peltonen, L, Jalanko, A, Varilo, T: Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 8:1913-1923, 1999

Shifman, S, Darvasi, A: The value of isolated populations. *Nat Genet* 28:309-310, 2001

Stoffers, DA, Ferrer, J, Clarke, WL, Habener, JF: Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat Genet* 17:138-139, 1997

Takara, M, Kouki, T, DeGroot, LJ: CTLA-4 AT-repeat polymorphism reduces the inhibitory function of CTLA-4 in Graves' disease. *Thyroid* 13:1083-1089, 2003

Ueda, H, Howson, JM, Esposito, L, Heward, J, Snook, H, Chamberlain, G, et al.: Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 423:506-511, 2003

Varilo, T, Paunio, T, Parker, A, Perola, M, Meyer, J, Terwilliger, JD, et al.: The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in

chromosomes of Finnish populations with different histories. *Hum Mol Genet* 12:51-59, 2003

Varilo, T, Peltonen, L: Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev* 14:316-323, 2004

Wahlund, S: *Demographic studies in the nomadic an the settled population of Northern Lapland*. Uppsala, 1932

Wang, XB, Kakoulidou, M, Giscombe, R, Qiu, Q, Huang, D, Pirskanen, R, et al.: Abnormal expression of CTLA-4 by T cells from patients with myasthenia gravis: effect of an AT-rich gene sequence. *J Neuroimmunol* 130:224-232, 2002

Wright, AF, Carothers, AD, Pirastu, M: Population choice in mapping genes for complex diseases. *Nat Genet* 23:397-404, 1999

Yamagata, K, Furuta, H, Oda, N, Kaisaki, PJ, Menzel, S, Cox, NJ, et al.: Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* 384:458-460, 1996a

Yamagata, K, Oda, N, Kaisaki, PJ, Menzel, S, Furuta, H, Vaxillaire, M, et al.: Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* 384:455-458, 1996b

Yanagawa, T, Mangklabruks, A, Chang, YB, Okamoto, Y, Fisfalen, ME, Curran, PG, et al.: Human histocompatibility leukocyte antigen-DQA1*0501 allele associated with genetic susceptibility to Graves' disease in a Caucasian population. *J Clin Endocrinol Metab* 76:1569-1574., 1993

Zimmet, P, Alberti, KG, Shaw, J: Global and societal implications of the diabetes epidemic. *Nature* 414:782-787, 2001