Genome Variation in Human Populations

Exploring the Effects of Demographic History and the Potential for Mapping of Complex Traits

ÅSA JOHANSSON
Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Rudbeck Laboratory, Dag Hammerskjölds väg 20, Uppsala, Friday, December 8, 2006 at 13:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

A major challenge in human genetics is to understand the genetic variation underlying common diseases. In this thesis, I focus on forces creating differences between individuals and genomic regions, methods for characterizing genomic variation, and the association between genomic and phenotypic variation. Genetic markers are widely used to locate genes associated with different phenotypes. In my first paper, I describe novel algorithms for automatic genotype determination of microsatellite markers, a procedure which is currently both time-consuming and error prone.

The co-segregation of genetic markers in a population leads to non-random association of alleles at different loci - linkage disequilibrium (LD). LD varies throughout the genome and differs between populations due to factors such as their demographic history. In my second paper, I discuss the increased power for mapping of human traits, that results from studying a population with appreciable levels of LD such as is found in the Swedish Sami population.

Lately, large-scale analyses of single nucleotide polymorphisms (SNPs) have become available and efforts have been made to identify a set of SNPs, which captures most of the genome variation in a population (tagSNPs). In my third paper, I describe the limitations of this approach when applied to data from an independent population sample of randomly ascertained SNPs. The transferability of tagSNPs between populations is poor, presumably due to variation in allele frequencies and the bias towards common SNPs used in most studies.

The level of genomic variation is influenced by population structure, recombination and mutation rate, as well as natural selection. During the exodus from Africa, humans have adapted to new environmental conditions. In my fourth paper, I describe a new method for identifying genomic regions carrying signatures of recent positive selection and apply this to an available dataset of millions of SNPs.

Keywords: genetics, evolution, microsatellite, SNP, selection, linkage disequilibrium, haplotype

Åsa Johansson, Department of Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, SE-75185 Uppsala, Sweden
© Åsa Johansson 2006

ISSN 1651-6206
ISBN 91-554-6722-9
urn:nbn:se:uu:diva-7293 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7293)
Publications

This thesis is based on the following papers, which are referred to in the text by their roman numerals:


IV **Johansson Å**, Gyllensten U. Identification of local selective sweeps in human populations since the exodus from Africa. *Manuscript.*

* Reprints were made with kind permission of Springer Science and Business Media.
Contents

Introduction.....................................................................................................7
Genome variation .......................................................................................7
  Linkage disequilibrium .........................................................................8
Genetic drift ............................................................................................9
Natural selection ....................................................................................9
Mapping of complex diseases .................................................................10
  Linkage analysis and whole genome association studies ....................12
  A haplotype map of the human genome ..............................................12
The Sami population ................................................................................13
  Origin of Sami .....................................................................................13
  Diet, lifestyle and disease in the Sami population ...............................15
  Potential for population-based epidemiology ....................................16
Present investigation ................................................................................18
  Paper I ..................................................................................................18
  Paper II ................................................................................................20
  Paper III ...............................................................................................24
  Paper IV ...............................................................................................27
Conclusions and further perspectives .....................................................33
Acknowledgements ..................................................................................36
References ...............................................................................................37
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>bp</td>
<td>Base Pair</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FR</td>
<td>$F_{ST}$ ratio</td>
</tr>
<tr>
<td>HDL</td>
<td>High-density lipoprotein</td>
</tr>
<tr>
<td>HLA</td>
<td>Human Leukocyte Antigen</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>LDL</td>
<td>Low-density lipoprotein</td>
</tr>
<tr>
<td>lod</td>
<td>Logarithm of the odds</td>
</tr>
<tr>
<td>LR</td>
<td>Length Ratio</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor Allele Frequency</td>
</tr>
<tr>
<td>mMAF</td>
<td>Minimum Minor Allele Frequency</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>QTL</td>
<td>Quantitative Trait Locus</td>
</tr>
<tr>
<td>s</td>
<td>Selection coefficient</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>STR</td>
<td>Short Tandem Repeat</td>
</tr>
<tr>
<td>YBP</td>
<td>Years Before Present</td>
</tr>
</tbody>
</table>
Introduction

Mapping genes underlying complex phenotypes is a major task in medical genetics. The development of high throughput genotyping instruments and sophisticated statistical tools has given a new dimension to the characterization of co-variation between genomic and phenotypic variation. This information reveals clues about the evolutionary history of humans and facilitates the understanding of non-communicable diseases. In this introduction, I first briefly describe the variation seen in the human genome. I will focus on two forces that influence the level of variation in different populations and genomic regions, genetic drift and natural selection. In the second part of the introduction, I describe strategies and efforts made to identify genes behind phenotypic variation. Finally, I introduce you to a population with a unique demographic history, ideal for epidemiological studies and for mapping of genes underlying complex traits.

Genome variation

Every human being is unique. The differences between individuals are due to the environment and to differences in the DNA sequence. The human genome consists of about 3 billion nucleotides (International Human Genome Sequencing Consortium 2004) and the number of nucleotide differences between two randomly chosen individuals is roughly 1 in every 1,300 base pair (bp) (Sachidanandam et al. 2001). However, the difference between humans is small compared to the more than 1 difference every 100 bp between us and our closest relative, the chimpanzee (Watanabe et al. 2004; Chimpanzee Sequencing and Analysis Consortium 2005).

The most abundant kind of variation between individuals is single nucleotide polymorphisms (SNPs). The estimated number of SNPs in the human population is about 11,000,000 (Kruglyak and Nickerson 2001), of which probably less than 1% cause alteration of any protein sequence. In addition, a large number of indels (insertion and deletion polymorphisms) and repeat elements can be found throughout the genome. The most common repeat elements are microsatellite markers or short tandem repeats. Microsatellites have a short core sequence of 1-6 units and the polymorphism is due to differences in the number of times the core is repeated in tandem. About 3% of the genome is comprised of simple sequence repeats (1-13 bases) of which
0.5% are dinucleotide repeats (Lander et al. 2001). Both SNPs and microsatellites are commonly used as genetic markers, i.e. loci that distinguish individuals without being functional polymorphisms.

With a few exceptions, most microsatellites does not influence any phenotype. Similarly, most of the SNP variation is neutral, while a fraction of SNPs are known to cause phenotypic differences. Each SNP normally has two variants and a low mutation rate of about $10^{-8}$ (Drake et al. 1998), whereas microsatellites have more alleles per locus and a higher mutation rate of up to $10^{-3}$ (Weber and Wong 1993). To date, at least 15,000 microsatellite markers have been validated for use in genotyping, whereas millions of SNPs are available. The co-segregation of genetic markers on a chromosome can lead to non-random association of alleles at different loci, or linkage disequilibrium (LD).

**Linkage disequilibrium**

There is a higher probability of two nearby loci to be inherited together than if they are far apart. This non-random segregation, leading to disequilibrium between markers, is referred to as LD. The pattern of LD in human genes has mostly been studied at short genomic distances using SNPs. Microsatellite markers generally show higher levels of LD, extending over longer physical distances, than markers with lower allelic variation. Due to the higher mutation rate of microsatellites, the LD between these markers reflects events that are more recent, whereas LD between SNPs might be older (Kauppi et al. 2003). The extent of LD varies throughout the genome. The main force breaking down LD is recombination, but a high mutation rate has a similar effect. There are also a number of factors that may increase LD, such as recent admixture between populations with different allele frequencies, inbreeding, a small population size over time, or a population bottleneck (Hill 1968; Chakraborty and Weiss 1988; Sheffield et al. 1998; Terwilliger et al. 1998; Shifman and Darvasi 2001). Some human populations have been found to have higher levels of LD than others. For example, the populations of Finland, Iceland and Sardinia have a history of isolation and subsequent expansion and show a higher level of LD than outbred European populations (Peterson et al. 1995; Dunning et al. 2000; Varilo et al. 2000; Mohlke et al. 2001; Angius et al. 2002; Varilo et al. 2003). Populations that have undergone certain kinds of demographic events, leading to higher LD, have been widely used in medical genetic studies. These populations have been very useful for the mapping of a number of recessive disease-causing genes (Norio 2003). Also, populations with a small and constant population size have been suggested to have increased LD, but in this case it is due to genetic drift rather than the founding from a limited number of individuals (Ohta 1982).
Genetic drift

The variation in allele frequencies from one generation to the next, caused by limited population size, is called genetic drift. Any polymorphism in a population has a certain probability of being eliminated or fixed. Since the probability of fixation of an allele equals its frequency, an allele is more likely to be fixed in a small population. Further, a small population will show larger random fluctuations in allele and genotype frequencies over time, with an average magnitude inversely proportional to the effective population size. Over time, the consequence of such fluctuations is some alleles being lost, resulting in decreased heterozygosity. On the other hand, new or rare mutations have higher chance to persist in an expanding population, compared to a population with constant population size.

Fluctuation in allele frequencies at linked sites results in fluctuation of haplotype frequencies and thereby increased LD. A small population is more affected by genetic drift and will consequently have increased LD relative to a large population. LD can also increase locally in the genome because of positive selection.

Natural selection

Even though most mutations in the genome are neutral (Kimura 1968), some mutations decrease or even increase the fitness of an individual. A mutation that decreases the fitness of an individual will have a higher probability of being eliminated from the population (negative or purifying selection), whereas a mutation that increases fitness will have higher chance of increasing in frequency or becoming fixed (positive selection). For some loci, high allelic variability is maintained through the higher fitness of heterozygous individuals, a phenomenon called balancing selection. One of the most well studied genetic systems, the histocompatibility locus of human leukocyte antigen (HLA), was early found to exhibit reduced homozygosity (Hedrick and Thomson 1983) due to balancing selection. HLA is the most polymorphic locus in the genome and plays a key role in the immune response.

Since the separation of the ancestors of modern humans and our closest extant relative, the chimpanzee, our genomes have diverged. Most of the differences are due to new mutations and genetic drift. In some cases, new mutations have led to phenotypes favorable to humans and some alleles have become fixed due to positive selection. New variants are being fixed either by random forces or by means of selection. One example is our language. This unique human ability for complex communication has been suggested to be affected by *FOXP1*, based on studies showing that a mutation in this gene causes a severe speech and language disorder (Lai et al. 2001). Inter-species comparisons of *FOXP1* revealed two amino acid changes on the
human lineage and, due to the lack of polymorphism in humans, a complete selective sweep was suggested (Enard et al. 2002).

Some adaptations to new environmental conditions have also been seen in different human populations. For instance, populations depending on agriculture often have a high tolerance to lactose and this is associated with a mutation in the lactase gene (Hollox et al. 2001; Enattah et al. 2002). The LCT gene is likely to have been under positive selection due to increased nutrition when consuming dairy products, which were introduced to humans during cattle domestication in Near East about 9,000 years ago (Bersaglieri et al. 2004).

While genetic drift has a large impact in small populations, the effect of natural selection increases proportionally to the effective population size. Due to the historically relatively small effective population size of humans, the fraction of polymorphism in the genome attributable to selection rather than genetic drift is expected to be low. Positive selection acting on a locus is expected to result in a more rapid fixation rate, or a change in allele frequency, and consequently less variation is seen around a site under selection. Selection increases the frequency of the beneficial allele but also of alleles at neighboring sites, resulting in increased LD in that region. Eventually, one large haplotype can be detected around the site under selection. This has been the case for the LCT gene, where more than 1 Mb of almost complete haplotype has been found to be associated with lactase persistence (Poulter et al. 2003).

Mapping of complex diseases

One of our most demanding tasks in medical genetics is mapping genes involved in the development of complex diseases. A complex disease is caused by a number of factors, both genetic and environmental. In addition, the main contributor can vary between populations. The low effect from each of many different factors makes it hard to identify the genes involved in the development of these diseases.

The mapping of genes involved in monogenic diseases has been quite successful. Unfortunately, most of non-communicable diseases in humans, such as epilepsy, migraine, multiple sclerosis, asthma, diabetes and heart- and lung diseases, are not monogenic. In the field of complex diseases, only a small fraction of the disease-causing genes have been identified. A strategy for mapping complex diseases has been to map genes for a particular trait, which predisposes to the disease. For example, obesity is a predisposing factor for a number of diseases such as diabetes, hypertension and coronary heart disease. Predisposition for diseases might be less complex than the disease itself. In addition, the classification of individuals into affected or
unaffected by a disease is not always trivial and the diagnosis might be subjective. Another issue is that information about a phenotype decreases if individuals are classified rather than measuring the phenotype as a quantitative variable. In the case of obesity, the measure of body mass index (BMI) is much more informative than dividing individuals into obese and non-obese (Figure 1).

To study a specific disease, the strategy is either to study the inheritance pattern in a pedigree or to compare variation in genetic markers between patients and controls. When studying quantitative traits, all individuals in a population are eligible and the cohort is not restricted to families with affected individuals. The variance of most quantitative characters is due to both environmental and genetic factors. For any two populations or sets of individuals, the variation might not be caused by the same factors, or the size of the effect of the factors might differ. Factors with high effect are easier to identify compared to minor factors. Presumably, using a population sample with a reduced number of factors will increase the power to identify individual factors (genetic or non-genetic). Still, most polymorphisms are shared between human populations and a genetic factor contributing to phenotypic variation in one population is likely to also contribute to the variation of the same trait in other populations.

The phenotypic variance of a trait depends on the genetic component and the environmental component. The heritability (h²) expresses the extent to which phenotypes are determined by genes transmitted from the parents. Presumably, the heritability is inversely correlated with the environmental variance and consequently minimizing the environmental variation in a population sample will increase the heritability. In addition, heritability of a character differs between populations, between groups of individuals and

---

Figure 1. Distribution of a quantitative trait. Most traits have a continuous distribution in a population. Classifying individuals using a threshold into affected or unaffected will result in loss of information as compared to using the empirical quantitative estimate of each individual.
over time for the same set of individuals. A high heritability also indicates the power for identifying the genes affecting a phenotype.

**Linkage analysis and whole genome association studies**

Linkage analysis is a method to find genomic regions segregating with a disease or a phenotypic trait in a pedigree. If a genetic marker is closely located to the locus affecting the trait value, it is more likely to be inherited together with the trait without recombination taking place in between the genetic marker and the causative allele. Consequently, the number of recombination in the pedigree is one main determinant of power in linkage analysis. Due to the higher polymorphism of microsatellite markers, these have traditionally been used in linkage analysis.

Association mapping, or LD mapping, is based on the association of an allele and the variation of a trait. A marker allele, which is associated with the variation of the trait value, can either be the causative allele or, more likely, in LD with the causative allele. Consequently, the strength of LD in the genomic region and the density of markers are both limiting factors in association mapping. Most genetic factors affecting phenotypic variation are likely to be SNPs. Consequently, the lower mutation rate of SNPs relative to microsatellites, makes SNPs more suited to association mapping. Until recently, association mapping has been used mainly to narrow down a genomic region identified through linkage analysis but lately, high throughput SNP typing instruments has been developed and whole genome association mapping is now performed including more than 500,000 SNPs. The large efforts made to characterize SNP variation for use in association mapping has resulted in a haplotype map of the human genome.

**A haplotype map of the human genome**

At the same time as the initial sequence of the human genome was published (Lander et al. 2001), a large number of SNPs were also identified and a SNP map, consisting of 1.42 million sites, was constructed (Sachidanandam et al. 2001). About two years later, the International HapMap Project was initiated (International HapMap Consortium 2003). The aim of this project was to determine the common DNA sequence variation in the human genome. This variation has been described as differences in allele frequencies and correlations between SNPs, using human populations from different continents. The large amount of data generated was intended to be useful in the mapping of human phenotypes. To date, the International HapMap Project includes more than 4 million SNPs in 270 individuals from 4 different populations and is one of the most important sources of information about the variation in the human genome.
It has been suggested that the pattern of recombination hotspots in the genome has resulted in a block-like structure of linkage disequilibrium and low haplotype diversity. This structure makes it possible to divide the genome into haplotype blocks or even to construct a haplotype map for each human population (Altshuler et al. 2005; Hinds et al. 2005). Each block can be described by a fraction of the included SNPs, which are called haplotype tagging SNPs (htSNPs) or tagging SNPs (tagSNPs). These SNPs should be able to capture most of the variation in the genome (Johnson et al. 2001). The identification of tagSNPs in a population may dramatically reduce the amount of genotyping needed to determine the variation of common SNPs in the human genome.

The Sami population

Sami are the indigenous people of the northernmost parts of Sweden, Finland, Norway and the Kola Peninsula of Russia (Figure 2). They speak a language belonging to the Finno-Ugric branch of the Uralic language family, with Finns, Karelians, and Estonians as their closest linguistic neighbors. The Sami where initially hunters (mainly of reindeer and moose), but over time they domesticated the reindeer and became reindeer herders. Today, the Sami population is estimated to be less than 100,000 individuals (Hassler et al. 2004). Some of the Sami are still reindeer herders and maintain a traditional lifestyle linked to the annual migration of reindeer between summer and winter grazing areas, while the rest have other occupations. The Sami population is not known to have experienced any dramatic population changes and, until recently, marriages between Sami and non-Sami families are believed to have been infrequent.

Origin of Sami

Sami are the extreme genetic outliers in Europe (Cavalli-Sforza and Piazza 1993) and their origin has for a long time been the focus of speculation. Archaeological findings on the west coast of Sweden have been dated to more than 10,200 YBP (years before present) (Nordqvist 2000) and radiocarbon dates also indicate that northern Fennoscandia was populated more than 9,000 YBP (Bergman 2004). However, whether or not those findings are from a pre-Sami culture is not known. By the 1950’s, the first studies of genetic markers in the Sami population (Allison et al. 1952; Allison et al. 1956; Beckman et al. 1959) were performed in order to investigate their origin. During the following decades, a large number of genetic markers were investigated such as blood groups, enzyme groups, and serum proteins (Beckman 1996). Most markers showed a difference between Sami and the general Swedish population and a higher similarity between Sami and Finns.
Some anthropometric studies agree with an influence to Sami from Mongolian populations. However, this relationship was not generally supported by the genetic studies, even though a few of the markers did show a similarity in allele frequencies between Sami and Asian populations (Beckman et al. 1988; Sikstrom and Nylander 1990; Beckman et al. 1993; Fan et al. 1993; Larsen et al. 2001; Evseeva et al. 2002).

Starting in the mid 1990’s, a number of studies based mainly on mitochondrial DNA (mtDNA) have been performed. Mitochondrial DNA has been widely used for population studies due to its maternal inheritance pattern and lack of recombination (Giles et al. 1980). Over 80% of all Sami exhibited one of two mtDNA motifs, one of which has been referred to as the Sami-specific motif (Sajantila et al. 1995). The Sami-specific lineage, referred to as haplogroup U5b1b1, is a subgroup of the more diverse haplogroup U5b1b, which is found throughout Europe (Tambets et al. 2004). U5b1b1 was consequently suggested to have arisen in Eastern Europe, whereas U5b1b was likely of southwestern European origin (Tambets et al. 2004). Similarly, the other major haplogroup, haplogroup V, was suggested to have an origin in Western Europe and an age of 16,000 years, compared to 8,000 years in Eastern Europe (Torroni et al. 2001). Other minor haplogroups that have been found in lower frequencies in Sami are H, D5 and Z (Tambets et al. 2004). Haplogroup H is frequent in European populations (Tambets et al. 2004), whereas both D5 and Z are likely to have Asian origins (Tambets et al. 2004; Ingman and Gyllensten 2006).

However, the history of the mitochondrial lineages in Sami agrees with one theory of the colonization of the European continent. During the early Upper Paleolithic, about 45,000 YBP, modern human hunter-gatherers started to colonize the European continent (Cavalli-Sforza et al. 1994).
Throughout the last glaciation, with its peak about 18,000-20,000 YBP (Mellars 1994), some groups of hunter-gatherers still populated small refuges of warmer climates. When the climate improved during the Upper Paleolithic of Mesolithic period, humans re-colonized the European continent. During the Neolithic period, farming was developed and a new spread of people, farmers, took place in Europe.

The most northern part of Europe is believed to have been populated some time after the last glaciation. The findings of an origin of both V and U5b1b in Western Europe about the same time as the last glaciation, suggests those haplogroups to have arisen from the hunter-gatherers who at this time still populated small refuges in Europe. During the recolonization of the European continent after the glaciation, both haplogroups could have spread eastwards, leading to the origin of the Sami haplogroups in Eastern Europe about 8,000 YBP.

One location that has been suggested to be the origin of the pre-Sami population is the area around Lake Ladoga and Lake Onega (Eriksson 1988), located in the east of Finland (Figure 2). The first settlers to this region arrived about 7,000 YBP and it has been suggested that during the same time period there was an influence from southeast Mongolian and Finno-Ugric people. This also agrees with the finding that some mitochondrial haplogroups with Asian origin are found in the Sami population.

One marker in Sami indicating a somewhat mixed ancestry is the lactase gene (LCT). Since the establishing of a farming culture in Near East during the Neolithic period, lactose persistence is believed to have spread in response to the selective advantage of being able to digest milk (Bersaglieri et al. 2004). Lactose tolerance in the Sami varies between 40 and 75 % for different subpopulations (Sahi 1994; Kozlov and Lisitsyn 1997), which is lower in comparison to the general Swedish population (91 %). However, Sami have been involved in reindeer herding only during the last 1,000 years (Kozlov and Lisitsyn 1997) and reindeer milk, which is low in lactose, has not had widespread use in Sami before the 1920’s (Haglin 1991). The high frequency of lactose tolerance in Sami, given their relatively short exposure to dairy products with a high lactose level, is surprising. This can be a reflection of strong genetic drift (Kozlov and Lisitsyn 1997) or recent admixture, but might also be an indicator of a large contribution of farming people to the Sami gene pool as they spread through Europe during the Neolithic period.

Diet, lifestyle and disease in the Sami population

The traditional Sami diet is high in animal products, mainly reindeer with the addition of fish and berries when available (Haglin 1991). Over the last decades, a large proportion of the Sami people have adapted to a diet more typical for Western cultures, but the traditional diet persists in Sami groups in-
volved with reindeer herding. The Sami diet is generally high in protein and fat and low in carbohydrate, fruits and vegetables (Haglin 1991, 1999; Nil- sensen et al. 1999). This diet might have arisen not only due to availability but also from the need of a high energy diet to carry out their work (Ross et al. 2006). Reindeer herding is a labor intensive occupation and even though motorized transports are available today, Sami have a higher physical activity compared to non-reindeer herding Sami and location-matched controls (Hermansen et al. 2002; Ross et al. submitted). However, considering the differences in dietary intake, Sami do not appear to have any differences in serum lipid profiles (Laitinen et al. 1996; Edin-Liljegren et al. 2004). It was shown earlier that Sami have a lower mortality from cardiovascular disease relative to other populations in the same area (Thelle and Forde 1979; Luoma et al. 1995a; Luoma et al. 1995b; Nayha 1997), but more recent studies have found that Sami have a more similar incidence of cardiovascular diseases (Hassler et al. 2005; Sjölander et al. in press). Reindeer meat has double the amount of α-tocopherol and quadruple the amount of selenium compared to both pork and beef, and a correlation has been seen between increased levels serum concentrations of α-tocopherol, albumin and selenium and decreased risk of coronary heart disease (Luoma et al. 1995b). Sami have been found to have a lipid profile suggested to decrease the risk of cardiovascular disease (Sampels et al. 2004) but in the most recent study of fatty acid composition, no difference was found between Swedish Sami and demographically matched controls (Edin-Liljegren et al. 2004). The results of more recent studies might reflect the fast transition in the Sami to a more Westernized lifestyle, partly in response to the high levels of radioactive fallout following the Chernobyl disaster (Ross et al. 2006).

Even though Sami were highly affected by radioactive fallout, both due to nuclear testing by the former Soviet Union (1950’s) and the Chernobyl disas- ter (1986), the overall risk for cancer is lower (40%) in Sami (Ross et al. 2006). One exception is stomach cancer were an elevated risk has been explained by their high intake of smoked and salted reindeer meat and low intake of foods containing fiber (Wiklund et al. 1990). On the other hand, reindeer meat contains a high amount of selenium, which has been suggested to be protective against cancer (Rayman 2005) and may explain some of the decreased overall risk.

Potential for population-based epidemiology

Studies of the Sami provide unusual opportunities to understand the causes of common diseases. While some of the Sami people are leading a traditional lifestyle, others have adapted a more Westernized lifestyle. This enables comparisons between Sami and non-Sami living in the same area but also between reindeer herding Sami and Sami with other occupations (Figure 3).
Figure 3. Comparing environmental and genetic factors. At least three population groups can be distinguished: reindeer herding Sami, Sami with other occupations, and non-Sami living in the same geographic area. The effect of environmental factors can be addressed by comparison between groups with the same genetic background but different lifestyle, and the effect of genetic factors can be addressed by comparison of non-reindeer herding Sami and non-Sami individuals.

These comparisons can facilitate the identification of genetic and environmental factors affecting human traits. For instance, environmental effects can be addressed by a comparison between groups with the same genetic background, and the effect of genetic factors can be similarly addressed by comparison of non-reindeer herding Sami and matched non-Sami controls. Sami maintaining a traditional lifestyle are also likely to be exposed to less environmental variation than the multicultural population of southern and central Sweden. Less environmental variance will result in higher heritability and more power to identify genetic components influencing a complex phenotype, which in the end will lead to better understanding of the development of common diseases.
Present investigation

Paper I: A novel method for automatic genotyping of microsatellite markers based on parametric pattern recognition

Aim
To develop accurate and efficient algorithms for converting the output from the microsatellite typing instrument into genotypes with the minimum number of errors.

Background
A whole genome scan normally includes about 400-800 microsatellite markers from several hundred individuals. Today, high-throughput techniques, including multiplex PCR (polymerase chain reaction) and capillary electrophoresis, will generate an enormous amount data each day. One limiting step in this procedure is the conversion of the capillary raw data into genotypes. The routine at our lab has been to:

1. Manually import data and define bins for each marker (decide which estimated length of the PCR fragments should represent each allele in a marker)
2. Use a program to assign each marker for each individual into the bins
3. Check each marker manually twice

This procedure is very time consuming, subjective and error prone. Errors and inconsistencies introduced during the genotyping will have a dramatic effect on the ability to map the genes underlying a phenotype. Resolving of raw data into genotypes is complicated by:

1. The Taq polymerase misreads the number of repeats in the marker, resulting in PCR fragments missing one or a few repeats (stutter bands)
2. For a fraction of the PCR fragments, a non-templated nucleotide is added to the 3’ end

Resolving these patterns is further complicated by the unknown fraction of fragments with different lengths, depending on which marker studying, PCR machine, batch of Taq polymerase, and the concentration of the DNA used. In addition, the length differences between alleles are in some cases only two
nucleotides (dinucleotides) making the pattern of the two alleles partly overlapping.

Results and Discussion

We have developed algorithms for the processing of raw data, fragment size analysis, automated allele binning and miscall determination. None of the steps need manual editing. Data is imported from the sequencing instrument and the complex pattern of each allele is resolved into genotypes. The data is checked for potential sources of errors such as overlapping spectra from the signals from the sequencing instrument, extra signals due to contamination, very low signal, and genotypes that do not agree with the expected fragment sizes. The algorithms were evaluated by comparing the results from 7,489 genotypes analyzed by our algorithms, to the same genotypes analyzed by the first version of GeneMapper™ software. All genotypes that disagree between the two applications were manually checked.

The comparison showed that our algorithms both have a higher number of genotypes assigned to a bin and makes fewer miscalls. This made our algorithms promising for use in analyzing genotype data. However, our method was slowed down by the combination of large datasets, use of a high technology programming language such as MATLAB on a 450 MHz Mac G3 computer and possibly, inefficient programming algorithms. Compared to the version of the GeneMapper™ software used at that time, the speed of our algorithms was sufficient. To date, a large number of new and more sophisticated versions of GeneMapper™ software are available and the need for alternative software has diminished.

During the last years, the field of gene mapping has gone through changes. Typing microsatellites is not the obvious choice when searching for disease genes. One explanation is the complications with high throughput techniques for microsatellite genotyping. In addition, data based on microsatellites are not suited for comparing or merging across studies, due to the variance in the performances of the PCR and the inconsistencies in binning of microsatellite alleles (Presson et al. 2006). The use of high throughput SNP typing, which enables analyzing genome wide linkage and association at the same time, has increased at the expense of microsatellites.
Paper II: Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations

Aim
To evaluate the extent of LD in the Sami population compared to populations with other demographic histories, and to evaluate the potential for mapping of complex traits in the Sami population.

Background
The idea of using small isolated populations for the mapping of complex traits may derive in part from the large number of rare disease genes successfully mapped in the Finnish population. The Finnish population has been partly isolated and the genetic diversity is low, relative to other European populations (Sajantila et al. 1996). Already in the early 1970’s, a large number of rare diseases were described in the Finnish population (Norio et al. 1973). To date, at least 36 rare diseases have been characterized (Norio 2003) and for most of them a single causative gene has been identified. One of the reasons for the success in mapping of disease genes in the Finnish population is the monogenic inheritance of these diseases. The large number of rare recessive diseases found in Finns is due to the population history. The Finnish population is believed to have undergone a population bottleneck approximately 4,000 YBP (Sajantila et al. 1996), which has resulted in the founding of the population from a small number of ancestors followed by rapid population growth. The rare recessive alleles existing in the founding population have thereby been able to maintain their frequencies.

Even though many recessive genes have been identified in the Finnish population, the success in identifying genes involved in complex phenotypes has been limited. LD is a major determinant of the power to map complex phenotypes and, while LD decays with time in an expanding population, a small population has the advantage that LD is instead generated over time by genetic drift. Terwilliger and colleagues (Terwilliger et al. 1998) proposed that drift mapping - linkage disequilibrium mapping in small, old populations of constant size, where the disequilibrium is the result of genetic drift rather than founder effect - might be more efficient for identifying common disease alleles. One population with a demographic history of small and constant population size over time is the Swedish Sami population.

Results and Discussion
In paper II, the genome-wide pattern of LD among nearly 400 microsatellite loci was examined in the Swedish Sami population and compared to the data
from 29 populations within the Human Genome Diversity Project (Cavalli-Sforza et al. 1991; Cann et al. 2002). The Human Genome Diversity Panel represents populations with range of different demographic histories. Due to the low-density set of markers used, LD was low in most populations. Restricting the analysis to samples with unrelated individuals revealed three populations with increased LD between markers in the interval of 0-6 Mb (Brahui from Pakistan, Yakut from Siberia and Sami from Sweden). The low number of populations with increased LD is not surprising since LD is not expected to extend over Mb distances in large or recently expanded populations. In addition, we studied LD between more closely spaced markers in Sami and the general Swedish population. We examined 88 microsatellite markers from 10 genomic regions of about 10 Mb each and found a significantly increased LD in Sami of up to 2.5 Mb. We also observed a somewhat lower heterozygosity in Sami compared to the general Swedish population.

The increased LD in Sami is not surprising. The history of at least partial separation from other populations, together with a small and constant population size over time made Sami a good candidate to be highly affected by genetic drift (Terwilliger et al. 1998). Genetic drift will not only increase LD but will also decrease diversity, which we found in Sami represented by a lower average number of alleles at the microsatellite loci. Fluctuations in allele frequencies, due to genetic drift, give a higher chance for alleles to become fixed. Many genetic markers used to study the genetic origin of Sami have shown large differences in Sami compared to other European populations but also large differentiation between different Sami groups (Beckman 1996). The later is consistent with either genetic drift, admixture or a combination of both. So far no comprehensive study of the level of admixture between Sami and the populations living in the same area has been performed. Both mitochondrial data (Ingman and Gyllensten 2006) and preliminary data of HLA haplotypes (Johansson et al. unpublished data), indicate a more similar frequency spectrum between Sami in the southern part of their range and other European populations as compared to more deviation frequencies between Sami in the northern part of the range and European populations. However, whether this reflects genetic drift or admixture, and in case of the latter if it represents recent or historical admixture, has not been investigated.

In the last couple of years, many studies based on admixture mapping have been published, supporting the power of this approach. The basic idea of admixture mapping is that admixture results in the generation of long haplotype blocks. Admixture mapping attempts to detect disease-associated variants that differ in frequency between populations. In the case of Sami, we do not know if our increase in LD is due to genetic drift or recent admixture, but either way it indicated a potential for the mapping of complex traits. Preliminary data based on pedigree analysis of about 400 individuals in the Sami population shows high heritability for a number of traits (Table 1).
Table 1. Heritability for a number of traits in Sami and two populations with similar population structure

<table>
<thead>
<tr>
<th>Trait</th>
<th>Sami (Sweden)</th>
<th>Kosraen (Micronesia)</th>
<th>Sardinian (Italy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fasting blood sugar</td>
<td>0.34(^1)</td>
<td>0.24</td>
<td>Na</td>
</tr>
<tr>
<td>BMI</td>
<td>0.44(^2)</td>
<td>0.45</td>
<td>0.43</td>
</tr>
<tr>
<td>Body Stature</td>
<td>0.90(^3)</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>0.30(^3)</td>
<td>0.29</td>
<td>0.25</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>0.16(^3)</td>
<td>0.25</td>
<td>0.19</td>
</tr>
<tr>
<td>HDL</td>
<td>0.46(^3)</td>
<td>Na</td>
<td>0.49</td>
</tr>
<tr>
<td>LDL</td>
<td>0.52(^3)</td>
<td>Na</td>
<td>0.43</td>
</tr>
<tr>
<td>Serum total cholesterol</td>
<td>0.49(^3)</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.41(^3)</td>
<td>0.21</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Abbreviations used: BMI-Body mass index; HDL-High density lipoprotein; LDL-Low density lipoprotein; Na-not available.

1) Heritability were calculated using the SOLAR software (Almasy and Blangero 1998)
2) Sex and age were used as covariates
3) Sex, age and BMI were used as covariates

The heritability in Sami is comparable to the estimates in other populations that have been chosen for gene mapping based to their demographic history. The highest heritability in Sami is seen for body stature. Even though this phenotype is known to have high heritability in human populations, the level in Sami is elevated. Sami are shorter (p<0.001) compared to non-Sami controls matched for age and place of living, with an average stature for females of 160.9 cm as compared to 164.1 cm, and for males 175.1 cm compared to 177.2 cm (Johansson et al. unpublished data). Interestingly, the standard deviation is larger in Sami compared to the controls (Figure 4) which, in agreement with the high heritability, might be explained by a recent admixture with the general Swedish population. Body stature is a well-studied character in humans, mainly due to the simplicity in measuring. Even though it is probably the most common variable in any study of quantitative traits, the success in identifying candidate genes for normal variation in body stature has been very limited. This indicates a complex inheritance pattern with many genes each contributing to a small effect, rather than a few genes with large effects. The use of a population such as Sami might result in fewer genetic effects due to the reduction in heterogeneity and thereby higher power to identify the underlying genetic effects. In addition, stature is a character which is particularly interesting to study in Sami due to the correlation between body size in a population and the latitude of the settlement or the average temperature; a relationship seen in many species (Bergmann 1847; Allen 1877; Roberts 1953).
Figure 4. Variation in stature among Sami and a non-Sami reference group from the same area.

The increased LD in Sami reflects genetic drift, admixture or a combination of both. Whichever, Sami is an interesting population for genetic studies not only because of the potential for mapping complex traits. The origin of Sami is still uncertain and so far, no comprehensive study has been performed to address this question using autosomal markers. The lower incidence of cardiovascular disease, which has been suggested in Sami, also makes markers associated with serum lipid levels interesting for further investigation. In addition, the ability to stratify the Sami population into reindeer-herding Sami who lead a traditional lifestyle and Sami who have adopted a Westernized lifestyle, makes it possible to estimate the relative effects of genetic and environmental factors contributing to the variation of phenotypic traits.
Paper III: Evaluation of the SNP tagging approach in an independent population sample – SNP discovery in the native population of Sami

Aim
To evaluate the use of tagSNPs by: 1) testing the transferability of tagSNPs from a large diverse population to the small Sami population, 2) testing the transferability of tagSNPs identified in a dataset of common SNPs, for their performance in a dataset of random sequence variation, and 3) comparing the number of tagSNPs required in Sami, relative to a large diverse population, to capture most of common SNPs.

Background
Since the determination of the human genome sequence, large efforts have been made to study genomic variation in humans (Sachidanandam et al. 2001; International HapMap Consortium 2003; Hinds et al. 2005). The total number of SNPs in the human population has been estimated to be 11 million and to date the SNP database (NCBI dbSNP Build 126) includes 5,546,513 SNPs that can be downloaded with genotype data. One major use of these data has been to study the correlation structure of common SNPs in the human genome and using this structure identify sets of tagSNPs that captures most of the genetic variation.

Results and Discussion
In our study, we have used SNP array-based discovery to randomly identify sequence variation in a 4.4 Mb region of chromosome 21 from unrelated Sami individuals. The two copies of chromosome 21 were physically separated in order to determine the complete phased haplotype for each chromosome fragment. As reference data we have used publicly available data from the HapMap populations: YRI - Yoruba from Nigeria, CEU - Utah residents with ancestry from northern and western Europe, CHB - Han Chinese form Beijing and JPT - Japanese from Tokyo (www.hapmap.org).

Our analysis did not reveal high transferability of tagSNPs between populations, or even between different sets of SNPs. The tagSNPs defined in any of the three HapMap populations (YRI, CEU or CHB+JPT) did not capture more of the variation in Sami as compared to randomly ascertained SNPs from the set of SNPs typed in HapMap. However, the power to detect new variants in Sami that have not been typed in the HapMap populations was, surprisingly, much lower than choosing the same number of tagSNPs randomly among all SNPs identified in Sami. There is higher average allele frequency for sites in Sami that overlap with sites typed in HapMap,
consistent with the bias towards high frequency SNPs in the HapMap data. In addition, among the overlapping sites there is a difference in allele frequency between sites that are tagged by any of the CEU tagSNPs and sites that are untagged. Untagged sites have both a lower average MAF and the differentiation from CEU is larger compared to tagged sites. A large deviation is seen in the distribution of allele frequency between our Sami data and the HapMap data. The observed distribution of mMAF in the Sami is in quite good agreement with the expectation (Figure 5) when applying Ewens sampling formula (Ewens 1972), \( S_\text{f} = S_2 \ln \left( \frac{1-f}{f} \right) \), where \( S_2 \) is the number of pairwise differences in the sample (2,543 in Sami), and \( S_\text{f} \) is the number of SNPs in the sample with a mMAF of \( f \). The SNP frequency distributions in the four HapMap population samples deviate substantially from expected distributions for large human populations (Kruglyak and Nickerson 2001), in particular for mMAF below 0.1 (Figure 5). This difference in allele frequencies can probably, at least partly, explain the low transferability of SNPs between datasets.

The poor transferability of tagSNPs might also be due to population differentiation. Untagged sites deviate more in allele frequencies compared to tagged sites, and the transferability is slightly higher from a population with European descent (CEU) compared to Asian or African (JPT+HCB or YRI). Even though Sami are genetic outliers in Europe, the overall differentiation is smaller between Sami and the European American population than to any of the other populations (Figure 6). In light of the small genetic distance between Sami and the European American population, the low transferability is not likely to be explained by differentiation alone.
In summary, we observe a dramatic decrease in performance of tagSNPs when they are applied to a different dataset than the one used to define them. Defining tagSNPs using a dataset such as HapMap, which is biased towards common variants, results in reduced coverage of the sequence variation. The explanation for our poor results is probably a combination of different SNPs and frequency distributions, population differentiation and inconsistency in the pattern of correlation structure between the Sami and the HapMap data. A lower correlation between SNPs that differ in allele frequencies is not surprising. The most commonly used measure of LD, \( r^2 \), is 1 only if the markers are completely correlated and have the same frequencies. The correlation will decrease when the frequencies of the markers differ. A more sensitive measure of LD between SNPs is attained by matching frequencies between SNPs (Eberle et al. 2006), or by using of a more efficient multipoint tagging approach (de Bakker et al. 2005). These methods could probably increase the performance of tagSNPs and also elevate the power of identifying the genetic factors behind human traits.

Nevertheless, we have estimated that by using a 500K SNP genotyping chip, we will capture more than 70% of the sequence variation in Sami among sites with a mMAF of 0.1. The combination of new methodologies for multipoint linkage and association analysis will, hopefully, allow for detecting a sufficient amount of the genomic variation to be able to identify the genetic determinants of complex traits in the Sami population.
Paper IV: Identification of local selective sweeps in human populations since the exodus from Africa

Aim
To develop and use a new method to identify genomic regions subjected to local positive selection in human populations.

Background
In the years since the human genome sequence has became available (Lander et al. 2001; Venter et al. 2001), much effort has been devoted to comparison of the human sequence to those of other species. Sequence elements that are conserved between humans and distantly related species have been taken as an indicator of negative (purifying) selection. Since the sequence of our closest relative, the chimpanzee, also become available (Watanabe et al. 2004; Chimpanzee Sequencing and Analysis Consortium 2005), it has also been possible to identify sequence elements with an accelerated rate of evolution, possibly reflecting positive selection. Since the speciation of human and chimpanzee, about 5 million years ago, selective forces might have acted episodically and, depending on the strength and time period of selection, the genomic signatures might have a different appearance. Over several decades, there has been considerable interest in identifying genes under selection in humans. In a recent review, Sabeti and colleagues (Sabeti et al. 2006) identified 44 different approaches, which can be grouped into 5 main categories, for identifying genes under selection in humans (Table 2).

Table 2. Approaches that have been used to identify signals of selection in the human genome.

<table>
<thead>
<tr>
<th>Feature of region under selection</th>
<th>Age of selective signature (years)</th>
<th>Selective signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>High proportion of functional mutations</td>
<td>&gt; 1,000,000</td>
<td>Many function-altering mutations identified by comparing non-synonymous to synonymous substitutions using at least two species.</td>
</tr>
<tr>
<td>Reduction in genetic diversity</td>
<td>&lt; 250,000</td>
<td>Excess of low frequency alleles, decreased heterozygosity or low ratio of polymorphism to divergence from an outgroup.</td>
</tr>
<tr>
<td>High-frequency derived alleles</td>
<td>&lt; 80,000</td>
<td>High frequency of derived relative to ancestral alleles in a population.</td>
</tr>
<tr>
<td>Differences between populations</td>
<td>&lt; 50,000-70,000</td>
<td>Differences in heterozygosity or allele frequencies between populations</td>
</tr>
<tr>
<td>Long haplotypes</td>
<td>&lt;30,000</td>
<td>Long haplotypes and reduced haplotype diversity or long range LD between high frequency alleles in a population.</td>
</tr>
</tbody>
</table>
Sabeti and colleagues (Sabeti et al. 2006) also reviewed genes which have been suggested as candidates for positive selection in different studies. The overlap in genes between different studies was quite small. For example, out of 39 genes previously identified based on function-altering mutation methods, there were only 4 genes in the top 1% candidate list for any of three genome wide scans that were found to be under selection based on a similar method (Bustamante et al. 2005; Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005). Similarly, the overlap between methods searching for genes subjected to more recent selection did not have a large overlap and more refined methods are required in order to distinguish between genes truly affected by selection and false positives.

Results and Discussion

In our manuscript, we describe a new method to identify genes that have undergone selection during the last 100,000 years, corresponding to the migration of humans out of Africa. The method combines different approaches (Table 2):

1. Differences in heterozygosity or allele frequencies between populations ($F_{ST}$)
2. Long haplotypes with reduced haplotype diversity

In order to identify population-specific events, we have evaluated each region in each population in relation to the genome average and to the other populations. The empirical data from Perlegen Sciences (Hinds et al. 2005), including 1.6 million SNPs from three populations (African American, European American and Han Chinese), was used to scan the genome for candidate loci under positive selection.

Only studying $F_{ST}$ revealed a number of interesting features. Average $F_{ST}$ is larger for non-genic compared to genic regions, which is probably due to genetic drift. However, when studying the tail of the distribution, including high $F_{ST}$ outliers, we find a larger fraction of genic compared to non-genic regions. One explanation for this can be that genes are more likely to be affected by positive selection than intergenic regions. There is also a large difference in $F_{ST}$ for different kinds of chromosomes (Figure 7). $F_{ST}$ is lowest for the autosomes (median $F_{ST}$: EA-HC = 0.0787, AA-EA = 0.0882 and AA-HC = 0.1134) followed by the Y-chromosome (median $F_{ST}$: EA-HC = 0.069, AA-EA = 0.138 and AA-HC = 0.150) and highest for the X chromosome (median $F_{ST}$: EA-HC = 0.0927, AA-EA = 0.1527 and AA-HC = 0.1725). When illustrating the difference between $F_{ST}$ of different chromosomes (Figure 7), the branches in the X- and Y-chromosome networks are almost twice as long as for the autosomes. One explanation can be that the smaller effective population size for sex chromosomes makes them more affected by
Figure 7. Genetic distance between three populations representing different continents. A) Average $F_{ST}$ for all blocks defined on autosomal chromosomes, B) Average $F_{ST}$ for all blocks defined on the X-chromosome and C) Average $F_{ST}$ for all blocks defined on the Y-chromosome.

Genetic drift and consequently $F_{ST}$ increases. Previously, the X chromosome has been associated with an increased density of candidate genes for positive selection (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005), mainly explained by a large number of sperm- and testis-associated genes. Consequently, one hypothesis that might explain part of the increased $F_{ST}$ for the X chromosome is that a larger number of genes have been subjected to positive selection.

Another interesting observation is that average $F_{ST}$ for autosomes in the three populations appear almost equidistant to each other, while for both sex chromosomes African Americans are more distantly related. Thus, the data for the sex chromosomes indicate a more recent common ancestry for the two non-African populations. This might reflect stochastic variation due to the lower number of markers on the sex chromosomes, but other factors including population history can have had an effect. A larger effective population size in Africa, compared to the other populations, increases the possible impact of selection relative to genetic drift. This will be more pronounced on a chromosome with a high density of genes under selection, as suggested for the X chromosome.

Our new method to detect loci under selection is based on two variables:

1. LR - a log transformed measure of the length of a haplotype blocks with reduced diversity relative the genome average and relative to the corresponding haplotype block in the two other population
2. FR - a log transformed measure of the differentiation for each haplotype block from the two other populations relative the genome average
By scanning the genome for genes under positive selection, we identified a list of candidate genes (top 1% in the distribution). The list includes a number of genes already proposed as candidates for positive selection such as *LCT, MCPH1* and *AIM1 (MATP)*, but also a large number of genes that have not earlier been suggested to be under positive selection. One category of genes that seems to be affected by positive selection are genes suggested to be involved in pigmentation, including *AIM1 (MATP), BNC2, OCA2, SLC24A5, RAB27A, DCT, EGFR* and *ATRN*. Skin pigmentation is one of the human characters with the highest variation between geographic regions. For most human traits about 85% of the variation is seen within local populations and only about 8% among major geographic regions (Lewontin 1972). In contrast, the variation in skin color is, to large extent (88%), seen between major geographic regions (Relethford 2002). This discrepancy from other characters might be explained by natural selection. There is evidence for a correlation between skin reflectance and the level of UV radiation (Jablonski and Chaplin 2000), and also a number of hypotheses for this adaptation. For example, UV exposure leads to the conversion of 7-dehydrocholesterol in the skin to pre-vitamin D₃ (Jablonski and Chaplin 2000). For this process, the required time of exposure to UV radiation is elevated with darker skin color (Holick et al. 1981). Vitamin D₃ is essential for normal growth, calcium absorption, and skeletal development (Neer 1975). Consequently, individuals exposed to low levels of UV radiation would preferably have pale skin to enable generation of vitamin D₃. In contrast to the advantages of pale skin, dark skin protects against the deleterious effects of UV radiation leading to skin cancer and degradation of folate acid. Folate acid deficiency causes various blood and neural disorders with profound effects on survival and fertility (Jablonski and Chaplin 2000). Presumably due to the advantages of both dark and pale skin pigmentation, a number of genes including MATP, OCA2 and SLC24A5, have been suggested to have been subjected to selection in different populations (Izagirre et al. 2006; Myles et al. 2006).

The overlap between ours and other studies aiming to identify selective events during the same time period is larger than expected by random, even though not complete. We do not observe an overlap between our candidate loci and loci suggested to have been under selection in the lineage leading to modern humans (Bustamante et al. 2005; Nielsen et al. 2005). However, this is not too surprising since selection acting on the common ancestor of all human populations is probably not different between single populations, and therefore not possible to identify using our method. Studying only population genetic data will not identify a selective event where the selective sweep was complete before the origin of modern humans. For example, this might be the case for *FOXP1*. This gene has been identified to have two fixed amino acid changes compared to chimpanzee, but no relevant polymorphism has been found in humans.
Approaches to identify positive selection have to distinguish between the relative contributions of genetic drift and natural selection. Most studies focus on the top 1% of candidates but it is not clear whether those are targets for selection or only extreme outliers of neutral evolution. In our simulations of the length of reduced haplotype diversity, we do observe that a selective sweep with a selection coefficient (s) of 0.01 will result in an average LR of 1.7 (Figure 8). However, when no selection is acting (s=0), about 10% of the simulated data exhibited an LR of 1.7. Consequently, it is not possible to distinguish between blocks that are neutrally evolving outliers and blocks that are under selection. The amount of selection in the human genome is fairly unknown but it has been suggested that as many as 3% of genes have been subjected to recent positive selection (Eberle et al. 2006). If we assume that 3% is the true fraction, it follows that roughly 660 of the about 22,000 genes in the human genome have been subjected to positive selection. Assuming that those 660 genes have a selection coefficient of s=0.01 and we use an LR of 3 as the threshold for being positively selected, 1% of the neutrally evolving genes (1%*97%*22,000=213) and 22% of the genes truly under selection (22%*660=145) will be suggested to be candidates for positive selection (Figure 8). Using more stringent criteria for the LR cutoff will result in a larger fraction of true/false positives, but consequently also a larger number of false negatives. For example, using a LR of 4.5 leaves 0.1% (21) of the neutral evolving genes and 5.9% (45) of the positively selected genes. Assuming a selection coefficient of 0.01 on 3% of the genes in our genome is probably an overestimation and consequently the ratio between false positives to truly selected genes is likely to be even larger.

Figure 8. Distribution of LR for neutrally evolving sites compared to sites with a selection coefficient of s=0.01
Similar to our estimations, the frequency of false positives among the top candidates for positive selection is probably high for most methods developed for scanning the genome for positive selection. At the same time, most of the genes that have undergone positive selection are not even detected. This is the most likely explanation to the incomplete overlap of candidate genes in different studies, where in some cases even the same dataset is used. Many selective events are weak or the signatures have been eradicated by genetic drift. Using a combined approach for searching for genes subjected to recent selection is one approach to reduce the number of false positives and the number of false negatives. Refined methods in combination with larger sample sizes are probably needed in order to construct a complete selection map of the human genome. Using our method, we identified 31 genes within 23 different genomic regions, which were statistically verified as being affected by positive selection. Even though those genes are not likely to be false positives, it certainly represents only a small fraction of all genes that has been subjected to positive selection in the history of human populations.
Conclusions and further perspectives

The focus in this thesis has been on variation in the human genome. I have discussed forces that affect the level of variation, methods to study the variation, and how to link it to the variance of a phenotype. Newly developed techniques have exponentially increased the amount of genotype and sequence data that can be generated. In addition, tools for evaluating the data have been refined. Nevertheless, there is still much unknown about the structure of the human genome and most applications need further improvements before genotype data can be evaluated more accurately.

The procedure for analyzing a large number of genotypes is related to problems such as genotyping errors. The more genotypes produced, the harder to control for errors. Approaches such as the one I have described for automated genotype calling of microsatellite alleles in Paper I, is one way to avoid introducing errors. Furthermore, the genotype determination process could be improved by incorporating the status of relatedness between individuals. Thereby, the probability of the observed genotypes can be more accurately assessed. When scanning the whole genome, another major issue to deal with is the problem of multiple testing.

When a large number of markers are tested for association, the probability of finding a neutral marker that co-segregates with a trait by chance is more likely than identifying a marker which is a minor genetic factor in a complex trait. Traditionally, a lod (logarithm of the odds) score of 3 has been used as the threshold for significance in genome-wide analyses using microsatellites. However, due to the small effect size of each genetic contributor, it has been difficult to reach a lod score of 3 when mapping genes underlying complex phenotypes. The use of hundreds of thousands of SNP markers will not make the multiple testing problem easier. Approaches to characterize the correlation structure of the human genome have efficiently reduced the amount of SNPs needing to be typed in order to capture most of the genome variation (Altshuler et al. 2005; Hinds et al. 2005). However, as we have seen in Paper III, there are many limitations to this approach including loss of information about the sites not yet explored in any population. In addition, novel powerful methods to evaluate LD and haplotype structure may give a more precise picture of the SNP architecture of the human genome that, in the end, will result in a more efficient selection of tagSNPs to cover variation in the human genome.
The large number of SNPs typed in multiple populations (Altshuler et al. 2005; Hinds et al. 2005) has also been used to identify regions with a deviating pattern of diversity, heterozygosity, LD and haplotype structure. Patterns that, in some cases, have been suggested to be signatures of positive selection. Methods to identify selective sweeps, such as the one I describe in Paper IV, have resulted in a number of candidate genes. The overlap between studies is small, reflecting the similarity in patterns between signatures of positive selection and genetic drift. Improved methods that combine different characteristics of selective sweeps, larger sample sizes, and a set of SNPs more representative of the variation in the genome, are probably needed before an complete picture of selection in the human genome can be obtained.

Another complication in the identification of selective sweeps is the complex pattern of genes contributing to a phenotype. Similar to the success in identifying genes causing rare recessive diseases, genes causing a monogenic phenotype, such as LCT and lactase tolerance, are more likely to identify as candidates for positive selection. A complex phenotype selected for in a population will result in a complex signature of selection including many genes or haplotypes at a single gene and probably small effects of the sweep in the region surrounding each gene. For example, in a chicken lineage that has undergone long-term controlled selection for body size, the major genetic architecture of the selective response is a network of four interacting loci (Carlborg et al. 2006). Similar events are likely to explain many of the natural adaptation to new environmental conditions in humans.

Identifying signatures of selective sweeps has long been an issue of curiosity, but it is also relevant for understanding disease in man. When searching for a correlation between genes and disease causing phenotypes, genes subjected to positive selection might be a good starting point. Many genes that have been under selection as a response to pathogens or other causes of illness might be the same genes that we are searching for in relation to human disease. A selective sweep increases the frequency of an allele and consequently, hitchhiking by nearby loci may result in the increase in frequency of alleles causing rare recessive diseases.

Even though we are far from resolving the complex pattern of genomic and phenotypic variation in humans, the future is promising. The amount of empirical data increases exponentially as the methods for genotyping and resequencing are being refined. In addition, the methods dealing with genotype/sequence data are becoming more advanced to handle genome-wide studies based on hundreds of thousands of SNPs. Also, the use of populations such as Sami, as I describe in Paper II, will be important in the mapping of complex traits. In populations like Sami, factors such as low diversity, high LD and reduced environmental variance can reduce the complexity of many multifactorial traits. The availability of population-based genealogical information in Sweden through digitized parish records makes it possible
to construct population pedigrees going as far back as about 1700 AD. This allows for more powerful studies of effective population size, migration and marriage patterns, but also for the mapping of genes behind human diseases.

The large amounts of SNP data produced represent an enormous asset for human genetic studies. We have already seen a number of publications based on those data and additional studies will be published in the next years. The improved methods for SNP genotyping as well as for genomic resequencing will further increase the amount of data available. New methods, such as those developed by Solexa (http://www.solexa.com/wt/page/index), are expected to enable resequencing of up to a billion bases in one run and to permit resequencing of an entire human genome at a cost of $100,000. Given the speed by which new genome technology develops, it is likely that within a few years resequencing complete human genomes rather than analysis of selections of SNPs will become the standard in individual studies, indicating a new era in human genomics.
Acknowledgements

The work presented in this thesis was performed at the Department of Genetic and Pathology, Faculty of Medicine, Uppsala University. It was supported by grants from The Foundation for Strategic Research (Genome Research Program), The Knut and Alice Wallenberg Foundation, The Linnaeus Centre for Bioinformatics, Mål 1 Sápmi, and The Swedish Medical Science Research Council.

I am grateful to a number of persons:

Ulf Gyllensten, my supervisor, for always listening and being open to new ideas. Your support and enthusiasm has meant a lot during the last years.

Anna, Geir, Malin, and Karolina for many valuable discussions through the years. Per-Ivan for lots of good ideas of how to deal with non-cooperating computers. Inger, Jenny and former members of the SSF Genotyping Center for assistance in the lab and for helping out with any problems related to genotyping. Max for kindly reviewing the text in this thesis. All present and ex-colleagues at the Rudbeck Laboratory, who have made it a great place for working. The members of the Sami communities for participating in our studies.

And finally, my family and friends just for being there.
Allen JA (1877) The influence of physical conditions in the genesis of species. Radical Review 1: 108-140
Beckman L (1996) Samerna en genetiskt unik urbefolkning. Solfjädern Offset AB, Umeå
Bergmann C (1847) Über die verhältnisse der warmeökonomie der thiere zu ihrer grosse. Gottingen Studien 1: 595-708
Chimpanzee Sequencing and Analysis Consortium T (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 201

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7293

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006