HUMAN GENETIC VARIATION IN THE BALTIC SEA REGION: FEATURES OF POPULATION HISTORY AND NATURAL SELECTION

Tuuli Lappalainen

Institute for Molecular Medicine Finland
University of Helsinki, Finland

and

Department of Biological and Environmental Sciences
Faculty of Biosciences

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biosciences of the University of Helsinki, in Auditorium XII, Main Building, Fabianinkatu 33, Helsinki, on May 15th 2009 at 12 noon

Helsinki 2009

SUPERVISORS **Päivi Lahermo**
Institute for Molecular Medicine Finland
University of Helsinki, Finland

Juha Kere
Institute for Biosciences and Nutrition
Karolinska Institutet, Stockholm, Sweden, and
Department of Medical Genetics
University of Helsinki, Finland

Kirsi Huoponen
Department of Medical Genetics
University of Turku, Finland

REVIEWERS **Antti Sajantila**
Department of Forensic Medicine
University of Helsinki, Finland

Kari Majamaa
Department of Neurology
University of Oulu, Finland

Jaakko Ignatius
Department of Clinical Genetics
University of Oulu, Finland

OPPONENT **Antti Sajantila**
Department of Forensic Medicine
University of Helsinki, Finland

CUSTOS **Minna Nyström**
Division of Genetics
Department of Biological and Environmental Sciences
University of Helsinki, Finland

ISBN 978-952-92-5418-7 (paperback)
ISBN 978-952-10-5468-6 (pdf)
http://ethesis.helsinki.fi
Helsinki University Print
Helsinki 2009

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to in the text by their Roman numerals. Additionally, some unpublished data are presented.

I **Lappalainen T***, Koivumäki S*, Salmela E, Huoponen K, Sistonen P, Savontaus M-L, Lahermo P (2006) Regional differences among the Finns: A Y-chromosomal perspective. *Gene* 376:207-215.

II **Lappalainen T**, Laitinen V, Salmela E, Andersen P, Huoponen K, Savontaus M-L, Lahermo P (2008) Migration waves to the Baltic Sea region. *Annals of Human Genetics* 72:337–348.

III **Lappalainen T**, Hannelius U, Salmela E, von Döbeln U, Lindgren CM, Huoponen K, Savontaus M-L, Kere J, Lahermo P (2009) Population structure in contemporary Sweden – A Y-chromosomal and mitochondrial DNA analysis. *Annals of Human Genetics* 73:61-73.

IV Salmela E*, **Lappalainen T***, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, Sistonen P, Savontaus M-L, Schreiber S, Kere J, Lahermo P (2008) Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* 3:e3519.

V **Lappalainen T**, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, Savontaus M-L, Schreiber S, Lahermo P, Kere J. Genomic landscape of positive natural selection in North European populations. *Submitted.*

*equal contribution

AUTHOR CONTRIBUTIONS

	I	II	III	IV	V
Study design	TL, SK, KH, MLS, PL	TL, VL, KH, MLS, PL	TL, UH, ES, CML, JK, PL	ES, TL, JK, PL	TL, ES, JK, PL
DNA samples and datasets	PS, MLS	PMA, PS, MLS	UH, UvD, JK	PMA, KDW, AF, PS, MLS, SS	PMA, KDW, PS, MLS, SS
Laboratory analysis	TL, SK	TL, VL	TL, UH	ES, TL, IF	TL, ES
Statistical analysis	TL, SK, ES, PL	TL, ES	TL, ES	ES, TL	TL, ES
Drafting the manuscript	TL, SK, PL	TL	TL	ES, TL	TL

The author initials are listed in the order of appearance in the manuscript. All authors have taken part in revising the manuscript draft. Abbreviations:

TL Tuuli Lappalainen
SK Satu Koivumäki
ES Elina Salmela
KH Kirsi Huoponen
PS Pertti Sistonen
MLS Marja-Liisa Savontaus
PL Päivi Lahermo
VL Virpi Laitinen
PMA Peter M. Andersen
UH Ulf Hannelius
UvD Ulrika von Döbeln
CML Cecilia M. Lindgren
IF Ingegerd Fransson
KDW Karin Dahlman-Wright
AF Andreas Fiebig
SS Stefan Schreiber

ABBREVIATIONS

AD	*Anno Domini*
AMOVA	analysis of molecular variance
BC	before Christ
BP	before present
CEPH	Centre d'Etude du Polymorphisme Humain
CNV	copy number variation
ddNTP	dideoxyribonucleotide triphosphate
DNA	deoxyribonucleic acid
EDAR	the ectodysplasin A receptor gene
EHH	extended haplotype homozygosity
FY	the Duffy blood group, chemokine receptor gene
Gb	gigabase
G6PD	the glucose-6-phosphate dehydrogenase gene
HG	haplogroup
HVS	hypervariable segment
IBS	identity by state
iHS	integrated haplotype score
indel	insertion/deletion
kb	kilobase
LCT	the lactase gene
LD	linkage disequilibrium
LRH	long-range haplotype
MALDI-TOF	matrix-assisted laser desorption/ionization time-of-flight
Mb	megabase
MDS	multidimensional scaling
mtDNA	mitochondrial DNA
PC(A)	principal component (analysis)
PCR	polymerase chain reaction
PPP2R2B	protein phosphatase 2, regulatory subunit B, beta isoform gene
RAB38	the RAB38, member RAS oncogene family gene
RFLP	restriction fragment length polymorphism
SLC45A2	the solute carrier family 45, member 2 gene
SNP	single nucleotide polymorphism
STR	short tandem repeat
TMRCA	the most recent common ancestor
UEP	unique evolutionary polymorphism
250K	250 000
500K	500 000

ABSTRACT

In this thesis, the genetic variation of human populations from the Baltic Sea region was studied in order to elucidate population history as well as evolutionary adaptation in this region. The study provided novel understanding of how the complex population level processes of migration, genetic drift, and natural selection have shaped genetic variation in North European populations.

Results from genome-wide, mitochondrial DNA and Y-chromosomal analyses suggested that the genetic background of the populations of the Baltic Sea region lies predominantly in Continental Europe, which is consistent with earlier studies and archaeological evidence. The late settlement of Fennoscandia after the Ice Age and the subsequent small population size have led to pronounced genetic drift, especially in Finland and Karelia but also in Sweden, evident especially in genome-wide and Y-chromosomal analyses. Consequently, these populations show striking genetic differentiation, as opposed to much more homogeneous pattern of variation in Central European populations. Additionally, the eastern side of the Baltic Sea was observed to have experienced eastern influence in the genome-wide data as well as in mitochondrial DNA and Y-chromosomal variation – consistent with linguistic connections. However, Slavic influence in the Baltic Sea populations appears minor on genetic level.

While the genetic diversity of the Finnish population overall was low, genome-wide and Y-chromosomal results showed pronounced regional differences. The genetic distance between Western and Eastern Finland was larger than for many geographically distant population pairs, and provinces also showed genetic differences. This is probably mainly due to the late settlement of Eastern Finland and local isolation, although differences in ancestral migration waves may contribute to this, too. In contrast, mitochondrial DNA and Y-chromosomal analyses of the contemporary Swedish population revealed a much less pronounced population structure and a fusion of the traces of ancient admixture, genetic drift, and recent immigration.

Genome-wide datasets also provide a resource for studying the adaptive evolution of human populations. This study revealed tens of loci with strong signs of recent positive selection in Northern Europe. These results provide interesting targets for future research on evolutionary adaptation, and may be important for understanding the background of disease-causing variants in human populations.

INTRODUCTION

1. Human population genetics

1.1 Background and scope

Population genetics aims at characterizing patterns and evolutionary changes of genetic variation in populations. Human population genetics examines these processes in *Homo sapiens*, aiming at understanding the history and current genetic diversity of our species. Knowledge of the genetic variation across the human genome is elementary for investigation of the processes that lie behind phenotypic variation, including disease. Many important research foci of medical genetics have stemmed from have population genetic processes – e.g. the distribution of linkage disequilibrium, the mutation process, and the evolution of both rare and common diseases. Additionally, variation of the genome provides a powerful tool for the study of human history. (Jorde *et al.* 1998, Cann 2001, Jorde *et al.* 2001, Cavalli-Sforza & Feldman 2003, Tishkoff & Verrelli 2003, Jobling *et al.* 2004, Cavalli-Sforza 2005, Garrigan & Hammer 2006)

The early population genetic analyses were based on blood group markers (e.g. Cavalli-Sforza *et al.* 1994). Mitochondrial genetics showed its strength in population genetic analysis in the late 1980s, and in the 1990s Y-chromosomal analysis emerged alongside it (Stoneking 1997, Cavalli-Sforza 1998). The analysis of these haploid markers focused mostly on population history, whereas studies of autosomal variation have also been motivated by understanding the patterns of genetic variation underlying human diseases (Cann 2001, Jorde *et al.* 2001, Jobling *et al.* 2004). In the 21st century, the analysis of genetic variation across the entire genome has rapidly become the mainstream of population genetic analysis.

1.2 Population genetic processes

Population genetics is based on the modern synthesis of evolutionary theory that formulated the theoretical basis of microevolution, i.e. the change of allele frequencies or their combinations in the course of generations. Several different processes may lie behind such a change: 1) mutation, 2) recombination, 3) genetic drift, 4) migration, 5) nonrandom mating, and 6) natural selection. Of these, mutation and recombination occur at the molecular level within cells, whereas the other processes take place in populations. In natural populations – including humans – all of these usually contribute to changes in allele and genotype frequencies and haplotype patterns. These processes

are briefly described below and summarized in Table 1. (Jobling *et al.* 2004, Hartl & Clark 2007, Nei 1987)

1.2.1 Mutation

Mutation is the source of all genetic variation, and is therefore essential for evolution. In addition to the mutational event itself, the term mutation is also used for rare genetic variants that occur with a frequency of under 1%, whereas more common variants are termed polymorphisms. There are several different types of mutations that create different classes of genetic polymorphism (see Section 2.2). The mutation rate depends on the type of the locus, but usually it is low enough to have little effect on allele frequencies.

1.2.2 Recombination

A new mutation always takes place in an existing chromosomal strand with a previous pattern of variation in adjacent loci, and the new variant remains associated to the surrounding variants – the haplotype – until this association is broken by recombination, which refers to the exchange of homologous strands of parental chromosomes in meiosis. However, recombination is rare, and progressively rarer with shorter physical distances, which leads to non-random association between nearby polymorphisms, called linkage disequilibrium (LD). Importantly, the recombination rate is not uniform across the human genome: it has been estimated that 88% of all recombination occur in 'hotspots', delimiting large haplotype blocks with little historical recombination (Reich *et al.* 2002, Schaffner *et al.* 2005, Slatkin 2008).

1.2.3 Genetic drift

There is always random variation in the reproductive success of individuals that causes the transmission of genes to the next generation of a population to be affected by coincidence. Thus, finite population size introduces random fluctuation of allele frequencies between generations, called genetic drift. It is stronger in small populations and leads to loss of genetic diversity: eventually all alleles drift to fixation, and the variation at that locus is lost and cannot be recovered without a new mutation or migration (Figure 1). Drift leads to the accumulation of genetic differences between populations with time, and is the main process behind human population differentiation.

Some population events are associated with particularly strong genetic drift. These include population bottlenecks, when the population size is temporarily reduced, and founder events, when a new population is founded by a small subset of the ancestral population. Allelic surfing occurs when alleles are randomly enriched in the advancing front of a spatially expanding population (Klopfstein *et al.* 2006). In all of these cases,

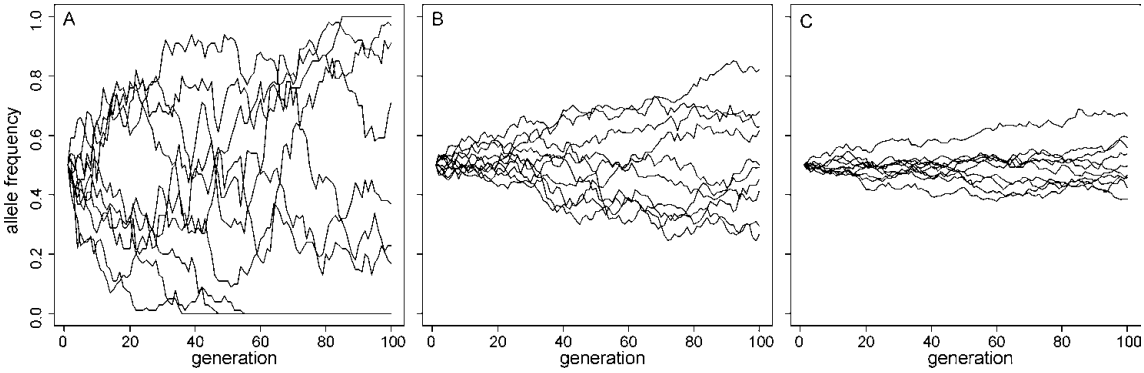

Figure 1. Genetic drift in a population of a constant size of a) 50, b) 500 and c) 2000 diploid individuals. Calculated with an allele frequency simulator described in V, unpublished.

randomly determined allele frequencies of a small population give rise to descending population frequencies, often leading to extreme genetic drift.

1.2.4 Migration

Novel populations are founded as people settle uninhabited regions, and the populations differentiate with time through the process of drift. Alone, such a process would create a hierarchical genealogy of populations that could be represented as a tree. However, populations are rarely isolated from each other, and gene flow via migration evens out allele frequency differences between populations. The relative importance of migration and drift is often difficult to determine: two population pairs may show different genetic distances despite the same time of split from an ancestral population if the extent of migration is different.

There are several population genetic models for migration. In human populations, recent analyses have suggested that the dominant pattern of migration may be isolation by distance (Novembre *et al.* 2008), a pattern in which migration gradually decreases with increasing geographical distance.

1.2.5 Nonrandom Mating

Inbreeding – non-random fusion of gametes – alone does not change allele frequencies but genotype frequencies, i.e. the combination of alleles of the same locus. In positive inbreeding, mating between similar individuals occurs more frequently than chance would suggest, and serves to increase the frequency of homozygotes, and vice versa in negative inbreeding. Mating can be selective relative to certain genes, or across the entire genome (Chaix *et al.* 2008).

The concept of non-random fusion of gametes can be extended to population units larger than the individual. In a population with a substructure, mating is more likely to occur within the subpopulations, and thus the heterozygosity relative to the entire population is lower than expected under panmixia, as first described by Sten Wahlund in 1928.

1.2.6 Natural selection

Natural selection – the different reproductive fitness of carriers of different alleles – is the force behind all evolutionary adaptation. Negative selection removes harmful variants, while positive selection increases the frequency of beneficial alleles. Balancing selection favours heterozygotes, thus maintaining variation that would otherwise be lost via drift.

The importance of selection in shaping the genetic variation of a species is one of the most classic debates of evolutionary genetics (see e.g. Nei 2005 for a review). According to the neutral theory, selection has a role mostly in removing deleterious mutations, while the selectionist theory states that positive selection is an important force in shaping genetic variation, and this has been supported by numerous examples. However, the proportion of the genome affected by positive selection remains unknown (Nielsen *et al.* 2007).

Table 1: Consequences of different population genetic processes

	Differences between populations	**Variation within a population**	**Affects**	**Strongest in**	**Importance in shaping variation of populations**
Mutation	Increases	Increases	Creates variation and sometimes changes allele frequencies across the genome	Large populations	Low
Recombination	Increases	Increases	Allelic combinations in haplotypes across the diploid genome	Large populations	Low
Genetic drift	Increases	Decreases	Allele frequencies across the genome	Small populations	Very high
Migration	Decreases	Increases	Allele frequencies across the genome	Depends on the population	Very high
Inbreeding	Increases	Decreases	Genotype frequencies of loci across the genome or at specific sites	Usually small populations	Varies
Natural selection	Increases or decreases	Decreases or maintains	Allele frequencies of specific loci	Large populations	Not known

1.3 The multidisciplinary study of human history

The scope of human population genetics touches the most ancient of questions: who are we, and where do we come from? This field of science is by no means the first to seek answers to these questions; in particular, archaeology, linguistics, and anthropology have a long tradition in the study of ancient human history. All these fields remain important today, with each of them having their characteristic scope, methods, source material and time scale (Jobling *et al.* 2004).

Archaeology relies on the material remains of human activity, and studies the past cultures, societies, and subsistence. It is able to reach back over one million years to the earliest preserved hominid artefacts. Linguistics traces the history of languages that is often related to the history of both cultures and biological populations. It has the narrowest temporal scope of up to only about 8000 years due to the rapid change of language (McMahon 2004). Physical anthropology studies the biological characteristics of humans and often particularly focuses on human adaptation to different environments, while paleoanthropology analyses the fossil record of the human lineage, thus characterizing the origin of our species (Wood 2000, Steegmann 2006). Finally, human population genetics, sometimes also called molecular anthropology, infers human history mostly from data of contemporary population genetic variation. It can be used for studying processes from the very recent to the ancient through an appropriate selection of genetic markers. Evolutionary genetics has no limit in temporal scope except for the age of life on Earth, but population genetics by definition studies intraspecies variation, which in the case of modern humans implies a time scale ranging from contemporary events to a few hundred thousand years back in time (e.g. Cann 2001, Cavalli-Sforza & Feldman 2003, Jobling *et al.* 2004, Garrigan & Hammer 2006). A further genetic approach makes use of ancient DNA extracted from prehistoric human remains (Jobling *et al.* 2004, Paabo *et al.* 2004).

The different disciplines studying human history are interrelated – for example a population migration may leave traces in the genome as well as in the anthropometric characteristics of populations, cultural remains, and the language of the descendants. Historical interpretation of population genetic observations is strongly dependent on archaeological and linguistic information. Thus, many prominent researchers have called for better integration of the different disciplines (Cavalli-Sforza *et al.* 1994, Cann 2001, Cavalli-Sforza & Feldman 2003, Diamond & Bellwood 2003) to form a field sometimes called archaeogenetics (Renfrew 2001). However, the underlying mechanisms behind the dispersal of culture, language, physical characteristics and genes are different, and providing factual evidence of a common historical event behind similar patterns observed by different disciplines has proven to be difficult. (Cavalli-Sforza *et al.* 1994, Cann 2001, Diamond & Bellwood 2003, McMahon 2004)

A particularly controversial feature of human diversity is ethnicity and its relationship to genetics. It is a complex and fluctuating concept that is formed via

politics, history, familial background and personal experiences, and its use in scientific contexts is controversial (Juengst 1998, Race, Ethnicity, and Genetics Working Group 2005, Lee *et al.* 2008). However, by analyzing a sufficient number of genetic polymorphisms, human populations defined by political, cultural and/or linguistic grounds can often be distinguished from each other even within a continent (e.g. Novembre *et al.* 2008), suggesting that such ethnic definitions may have some validity also in a biological sense. Being a difficult concept even in modern societies, the question of ethnicity of populations or cultures of the past is impossible to answer – there are no methods to connect historical cultures, assumed languages and observed genetic features to ethnicities, or to define ethnic units of the past (McMahon 2004), because ethnicity is inherently dependent on the subjective experiences of individuals and is imperfectly reflected in their material culture, language or genes.

2. From genotypes to history – population genetic analysis

2.1 The structure of the human genome

The three billion base pairs of the nuclear human genome are divided into 22 pairs of autosomal chromosomes, the X chromosome, of which females have two and males one copy, and the Y chromosome, present as a single copy only in males. Additionally, mitochondria have their own small circular DNA molecule, mitochondrial DNA (mtDNA). Of each autosomal chromosome pair, one is inherited from the mother and one from the father, and the homologues recombine in every meiosis. The X chromosome recombines only in females, except for the small pseudoautosomal regions close to the telomeres of the X and Y chromosomes that recombine in the male meiosis. The Y chromosome, except for the pseudoautosomal regions, and the mitochondrial DNA generally do not recombine – although rare cases of recombination or paternal inheritance in mtDNA have been reported (see e.g. Pakendorf & Stoneking 2005 for a review). In this thesis, 'Y chromosome' is used to refer to the non-recombining element, if otherwise not specified. (Table 2)

2.2 Types of genetic polymorphism

The spectrum of DNA sequence variation ranges from single base pair variants to changes in the copy number of entire chromosomes, and a full understanding of this spectrum as well as of the evolution, organization and function of different types of

Table 2. Characteristics of the autosomes, X chromosome, Y chromosome and mtDNA

	Autosomes	**X chromosome***	**Y chromosome***	**mtDNA**
Location	Nucleus	Nucleus	Nucleus	Mitochondria
Inheritance	♀ & ♂ Biparental	♀ Biparental ♂ Maternal	♀ Not applicable ♂ Paternal	♀ & ♂ Maternal
Recombination	Every meiosis	Every meiosis in females	Never	Never
Copy number per cell	♀ & ♂ 2 × 22	♀ 2 ♂ 1	♀ 0 ♂ 1	♀ & ♂ from hundreds to thousands
Effective population size	1 (reference)	3/4	1/4	1/4
Types of polymorphisms	All	All	All	SNPs, small insertions/deletions
Total length (NCBI Build 36.1)	2.87 Gb	149 Mb	57.8 Mb	16.6 kb

** Pseudoautosomal regions behave like autosomes*

polymorphism is yet to be achieved. Different types of DNA polymorphisms are also behind variation of serum proteins, the analysis of which was the first tool to study human genetic diversity (Cavalli-Sforza *et al.* 1994). The most important and commonly analyzed types of DNA polymorphism are reviewed below.

The smallest units of variation are single nucleotide polymorphisms (SNPs), created by point mutations that affect a single base of the genome. They are numerically the most common type of variation: there are 6.6 million validated SNPs in the genome (dbSNP build 129), and the total number of common SNPs (minor allele frequency ≥ 0.05) is estimated to be 9-10 million (International HapMap Consortium *et al.* 2007). The rate of mutation from one base to another, approximately 2.5×10^{-8} per base per generation (Matise *et al.* 2007), is so low that the vast majority of SNPs are a result of a unique mutational event in the past. Most SNPs are non-functional, but many affect protein structure or gene expression, or have another functional impact (Hinds *et al.* 2005, International HapMap Consortium 2005, International HapMap Consortium *et al.* 2007, Stranger *et al.* 2007). At present, due to their abundance and ease of high-throughput genotyping, SNPs are the most commonly used genetic markers for gene mapping and for analyses of population genetic variation.

The previous standard markers for an analysis of genetic variation were microsatellites, or short tandem repeats (STRs): variations in the number of a few base

pair repeats. The mutation rate of these loci is much higher than that of SNPs, about 1.5 × 10^{-3}, creating frequent recurrent and backmutations (Butler 2006). As a result of the high mutation rate, microsatellites are highly polymorphic and thus informative as markers, but reliable high-throughput genotyping is technically more challenging, and their coverage of the genome is uneven (NIH/CEPH Collaborative Mapping Group 1992). They are still used in genetic analyses especially in forensics (Butler 2006) and also in population genetics.

Structural variation refers to larger changes in the genome, and includes balanced variations, where a fragment of a chromosome has become inverted or translocated into another place, and copy number variations (CNVs), where the number of a particular genomic segment differs between individuals. Usually, only variations of over 1 kb have been included in this category, although the threshold is arbitrary (Hurles *et al.* 2008). Recently, large-scale genotyping of structural variation in the genome has become possible, leading to increasing understanding of its importance for genome organization and function. Genotyping and analysis of CNVs remains challenging, which makes them impractical as genetic markers in population genetic or gene mapping studies, but they have been suggested to be a major source of phenotypic variation in humans (Hurles *et al.* 2008, McCarroll *et al.* 2008).

2.3 Human genetic variation

2.3.1 Autosomal and X-chromosomal variation

Much of the knowledge of the patterns of SNP variation in humans stems from the HapMap project that has catalogued the variation of millions of SNPs in four populations (International HapMap Consortium 2005, International HapMap Consortium *et al.* 2007), and a similar analysis by Perlegen Sciences (Hinds *et al.* 2005). In addition to these international efforts, other large datasets have become available via the development of technology for high-throughput genotyping of hundreds of thousands of SNPs across the entire genome. The majority of the genome-wide datasets originate from genetic association studies that search for common genetic variants predisposing to complex disease (see, for example, Balding 2006, Wellcome Trust Case Control Consortium 2007, Bodmer & Bonilla 2008). Recently, the development of sequencing technology has allowed large-scale resequencing of entire genomes (Mardis 2008, Shendure & Ji 2008), which will add enormously to our knowledge of the variation in the human genome. In particular, the importance of rare variants is now becoming acknowledged, after the early focus on common variation (Bodmer & Bonilla 2008).

The HapMap data have provided detailed information of the pattern of linkage disequilibrium (LD) in human populations, and uncovered the redundancy of much of

the common variation in the genome: over 80% of the over 3 million common SNPs analyzed in HapMap II are well correlated with other SNPs, and thus genotyping only a subset of these variants, so-called tagging SNPs, will provide information on most of the genome (International HapMap Consortium *et al.* 2007). The haplotype block boundaries have proven to be relatively uniform across the populations due to shared history as well as common recombination hotspots (International HapMap Consortium 2005, Gonzalez-Neira *et al.* 2006, International HapMap Consortium *et al.* 2007, Jakobsson *et al.* 2008), although the extent of LD varies between populations (Jakobsson *et al.* 2008). In addition to linkage between SNPs, copy number polymorphisms are also often linked to SNPs (McCarroll *et al.* 2008).

Population-based association studies have led to increased interest in population genetics because unknown population structure has been shown to be an important confounding factor in association studies (Freedman *et al.* 2004, Marchini *et al.* 2004): if the case and control populations differ in their ancestry, the association analysis may discover loci with frequency differences between populations rather than those associating to disease. However, various methods to correct for population structure have been developed (see Tian *et al.* 2008a for a review).

2.3.2 Mitochondrial DNA and Y-chromosomal variation

The basic structure and types of variation in the non-recombining proportion of the Y chromosome resemble those of the other chromosomes, but its paternal inheritance and lack of recombination have led to an enrichment of tandem repeats and genes with male-specific functions (Jobling & Tyler-Smith 2003). In contrast, mitochondrial DNA differs from the nuclear genome in many respects. Mitochondria probably descend from an aerobic bacterium that became an organelle of the eukaryotic cell though endocytosis, and thus also its genome shares many properties of prokaryotic DNA. The circular 16 569 base pairs of human mtDNA contain 37 densely packed intronless genes and a short regulatory region, the D-loop. The mitochondrial genes are necessary in oxidative phosphorylation, the main function of the mitochondria, as well as in DNA replication and protein synthesis. There are no major repetitive elements, insertions or deletions. The mutation rate of mtDNA is on average several orders of magnitude higher than that of the nuclear genome, although there is large variation between different parts of mtDNA. (Pakendorf & Stoneking 2005, Wallace 2005, Torroni *et al.* 2006)

The evolutionary history of mitochondrial DNA and the Y chromosome differ from autosomes and the X chromosome in many respects. The lack of recombination results in inheritance of these marker systems as two haplotype blocks that are altered only via mutation. The Y chromosome and mtDNA are also unique in their uniparental inheritance, thus forming historical paternal and maternal lineages. The effective

population size of mtDNA and the Y chromosome is ¼ compared to the autosomes, since only one copy of these molecules is passed on to the next generation per four copies of each autosomal chromosome. Thus, genetic drift is stronger and differences between populations higher than for autosomal markers. (Jobling & Tyler-Smith 2003, Tishkoff & Verrelli 2003, Garrigan & Hammer 2006, Underhill & Kivisild 2007)

Most of the known SNPs and structural variations of the Y chromosome and the coding region of mtDNA are unique evolutionary polymorphisms (UEPs): results of a unique mutational event in the human history. The phylogeny of these markers is a perfect tree whose hierarchical structure corresponds to the historical accumulation of mutations. The ease of reconstructing the phylogeny is the main advantage of mtDNA and Y-chromosomal analysis when compared to the complex networks of recombining markers. The hierarchical trees have standardized nomenclature systems of haplogroups that are haplotype groups carrying specific motifs of UEPs (Figure 2, Figure 3). Haplogroups can be grouped into macrohaplogroups and divided into subhaplogroups (Macaulay *et al.* 1999, Torroni *et al.* 2006, Underhill & Kivisild 2007, Karafet *et al.* 2008). The Y-chromosomal classification and nomenclature system is being systematically maintained and updated, and thus the names of the haplogroups corresponding to particular polymorphisms have changed several times. In this study, the old nomenclature from the year 2002 is used, and the conversion of the names used in this study to the most recent phylogeny is given in Table 3 (Y Chromosome Consortium 2002, Karafet *et al.* 2008).

Each haplogroup is a result of a mutation that has been inherited by all the descendants of a single individual in a paternal or maternal lineage. Thus, each haplogroup has its characteristic frequency pattern across the world that is indicative of the historical distribution of the carriers of the polymorphism (Figure 2, Figure 3). In addition to the perfect tree of haplogroups, Y-chromosomal microsatellites and SNPs in the D-loop of mtDNA (in addition to some other polymorphisms) have a very high mutation rate, resulting in frequent recurrent mutations during human history. These polymorphisms are efficient for analyzing local or regional population structure within a shorter time span, and also for analysis of patterns of variation within haplogroups: the time and place of a unique mutation can be determined by analyzing haplotype variation within the haplogroup, since a longer time span implies more time for subsequent mutations to accumulate.

The patterns of mtDNA and Y-chromosomal variation show interesting differences (see Underhill & Kivisild 2007 for a review). In general, mtDNA variation is more evenly distributed across ethnic and linguistic barriers, whereas Y-chromosomal variation is more localized, and corresponds better to linguistic variation. Some of the differences between mtDNA and the Y chromosome have been explained by differences in male and female population histories. One such difference arises by the common practice of patrilocality, in which females tend to move close to their husband's home, resulting in a higher migration rate of females. Furthermore, male reproductive success

varies more than that of females, which in practice results in a smaller effective population size for the Y chromosome compared to mtDNA, although theoretically the effective population sizes are the same. (Oota *et al.* 2001, Cavalli-Sforza & Feldman 2003, McMahon 2004, Underhill & Kivisild 2007, Hammer *et al.* 2008)

The advantage of the study of haploid markers lies in the possibility of estimating the temporal scale of events and distinguishing different layers of migratory waves with a relatively high degree of precision. However, despite the many applications and ease of mitochondrial DNA and Y-chromosomal analysis, they represent only two loci in the human genome. The evolution of each individual locus is always affected by stochastic events, and possibly also natural selection, although the importance of such selection in shaping mtDNA and Y-chromosomal variation is still debated (Jobling & Tyler-Smith 2003, Kivisild *et al.* 2006, Meiklejohn *et al.* 2007). Consequently, the story of human history told by mtDNA and the Y chromosome may not be devoid of bias, and relying on them alone is risky (Jobling & Tyler-Smith 2003, Garrigan & Hammer 2006, Underhill & Kivisild 2007).

Table 3. Conversion of the Y-chromosomal haplogroup (HG) nomenclature between those used in this study (HG 2002: Y Chromosome Consortium 2002) and the most recent phylogeny (HG 2008: Karafet et al. 2008).

polymorphism	HG 2002	HG 2008
-	Y*	B*
SRY-1532	A	A
M216	C	C
YAP, M203	DE	DE
P14	F*	F*
M201	G	G
M170	I	I
M253	I1a	I1
P37	I1b	I2a
M223	I1c	I2b
12f2	J	J

polymorphism	HG 2002	HG 2008
M9	K*	K*
LLY22g	N	N1
N43	N2	N1b
Tat	N3	N1c
M175	O	O
92R7, M45	P*	P*
P36	Q	Q1
M207	R	R
SRY-1532	R1a	R1a
M17	R1a1	R1a1
P25	R1b	R1b1

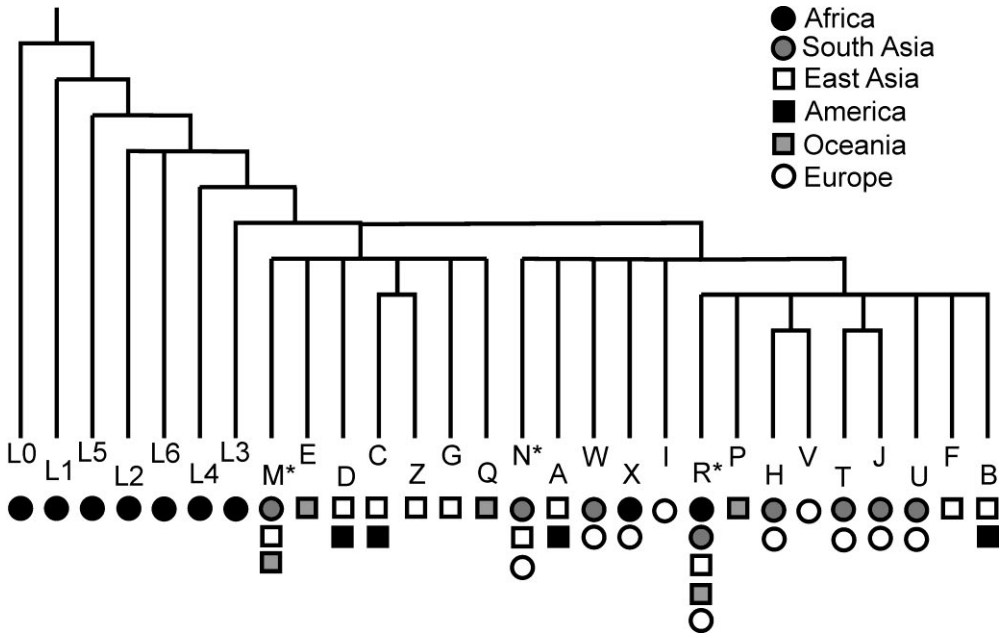

Figure 2. Mitochondrial DNA haplogroup tree – the main haplogroups and their continental distributions. (Underhill & Kivisild 2007)

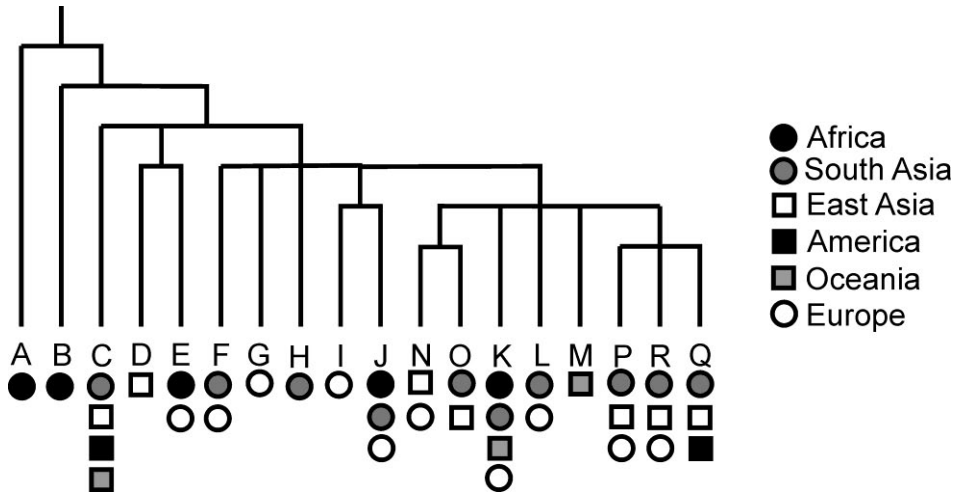

Figure 3. Y-chromosomal haplogroup tree – the main haplogroups and their continental distributions. (Underhill & Kivisild 2007)

2.3.3 Patterns of human genetic variation

Autosomal, X-chromosomal, mitochondrial DNA and Y-chromosomal markers, as well as blood group polymorphisms, have been used for analysing the patterns of population genetic variation. It has been shown that the genetic diversity of humans is lower than among many other species (e.g. Jorde *et al.* 2001 and references therein). This is likely caused by the relatively recent origin of our species less than 200 000 years ago in Africa (Cann *et al.* 1987, Cavalli-Sforza & Feldman 2003, Tishkoff & Verrelli 2003, Garrigan & Hammer 2006, Relethford 2008). The consensus is that modern humans colonized the other continents via migrations out of Africa, and replaced the ancestral human populations such as the Neanderthals, but a small degree of admixture has not been ruled out (Garrigan & Hammer 2006, Green *et al.* 2006, Relethford 2008). The decreasing diversity of human populations with increasing distance from Africa supports serial bottlenecks during the dispersal out of Africa. Furthermore, the recent origin is consistent with the small proportion of genetic difference between human populations: it has been estimated that slightly less than 90% of human genetic variation is between individuals, only a few percent between populations within continents, and less than ten percent of the variation is explained by continental grouping of individuals. Much of the variation between populations appears to follow geographic clines, lacking strong genetic clustering on linguistic or ethnical grounds but exhibiting small genetic borderlines following geographical barriers. (e.g. Barbujani *et al.* 1997, Rosenberg *et al.* 2002, Rosenberg *et al.* 2005, Conrad *et al.* 2006, Jakobsson *et al.* 2008, Li *et al.* 2008, Novembre *et al.* 2008).

2.4. Analysis of positive natural selection

Positive natural selection is the force behind evolutionary adaptation, and is of major interest for elucidating the background of phenotypic variation between human populations. However, not all phenotypic variation need be adaptive: genetic drift can also affect phenotypic traits (Roseman & Weaver 2007, Betti *et al.* 2009). Positive natural selection leads to an increase in the frequency of the beneficial variant and the haplotype surrounding it, eventually leading to fixation, a process often referred to as "selective sweep". Selection may commence for example when a new variant enters a population through mutation or migration from another population, or when an environmental change makes an existing neutral polymorphism advantageous.

2.4.1 Signatures of positive selection

The process of positive selection leaves a characteristic trace in the variation of the affected genomic region, and there are several statistical tests for detecting these signatures, most focusing on one or two characteristic signs of selective sweeps. Many

classical tests are based on comparisons to other species (see e.g. Nielsen 2005, Sabeti *et al.* 2006, Anisimova & Liberles 2007, Nielsen *et al.* 2007 for reviews); the most important tests focusing on variation within populations are summarized below and in Table 4.

A selective sweep leads to fixation of a single haplotype, thus eliminating pre-existing variation surrounding the selected site – with the exception of rare recombination and mutation events. This creates a characteristic pattern of a relatively high number of rare alleles. Many classical tests for detecting selection, such as Tajima's D (Tajima 1989), attempt to detect this pattern. Some tests also consider the ancestral state of the alleles: regions affected by recent natural selection are likely to be enriched in high-frequency or fixed derived alleles. However, these tests may be sensitive to demographic factors and ascertainment bias, since the full allele frequency spectrum is never captured by studies based on SNP genotyping. (Carlson *et al.* 2005, Nielsen 2005, Williamson *et al.* 2005, Kelley *et al.* 2006, Sabeti *et al.* 2006, Nielsen *et al.* 2007, Williamson *et al.* 2007)

Another group of tests of selective sweeps concentrates on the pattern of haplotype variation and linkage disequilibrium in the region surrounding the selected locus. During a selective sweep, a haplotype surrounding the selected variant rises to high frequency rapidly, leaving little time for recombination to break the haplotype, while the other haplotypes at the same locus have a normal pattern of variation. Detection of such extraordinary haplotypes, first suggested by Sabeti *et al.* (Sabeti *et al.* 2002), has been the basis of many powerful methods to detect the selection of variants that have not yet reached fixation (Sabeti *et al.* 2006, Voight *et al.* 2006, Wang *et al.* 2006, Sabeti *et al.* 2007). Recently, this approach has been modified to detect past positive selection of already fixed haplotypes by analyzing population differences (Kimura *et al.* 2008, Sabeti *et al.* 2007, Tang *et al.* 2007) or increased linkage disequilibrium in a recently selected region (O'Reilly *et al.* 2008). These tests have the advantage of being less sensitive to ascertainment bias, and they are easily applicable on a genome-wide scale.

Differentiation between populations across the genome is caused by population history, but recent positive selection has been suggested to underlie those loci with clearly outlying values of allele frequency differences (Akey *et al.* 2002, Beaumont & Balding 2004, Weir *et al.* 2005, Myles *et al.* 2008, Oleksyk *et al.* 2008). This is obviously true for loci that are beneficial only in some environments, creating local selective pressures, but also for situations when a globally beneficial variant is still in the process of spreading throughout all the continents. However, recent research has indicated that neutral population processes, too, especially allelic surfing, may be behind extreme differentiation of individual loci, making it unreliable as sole evidence of selection (Klopfstein *et al.* 2006, Hofer *et al.* 2009). Allelic surfing may also mimic other features of natural selection, creating false positives in LD based tests, too (Nielsen *et al.* 2007).

Most of the genome-wide scans for positive natural selection are based on empirical analysis – i.e. the distribution of the selected test statistic is calculated throughout the genome, and the loci in the tail of the distribution are inferred to be affected by selection. The complication is that simulation studies have demonstrated that this approach leads to a high number of false negatives, and probably also some false positives, too (Kelley *et al.* 2006). Furthermore, since the extent of selection affecting the human genome is unknown, defining the threshold for the outliers of the empirical distribution is arbitrary, and assigning statistical significance – instead of simply describing how rare similar patterns are in the genome – is not possible (Kelley *et al.* 2006, Teshima *et al.* 2006, Nielsen *et al.* 2007). A more desirable approach would be to calculate a proper null distribution of genetic variation without selection, and compare the observed patterns with that. Despite relatively promising results from a few studies (Kim & Stephan 2002, Nielsen *et al.* 2005, Williamson *et al.* 2007), calculation of the null distribution may be affected by deficient modelling of demography and other factors.

Despite the major effort directed at unraveling the patters of natural selection and the several success stories (see below), the current methods probably create a biased and to some extent also erroneous picture of the traces of positive selection in the human genome (Nielsen *et al.* 2007). The overlap between the loci discovered by different studies is far from perfect (Biswas & Akey 2006, Nielsen *et al.* 2007, Oleksyk *et al.* 2008). The power of different statistics is affected by several factors, for example the demographic history of the studied population, the temporal scheme and strength of selection, the recombination pattern of the surrounding region, and whether the selection commences via a new mutation or from older variation (Teshima *et al.* 2006, Sabeti *et al.* 2007, O'Reilly *et al.* 2008). Consequently, the tests are often best suited to finding signs of strong, recent selection of a variant that emerged from a new mutation in a population of a stable size. Furthermore, few simulations of the performance of different tests include more complex features of genomic variation, such as evolution of recombination hotspots. There is still much work to be done developing new statistical methods and evaluating the old ones to obtain a more complete picture of positive selection in the human genome. Additionally, functional studies are required to verify the findings of genetic studies (Nielsen *et al.* 2007).

2.4.2 Observed patterns of selection in the human genome

For decades, the study of natural selection in the human genome was limited to candidate genes, which yielded several interesting examples of genes affected by positive selection (see e.g. McVean & Spencer 2006, Sabeti *et al.* 2006 for reviews). Recently, the availability of genome-wide datasets from the HapMap project, Perlegen Sciences and from genome-wide SNP chips has provided material for scanning the

Table 4. Effects of selective sweeps in the genomic region surrounding the beneficial variant (Nielsen 2005, Biswas & Akey 2006, McVean & Spencer 2006, Sabeti et al. 2006, Nielsen et al. 2007, O'Reilly et al. 2008)

	Effect of a selective sweep on genetic variation			
	Selected variant still segregating	**Selected variant reached fixation**	**Time scale for humans (years)**	**Most common methods***
Haplotype spectrum	Long high-frequency haplotypes carrying the selected allele, other haplotypes of normal variability	Increased linkage disequilibrium	< 30 000	LRH, iHS, XP-EHH, LDD, Ped/Pop etc.
Population differentiation	Increases	Decreases	< 50-75 000	F_{ST}, p_{excess}
Ancestral/derived alleles	Excess of high-frequency derived alleles	Excess of high-frequency derived alleles	< 80 000	Fay and Wu's H, Fu and Li's F
Allele frequency spectrum	Excess of both high- and low-frequency alleles	Excess of rare alleles	< 250 000	Tajima's D, Fu and Li's F
Number of segregating sites	Slightly decreases	Strongly decreases	< 250 000	Tajima's D, HKA, Fu and Li's F
Genetic differences between species	NA	Increased	> 6 million	HKA

* *Abbreviations and symbols: Long-range-haplotype (LRH), integrated haplotype score (iHS), cross-population extended haplotype homozygosity (XP-EHH), linkage disequilibrium decay (LDD), Hudson-Kreitman-Aguadé (HKA).*

entire genome for signs of selection. These studies have characterized several genes affected by recent selection acting on, for example, nutrition (*LCT*, Bersaglieri *et al.* 2004), pathogen resistance (*FY*, Hamblin *et al.* 2002; *G6PD*, Verrelli *et al.* 2006), skin pigmentation (*SLC45A2*, International HapMap Consortium 2005) and hair morphology (*EDAR*, Sabeti *et al.* 2007). Several studies have observed an enrichment of positively selected genes in gene ontology categories such as gametogenesis, immunological functions, sensory perception and steroid metabolism (Bustamante *et al.* 2005, Voight *et al.* 2006), providing interesting information on the systemic targets of human adaptation.

Many genes that have been influenced by natural selection are also important for human disease. Genes that contribute to Mendelian diseases have been shown to be more often under negative selection (Barreiro *et al.* 2008, Blekhman *et al.* 2008), and enrichment of genes affecting complex diseases has been suggested for loci under

positive selection (Bustamante *et al.* 2005, Nielsen *et al.* 2007). At least for some genes, this may be due to false positive associations due to increased population differences in the loci under selection (Freedman *et al.* 2004, Lange *et al.* 2008, Tian *et al.* 2008a). However, this is unlikely to be the full explanation. Most complex diseases have negative fitness effects, and thus it should be unlikely for high-frequency predisposing variants to be found in populations, and yet this is often the case – possibly due to natural selection. The observed pattern can arise from balancing selection – such as for many variants providing malaria resistance – or a change in the direction of selection, as in the famous "thrifty gene" hypothesis, according to which the advantage of high metabolic efficiency during most of human history is behind our contemporary susceptibility to diabetes and obesity (Nielsen *et al.* 2007).

3. Population history and genetic variation in Northern Europe

3.1 Europe

3.1.1 History

Anatomically modern humans arrived in Europe about 45 000-40 000 BP, probably mainly from the Middle East. The continent had already been inhabited by Neanderthals, who disappeared about 30 000 years ago after some 10 000 years of coexistence with modern humans (Mellars 1997, Mellars 2004, Mellars 2006). It is still debated whether the species interbred, thus leaving a Neanderthal contribution to the gene pool of modern Europeans, but genetic evidence suggests that the possible admixture was minor (Currat & Excoffier 2004, Green *et al.* 2006, Noonan *et al.* 2006).

Palaeolithic humans lived in small, mobile groups, whose subsistence was based on gathering and hunting the large game of Ice Age Europe. Northern Europe remained uninhabited due to the continental ice sheet, and during colder periods the human populations of Central Europe retreated to refugia in the south, where many other animal and plant species also survived. The end of the Ice Age around 12 000 BP marked the transition to the Mesolithic period, characterized by human migrations northward and more diverse subsistence strategies, with a heavier reliance on marine resources in coastal areas. (Mithen 1997, Peregrine 2001)

In Southern and Central Europe, the emergence of Neolithic traditions around 8000 BP was defined by the adoption of agriculture, ceramic traditions and a sedentary lifestyle. Agriculture spread to Europe from the Near East, where domestication of plants and animals had begun a few thousand years earlier, but it is still unknown whether the transition was brought to Europe by new immigrants or by cultural

diffusion – this may have varied between different parts of Europe. However, hunting and gathering remained important for several millennia, and in northernmost Europe the first Neolithic cultures adopted ceramics while still retaining their ancestral hunter-gatherer lifestyle. (Sherratt 1997b, Whittle 1997, Peregrine 2001)

Metal was introduced to South-Eastern Europe about 4500 BC and to Western Europe around 2500 BC; Bronze Age Europe was often characterized by hierarchical communities with extensive trade networks. The taming of the horse in the East European steppe introduced mobile pastoralism, and agriculture begun to gain a hold in northernmost Europe via the Neolithic Corded Ware culture. While bronze often had a symbolic rather than practical function, iron – introduced about 800 BC – was a more useful material for tools. The centralization of communities and development of social stratification continued, culminating in the formation of the Roman Empire. (Harding 1997, Sherratt 1997a, Peregrine 2001)

3.1.2 Languages

Most of the European languages belong to the Indo-European family. Its origins are still under debate: some linguists and archaeologists favour the hypothesis of an ancient spread from Anatolia via the development of agriculture, while others claim that Indo-European languages gained their dominance thousands of years later through the Kurgan culture and the taming of the horse in Eastern Europe (Diamond & Bellwood 2003). The Indo-European language family has several branches, including for example the Baltic languages in Latvia and Lithuania, Germanic languages in Scandinavia, Germany and Britain, Slavic languages in Eastern Europe, and Romance languages in the southwest. Languages belonging to the Finno-Ugric family are spoken in Hungary, the Baltic Sea region, the Volga-Ural region and in Siberia. Their origin is no better known than that of Indo-European languages: There have been controversial suggestions that the Finno-Ugric languages represent the most ancient linguistic strata in Northern Europe (Wiik 2002), but this hypothesis has been widely rejected by linguists (Häkkinen 2009 and references therein). The classical view has been that the Finno-Ugric languages were carried to the Baltic Sea region during the Comb Ceramic culture around 4000 BC from the Volga-Ural region, but this has recently been challenged by claims of a much more recent arrival of the Finno-Ugric language to the Baltic Sea region during the Bronze Age around 1800 BC (Aikio & Aikio 2001, Häkkinen 2009 and references therein).

3.1.3 Genetic variation

The genetic background of Europeans has been one of the main research foci of population genetic research. Generally, Y-chromosomal haplogroups show much stronger differences between regions and populations than mtDNA variation, which is

relatively uniform across Europe. Recently, genome-wide studies have yielded information on population differentiation in Europe, escaping the problem of using only a few loci. The most important findings of these analyses are outlined below.

Both mitochondrial DNA and Y-chromosomal variation have been associated with post-Ice Age migrations from different refugia. Several mitochondrial DNA haplogroups (V, U5b, H1, H3) have a diversity and frequency pattern suggesting an Iberian origin, and they are common throughout Europe (Torroni *et al.* 1998, Achilli *et al.* 2004, Loogvali *et al.* 2004, Achilli *et al.* 2005, Pereira *et al.* 2005). A similar origin has been suggested for Y-chromosomal haplogroups R1b and possibly also I1a, which harbour strong frequency gradients within Europe (Semino *et al.* 2000, Rootsi *et al.* 2004). A reverse frequency pattern from east to west has been observed in some mtDNA (H2, U4) (Malyarchuk *et al.* 2002, Loogvali *et al.* 2004) and Y-chromosomal (R1a, N3) haplogroups (Rootsi *et al.* 2007, Balanovsky *et al.* 2008). These have been associated with the eastern refugia in Ukraine and Siberia, the with Finno-Ugric migrations, and/or with the expansion of the Slavs. Additionally, many haplogroups have a frequency cline from the Near East to Europe (Di Giacomo *et al.* 2004, Balanovsky *et al.* 2008), which has often been interpreted as a trace of Neolithic migrations. Altogether, these frequency clines observed in the mtDNA and Y-chromosomal variation correspond relatively closely to the results from the early studies using classical blood group markers (Cavalli-Sforza *et al.* 1994, Rosser *et al.* 2000, Semino *et al.* 2000, Richards *et al.* 2002).

The question of the relative contribution of the ancient European Palaeolithic populations and the Neolithic migrants from the Near East to the modern European gene pool has been studied intensively. However, no consensus has been reached, and the estimates of the proportion of the Neolithic contribution have ranged from 20% to 100%. Analyses of ancient DNA support a major Palaeolithic component (Haak *et al.* 2005 and references therein), and Y-chromosomal variation has indicated a bigger Neolithic contribution than mtDNA variation (Chikhi *et al.* 2002), perhaps suggesting different male and female histories. A common pattern in genetic variation in Europe is the decrease of genetic diversity towards the north, which has been interpreted as a sign of migrations from the south which have caused serial bottlenecks (Lao *et al.* 2008, Novembre *et al.* 2008).

The early findings of clinal patterns of variation in Europe were often interpreted as distinct migration waves (see e.g. Cavalli-Sforza *et al.* 1994 and references above). However, recent research has shown that clinal patterns in principal component analysis are easily produced with a simple isolation-by-distance process of spatial variation (Novembre & Stephens 2008). Accordingly, many recent studies of population structure in Europe using genome-wide data have yielded a striking resemblance between geographical and genetic distances between individuals and populations (Heath *et al.* 2008, Lao *et al.* 2008, Novembre *et al.* 2008). Some outliers – such as the Finns (Lao *et al.* 2008) – can still be observed, but no major genetic borderlines have been observed.

3.2 The Baltic Sea region

3.2.1 History

Soon after the ice sheet retreated from Northern Europe around 12 000 BP, the Baltic Sea region was inhabited via several routes: The majority of the first inhabitants arrived in Scandinavia from Central Europe via Denmark, and in the eastern side of the Baltic Sea from the south, southeast and east. Additionally, the ice-free Norwegian coast provided a migration route to northern Fennoscandia. The early populations were Mesolithic hunter-gatherers, who relied heavily on marine resources. The adoption of agriculture and ceramics began around 4000 BC in southern Scandinavia via influences from Central Europe. On the eastern side of the Baltic Sea, the first ceramic culture, the Comb Ceramic, arrived from the east, first in 4200 BC but more forcefully in 3200 BC. A major cultural change was brought from Central to Northern Europe in 2300 BC by the Corded Ware (Battle-Axe) culture, whose spread may have been accompanied by population migration. Despite some attempts at agriculture during this period, hunting and gathering prevailed well into the metal ages in the northernmost Baltic Sea region. The archaeological cultures of northern Fennoscandia remained distinct from the southern development, with derivatives of the Comb Ceramic culture. (Siiriäinen 2003)

The Corded Ware culture was followed by the flourishing and rich Scandinavian bronze culture in approximately 1800 BC, also spreading to coastal Finland. Baltic countries, northern Fennoscandia, Eastern Finland and Karelia were influenced by the eastern bronze culture with its origins in Central Russia. During the Bronze Age, agriculture was properly introduced in Finland, both from Scandinavia and from the east (Siiriäinen 2003). In the Iron Age, beginning in 500 BC, the strong cultural contacts between southern Scandinavia and northern Germany prevailed, now possibly associated with an early Scandinavian language. South-western Finland and the Baltic countries showed close ties, which has been suggested to imply the emergence of Baltic Finnic languages. In Eastern Finland, northern Fennoscandia and Karelia, the tradition of eastern contacts continued, and a possible association with the Sami has been suggested. Petty chieftains appeared and this development continued throughout the Iron Age; the first centralized nations emerged at the turn of the first millennium. In the 8th and 9th centuries, the Vikings spread Scandinavian influence over much of Northern Europe, along the Atlantic coast as well as into Russia. Alongside with the entirety of Eastern Europe, the Baltic Sea region, too, was affected by the expansion of the Slavs in 600-900 AD. In the Middle Ages, German merchants and clergymen had a profound effect on the urban life of Northern Europe, especially in the Baltic countries, and during the later centuries, Sweden and Russia controlled large areas around the Baltic Sea. (Myhre 2003)

3.2.2 Genetic variation

European haplogroups and sequence motifs prevail in the mtDNA variation in all the populations of the Baltic Sea region with few differences between the populations (Torroni *et al.* 1996, Finnila *et al.* 2001, Helgason *et al.* 2001, Pliss *et al.* 2006, Hedman *et al.* 2007). Of the Y-chromosomal haplogroups, West European R1b is common particularly in the south-western part of the region, in a similar manner to I1a, which reaches its highest frequencies in Scandinavia but has been suggested to have West European roots. R1a, common in Eastern and Central Europe, is common in all the populations, although less so in Finland. (Rootsi *et al.* 2004, Kayser *et al.* 2005, Dupuy *et al.* 2006, Karlsson *et al.* 2006, Balanovsky *et al.* 2008). Recent genome-wide studies have supported the mainly European background of the populations of the Baltic Sea region (Heath *et al.* 2008, Jakkula *et al.* 2008, Lao *et al.* 2008, Novembre *et al.* 2008, McEvoy *et al.* 2009).

Eastern influences on the Baltic Sea region are most evident in the Y-chromosomal variation, although first observed by blood group markers (Guglielmino *et al.* 1990) and seen also in early genome-wide studies (Bauchet *et al.* 2007). Haplogroup N3 shows an interesting frequency pattern, being common on the eastern side of the Baltic Sea, in the Volga-Ural region and in Siberia, but despite numerous studies, the origin and historical association of the haplogroups remain unclear (Zerjal *et al.* 2001, Derenko *et al.* 2007, Rootsi *et al.* 2007, Mirabal *et al.* 2009). In mtDNA variation, eastern contacts have been most studied among the Sami, who appear to harbour relatively recent contacts with the Volga-Ural region (Meinila *et al.* 2001, Tambets *et al.* 2004, Ingman & Gyllensten 2007). Even though the eastern influence on the genetic variation is consistent with the eastern origins of the Finno-Ugric languages, as a whole, geography has been shown to be a more important determinant of Y-chromosomal variation than language in the Baltic Sea region (Zerjal *et al.* 2001).

Many populations of the region – especially the Sami, Finns and Karelians – show a decrease in genetic diversity and differentiation from the other Europeans, and there is still an ongoing debate to what extent this is caused by genetic drift or by a major eastern component among these populations (Cavalli-Sforza *et al.* 1994, Sajantila *et al.* 1995, Lahermo *et al.* 1996, Sajantila *et al.* 1996, Lahermo *et al.* 1999, Kaessmann *et al.* 2002, Hedman *et al.* 2004, Tambets *et al.* 2004, Service *et al.* 2006, Jakkula *et al.* 2008, Lao *et al.* 2008, McEvoy *et al.* 2009). In the case of the Sami, it has been suggested that their extreme differentiation from the neighbouring populations is mostly explained by their small population size, bottleneck effects and isolation, rather than a dramatically different origin compared to the other populations of the Baltic Sea region (Tambets *et al.* 2004).

3.3 Finland

3.3.1 History

The main features of Finnish population history are outlined above; however, in addition to these, several smaller events have shaped the population structure in Finland. In the Early Middle Ages, the coastal regions of Finland experienced a wave of immigrants from Sweden, and it is believed that the current Swedish-speaking minority in Finland descends from these settlers. The population in Eastern and Northern Finland remained very sparse and scattered until the 16th century, when a major migration wave from Southern Savo, encouraged by the King Gustav Vasa, led to the settlement of these regions. The majority of the current population descends from these settlers, who were very few in number especially in the northernmost regions. Until the rapid population growth and internal migrations beginning in the late 19th century, the Finnish population consisted of small communities with very little migration occurring over longer distances. (Pitkänen 2007)

3.3.2 Genetic variation

The small population size, historical founder and bottleneck effects, and pronounced local isolation have had pronounced effects on the genetic variation of the Finns. Large genetic differences between villages that partly average out at the regional level were first observed by H.R. Nevanlinna in 1972 (Nevanlinna 1972), but later seen for example in a Y-chromosomal study, where a small county showed extreme differences compared to larger population units (Palo *et al.* 2008). Consistent with this, a recent genome-wide analysis showed extreme local differentiation in Northern Finland (Jakkula *et al.* 2008). The different population history of Eastern and Western Finland has created regional differences observed in autosomal and Y-chromosomal variation (Workman *et al.* 1976, Kittles *et al.* 1998, Hedman *et al.* 2004, McEvoy *et al.* 2009), but similar patterns are less evident in mitochondrial DNA (Meinila *et al.* 2001, Hedman *et al.* 2007). The special features of Finnish history are also responsible for the Finnish disease heritage, the enrichment of about 30 rare Mendelian diseases in the Finnish population (Norio 2003a, Norio 2003b, Norio 2003c).

3.4 Sweden

3.4.1 History

Just as in Finland, there are long-standing prehistoric cultural differences between Northern and Southern Sweden. The northern part of the country retained a hunter-

gatherer type of subsistence for millennia after agriculture was established in southern Sweden, and the material remains show a strong cultural connection to other parts of northern Fennoscandia and possibly to the Sami. In Central and Southern Sweden, the differences are less pronounced, although the distinction between the southern Götaland and central Svealand predates historical time, and it was not until the 12th century that they were united under one ruler. (Lindkvist 2003, Lindqvist 2006)

Nonetheless, the formation of the country took centuries more. The southernmost parts of Sweden were originally Danish, and Sweden gained control of the area only after centuries of warfare. Finland was a part of Sweden from the 13th century to 1809, and especially in the 18th century many Finns migrated to Central Sweden. During the 17th century, Sweden reigned over large regions across the Baltic Sea, but these conquests probably left few permanent marks in the Swedish population. Norway and Sweden formed a union in the 19th century, and the western Swedish counties particularly have had substantial Norwegian influence. (Lindkvist 2003, Lindqvist 2006)

During the past decades, immigration to Sweden from all over the world has been substantial, and in 2007, 13.4% of the population in Sweden was of foreign origin. In particular, the biggest cities of Stockholm, Malmö and Gothenburg (Göteborg) harbour large immigrant communities. The biggest immigrant groups are from Finland, Scandinavia and other West European countries, the Balkans and the Middle East (Figure 4).

3.4.2 Genetic variation

The genetic variation of the Swedish population has been less studied compared to, for example, Finland. The Y-chromosomal variation appears to follow a general European or Scandinavian pattern, but the internal differences are slight, with increased Danish influence in the south, and a strong divergence of the Västerbotten area in north-eastern Sweden (Holmlund *et al.* 2006, Karlsson *et al.* 2006). Additionally, autosomal studies have indicated a very slight population structure overall (Hannelius *et al.* 2008), with some differences between the river valleys of Northern Sweden (Einarsdottir *et al.* 2007).

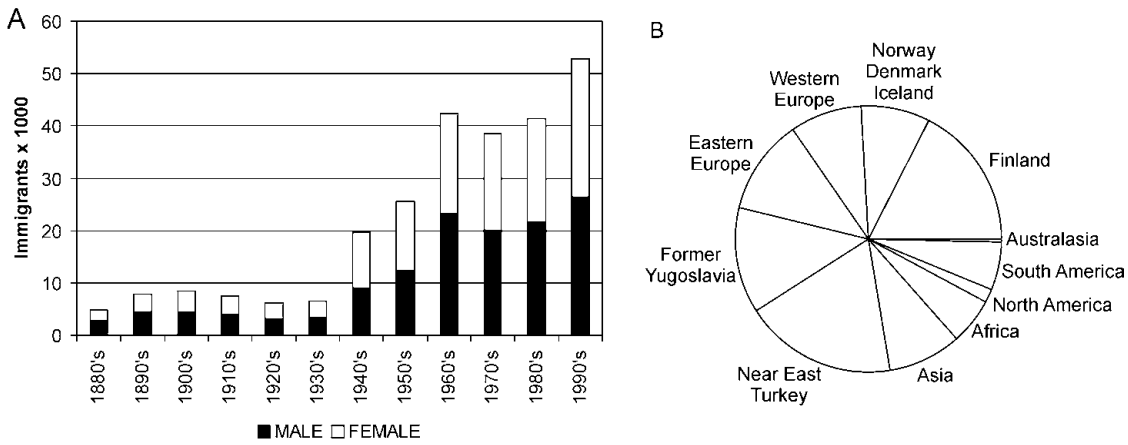

Figure 4. Immigration to Sweden (Statistiska Centralbyrån 2008). From III.

AIMS OF THE STUDY

The aim of the study was to characterize population genetic variation in the Baltic Sea region from many perspectives; specifically to analyze:

1. Population differentiation, migratory waves and genetic diversity in the Baltic Sea region using Y-chromosomal and mitochondrial DNA (II) as well as genome-wide (IV) markers.
2. The historical population structure within Finland using Y-chromosomal (I) and genome-wide variation (IV)
3. The population structure in contemporary Sweden based on Y-chromosomal and mitochondrial DNA variation (III)
4. Loci across the genome that have been affected by recent positive natural selection in North European populations (V)

MATERIAL AND METHODS

1. Samples and datasets

This study was based on a large sample collection from several populations of the Baltic Sea region. The samples have been collected with informed consent according to the guidelines of the declaration of Helsinki (1964), and the use of the samples has been approved by the local ethics committees. Additionally, genome-wide datasets from Germany, Britain, and the HapMap project were used. The samples and datasets used in this study are summarized in Figure 5, Figure 6 and Table 5.

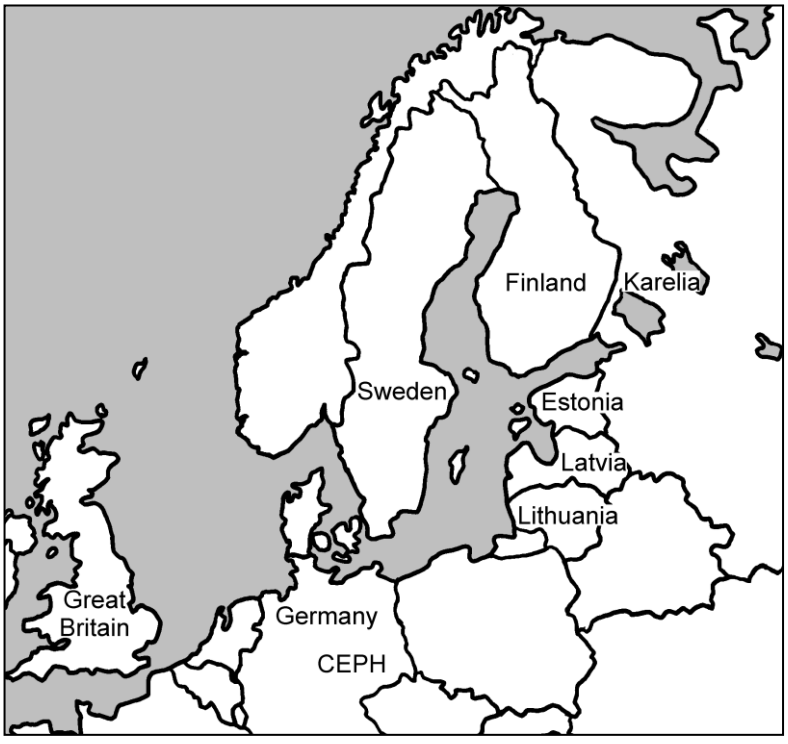

Figure 5. Map of the studied European populations. See the original publications and Figure 6 for geographical distributions of the sampled areas, and Table 5 for details of the sample sets.

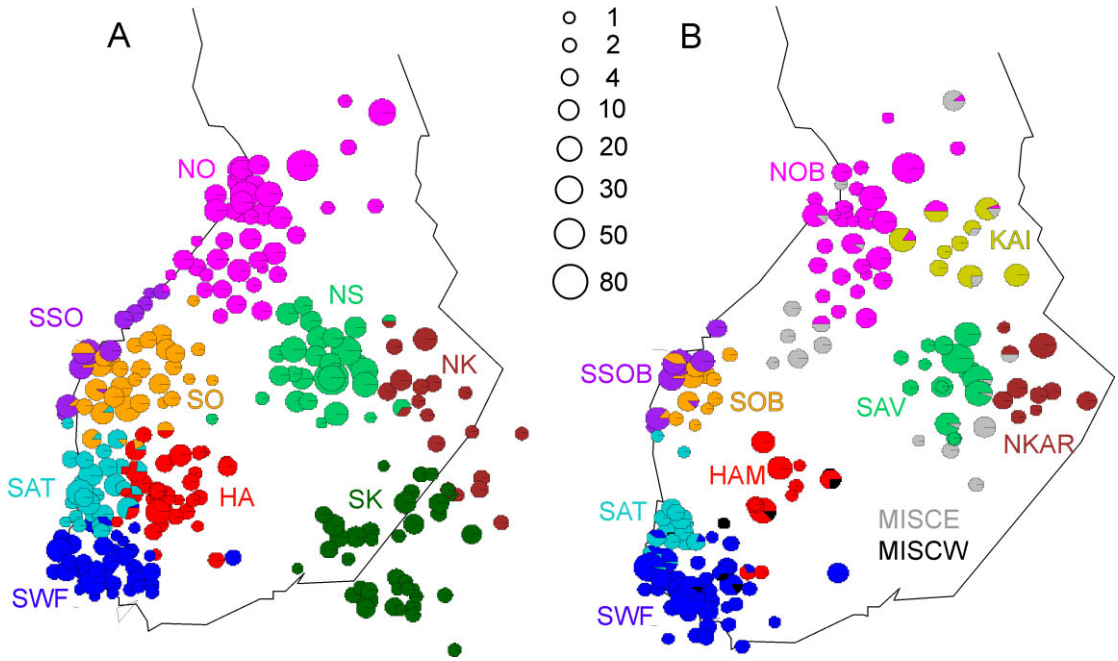

Figure 6. Birth counties of the grandparents of the Finnish samples in the Y-chromosomal study (I) (a), and in the genome-wide study (IV) (b). The size of the circles corresponds to the number of grandparents from each county in a logarithmic scale, and the colours denote the regional classification indicated with the abbreviations as in the original publications. The abbreviations are: South-Western Finland (SWF), Satakunta (SAT), Häme (HA/HAM), Southern Ostrobothnia (SO/SOB), Swedish-Speaking Ostrobothnia (SSO/SSOB), Northern Ostrobothnia (NO/NOB), Kainuu (KAI), Northern Savo (NS/SAV), Northern Karelia (NK/NKAR), Southern Karelia (SK), Miscellaneous East (MISCE) and Miscellaneous West (MISCW).Unpublished.

Table 5. Samples and datasets used in this study (Hannelius et al. 2005, International HapMap Consortium 2005, Krawczak et al. 2006, Wellcome Trust Case Control Consortium 2007).

Population	**Abbreviation**	**Used in study (sample size)***	**Type**	**Geographical location**	**Ancestry and ascertainment**	**Acknowledgement**
Finland	FIN FIE + FIW	I,II (536) IV,V (280)	Genomic DNA	See Figure 6	Birth county of grandparents	Pertti Sistonen, Marja-Liisa Savontaus, Antti Sajantila, Päivi Lahermo
Sweden I	SWE	II (307) IV,V (113)	Genomic DNA	Eastern Sweden	Ethnic Swedes	Peter Andersen
Sweden II	SWE	III (1703)	Whole genome amplified DNA	Whole country	Birth hospital	Ulf Hannelius, Ulrika von Döbeln, Juha Kere
Karelia	KAR	II (512)	Genomic DNA	Aunus, Viena & Tver Karelia	Ethnic Karelians, town of residence	Tuula Koski
Estonia	EST	II (114)	Genomic DNA	Whole country	Ethnic Estonians	Richard Villems
Latvia	LAT	II (117)	Genomic DNA	Whole country	Ethnic Latvians	Richard Villems
Lithuania	LIT	II (163)	Genomic DNA	Whole country	Ethnic Lithuanians	Richard Villems
Germany	GER	IV, V (256)	Data	Kiel Province	Residents in Kiel province	Stefan Schreiber
Great Britain	BRI	IV (296) V (700)	Data	Great Britain	Region of birth; no recent immigrants	Wellcome Trust Case Control Consortium
CEPH	CEU	IV, V (58)	Data	Utah, USA	Utah residents of European background	HapMap Consortium, Affymetrix
China	CHB	IV, V (45)	Data	Beijing, China	Han Chinese	HapMap Consortium, Affymetrix
Japan	JPT	IV, V (42)	Data	Tokyo, Japan	Japanese	HapMap Consortium, Affymetrix
Nigeria	YRI	IV, V (56)	Data	Ibadan, Nigeria	Yoruba	HapMap Consortium, Affymetrix

** The sample size corresponds to the biggest sample size used in the study – in some of the analyses the sample size may be smaller*

2. Genotyping

2.1 Markers

The markers used in this study are listed in Table 6 – for a detailed account, see the original publications.

Table 6: Summary of genetic markers used in this study

Study	Mitochondrial DNA	Y chromosome	Genome-wide
I	-	Haplogroup analysis:12 SNPs, 2 indels Haplotype analysis: 9 microsatellites	-
II	Haplogroup analysis: 17 coding region SNPs Haplotype analysis: HVS1 sequence	Haplogroup analysis: 17 SNPs, 1 indel Haplotype analysis: 9 microsatellites	-
III	Haplogroup analysis: 32 SNPs	Haplogroup analysis: 10 SNPs	-
IV	-	-	Affymetrix 250K Sty SNP array
V	-	-	Affymetrix 250K Sty & 250K Nsp SNP arrays

2.2 SNP genotyping (I-V)

2.2.1 RFLP and allele-specific PCR (I,II)

Most of the Y-chromosomal SNPs in I and some of the mtDNA SNPs in II were genotyped by the traditional RFLP method, where the segment flanking the locus is amplified by PCR, digested with a restriction enzyme whose restriction site is modified by the studied SNP, and the resulting fragments are separated using an agarose gel. Two insertion-deletion polymorphisms and one SNP were genotyped by allele-specific PCR, where the primers are designed to overlap the SNP site so that PCR product is obtained only from one allele. Genotype calling was done manually. All the genotyping reactions had negative controls and positive controls when necessary, and the samples were either re-genotyped or excluded from analysis if the genotypes could not be defined.

2.2.2 Sequenom (II,III)

All of the SNP genotyping in III and most in II was done on the Sequenom MALDI-TOF platform (San Diego, CA). In this method, the SNP regions are amplified in a multiplex PCR reaction, and the SNP allele is captured by a short extension of a primer aligning adjacent to the SNP site. The masses of the resulting fragments – which depend on the SNP allele – are defined by MALDI-TOF mass spectroscopy. Genotype-calling was done by the Sequenom Typer 3.1 and 4.0 software. All the Sequenom reactions were done with several negative and positive controls, and the call rates, consistency with the known phylogeny, and correct genotyping of the control and duplicate samples were checked.

2.2.3 The Affymetrix SNP array (IV, V)

The genome-wide SNP genotyping by the Affymetrix (Santa Clara, CA) 250K Sty array was done by the Bioinformatics and Expression Analysis core facility in the Karolinska Institutet, Stockholm, Sweden. Genomic DNA was digested with a restriction enzyme, labelled and hybridized to chips with probes of the sequence of each SNP allele and flanking region. Genotype-calling was done automatically by the GTYPE software using the BRLMM algorithm. The data used in the analyses passed a stringent quality control according to common standards of genome-wide SNP genotyping.

2.3 Microsatellite genotyping (I, II)

Y-chromosomal microsatellite loci were amplified in PCR with one fluorescent primer, and the pooled PCR products were separated according to size by an Applied Biosystems 3730 sequencer (Carlsbad, CA). Genotype-calling was done using the GeneMapper software by Applied Biosystems. All the genotyping was done with negative and positive controls and duplicate samples, and the call rates and correct genotyping of the controls and duplicates were checked.

2.4 Sequencing (II)

Sequencing of the hypervariable segment I of the mitochondrial DNA was performed by standard Sanger sequencing in the forward direction, and also in the reverse direction when necessary. The region was amplified by PCR, the excess nucleotides and primers were removed, and the sequencing reaction was performed with fluorescently labelled ddNTPs. The fragments were separated by Applied Biosystems capillary electrophoresis, and the chromatograms were read using the Staden Package software

(http://staden.sourceforge.net/). All the PCR reactions were performed with negative controls.

3. Population genetic analysis

The challenge of population genetic studies is rarely the genotyping but the statistical analysis. Data management, formatting and analysis in this study has been performed using R (R Development Core Team 2008, http://www.R-project.org), Perl, Matlab (Math-Works, Inc. Natick, MA) and MS Office, in addition to some specific population genetic software referred to below. The most important or nonstandard statistical analysis methods are briefly discussed below; for a detailed account of these and some additional methods, see the original publications.

3.1 Differences between populations

3.1.1 Principal component analysis and multidimensional scaling

Principal component analysis or PCA (I-III) and multidimensional scaling or MDS (IV) are methods for displaying complex datasets in fewer dimensions in order to extract and visualize the most important trends. The first principal component (PC) is an eigenvector fitted to the correlation or covariance matrix of the data (e.g. haplogroup frequencies of populations) that explains most of the observed variation. The following PCs are always perpendicular to the preceding component. The eigenvalues of the PCs express how much of the variation they account for. Another method for visualizing complex data, classical multidimensional scaling, takes the data as a matrix of dissimilarities, such as genetic distances between individuals or populations, and produces an output of distances in the desired number of dimensions so that the deviations from the original distances are minimized. In I-III, PCA was calculated from the covariance matrices of haplotype frequencies in Matlab, and in IV, MDS was calculated from the identity by state (IBS) distance matrices in R.

Both of these methods are useful for finding trends behind complex datasets. However, the results are often very dependent on the data included in the analysis, which makes it difficult and risky to draw conclusions based on PCA or MDS plots alone. In PCA, complex clinal patterns easily interpreted as migration waves have been shown to arise even in the presence of a pattern of simple isolation by distance (Novembre & Stephens 2008). Similarly, while the methods often produce visually attractive plots, the statistical testing of clustering patterns is difficult, and they fail to

consider the uncertainty of the input data: allele frequencies or genetic distances calculated from a population of 5 samples are treated equally to those obtained from 1000 samples.

3.1.2 Allele frequency-based measures

F-statistics, originally developed by Sewall Wright, is a classical way of partioning genetic variance into components representing different levels of population hierarchy – individuals, populations, and groups of populations – and it is based on estimating the decrease of heterozygosity due to non-random mating at each level studied. The measure of deviation from panmixia due to population subdividision relative to the total genetic variance, F_{ST}, is a commonly used measure of genetic distance between two populations. Several modifications of Wright's F_{ST} have been developed to account not only for the heterozygosity of a single locus but also nucleotide differences between multiple markers, as in Analysis of Molecular Variance (AMOVA, see e.g. Weir & Cockerham 1984, Excoffier *et al.* 1992). Another important derivative of F_{ST} is the adjustment for the stepwise mutation model of microsatellite loci (R_{ST}, Slatkin 1995). In I-IV, population-based F-statistics and AMOVA were calculated in Arlequin (Schneider *et al.* 2000), with the significance estimated by permutation. An R script was written to calculate F_{ST} for each SNP of the genome-wide dataset in V (Weir & Cockerham 1984, Akey *et al.* 2002).

In order to compare the extent of eastern influence among the North European populations studied for the genome-wide dataset, the number of markers whose frequency deviates from the HapMap European frequency towards or away from the HapMap Asian frequencies was calculated. If all the North European populations had diverged from a common proto-European ancestral population merely by drift, there should be no reason why different numbers of markers should drift to a particular direction in each population, even though the extent of drift can obviously be different. The significance of the differences between populations was calculated by a standard χ^2 test.

3.1.3 Individual-based analyses

The Structure software (Pritchard *et al.* 2000) is based on a Bayesian algorithm that assigns individuals to a given number of clusters according to genotype data of a large number of unlinked loci, so that deviation from the Hardy-Weinberg equilibrium is minimized. In the admixture model, each individual is assigned jointly to several clusters with varying proportions. By running Structure with several different numbers of clusters and comparing their posterior probabilities, it is possible to estimate which number of clusters corresponds best to the data. However, in strongly admixed populations and with a pattern of isolation by distance rather than discrete population

units, the inference of the correct number of clusters and its biological interpretation is often difficult. However, Structure is one of the few methods where no prior population assignment is needed.

A more straightforward method for estimating differences between populations from large datasets is provided by calculating the distribution of identity by state between all individual pairs between two populations. In IV, that was calculated using the R package GenABEL (Aulchenko *et al.* 2008), and the statistical significance of differences between population pairs was calculated by a Mann-Whitney U test.

3.2 Measures of genetic diversity

The extent of genetic diversity in a population can be calculated in a variety of ways. One classical statistic, used in I, is haplotype diversity (Nei 1987), which is defined as the probability that two randomly selected haplotypes are different. This statistical method does not take genetic distance between haplotypes into account, unlike the calculation of average number of pairwise differences between haplotypes (Tajima 1983, Tajima 1989), used in II and III. For genome-wide data, calculation of average identity by state (IBS) over all the markers between two individuals is analogous to the average number of pairwise differences, and the mean or median of IBS values for all individual pairs within a population is the higher, the more similar the individuals are genetically to each other. The significance of the differences of the IBS distributions in IV was estimated by a Mann-Whitney U test.

In the genome-wide analysis of IV, the number of monomorphic markers in the populations and the distribution of minor allele frequencies were calculated in order to estimate the extent of the fixation of rare alleles due to genetic drift. Additionally, high genetic drift via small population size or bottleneck events leads to a random loss of haplotypes, which increases linkage disequilibrium (LD). D', a common measure of LD (Lewontin 1964), was calculated for all SNP pairs within 100 markers from each other, and plotted as a function of physical distance. Significance of the differences between populations was calculated by a Mann-Whitney U test.

3.3 Correlation analyses

In III and IV, the patterns of genetic variation within Sweden and Finland, respectively, were analyzed by calculating the correlation between matrices of genetic and geographic distances with a Mantel test (Mantel 1967) in Arlequin and the R package ade4 (Chessel *et al.* 2004). A related method, spatial autocorrelation analysis, was used for detection of geographical trends in the haplotype data in III, using the AIDA software (Bertorelle & Barbujani 1995). It calculates the correlation between the geographical and haplotype distances of individual sample pairs, thus avoiding

predefined classification of subpopulations. In III, geographical and population historical trends behind haplogroup frequencies were investigated by calculating the Pearson's correlation between the latitude or proportion of immigrants and haplogroup frequencies. The significance of all of these correlations was determined by permutation.

An additional test, not included in the original publications, was designed to test the correlation between the genetic and geographic distances relative to different geographical directions in Finland, calculated from the genome-wide data. The geographical locations of each individual were projected onto 2 dimensions using the Bonne projection (McIlroy *et al.* 2005), and these coordinates were projected onto a line with a specified direction. The Euclidean distances were calculated between the projected points, thus obtaining geographic distances between the locations relative to the right angle of the projection line. The Spearman correlation between these distances and the genetic distances of all individual pairs was calculated, and this was repeated for the 40 different angles of the projection line.

3.4 Median-joining network analysis

An important method for visualization of haplotype data is the construction of a phylogenetic network of haplotypes, which allows inspection of their population and allele frequency distributions. For haplotypes without recombination or recurrent mutations, the analysis produces a perfect tree instead of a network. In this study, median-joining networks of Y-chromosomal (I,II), mtDNA (II), and autosomal haplotypes (V) were constructed with the Network software (fluxus-engineering.com, (Bandelt *et al.* 1995, Bandelt *et al.* 1999, Polzin & Daneschmand 2003).

Calculation of the time to the most recent common ancestor (TMRCA) of the observed haplotype variation is possible if the mutation rate is known. Simple models are based on the calculation of the number of mutations and converting this number into years based on the mutation rate and generation length. In II, the time to TMRCA was calculated for the most common Y-chromosomal and mtDNA haplogroups in the Baltic Sea region by the Network software by using previously estimated mutation rates (Forster *et al.* 1996, Zhivotovsky *et al.* 2004).

3.5 Tests of positive natural selection (V)

3.5.1 Genome-wide analysis

For each marker of haplotype phased (Browning & Browning 2007) genome-wide data, two test statistics were calculated separately in each population to detect positive

selection: the single-SNP long-range haplotype test (LRH, Sabeti *et al.* 2007), and the integrated haplotype score test (iHS, Voight *et al.* 2006). Both of these tests are based on calculating the extended haplotype homozygosity (EHH) statistic for both alleles of each SNP until it decays under a threshold, and comparing the EHH of the alleles. If one of the alleles has been increasing in frequency due to natural selection, recombination has had less time to break the surrounding haplotype than for alleles of similar frequency evolving neutrally. Thus, the selected allele should be surrounded by a longer haplotype. The LRH is based on the EHH between the SNP studied and one SNP on each side, where the total homozygosity of the haplotype has decreased below 0.05. The iHS is calculated from the landscape of EHH decay, and is thus based on several markers. The statistics were calculated by the Sweep software (Sabeti *et al.* 2007).

The single-SNP iHS and LRH values were standardized in frequency bins (Voight *et al.* 2006, Sabeti *et al.* 2007) to obtain comparable values across the genome. These values, in addition to the genome-wide F_{ST} values, were analyzed in overlapping 200 kb windows in order to be able to extract the regions with several SNPs with outlying values. The windows were classified into extreme and suggestive outliers, and the most likely candidate regions for positive natural selection were those windows that fell into the best category based on the iHS or LRH while also having a suggestive signal in LRH or iHS, respectively, or in F_{ST}.

The performance of the iHS and LRH with different sample sizes was assessed by calculating these statistics for chromosomes 1-3 in randomly-selected British samples of seven different sizes. The correlation between the standardized values of each marker was calculated between the largest sample of 700 individuals and all the other tested sample sizes. Since the sample size was observed to affect the reliability of the statistics, the iHS and LRH values from populations with smaller sample sizes were downscaled with the correlation value of respective sample size before the windowing analysis described above.

3.5.2 Simulations

Coalescent simulations by the SelSim software (Spencer & Coop 2004) were used for investigating the performance of iHS and LRH tests for datasets of different SNP densities and sample sizes. Genomic regions were simulated with a neutral model and a single selection scenario, and different marker densities and sample sizes were collected from the simulation results, adjusting for ascertainment bias by matching the allele frequency distribution to that observed in real data (Voight *et al.* 2006). iHS and LRH calculations were performed as described above, and the power was calculated by adjusting the false discovery rate of each analysis to approximately 1%.

In order to analyze the extent of genetic differentiation between closely related populations due to natural selection, simulations of allele frequencies were performed using realistic demographic models and various scenarios of natural selection. The

simulations were done with two population pairs, one corresponding to two North European populations, and one corresponding to an East Asian and a European population. The demographic models – including variable population sizes and migration – were adjusted to match the observed allele frequency differences between populations.

RESULTS AND DISCUSSION

1. Genetic variation in the Baltic Sea region

Genetic variation of the populations in the Baltic Sea region was investigated using mitochondrial DNA and Y-chromosomal data (II, also I and III) as well as genome-wide SNP data (IV). Together, these studies provide a comprehensive picture of the genetic variation and population history in Northern Europe, especially in the Baltic Sea region. However, the coverage of the sample sets varied between the studies: Finnish and Swedish samples were analyzed in both II and IV as well as in I and III, respectively, while the mtDNA and Y-chromosomal analysis also covered Karelia and the Baltic countries, and the genome-wide analysis included German, British and European-American samples.

1.1 Y-chromosomal variation (I, II, III)

The Y-chromosomal variation in the Baltic Sea region has intriguing differences between the populations, and the haplogroup frequency variation is unusually large for such a small geographic region. There are four abundant haplogroups: N3, I1a, R1a1 and R1b (Figure 7), whose frequency distributions and diversity patterns are reviewed below.

Haplogroup N3 had high frequencies on the eastern side of the Baltic Sea, which is consistent with earlier studies (Lahermo *et al.* 1999, Raitio *et al.* 2001, Zerjal *et al.* 2001, Laitinen *et al.* 2002, Karlsson *et al.* 2006). The haplogroup has been suggested to originate from Mongolia or Northern China, but the subsequent migration routes carrying the haplogroup westward remain unclear (Zerjal *et al.* 1997, Derenko *et al.* 2007, Rootsi *et al.* 2007, Mirabal *et al.* 2009). In the Baltic Sea region, N3 clearly marks an eastern influence, but a more accurate origin or temporal scale is difficult to denote. The microsatellite variation within the haplogroup showed distinct haplotype clusters for the Finno-Ugric and Baltic-speaking populations, as suggested earlier (Zerjal *et al.* 2001, Roewer *et al.* 2005). This, in addition to the high haplotype diversity of both of the clusters, suggests two different migration waves along which N3 was carried to the Baltic Sea region. However, it is unclear to which extent the higher N3 frequency in Eastern compared to Western Finland actually marks increased eastern immigration, and how much of the high frequency is due to genetic drift – the haplotype diversity was

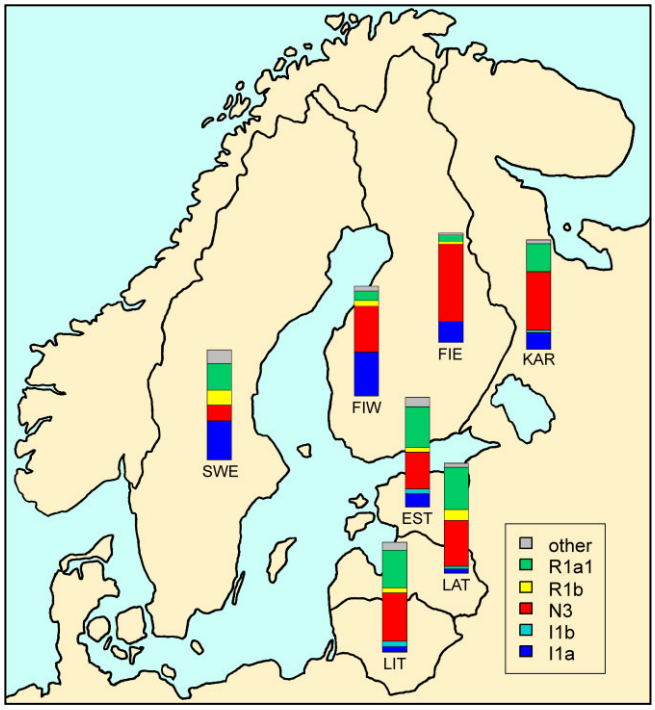

Figure 7. Y-chromosomal haplogroup frequencies in the populations studied. The abbreviations are: Sweden (SWE), Western Finland (FIW), Eastern Finland (FIE), Karelia (KAR), Estonia (EST), Latvia (LAT), Lithuania (LIT). Based on data from II.

lower in Eastern Finland, suggesting at least some effect of drift. Haplotype comparisons (www.yhrd.org) and a more detailed analysis of the N3 frequencies in Sweden (III) suggested westward diffusion of the haplogroup from Finland to northern and central Sweden, and also from the Baltic countries towards Poland.

Haplogroup I1a is known to have its highest frequencies in Scandinavia, but its origins have been suggested to lie in Western Europe (Rootsi *et al.* 2004). An important finding of I was that I1a actually reaches an equally high frequency in Western Finland and is common also in Eastern Finland, easily interpreted as a sign of Scandinavian migration to Finland. Interestingly, the high diversity of the haplogroup in Eastern Finland and Baltic countries, and the haplotype comparisons of I1a in II complicate this pattern, suggesting that its presence in Finland and the Baltic countries may be due to migration also from the south rather than solely from Sweden.

The most common Y-chromosomal haplogroup of Eastern Europe is R1a1 that is especially common among Slavic populations (Balanovsky *et al.* 2008). However, the high frequencies of the haplogroup also in Poland, Germany and Scandinavia (Kayser *et al.* 2005, Dupuy *et al.* 2006, Karlsson *et al.* 2006) and particularly the haplotype comparisons suggest that in the Baltic Sea region, its high frequencies in most of the populations are unlikely to be due to Russian influence but, instead, stem from migrations from northern parts of Central Europe, similarly to I1a. This is supported by

the low frequencies of haplogroup I1b, which is common in Russia (Rootsi *et al.* 2004). The northward expansion of both R1a1 and I1a from Central Europe may be a result of population migrations during the Neolithic, or later periods. In Sweden, the high frequencies of the haplogroup in the western parts of the country observed in III are likely to be due to ancient influence from Norway where the haplogroup is very common (Dupuy *et al.* 2006).

Haplogroup R1b reaches very high frequencies in Western Europe with a rapid decline eastward (Rosser *et al.* 2000, Semino *et al.* 2000, Kayser *et al.* 2005), and among the populations studied it showed a decreasing frequency cline towards the northeast. The Central European roots of R1b are also evident in the geographical cline in Sweden with highest frequencies in the southern parts of the country (III and Karlsson *et al.* 2006).

Estimating the age of haplogroups is important for connecting genetic patterns to historical phenomena. However, it is dependent on the correct estimation of the mutation rate, which has proven to be difficult. Rates calculated from pedigrees are 3-4 times higher than evolutionary rates (Parsons *et al.* 1997, Howell *et al.* 2003, Dupuy *et al.* 2004, Zhivotovsky *et al.* 2004, Zhivotovsky *et al.* 2006), and it is unclear which should be used for the calculation of the most recent common ancestor for major haplogroups in large geographic regions. It has recently been suggested (Pontikos 2008) that the widely used evolutionary rate of the Y chromosome (Zhivotovsky *et al.* 2004) is strongly underestimating the effective mutation rate due to not accounting for population growth and the bias of analyzing the biggest haplogroups that have grown at rates exceeding the general growth rate of the population. These analyses have not been published in a peer-reviewed journal, but they appear to correctly point out at least some problems of the commonly used models. Thus, the appropriate mutation rate to use for analyzing the temporal scale of the Y-chromosomal haplogroup variation may be a few times lower than was used in II – close to the pedigree rate. The same bias should apply to mitochondrial DNA, too. If the revised rates (Pontikos 2008) were used instead, TMRCAs for the main Y-chromosomal haplogroups I1a, N3 and R1a1 would be in the order of 3000-4000 years before present. These dates would imply that instead of the proposed Neolithic arrival of these haplogroups, their upper age limit would be in late Neolithic or early Bronze Age. Interestingly, the revised age of N3 variation in the Baltic Sea region would actually correspond nicely with the recently suggested Bronze Age arrival of the Finno-Ugric language (Häkkinen 2009). However, given the current uncertainty of the appropriate mutation rates, all time estimates should be used with great caution.

1.2 Mitochondrial DNA variation (II, III)

Mitochondrial DNA variation in the Baltic Sea region followed a general European pattern, and had much smaller differences between the populations studied than the Y-chromosomal variation. However, several interesting patterns were observed. Haplogroups H1 and U5b reached high frequencies in the Baltic Sea region, a feature similar to South-Western Europe despite being rarer in Central Europe (Achilli *et al.* 2004, Achilli *et al.* 2005, Pereira *et al.* 2005, Torroni *et al.* 2006). The presence of these haplogroups – with probable origins in the Iberian refugium – throughout Europe strongly supports a high contribution of West Europeans in the settlement of the continent (Torroni *et al.* 2006), but the enrichment in these haplogroups in the extremites of the European continent is not explained by that alone. However, additional data from other parts of Europe are needed to explain the observed frequency pattern.

The mitochondrial DNA results showed relatively high frequencies of haplogroups abundant among the Sami, such as U5b1b, Z, and D5, among the Karelians and to a smaller degree also among the Finns and Swedes, supporting earlier results of admixture between the these populations (Sajantila *et al.* 1995, Lahermo *et al.* 1996, Finnila *et al.* 2001, Meinila *et al.* 2001, Tambets *et al.* 2004, Hedman *et al.* 2007). Interestingly, the eastern elements in the Sami mtDNA variation have been associated to the Volga-Ural region (Ingman & Gyllensten 2007), and since these haplogroups were observed to be present among the neighbouring populations as well, it is plausible to postulate some degree of eastern influence in the entire Baltic Sea region. Further support is provided by the comparatively high frequency and diversity of haplogroup U4, which is most common in the Volga-Ural region (Bermisheva *et al.* 2002, Pimenoff *et al.* 2008). These observations are consistent with the eastern influences observed in the Y-chromosomal variation of the populations of the Baltic Sea region.

1.3 Genome-wide variation (IV)

In the genome-wide analysis of over 200 000 SNPs, genetic differences between the Germans, British and the European-American individuals were small, although statistically significant, while the differentiation of Swedes, Western Finns and most of all Eastern Finns was much more pronounced (Figures 8-10). This is consistent with other studies showing relatively small differences within Central Europe as opposed to the increased differentiation of the Finns (Cavalli-Sforza *et al.* 1994, Seldin *et al.* 2006, Bauchet *et al.* 2007, Heath *et al.* 2008, Jakkula *et al.* 2008, Lao *et al.* 2008, Novembre *et al.* 2008, Price *et al.* 2008, Tian *et al.* 2008b, McEvoy *et al.* 2009).

High linkage disequilibrium, increased similarity within the population (Figure 9), and various other measures showed a decrease in genetic diversity especially in Eastern Finland, and to a lesser degree also in Western Finland and Sweden. This is best

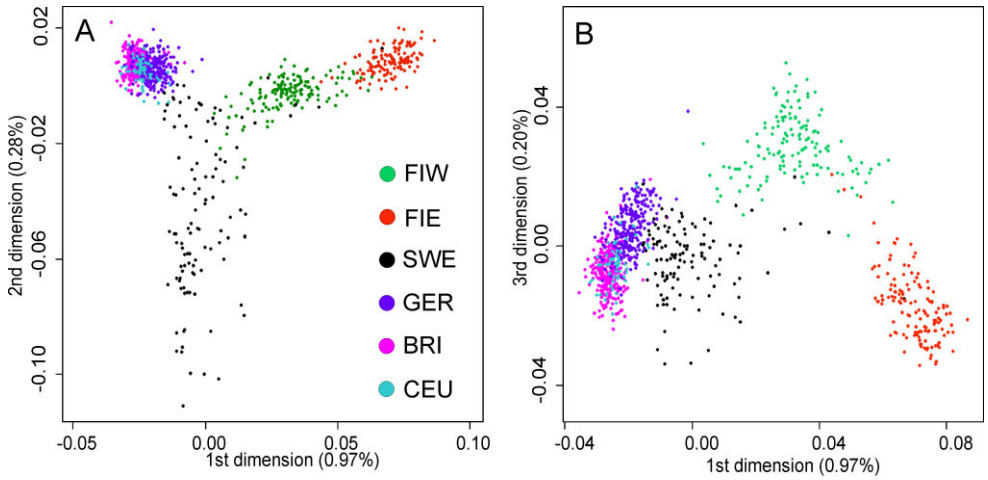

Figure 8: A multidimensional scaling plot of the identity-by-state distances between European individuals in the 1st and 2nd (a) and 1st and 3rd (b) dimensions. The abbreviations are: Western Finland (FIW), Eastern Finland (FIE), Sweden (SWE), Germany (GER), Great Britain (BRI) and CEPH (CEU). Adopted from IV.

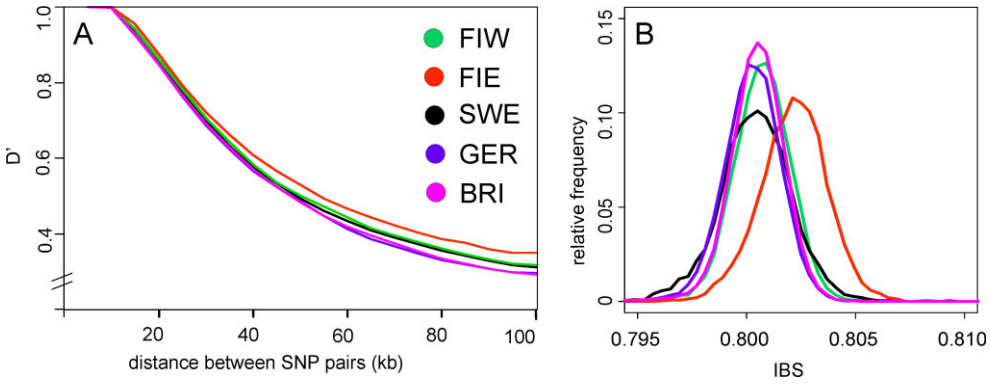

Figure 9: Linkage disequilibrium (D') in the populations studied as a function of intermarker distance (a), and the distribution of pairwise identity by state within each population (b). Abbreviations as in Figure 8. Adopted from IV.

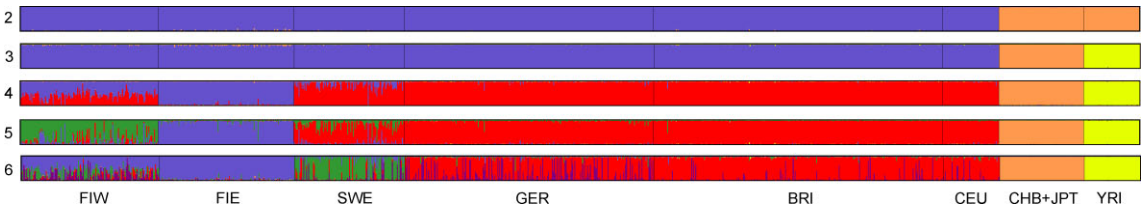

Figure 10: Structure analysis with 2-6 clusters. Each individual is represented by a thin vertical line, and the colours denote proportions of different clusters. Abbreviations as in Figure 8. The analysis is based on data in IV, but has been recalculated to include all the HapMap populations.

explained by a pronounced genetic drift caused by the later foundation of these populations, smaller population size, isolation, and population bottlenecks (see section 3.3). Similar phenomena have been observed in other studies (Sajantila *et al.* 1995, Sajantila *et al.* 1996, Kittles *et al.* 1999, Lahermo *et al.* 1999, Hedman *et al.* 2004, Jakkula *et al.* 2008) and in the Y-chromosomal analyses of this study. Genetic drift isprobably one of the main contributors to the increased differentiation of Finns and Swedes, too.

The comparison of the European populations to the HapMap data from Asia revealed signs of increased eastern influence especially among the Eastern Finns: they were more similar to the Asians than the other populations ($p < 10^{-14}$), showed a slight Asian component in the Structure analysis (Figure 10), and there was an increased number of markers with frequency deviation towards the Asian populations ($p < 10^{-5}$). This strongly supports the Y-chromosomal and mitochondrial observations of increased eastern influence in Finland in this and other studies (Guglielmino *et al.* 1990, Cavalli-Sforza *et al.* 1994, Lahermo *et al.* 1999, Zerjal *et al.* 2001). Unfortunately, the lack of data from more relevant reference populations from the east, and also from the Sami population, makes it impossible to fully analyze the extent and origins of eastern contribution among the Finns.

1.4 Summary

The population history of the Baltic Sea region is an interplay of a variety of migrations and genetic drift. Autosomal, Y-chromosomal and mitochondrial DNA results show that the majority of genetic variation in the Baltic Sea region is shared with Central European populations. In particular, the current area of northern Germany and Poland appear important in settling the northern regions. Additionally, populations on the eastern side of the Baltic Sea, most of all the Eastern Finns, show clear signs of eastern influence. The exact origin, temporal scheme and magnitude of the eastern gene flow remain unclear, although mtDNA and Y-chromosomal as well as linguistic evidence points to the Volga-Ural region. Admixture with the Saami has been strongest among the Karelians, but also among the Finns and Swedes. Slavic influence appears very slight in most populations of the Baltic Sea region.

The late settlement of the Baltic Sea region after the Ice Age, low population densities since then in combination with incidental population crises and the relative isolation have led to strong genetic drift, which has had a profound effect on the genetic variation of the populations. In the Baltic Sea region, diversity values decrease towards the northeast, being the lowest in Karelia and Eastern Finland, where all the aforementioned factors have been even more pronounced than elsewhere. Drift is likely the main cause behind the overall genetic differentiation of the populations compared to Central Europeans.

2. The population structure in Finland

The population structure in Finland was analyzed both from the Y-chromosomal (I) and genome-wide (IV) perspectives. The sample set covered large areas of Western and Eastern Finland, but lacked coverage in the middle of the country, in the south, and in the north (Figure 6, p. 36). The larger sample set in the Y-chromosomal analysis provided a more even coverage of the provinces studied than the genome-wide analysis, whose samples often originate from only a part of a province. The samples were classified according to the birth place of the grandparents, and thus the sample set does not represent the modern, admixed population, but rather the historical population structure.

2.1 Differences between Western and Eastern Finland

The differences in Y-chromosomal variation between Eastern and Western Finland were substantial, accounting for as much as 9% of the entire variation ($p < 0.001$). The most common haplogroup of the Finns, N3, was almost twice as common in Eastern Finland as in the west, whereas the reverse was true for haplogroup I1a (Figure 11). STR variation followed the same pattern of east-west differentiation, mostly due to the different haplogroup frequencies, but also due to haplotype structure within haplogroups, especially N3. Additionally, genetic diversity in Eastern Finland was clearly lower than in the west.

In the genome-wide data, the same pattern of pronounced genetic differences between Eastern and Western Finland and decreased diversity in the east was observed (Figures 8-10, Figure 12). Notably, the genetic distance between Eastern and Western Finland ($F_{ST} = 0.0032$, $p<0.001$) was higher than for many other European population pairs separated by larger geographic distances. An additional analysis of correlation between genetic and geographic distances relative to different directions showed a larger correlation in the east-west direction, implying that genetic variation is better captured by an east-west than a north-south geographical cline (Figure 13).

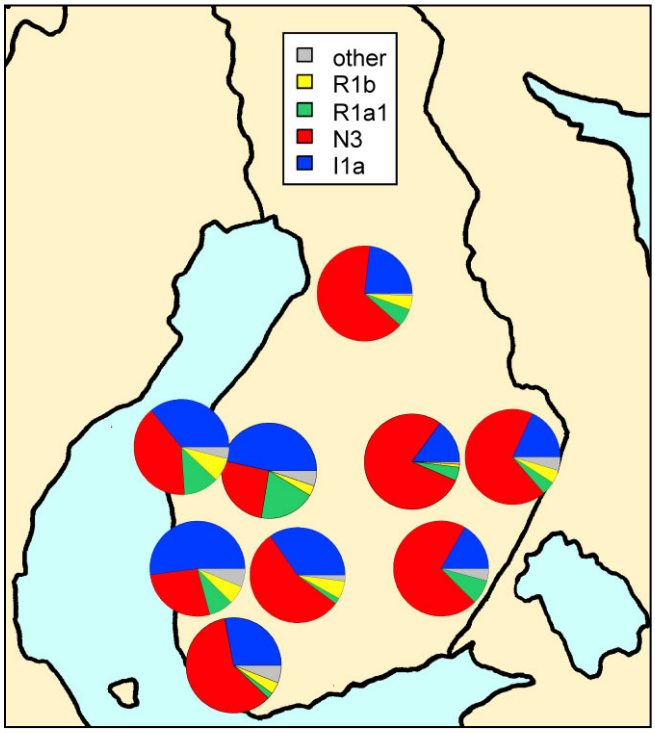

Figure 11. The frequencies of the most common Y-chromosomal haplogroups in nine Finnish provinces. See Figure 6 (p.36) for definitions of the studied regions. Based on data from I.

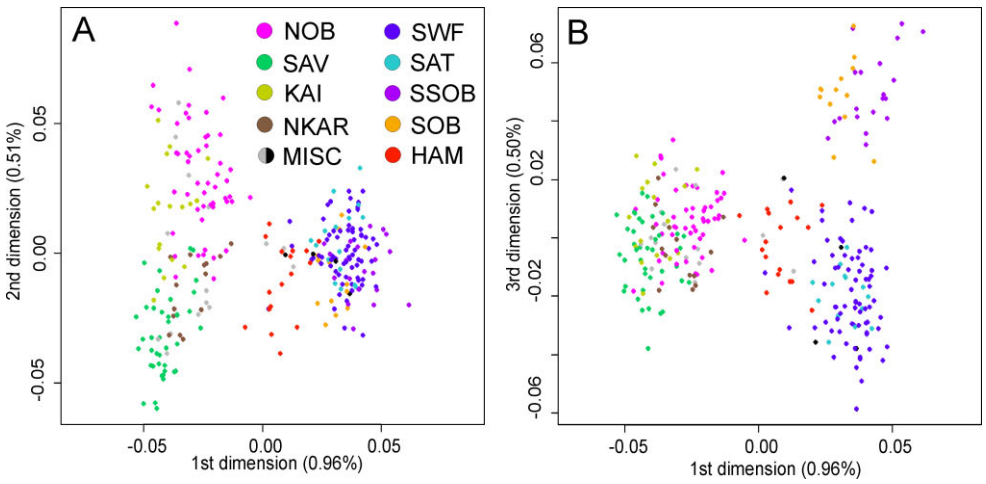

Figure 12. A multidimensional scaling plot of identity-by-state distances within Finland in the 1st and 2nd dimensions (a), and in the 1st and 3rd dimensions (b). Abbreviations: Northern Ostrobothnia (NOB), Northern Savo (SAV), Kainuu (KAI), Northern Karelia (NKAR), Miscellaneous (MISC), South-Western Finland (SWF), Satakunta (SAT), Swedish-Speaking Ostrobothnia (SSOB), Southern Ostrobothnia (SOB), Häme (HAM). Adopted from IV.

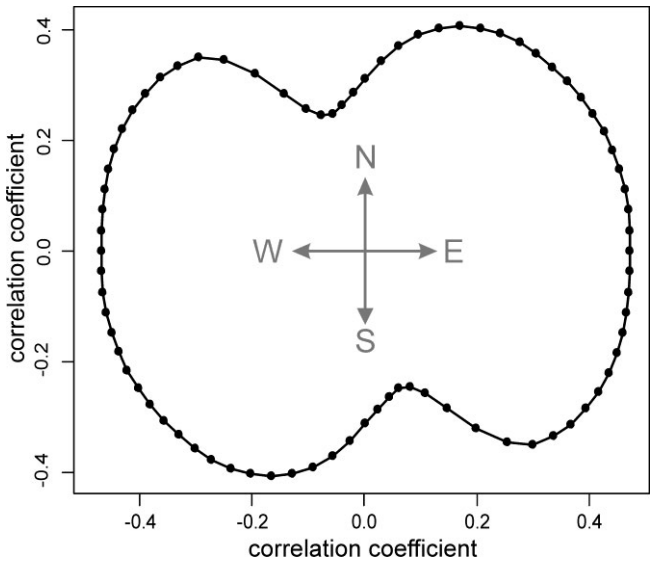

Figure 13. Correlation between genetic and geographic distances relative to different geographical directions (as indicated by the letters) based on the genome-wide data from IV. The genetic distances have been calculated from the IBS matrix of all the Finnish individuals studied. The geographic distances have been calculated relative to 40 different directions by projecting the individual coordinates to lines of 40 different angles. The correlations in the plot represent Mantel tests between the genetic and geographic distances, plotted according to the right angle of the projection lines. See Methods (p. 43) for further details. Unpublished.

2.2 Differences between provinces

The Y-chromosomal haplogroup frequencies varied not only between Eastern and Western Finland but also between individual counties. However, historical interpretation of these differences is difficult. Some of the differences may stem from differences in population history – such as admixture in South-Western Finland due to the old capital of Turku. However, many of the observed patterns may equally well be due to genetic drift or sampling error, the effects of which are difficult to exclude with the available dataset.

Even though the number of samples of the genome-wide analysis limited the analysis methods, the MDS and Structure analysis of the Finnish individuals showed interesting differences between the provinces. The eastern subpopulations show a clinal pattern from the southeast to northwest, and the western cluster is divided into south-western and Ostrobothnian clusters, consistent with the geographically discontinuous sample selection in the west. Importantly, Northern Ostrobothnians clustered invariably with the eastern samples, which is consistent with their historical background – however, the samples in this study were collected mostly from the north-eastern parts,and samples from the coastal regions could have clustered differently. In the MDS

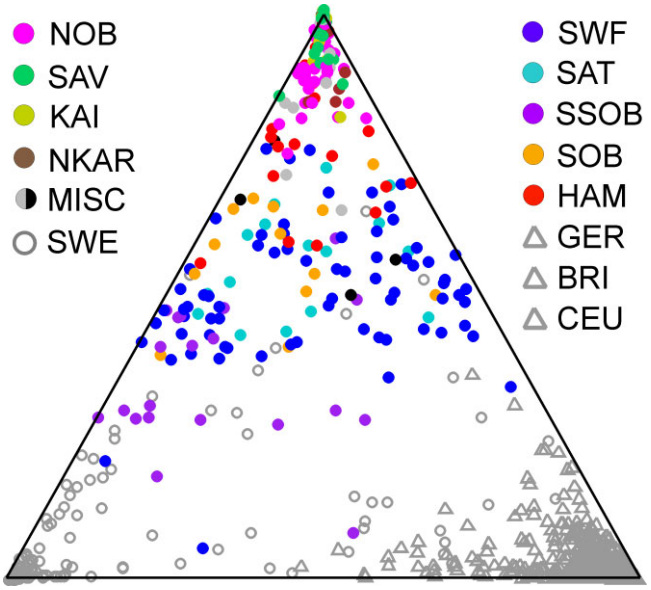

Figure 14. Proportions of the different clusters of the Structure analysis among the Finnish counties. Modified from IV. See Figure 15 for abbreviations.

plot of the Finns alone, the Finnish and Swedish-speaking Ostrobothnians were inseparable – again, however, especially the Finnish-speaking samples do not represent their entire province. Interestingly, in the MDS and Structure analyses with the Swedes included, the Swedish-speaking Ostrobothnians show a clear Swedish affinity (Figure 14, MDS analysis not shown). This is consistent with earlier results (Virtaranta-Knowles *et al.* 1991, Hannelius *et al.* 2008), and suggests that the Swedish-speaking Finns are genetically between Finland and Sweden. However, since the sample size in this study was small and only two Swedish-speakers from Southern Finland were sampled, these results should be considered preliminary.

2.3 Summary

The observations of this study are consistent with several other Y-chromosomal and autosomal studies (Workman *et al.* 1976, Kittles *et al.* 1998, Lahermo *et al.* 1999, Hedman *et al.* 2004, Hannelius *et al.* 2008, Jakkula *et al.* 2008, McEvoy *et al.* 2009) indicating that the population structure in Finland is unusually strong compared to other European populations. Minor small-scale population structure has also been observed among the Swiss and Germans and in the UK (Heath *et al.* 2008, Novembre *et al.* 2008, McEvoy *et al.* 2009). The geographical differentiation is even stronger in Northern Finland than in the regions analyzed in this study due to serial bottleneck effects (Jakkula *et al.* 2008).

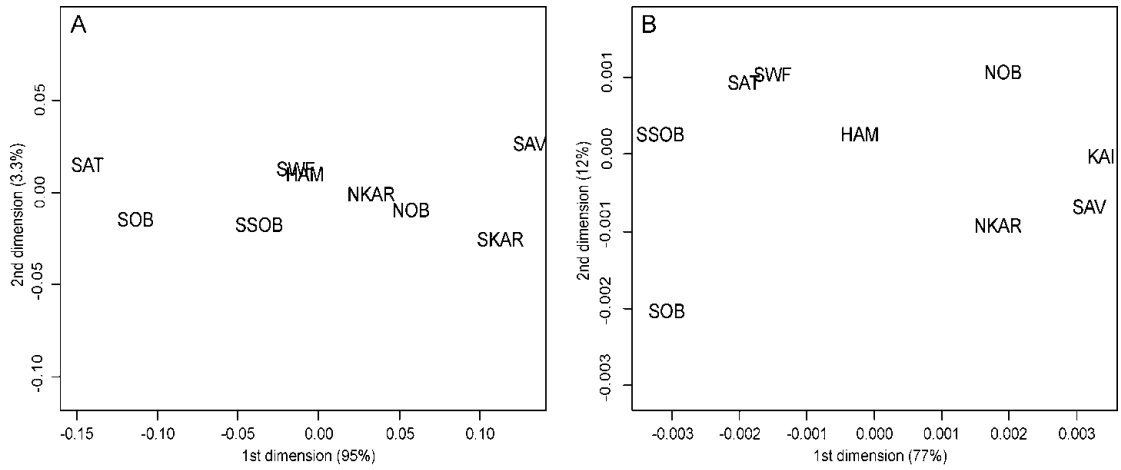

Figure 15. A multidimensional scaling plot of genetic distances between Finnish counties based on Y-chromosomal (a) and genome-wide (b) data. R_{ST} for the Y-chromosomal data from I and F_{ST} for the genome-wide data in IV were calculated in Arlequin. Multidimensional scaling was performed in R. Abbreviations: South-Western Finland (SWF), Satakunta (SAT), Häme (HAM), Southern Ostrobothnia (SOB), Swedish-Speaking Ostrobothnia (SSOB), Northern Ostrobothnia (NOB), Kainuu (KAI), Northern Savo (SAV), Northern Karelia (NKAR), Southern Karelia (SKAR). Unpublished.

These results are concordant with the known population history. While the coastal regions in Western Finland have been settled for several millennia, the later settlement, small population size and bottleneck events have led to pronounced genetic drift compared to the older, denser and more stable populations of Central Europe. Drift has been even stronger in Northern and Eastern Finland, where the late foundation of the populations after the 16th century created pronounced founder effects (Norio 2003b, Pitkänen 2007). Additionally, the low population densities have led to local isolation of small breeding units with local accumulation of genetic differences (Nevanlinna 1972).

It is also possible that different migratory waves to Finland have had quantitatively different effects in Eastern and Western Finland. As discussed previously, the east Eurasian influences observed in Finland appear stronger in Eastern Finland, while Scandinavian and possibly also Baltic influences are stronger in the west. The partly different origins of the populations in different parts of Finland would be congruent with the several archaeological cultures in the Finnish history that have affected only a part of the country, including the Corded Ware and Bronze Age cultures and coastal immigrations during the Iron and early Middle Ages (Myhre 2003, Siiriäinen 2003). However, with the currently available genetic methods it is difficult to calculate the dating and magnitude of such possible migrations or to estimate the relative extent of migrations and drift behind the genetic differences between Eastern and Western Finland.

Both the Y-chromosomal and genome-wide results support the large genetic difference between Eastern and Western Finland. However, for some counties the

results differ between the Y-chromosomal and genome-wide analyses (Figure 15). Part of the differences is probably explained by different sampling within the provinces. Furthermore, Y-chromosomal results are based on a single locus that is strongly affected by genetic drift, and thus the Y-chromosomal haplogroup and haplotype frequencies may be imperfect representations of the true differences between populations. Generally, genome-wide analysis provides a more reliable picture of the local patterns of variation.

3. The population structure in Sweden (III)

The population samples used in the mitochondrial DNA and Y-chromosomal analysis of the Swedes covered all the individuals born in Sweden during one week in 2003. This enabled an analysis of the population structure as it is today, affected by ancient as well as recent events. The only available information on the background of the samples was the birth hospital. (Hannelius *et al.* 2005)

3.1 Mitochondrial DNA and Y-chromosomal results

Particularly the Y-chromosomal but also the mtDNA variation showed signs of recent immigration to Sweden. Four Y-chromosomal haplogroups (I1b, R1b, F* and K*) were observed to have a statistically significant positive correlation with the proportion of recent immigrants, all of them known to occur in high frequencies in the countries that are among the origins of the large immigrant groups in Sweden (I, Semino *et al.* 2000, Underhill *et al.* 2001, Jobling & Tyler-Smith 2003, Rootsi *et al.* 2004). The most common Y-chromosomal haplogroup, I1a, had a strongly negative correlation with the proportion of immigrants, most probably due to replacement by foreign haplogroups. Mitochondrial DNA haplogroups showed less signs of immigration, which is expected due to the more uniform distribution of mtDNA haplogroups across Europe. Similarly, principal component analyses of Y-chromosomal and mtDNA variation showed correlation with the proportion of immigrants, with the notable differentiation of Malmö and Gothenburg in the Y-chromosomal analysis.

Several mtDNA and Y-chromosomal haplogroups showed signs of ancient minorities in Sweden and admixture with neighbouring populations. The mtDNA haplogroup most characteristic for the Sami population, U5b1b (Tambets *et al.* 2004), had a strong frequency cline, being more abundant in the north. Y-chromosomal haplogroup N3 and mitochondrial DNA H1f in many counties reveal the presence of

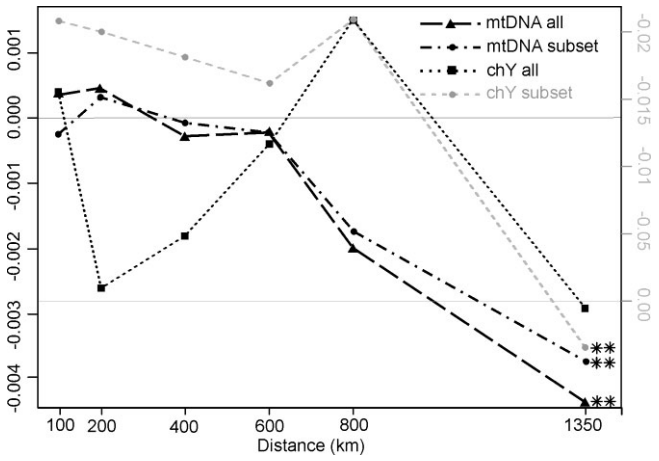

*Figure 16. Results of the spatial autocorrelation analysis in Sweden for the full mtDNA and Y-chromosomal datasets and with the counties of lowest diversity excluded. The grey line corresponds to the axis on the right. The asterisks denotes statistical significance with ** corresponding to 99% confidence level. From III.*

ancient or recent Finnish influence (I, II, Loogvali *et al.* 2004). Norwegian influence may contribute to the high frequency of Y-chromosomal R1a1 in the western parts of Sweden (Dupuy *et al.* 2006), and the abundance of R1b in the south is consistent with Danish and/or Central European admixture (Semino *et al.* 2000, Karlsson *et al.* 2006).

Mitochondrial DNA variation within Sweden showed a consistent trend of correlation between the geographic and genetic distances of the counties ($r = 0.28$, $p=0.019$). The same was true for the Y-chromosomal variation but only after the counties of the lowest diversity were removed from the analysis ($r=0.38$, $p=0.011$). However, excluding the cities with the largest immigrant populations did not have the same effect. This is also illustrated by the spatial autocorrelation analysis (Figure 16), which makes use of all the information of the birth hospitals without any additional grouping into counties. These results suggest that isolation by distance may be the main process determining the distribution of genetic variation within Sweden, and deviation from that pattern appears to be mostly due to local drift rather than immigration.

3.2 Summary

The results demonstrate how both ancient and recent events affect the genetic structure of a country. In general, the pattern of genetic variation in Sweden appears clinal, with a much less pronounced population structure than for example in Finland when compared to the Y-chromosomal analysis in I and other studies (Hannelius *et al.* 2008). However, the different sampling strategies of the Finnish and Swedish sample sets exaggerate the difference: the Finnish samples were ascertained according to the birth place of the

grandparents and filtered to exclude individuals with admixed background, whereas the sampling of the Swedes captures the contemporary population as it is.

The use of whole-genome amplified samples provided access to an unbiased collection of population samples that would have been difficult and expensive to obtain by other means. This study demonstrated the possibilities of population genetic use of existing sample collections even when only a limited amount of information on the background of the sample donors is available. Mitochondrial DNA and Y-chromosomal SNPs had sufficient power to find differences within the country, unlike for 34 autosomal SNPs genotyped from the same sample set (Hannelius *et al.* 2008), which demonstrates the power of haploid markers for detection of population structure due to the more pronounced spatial variation of haploid polymorphisms.

It is intriguing that in the MDS plot of the genome-wide data (Figure 8, p. 50), the Swedish individuals appear very scattered in the second dimension, intuitively suggesting that variation within the country might be substantial. However, additional comparisons of the IBS and MDS distances revealed that this is mostly an artefact of the MDS. Additionally, other analyses indicated that the Swedes are not particularly diverse; on the contrary, they show more signs of genetic drift than the Central European populations, although less than the Finns (Figure 9). Unfortunately, the lack of detailed information on the origins of the Swedish samples in the that study hinders any further analysis of population structure based on that data.

4. Natural selection in Northern Europe

The genome-wide analysis of loci under recent positive selection in North European populations yielded very strong signs of natural selection in 60 genomic regions containing 121 genes (Table 7). Interestingly, the genes in the selected regions show an overrepresentation of association with human diseases according to the Genetic Association Database and National Human Genome Research Institute catalogue of genome-wide association studies (χ^2 $p < 10^{-4}$). This result supports the earlier suggestions that positive selection is an important force behind human diseases (Bustamante *et al.* 2005, Nielsen *et al.* 2007, Blekhman *et al.* 2008).

With this dataset, it is impossible to determine which genomic variants within these regions associate with increased fitness, but some of the regions contain genes with particularly interesting functions. Figure 17 shows examples of median-joining networks of two genes with strong signs of selection, RAB38 and PPP2R2B. Both of the genes have interesting functions: a mutation in RAB38 causes oculocutaneous albinism in mice and it is expressed in melanocytes (Loftus *et al.* 2002), suggesting a

potential role in pigmentation also in humans. PPP2R2B has been suggested to be involved in adenovirus replication (Ben-Israel *et al.* 2008).

Coalescent simulations were used to show that both SNP density and sample size affect the performance of iHS and LRH tests. While the best power is obtained with SNP density corresponding to HapMap data, genome-wide data was shown to provide a good resource for analysis of natural selection. However, scanning for signs of positive natural selection is a challenging task. No statistical method is suitable for covering the full spectrum of natural selection (e.g. Nielsen *et al.* 2007). iHS and LRH statistics are only appropriate for detecting relatively recent selection where the haplotype has not yet reached fixation, and their performance is affected by, for instance, demographic history and the frequency of the selected haplotype (Voight *et al.* 2006, Sabeti *et al.* 2007). All these factors and true selective differences probably contribute to the differences between the populations of this study as well as between different studies. Furthermore, the precise identification of the genes and specific variants behind the selective advantage remains challenging, and will require genomic sequence data as well as functional characterization of the genes.

Despite the limitations, this analysis discovered dozens of regions with strong evidence of natural selection in Europe, and these loci will be interesting targets for follow-up studies. Genome-wide data provides an important source of large samples from multiple populations, enabling a much better coverage of the world populations that the densely genotyped datasets from, for example, the HapMap project. Selection scans create new perspectives for population genetic studies, enabling the analysis of not only neutral population processes such as migration and drift, but also the adaptive evolution of human populations in different regions of the world.

Table 7. The genomic regions showing the strongest signs of positive natural selection in the 250K dataset. ++ denotes extreme outliers and + a suggestive signal in the selection tests. The underlined genes associate with a human disease or trait (see the text for details). The positions are according to human genome Build 36. Abbreviations: integrated haplotype score (iHS), long-range haplotype (LRH), Population abbreviations as in Figure 17. Modified from V.

Region (Mb)	iHS	LRH	F_{ST}	BRI	GER	FIW	FIE	SWE	New*	Genes
1:55.7-56.0	++	++		++	++	+		+	+	-
1:160.1-160.4	++	+		+	+	++		+	+	ATF6,OLFML2B,NOS1AP
1:217.8-218.2	++	++	++	+	++					SLC30A10
2:105.4-105.8	++	+	++	++	++	++	++	++	+	FHL2,NCK2
2:152.2-152.5	++	++		++	++			+		NEB,ARL5A,CACNB4
2:178.1-178.6	++	+	+	++	++	+		+		AGPS,TTC30B,-A,PDE11A
3:10.2-10.5	+	++	++	++	++	+	+		+	IRAK2,TATDN2,C3orf42,GHRL, GHRLOS,SEC13,ATP2B2
3:59.3-59.5	++	+		+	++	+	++		+	-
3:141.5-141.8	++	++		++	++	++	++	++		CLSTN2
4:24.4-24.7	+	++		++	+	+		+	+	SOD3,CCDC149,LGI2
4:41.6-42.0	++	++		++	++	++	++	++		TMEM33,WDR21B,SLC30A9,BEND4
4:141.5-141.8	+	++		++	+	++			+	SCOC,CLGN,ELMOD2,UCP1,TBC1D9
5:80.2-80.5	++	+		+	++	+		+	+	MSH3,RASGRF2
5:115.6-115.8	++	+		++	+	+	++	+	+	COMMD10
5:134.7-135.0	++	+		+	++	++	+		+	H2AFY,C5orf20,TIFAB,NEUROG1, CXCL14
5:142.0-142.6	++	++		++	++	++	+	+		FGF1,ARHGAP26
5:145.9-146.2	++	++		++	+		+			PPP2R2B
6:46.8-47.1	++	+		++	++	+	+		+	PLA2G7,MEP1A,GPR116,-110
7:19.3-19.6	++	+		+	++			+	+	-
7:36.7-37.3	++	++		++	+	++	++	+		AOAH,ELMO1
8:139.6-139.9	++	+		+	++			+		COL22A1
10:43.6-44.0	++	+		++	+	+	+	+	+	HNRNPA3P1
10:65.1-65.4	++	+		++	++	+		+		-
10:84.5-84.8	++	+		++	+	++	+	++	+	NRG3
11:87.4-87.6	++		+		+		+	++	+	RAB38
11:116.0-116.3	++	+		++	+	+	+	++		BUD13,ZNF259,APOA5,-A4,-C3, -A1,KIAA0999
12:2.8-3.0	+	++		++	+	+	+			ITFG2,NRIP2,FOXM1,C12orf32, TULP3,TEAD4
12:66.8-67.1	++	+	+	++	+	++	++	+	+	IFNG,IL26,IL22,MDM1
14:60.8-61.3	++	++		++	++	++	+	+		TMEM30B,PRKCH,HIF1A,SNAPC1
14:68.4-68.6		++	+	++	++	+			+	ACTN1,WDR22
16:77.1-78.0	++	++		++	+	+	+			WWOX
16:78.3-78.6	++	++		++	+					-
16:81.3-81.6	++	+	+	++	++	++	++	+	+	-
16:81.7-82.0	++	+		++	+	+	+		+	CDH13
17:60.5-61.0	++	++		++	++	++	++	++	+	RGS9,AXIN2
18:7.2-7.8	++	++		++	+	++	++	+		PTPRM
20:16.1-16.3	+	++		++	+		+		+	KIF16B
22:44.9-45.2	+	++		++	+	+	+	++		PPARA,C22orf40,PKDREJ,TTC38, CN5H6.4,GTSE1,TRMU,CELSR1

* According to (Oleksyk *et al.* 2008)

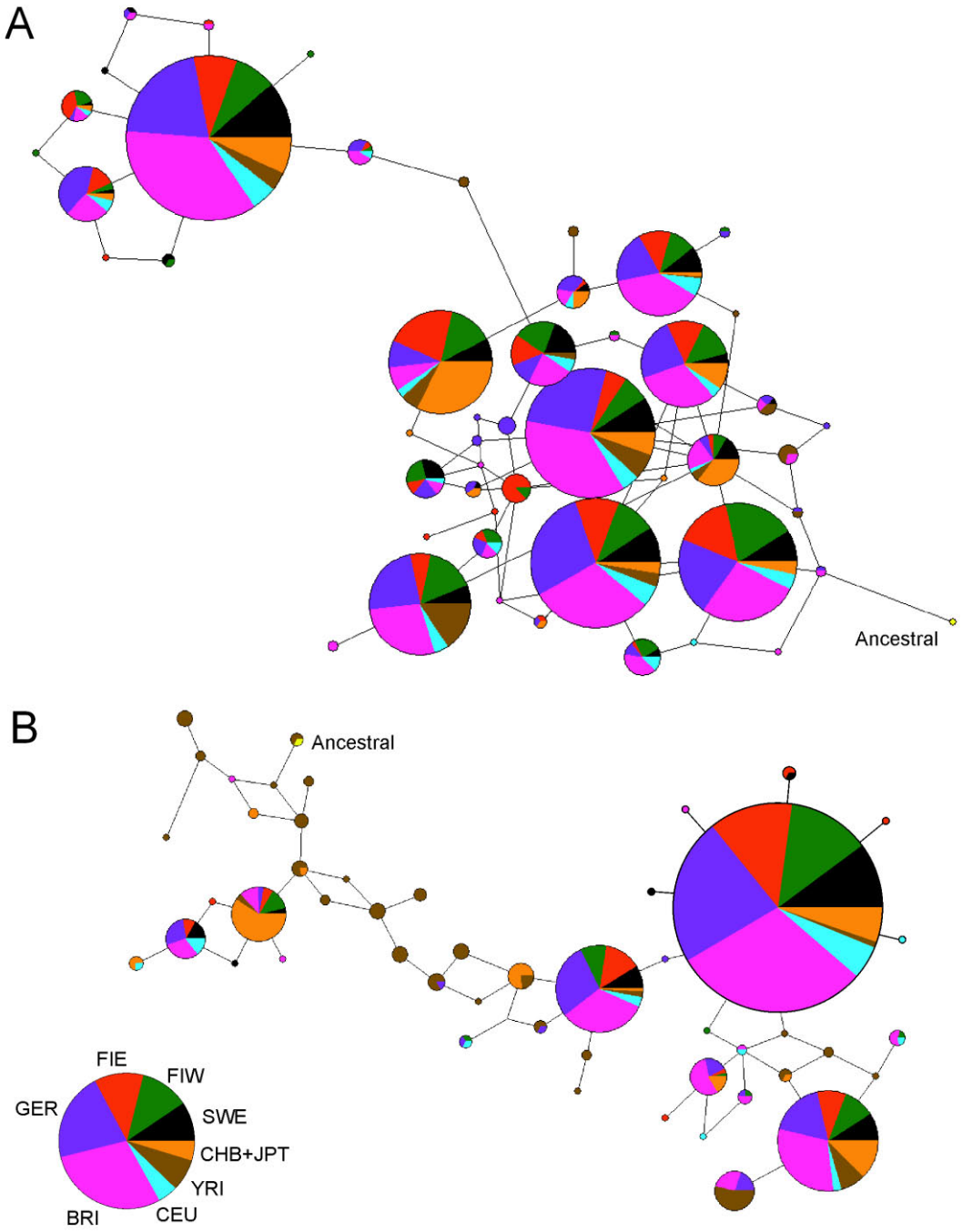

Figure 17. Median-joining networks of haplotypes in the regions Chr11:87,480,000-87,590,000 containing 15 SNPs in the RAB38 gene (a), and Chr5:145,970,000-146,030,000 containing 13 SNPs in the PPP2R2B gene (b). Nodes denote the haplotypes, with their size corresponding to the overall frequency – the legend showing the colour codes of the population frequencies also shows the relative sizes of the study samples to assist the interpretation of haplotype frequency differences. The branches connecting the haplotypes denote the SNPs differing between haplotypes. The ancestral (chimpanzee) haplotype is marked in yellow. Abbreviations: Western Finland (FIW), Eastern Finland (FIE), Sweden (SWE), Germany (GER), Great Britain (BRI), CEPH (CEU), Yoruba from Ibadan, Nigeria (YRI), Han Chinese from Beijing, China (CHB) and Japanese from Tokyo, Japan (JPT). From V.

5. Marker and sample selection in population genetic studies

5.1 Haploid *versus* autosomal markers

Only a few years ago, the cost of genotyping – together with the requirements of DNA quality and quantity – kept large-scale analysis of polymorphisms across the genome well beyond the reach of most population genetic studies. For a long time, the most efficient approach for population genetic analysis was provided by mitochondrial DNA and the Y chromosome with their greater differentiation across populations: only tens of markers are needed to find genetic differences even between closely related subpopulations, while autosomal studies with limited numbers of markers easily lack power (Rosenberg *et al.* 2005). MtDNA and Y-chromosomal markers also have the advantage of well characterized phylogeography, enabling separation of historical layers and migration routes.

However, focusing on merely two loci of the genome is not devoid of problems. Population history affects the entire genome, but in addition, each individual region is also inevitably affected by stochastic processes, i.e. pure chance. Consequently, each region has a more or less different history, which may or may not represent the history of the population without serious bias. The strong genetic drift in mtDNA and the Y chromosome due to the smaller effective population size is a double-edged sword: the advantage is the strong population structure mentioned above, but because the allele frequencies are heavily affected by drift, interpretations based on them may be unreliable. Additionally, individual loci may be affected by natural selection, in which case they are unreliable for studying neutral processes such as migration. The effect of natural selection for mtDNA and Y-chromosomal variation is still under debate (Jobling & Tyler-Smith 2003, Kivisild *et al.* 2006, Meiklejohn *et al.* 2007). However, the absence of recombination makes these loci even more vulnerable to selection, since natural selection acting on a single variant affects the variation of the entire locus. Furthermore, mtDNA and the Y chromosome contain only a tiny fraction of human genes, thus holding few answers in the quest for the genetic factors behind phenotypic differences among humans and the adaptive evolution of human populations.

Thus, the use of large numbers of autosomal markers across the genome has many advantages over Y-chromosomal and mitochondrial DNA analysis in the inference of population history. Since possible natural selection and stochastic processes affect each locus in a different manner, averaging over several loci leaves only the traces of those population historic processes that have affected the entire genome. Thus, using large numbers of markers provides a more accurate picture of relationships between populations and the extent of genetic diversity. An additional advantage is the relatively straightforward application of the observed population structure for purposes of population-based association studies. However, there is still a lack of statistical

methods taking full advantage of genome-wide data especially on the haplotype level. Haplotype blocks of the genome accumulate mutations in a hierarchical manner similar to completely non-recombining loci, and they could be analyzed in a similar manner to mtDNA and the Y chromosome. Such an approach could provide the best of both haploid and genome-wide approaches: a large number of loci, and the ability to disentangle different historical layers and migrations. A powerful approach to obtain estimates of temporal scale would be to compare the lengths of haplotypes of different origins. In the near future, genomic sequencing will provide yet another new source of data for population genetic analysis.

The number and type of loci needed for population genetic analysis depends on the study. Genome-wide coverage is essential for studies that aim to analyze the distribution of different phenomena across the genome, such as scans for natural selection. However, in studies that analyze the data averaging over the studied loci, a few thousand informative autosomal markers should be sufficient to separate individuals and populations from each other. Even though the costs of genotyping are decreasing, these approaches are still expensive to perform for very large numbers of samples. Hence, mtDNA and the Y chromosome still remain a cost-efficient way to obtain at least an initial view of the structure of a population. Additionally, at the time of writing, they are still the best available method for obtaining information on the different historical strata in populations, although the situation is bound to change soon.

5.2 Marker ascertainment bias

Another important issue in marker selection is their ascertainment, which may easily introduce serious bias in population genetic studies. If markers have been discovered and selected based on a different sample set than the final study sample, the markers are unlikely to fully capture the diversity of the studied population. Consequently, if marker discovery is done in a geographically limited sample set but the markers are used to characterize genetic variation from a variety of populations, the markers are efficient for capturing the variation in the populations closely related to the ascertainment samples but not in others. This may lead to underestimation of genetic diversity or population structure in some of the populations, of which there are also several real examples (Jobling & Tyler-Smith 2003, Romero *et al.* 2009).

Marker discovery is usually done on a much smaller set of samples than the final genotyping, and thus the allele frequency spectrum is biased towards common alleles (Eberle & Kruglyak 2000, International HapMap Consortium 2005, International HapMap Consortium *et al.* 2007). This is a problem for selection tests based on comparing the allele frequency spectrum. Thus, selection tests used for the current genome-wide data sets need to be insensitive to this bias in marker selection, such as the EHH-based statistics used in this study.

Sequence data is completely free of ascertainment bias, and the high mutation rate of microsatellite loci keeps the marker informativeness relatively uniform across continents (Romero *et al.* 2009); these are the types of data underlying diversity analyses in most of the mitochondrial DNA and Y-chromosomal analyses of this study. In contrast, SNP and structural variation analyses in particular are more sensitive to ascertainment bias (Romero *et al.* 2009). The markers in commercial genome-wide arrays are usually collected from various different sources, and thus the extent of ascertainment bias is not well known. For the small geographical regions analyzed in this study, this is probably a minor problem. However, marker selection may very well affect the exact values of, for example, F_{ST} (Clark *et al.* 2005), and rare alleles with a more limited geographical distribution might provide a better resolution of the local population structure (Novembre *et al.* 2008). However, it has been shown that within Europe, genome-wide markers of different minor allele frequencies show very similar patterns of variation, thus suggesting that ascertainment bias has little effect (Heath *et al.* 2008). In any case, possible bias is probably such that the observed patterns are true, whereas some minor phenomena may go unnoticed.

5.3 Sampling for population genetic studies

Population genetic analysis requires a sample set representative of the studied population, and the results can be safely generalized only to the region that the sampling covers. Thus, obtaining samples with well-ascertained ancestry is as crucial as correct phenotypes in genetic epidemiology. Many studies that analyze ancient population history use samples carefully selected according to familial background, such as the Finnish sample set of this study. Collecting such information often requires sample collection done especially for population genetic purposes, which adds to the cost and difficulty of yielding adequate sample collections for all interesting research questions. This is true especially for remote indigenous populations (Cavalli-Sforza 2005).

The whole approach of collecting samples ascertained according to linguistic, ethnic and national criteria has been criticized, since such sampling reflects the historical population rather than the current variation. Also, genetic clustering of populations has been suggested to arise simply from the clustered sampling units (Serre & Paabo 2004) – however, this has been refuted in a later study suggesting that small geographical barriers create true genetic discontinuities that prevail alongside clinal patterns (Rosenberg *et al.* 2005). In any case, the common requirement of non-admixed ancestry in population genetic studies excludes an increasing proportion of the world's population, and ethnically and linguistically selected samples may actually fail to reflect the true patterns of variation in diverse and admixed populations (McMahon 2004). In this study, the Swedish sample set of III was collected to reflect the contemporary population without bias by including all individuals born within a certain time span.

Similar approaches may become more common in the future through the development of biobanks and other large sample collections with little information of the donor background.

Irrespective of the sampling approach, there has been growing emphasis on the ethical, legal and societal issues of sample collection from human subjects. Population genetics rarely focuses on individuals or reveals anything of phenotypic significance, and thus the ethical problems from an individual's point of view are often minor for disease-oriented research. However, because the results concern the entire population, including individuals who did not participate, there has been debate concerning the need for societal engagement instead of mere informed consent from the participating individuals (TallBear 2007). Major attempts to catalogue human genetic variation, such as the HapMap project (International HapMap Consortium 2005), have made an effort of provide information on the research in the local language, and engage the societies and their leaders in the decision making – having learned from the controversy surrounding the Human Genome Diversity Project (Harding & Sajantila 1998, TallBear 2007). However, even that approach is controversial due to practical and theoretical reasons: some researchers claim that group consent suggests that group classifications are supported by scientific evidence. (Juengst 1998, Greely 2001, Cavalli-Sforza 2005, Race, Ethnicity, and Genetics Working Group 2005, Rotimi *et al.* 2007, TallBear 2007, Lee *et al.* 2008)

6. Population genetics and society

6.1 Population genetics in the public eye

Population history has always intrigued people due to its close relationship to questions related to ethnic, national and even personal identity. Population genetics receives a lot of public attention, and is often perceived as a novel and exciting approach that offers new perspectives on ancient questions. However, in many respects, the relationship between population geneticists and the public is radically different compared to the traditional fields studying human history, such as linguistics, history and archaeology. The questions of most interest to the public are often of local or national scale, while scientists earning their merits in international arenas can rarely engage in local issues. Furthermore, even though the humanities have their own specific methodology, popularization of those fields is usually easier than of population genetics with its complex scientific framework that is often quite alien to the general public.

6.2 Genetic ancestry testing

During the recent years, many commercial services have emerged to provide genetic analysis directly to the public. Many of these services analyze genetic risk factors of monogenic and common diseases, but another focus is genetic ancestry testing, seeking to shed light on genetic roots and the population affiliation of customers. Y-chromosomal and mitochondrial DNA analyses have dominated the field of historically focused services, but recently genome-wide analysis has also emerged as a powerful new method. These services are gaining ground particularly in North America, but many European countries, including Finland, also have active communities on the Internet engaged in digging into the history of their family via genetic information.

Scientific research forms the theoretical and practical framework for ancestry testing, and provides reference data – the data from this study have also been used by the ancestry testing community. On the other hand, the data sets acquired through commercial testing can provide a novel resource for professional scientists, thus avoiding the difficult and costly sampling and genotyping. This is the approach of the Genographic project that has produced not only ancestry information for customers but also data for professional scientists, subsequently published in peer-reviewed articles (Behar *et al.* 2007).

The public interest towards population history is generally a positive thing. However, ancestry testing has its problems, some of which are little discussed in the public domain (TallBear 2007, Bandelt *et al.* 2008). Many consider ancestry testing to be innocent and harmless compared to genetic testing related to disease risks, but actually very few things agitate people as much as questions of national, ethnic and personal identity. Little research has been done on the subject, but genetic ancestry information that is in conflict with the earlier personal view may cause even more distress than knowledge of genetic predisposition to a complex disease (Bolnick *et al.* 2007). Many people have the tendency to regard genetic information as more profound and true than other kinds of historical information – but in the case of personal identity, this view is obviously flawed (Bolnick *et al.* 2007, TallBear 2007). Additionally, there are further ethical concerns related to confidentiality and consent issues (Bandelt *et al.* 2008). Thus, it is essential that people engaged in ancestry testing have the means to interpret their results in the correct context. One of the ways to achieve this is to increase public understanding of population genetics.

CONCLUSIONS AND FUTURE PROSPECTS

In the course of this study, population genetic analysis has expanded from mitochondrial and Y-chromosomal analysis to cover the whole genome. The early enthusiasm of gene-language correlations among population geneticists has been replaced by a more realistic view of the complexity of human history, and an understanding of the difficulty of assessing the interplay between the biological, cultural and linguistic aspects. MtDNA and the Y chromosome are still useful for analyses of different historical strata, and are an efficient way to perform initial analyses of population structure. However, the genome-wide approach has enabled a whole new level of accuracy when interpretations do not depend on only few loci. Similarly, the analysis of natural selection has advanced from candidate gene studies to scanning the entire genome, although full characterization of novel findings still requires detailed gene-based functional analysis.

This study, alongside many others, has showed that patterns of variation observed in genome-wide data are quite congruent with earlier studies – for example, the extensive genetic drift and eastern influence in Finland were first observed decades ago. This study has confirmed that even the easternmost populations of the Baltic Sea region have their genetic roots mainly in Central Europe, with a smaller degree of eastern influence – possibly from the Volga-Ural region. The small population size, in addition to bottleneck and founder events, has shaped the genetic variation of North European populations extensively and had a profound impact on the population structure, especially within Finland and also, to a lesser extent, in Sweden. As a result of these processes and a variety of regional and local migrations, each population has its unique gene pool.

Genome-wide analysis has opened whole new fields of research now available for study, and the full potential of the extensive datasets has not yet been utilized. Analyzing population differentiation and diversity levels is a very straightforward approach which does not actually require coverage of the entire genome – a few thousand markers are adequate for that purpose. However, the future will probably show development of more advanced haplotype-based methods to study population history in unparalleled detail. Accumulating genome-wide data will provide a source of reference data from an expanding number of human populations. Additionally, genomic sequencing is leading the field towards population genomics, offering an extremely valuable data source free from ascertainment bias, particularly for the analysis of natural selection across the human genome.

The recent understanding of the mostly clinal nature of genetic variation in Europe also brings new perspectives to mtDNA and Y-chromosomal analysis, where the

aim has often been to distinguish and date migration waves with great precision. In reality, much of the observed gene flow is likely a result of slow diffusion. On the other hand, genome-wide analysis currently lacks a methodology for distinguishing different historical layers of migrations. Thus, the importance of different historical processes and migratory patterns shaping the genetic variation of modern human populations still remains unknown.

The motives for the analysis of the genetic variation of human populations are twofold, both reflected in this study. Comprehension of the sources of human variation, especially evolutionary adaptations, is essential for understanding the genetic causes underlying human diseases and other phenotypic variation. Another research focus of population genetic research is aimed at understanding human history in collaboration with other historical sciences: ideally, an understanding of population genetic processes and human history is gained through a synthesis of the molecular history of multiple genetic loci as well as historical information. Ultimately, contemporary diversity is formed through a complex history of interactions between humans and their physical, biological, social and cultural environment, and the future will hopefully hold a more complete understanding of different aspects of human population history.

ACKNOWLEDGEMENTS

This work has been performed at the Finnish Genome Center, now a part of the Institute for Molecular Medicine Finland, and I would like to thank its current and previous leaders for providing the excellent research facilities without which this work would not have been possible. I would also like to thank the Genome Informatics Unit for the bioinformatics services that were equally important for this work.

The funding for this study has been provided by the Emil Aaltonen Foundation, the Finnish Cultural Foundation, the Research Foundation of the University of Helsinki, the Sigrid Juselius Foundation, the Academy of Finland, the Swedish Research Council, the Betty Väänänen Foundation, and the Research Foundation of the Finno-Ugric Cultural Research, and I thank them for their support.

The original reviewers of this thesis, Professors Antti Sajantila and Kari Majamaa, are gratefully acknowledged for their valuable comments and suggestions, and I am grateful to Professor Jaakko Ignatius for agreeing to be an additional reviewer at such short notice. I also thank Maaria and Damon Tringham for carefully revising the English language of my thesis.

I have had the privilege of being guided by many wonderful senior scientists, and my most heartfelt thanks goes to Päivi Lahermo, Juha Kere, Kirsi Huoponen, and also to Marja-Liisa Savontaus, who has taken part in this project despite not being my official supervisor. You all have given me your time and support whenever I have needed it, still allowing me to work independently, and from you I've learned to have my feet on the ground, or head in the clouds, or both at the same time. It has been a pleasure to work with you all.

Of all my many co-authors and collaborators, I owe most to Elina Salmela. Without your scientific input this would be a poorer study, without all the things I have learned from you I would be a poorer scientist, and without your friendship the whole journey would have been a lot less interesting. Being able to share the ups and downs of research with you has been hugely important. I'm also grateful to my other co-authors Satu Koivumäki, Virpi Laitinen and Ulf Hannelius for kindly sharing their data, samples and expertise in our joint projects.

This study would not have been possible without the excellent sample and data sets, of which I am indebted to Pertti Sistonen, Peter Andersen, Ulrika von Döbeln, Tuula

Koski, Stefan Schreiber, and many others. I would also like to thank Ingegerd Fransson and people at BEA for genotyping, Anu Puomila and Riitta Lehtinen for help with the Finnish DNA samples, and Ella Granö for genotyping and sequencing lots of data.

I owe many thanks to the people at the Finnish Genome Center for all the help throughout the years – Sirkku, Anu and others at the lab; Timo, Tomi and the other guys from GIU for all the computer-related help; Anne, Kyösti, Kari and others for data handling; Jouko, Riitta and Susanna for keeping all the practical things running, and everyone else for all the help and companionship. I'd also like to thank all the people at Juha's research group in Helsinki and Stockholm as well as the people in Päivi Saavalainen's group for inspiring discussions and cheerful company.

All my friends from childhood, school, and university I thank for the fun, talk, travel and partying. Luckily you guys keep reminding me that there's life outside Biomedicum. I owe most of all to my parents and little bro – without learning from you early in life how to be curious about the world I would not be a scientist. And, finally, I thank Robert for not allowing science to be the sole content of my life.

REFERENCES

Achilli A, Rengo C, Battaglia V *et al.* (2005) Saami and Berbers--an unexpected mitochondrial DNA link. *Am.J.Hum.Genet.* 76:883-886.

Achilli A, Rengo C, Magri C *et al.* (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am.J.Hum.Genet.* 75:910-918.

Aikio A & Aikio A (2001) Heimovaelluksista jatkuvuuteen - suomalaisen väestöhistorian tutkimuksen pirstoutuminen (From tribal migrationsto continuity - fragmentation of the study of Finnish population history). *Muinaistutkija* 4:2-20.

Akey JM, Zhang G, Zhang K, Jin L & Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805-1814.

Anisimova M & Liberles DA (2007) The quest for natural selection in the age of comparative genomics. *Heredity* 99:567-579.

Aulchenko Y, Struchalin M, Ripke S & Johnson T (2008) genABEL: genome-wide SNP association analysis. Version 1.3-5.

Balanovsky O, Rootsi S, Pshenichnov A *et al.* (2008) Two sources of the Russian patrilineal heritage in their Eurasian context. *Am.J.Hum.Genet.* 82:236-250.

Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat.Rev.Genet.* 7:781-791.

Bandelt HJ, Forster P & Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol.Biol.Evol.* 16:37-48.

Bandelt HJ, Forster P, Sykes BC & Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753.

Bandelt HJ, Yao YG, Richards MB & Salas A (2008) The brave new era of human genetic testing. *Bioessays* 30:1246-1251.

Barbujani G, Magagni A, Minch E & Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. *Proc.Natl.Acad.Sci.U.S.A.* 94:4516-4519.

Barreiro LB, Laval G, Quach H, Patin E & Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nat.Genet.* 40:340-345.

Bauchet M, McEvoy B, Pearson LN *et al.* (2007) Measuring European population stratification with microarray genotype data. *Am.J.Hum.Genet.* 80:948-956.

Beaumont MA & Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol.Ecol.* 13:969-980.

Behar DM, Rosset S, Blue-Smith J *et al.* (2007) The Genographic Project public participation mitochondrial DNA database. *PLoS Genet.* 3:e104.

Ben-Israel H, Sharf R, Rechavi G & Kleinberger T (2008) Adenovirus E4orf4 protein downregulates MYC expression through interaction with the PP2A-B55 subunit. *J.Virol.* 82:9381-9388.

Bermisheva M, Tambets K, Villems R & Khusnutdinova E (2002) Diversity of mitochondrial DNA haplotypes in ethnic populations of the Volga-Ural region of Russia. *Mol.Biol.(Mosk)* 36:990-1001.

Bersaglieri T, Sabeti PC, Patterson N *et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am.J.Hum.Genet.* 74:1111-1120.

Bertorelle G & Barbujani G (1995) Analysis of DNA diversity by spatial autocorrelation. *Genetics* 140:811-819.

Betti L, Balloux F, Amos W, Hanihara T & Manica A (2009) Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc.Biol.Sci.* 276:809-814.

Biswas S & Akey JM (2006) Genomic insights into positive selection. *Trends Genet.* 22:437-446.

Blekhman R, Man O, Herrmann L *et al.* (2008) Natural selection on genes that underlie human disease susceptibility. *Curr.Biol.* 18:883-889.

Bodmer W & Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat.Genet.* 40:695-701.

Bolnick DA, Fullwiley D, Duster T *et al.* (2007) Genetics. The science and business of genetic ancestry testing. *Science* 318:399-400.

Browning SR & Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am.J.Hum.Genet.* 81:1084-1097.

Bustamante CD, Fledel-Alon A, Williamson S *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437:1153-1157.

Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *J.Forensic Sci.* 51:253-265.

Cann RL (2001) Genetic clues to dispersal in human populations: retracing the past from the present. *Science* 291:1742-1748.

Cann RL, Stoneking M & Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325:31-36.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ & Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15:1553-1565.

Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present and future. *Nat.Rev.Genet.* 6:333-340.

Cavalli-Sforza LL (1998) The DNA revolution in population genetics. *Trends Genet.* 14:60-65.

Cavalli-Sforza LL & Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nat.Genet.* 33 Suppl:266-275.

Cavalli-Sforza LL, Piazza A & Menozzi P (1994) History and Geography of Human Genes. Princeton University Press, Princeton.

Chaix R, Cao C & Donnelly P (2008) Is mate choice in humans MHC-dependent? *PLoS Genet.* 4:e1000184.

Chessel D, Dufour AB & Thioulouse J (2004) The ade4 package-I- One-table methods. *R News* 4:5-10.

Chikhi L, Nichols RA, Barbujani G & Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. *Proc.Natl.Acad.Sci.U.S.A.* 99:11008-11013.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH & Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496-1502.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA & Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat.Genet.* 38:1251-1260.

Currat M & Excoffier L (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* 2:e421.

Derenko M, Malyarchuk B, Denisova G, Wozniak M, Grzybowski T, Dambueva I & Zakharov I (2007) Y-chromosome haplogroup N dispersals from south Siberia to Europe. *J.Hum.Genet.* 52:763-770.

Di Giacomo F, Luca F, Popa LO *et al.* (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum.Genet.* 115:357-371.

Diamond J & Bellwood P (2003) Farmers and their languages: the first expansions. *Science* 300:597-603.

Dupuy BM, Stenersen M, Egeland T & Olaisen B (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum.Mutat.* 23:117-124.

Dupuy BM, Stenersen M, Lu TT & Olaisen B (2006) Geographical heterogeneity of Y-chromosomal lineages in Norway. *Forensic Sci.Int.* 164:10-19.

Eberle MA & Kruglyak L (2000) An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet.Epidemiol.* 19 Suppl 1:S29-35.

Einarsdottir E, Egerbladh I, Beckman L, Holmberg D & Escher SA (2007) The genetic population structure of northern Sweden and its implications for mapping genetic diseases. *Hereditas* 144:171-180.

Excoffier L, Smouse PE & Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479-491.

Finnila S, Lehtonen MS & Majamaa K (2001) Phylogenetic network for European mtDNA. *Am.J.Hum.Genet.* 68:1475-1484.

Forster P, Harding R, Torroni A & Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am.J.Hum.Genet.* 59:935-945.

Freedman ML, Reich D, Penney KL *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nat.Genet.* 36:388-393.

Garrigan D & Hammer MF (2006) Reconstructing human origins in the genomic era. *Nat.Rev.Genet.* 7:669-680.

Gonzalez-Neira A, Ke X, Lao O *et al.* (2006) The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* 16:323-330.

Greely HT (2001) Informed consent and other ethical issues in human population genetics. *Annu.Rev.Genet.* 35:785-800.

Green RE, Krause J, Ptak SE *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330-336.

Guglielmino CR, Piazza A, Menozzi P & Cavalli-Sforza LL (1990) Uralic genes in Europe. *Am.J.Phys.Anthropol.* 83:57-68.

Haak W, Forster P, Bramanti B *et al.* (2005) Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* 310:1016-1018.

Häkkinen J (2009) Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Journal de la Société Finno-Ougrienne* 92.

Hamblin MT, Thompson EE & Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am.J.Hum.Genet.* 70:369-383.

Hammer MF, Mendez FL, Cox MP, Woerner AE & Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet.* 4:e1000202.

Hannelius U, Lindgren CM, Melen E, Malmberg A, von Dobeln U & Kere J (2005) Phenylketonuria screening registry as a resource for population genetic studies. *J.Med.Genet.* 42:e60.

Hannelius U, Salmela E, Lappalainen T *et al.* (2008) Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. *BMC Genet.* 9:54.

Harding A (1997) Reformation in Barbarian Europe, 1300-600 BC. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Oxford University Press, Oxford. pp. 304-336.

Harding RM & Sajantila A (1998) Human genome diversity--a project? *Nat.Genet.* 18:307-308.

Hartl DL & Clark AG (2007) Principles of Population Genetics. Sinauer and Associates, Sunderland, MA.

Heath SC, Gut IG, Brennan P *et al.* (2008) Investigation of the fine structure of European populations with applications to disease association studies. *Eur.J.Hum.Genet.* 16:1413-1429.

Hedman M, Brandstatter A, Pimenoff V, Sistonen P, Palo JU, Parson W & Sajantila A (2007) Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic Sci.Int.* 172:171-178.

Hedman M, Pimenoff V, Lukka M, Sistonen P & Sajantila A (2004) Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. *Forensic Sci.Int.* 142:37-43.

Helgason A, Hickey E, Goodacre S, Bosnes V, Stefansson K, Ward R & Sykes B (2001) mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am.J.Hum.Genet.* 68:723-737.

Hinds DA, Stuve LL, Nilsen GB *et al.* (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072-1079.

Hofer T, Ray N, Wegmann D & Excoffier L (2009) Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Ann.Hum.Genet.* 73:95-108.

Holmlund G, Nilsson H, Karlsson A & Lindblom B (2006) Y-chromosome STR haplotypes in Sweden. *Forensic Sci.Int.* 160:66-79.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM & Herrnstadt C (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am.J.Hum.Genet.* 72:659-670.

Hurles ME, Dermitzakis ET & Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends Genet.* 24:238-245.

Ingman M & Gyllensten U (2007) A recent genetic link between Sami and the Volga-Ural region of Russia. *Eur.J.Hum.Genet.* 15:115-120.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299-1320.

International HapMap Consortium, Frazer KA, Ballinger DG *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851-861.

Jakkula E, Rehnstrom K, Varilo T *et al.* (2008) The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population. *Am.J.Hum.Genet.* 83:787-794.

Jakobsson M, Scholz SW, Scheet P *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998-1003.
Jobling MA & Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat.Rev.Genet.* 4:598-612.
Jobling MA, Hurles M & Tyler-Smith C (2004) Human Evolutionary Genetics. Garland Science, New York.
Jorde LB, Bamshad M & Rogers AR (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays* 20:126-136.
Jorde LB, Watkins WS & Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Hum.Mol.Genet.* 10:2199-2207.
Juengst ET (1998) Group identity and human diversity: keeping biology straight from culture. *Am.J.Hum.Genet.* 63:673-677.
Kaessmann H, Zollner S, Gustafsson AC *et al.* (2002) Extensive linkage disequilibrium in small human populations in Eurasia. *Am.J.Hum.Genet.* 70:673-685.
Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL & Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18:830-838.
Karlsson AO, Wallerstrom T, Gotherstrom A & Holmlund G (2006) Y-chromosome diversity in Sweden - a long-time perspective. *Eur.J.Hum.Genet.* 14:963-970.
Kayser M, Lao O, Anslinger K *et al.* (2005) Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Hum.Genet.* 117:428-443.
Kelley JL, Madeoy J, Calhoun JC, Swanson W & Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980-989.
Kim Y & Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765-777.
Kimura R, Fujimoto A, Tokunaga K & Ohashi J (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* 2:e286.
Kittles RA, Bergen AW, Urbanek M, Virkkunen M, Linnoila M, Goldman D & Long JC (1999) Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am.J.Phys.Anthropol.* 108:381-399.
Kittles RA, Perola M, Peltonen L *et al.* (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am.J.Hum.Genet.* 62:1171-1179.
Kivisild T, Shen P, Wall DP *et al.* (2006) The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172:373-387.
Klopfstein S, Currat M & Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Mol.Biol.Evol.* 23:482-490.
Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE & Schreiber S (2006) PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* 9:55-61.
Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L & Savontaus ML (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am.J.Hum.Genet.* 58:1309-1322.

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

Laitinen V, Lahermo P, Sistonen P & Savontaus ML (2002) Y-chromosomal diversity suggests that Baltic males share common Finno-Ugric-speaking forefathers. *Hum.Hered.* 53:68-78.

Lange EM, Sun J, Lange LA *et al.* (2008) Family-based samples can play an important role in genetic association studies. *Cancer Epidemiol.Biomarkers Prev.* 17:2208-2214.

Lao O, Lu TT, Nothnagel M *et al.* (2008) Correlation between genetic and geographic structure in Europe. *Curr.Biol.* 18:1241-1248.

Lee SS, Mountain J, Koenig B *et al.* (2008) The ethics of characterizing difference: guiding principles on using racial categories in human genetics. *Genome Biol.* 9:404.

Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49:49-67.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100-1104.

Lindkvist T (2003) Kings and Provinces in Sweden. In Helle K & Jansson T: Combredge History of Scandinavia. Combridge University Press, Cambridge.

Lindqvist H (2006) A History of Sweden. Nordstedts Förlag, Stockholm.

Loftus SK, Larson DM, Baxter LL *et al.* (2002) Mutation of melanosome protein RAB38 in chocolate mice. *Proc.Natl.Acad.Sci.U.S.A.* 99:4471-4476.

Loogvali EL, Roostalu U, Malyarchuk BA *et al.* (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol.Biol.Evol.* 21:2012-2021.

Macaulay V, Richards M, Hickey E *et al.* (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am.J.Hum.Genet.* 64:232-249.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Wozniak M & Miscicka-Sliwka D (2002) Mitochondrial DNA variability in Poles and Russians. *Ann.Hum.Genet.* 66:261-283.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27:209-220.

Marchini J, Cardon LR, Phillips MS & Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nat.Genet.* 36:512-517.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annu.Rev.Genomics Hum.Genet.* 9:387-402.

Matise TC, Chen F, Chen W *et al.* (2007) A second-generation combined linkage physical map of the human genome. *Genome Res.* 17:1783-1786.

McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat.Genet.* 40:1166-1174.

McEvoy BP, Montgomery GW, McRae AF *et al.* (2009) Geographical structure and differential natural selection amongst North European populations. *Genome Res.* .

McIlroy D, Brownrigg R & Minka TP (2005) mapproj: Map Projections. R package version 1.1-7.1.

McMahon R (2004) Genes and languages. *Community Genet.* 7:2-13.

McVean G & Spencer CC (2006) Scanning the human genome for signals of selection. *Curr.Opin.Genet.Dev.* 16:624-629.
Meiklejohn CD, Montooth KL & Rand DM (2007) Positive and negative selection on the mitochondrial genome. *Trends Genet.* 23:259-263.
Meinila M, Finnila S & Majamaa K (2001) Evidence for mtDNA admixture between the Finns and the Saami. *Hum.Hered.* 52:160-170.
Mellars P (2006) Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* 313:796-800.
Mellars P (2004) Neanderthals and the modern human colonization of Europe. *Nature* 432:461-465.
Mellars P (1997) The Upper Palaeolithic Revolution. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Oxford University Press, Oxford. pp. 42-79.
Mirabal S, Regueiro M, Cadenas AM *et al.* (2009) Y-Chromosome distribution within the geo-linguistic landscape of northwestern Russia. *Eur.J.Hum.Genet.* .
Mithen SJ (1997) The Mesolithic Age. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Ofxord University Press, Oxford. pp. 79-136.
Myhre B (2003) The Iron Age. In Helle K: Cambridge History of Scandinavia. Cambridge University Press, Cambridge. pp. 60-94.
Myles S, Tang K, Somel M, Green RE, Kelso J & Stoneking M (2008) Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann.Hum.Genet.* 72:99-110.
Nei M (2005) Selectionism and neutralism in molecular evolution. *Mol.Biol.Evol.* 22:2318-2342.
Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.
Nevanlinna HR (1972) The Finnish population structure. A genetic and genealogical study. *Hereditas* 71:195-236.
Nielsen R (2005) Molecular signatures of natural selection. *Annu.Rev.Genet.* 39:197-218.
Nielsen R, Hellmann I, Hubisz M, Bustamante C & Clark AG (2007) Recent and ongoing selection in the human genome. *Nat.Rev.Genet.* 8:857-868.
Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG & Bustamante C (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566-1575.
NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. *Science* 258:67-86.
Noonan JP, Coop G, Kudaravalli S *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113-1118.
Norio R (2003a) Finnish Disease Heritage I: characteristics, causes, background. *Hum.Genet.* 112:441-456.
Norio R (2003b) Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Hum.Genet.* 112:457-469.
Norio R (2003c) The Finnish Disease Heritage III: the individual diseases. *Hum.Genet.* 112:470-526.
Novembre J, Johnson T, Bryc K *et al.* (2008) Genes mirror geography within Europe. *Nature* 456:274.
Novembre J & Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat.Genet.* 40:646-649.

Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ & Smith MW (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* 3:e1712.

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T & Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat.Genet.* 29:20-21.

O'Reilly PF, Birney E & Balding DJ (2008) Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.* 18:1304-1313.

Paabo S, Poinar H, Serre D *et al.* (2004) Genetic analyses from ancient DNA. *Annu.Rev.Genet.* 38:645-679.

Pakendorf B & Stoneking M (2005) Mitochondrial DNA and human evolution. *Annu.Rev.Genomics Hum.Genet.* 6:165-183.

Palo JU, Pirttimaa M, Bengs A *et al.* (2008) The effect of number of loci on geographical structuring and forensic applicability of Y-STR data in Finland. *Int.J.Legal Med.* 122:449-456.

Parsons TJ, Muniec DS, Sullivan K *et al.* (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat.Genet.* 15:363-368.

Peregrine PN (2001) Introduction. In Peregrine PN & Ember M: Encyclopedia of Prehistory, Volume 4: Europe. Kuewer Academic / Plenum Publishers, New York. pp. xvii-xxi.

Pereira L, Richards M, Goios A *et al.* (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* 15:19-24.

Pimenoff VN, Comas D, Palo JU, Vershubsky G, Kozlov A & Sajantila A (2008) Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. *Eur.J.Hum.Genet.* 16:1254-1264.

Pitkänen K (2007) Suomen väestön historialliset kehityslinjat. In Koskinen S, Martelin T, Notkola IL *et al*: Suomen väestö. Gaudeamus Helsinki University Press, Tampere. pp. 41-76.

Pliss L, Tambets K, Loogvali EL *et al.* (2006) Mitochondrial DNA portrait of Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. *Ann.Hum.Genet.* 70:439-458.

Polzin T & Daneschmand SV (2003) On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters* 31:12-12-20.

Pontikos D (2008) How Y-STR variance accumulates: a comment on Zhivotovsky, Underhill and Feldman (2006). In: Dienekes' Antropology Blog, http://dienekes.blogspot.com/2008/07/how-y-str-variance-accumulates-comment.html.

Price AL, Butler J, Patterson N *et al.* (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* 4:e236.

Pritchard JK, Stephens M & Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.

R Development Core Team (2008) R: A language and environment for statistical computing. Version 2.6.2.

Race, Ethnicity, and Genetics Working Group (2005) The use of racial, ethnic, and ancestral categories in human genetics research. *Am.J.Hum.Genet.* 77:519-532.

Raitio M, Lindroos K, Laukkanen M, Pastinen T, Sistonen P, Sajantila A & Syvanen AC (2001) Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Res.* 11:471-482.

Reich DE, Schaffner SF, Daly MJ *et al.* (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat.Genet.* 32:135-142.

Relethford JH (2008) Genetic evidence and the modern human origins debate. *Heredity* 100:555-563.

Renfrew C (2001) From molecular genetics to archaeogenetics. *Proc.Natl.Acad.Sci.U.S.A.* 98:4830-4832.

Richards M, Macaulay V, Torroni A & Bandelt HJ (2002) In search of geographical patterns in European mitochondrial DNA. *Am.J.Hum.Genet.* 71:1168-1174.

Roewer L, Croucher PJ, Willuweit S *et al.* (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum.Genet.* 116:279-291.

Romero IG, Manica A, Goudet J, Handley LL & Balloux F (2009) How accurate is the current picture of human genetic variation? *Heredity* 102:120-126.

Rootsi S, Magri C, Kivisild T *et al.* (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. *Am.J.Hum.Genet.* 75:128-137.

Rootsi S, Zhivotovsky LA, Baldovic M *et al.* (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur.J.Hum.Genet.* 15:204-211.

Roseman CC & Weaver TD (2007) Molecules versus morphology? Not for the human cranium. *Bioessays* 29:1185-1188.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK & Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1:e70.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA & Feldman MW (2002) Genetic structure of human populations. *Science* 298:2381-2385.

Rosser ZH, Zerjal T, Hurles ME *et al.* (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am.J.Hum.Genet.* 67:1526-1543.

Rotimi C, Leppert M, Matsuda I *et al.* (2007) Community engagement and informed consent in the International HapMap project. *Community Genet.* 10:186-198.

Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832-837.

Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science* 312:1614-1620.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913-918.

Sajantila A, Lahermo P, Anttinen T *et al.* (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res.* 5:42-52.

Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C & Paabo S (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc.Natl.Acad.Sci.U.S.A.* 93:12035-12039.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ & Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576-1583.

Schneider S, Roessli D & Excoffier L (2000) Arlequin: A software for population genetics data analysis. Version 2.000.

Seldin MF, Shigeta R, Villoslada P *et al.* (2006) European population substructure: clustering of northern and southern populations. *PLoS Genet.* 2:e143.

Semino O, Passarino G, Oefner PJ *et al.* (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* 290:1155-1159.

Serre D & Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* 14:1679-1685.

Service S, DeYoung J, Karayiorgou M *et al.* (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat.Genet.* 38:556-560.

Shendure J & Ji H (2008) Next-generation DNA sequencing. *Nat.Biotechnol.* 26:1135-1145.

Sherratt A (1997a) The Emergence of Élites: Earlier Bronze Age Europe, 2500-1200 BC. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Oxford University Press, Oxford. pp. 244-277.

Sherratt A (1997b) The Transformation of Early Agrarian Europe: The Later Neolithic and Copper Ages, 4500-2500 BC. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Oxford University Press, Oxford. pp. 167-202.

Siiriäinen A (2003) The Stone and Bronze Ages. In Helle K: The Cambridge History of Scandinavia. Cambridge University Press, Cambridge.

Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat.Rev.Genet.* 9:477-485.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457-462.

Spencer CC & Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 20:3673-3675.

Statistiska Centralbyrån (2008) In: Statistiska Centralbyrån / Statistics Sweden, http://www.scb.se/.

Steegmann AT (2006) Physiological anthropology: past and future. *J.Physiol.Anthropol.* 25:67-73.

Stoneking M (1997) The human genome project and molecular anthropology. *Genome Res.* 7:87-91.

Stranger BE, Nica AC, Forrest MS *et al.* (2007) Population genomics of human gene expression. *Nat.Genet.* 39:1217-1224.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437-460.

TallBear K (2007) Narratives of race and indigeneity in the Genographic Project. *J.Law Med.Ethics* 35:412-424.

Tambets K, Rootsi S, Kivisild T *et al.* (2004) The western and eastern roots of the Saami--the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *Am.J.Hum.Genet.* 74:661-682.

Tang K, Thornton KR & Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biol.* 5:e171.

Teshima KM, Coop G & Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16:702-712.

Tian C, Gregersen PK & Seldin MF (2008a) Accounting for ancestry: population substructure and genome-wide association studies. *Hum.Mol.Genet.* 17:R143-50.

Tian C, Plenge RM, Ransom M *et al.* (2008b) Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* 4:e4.

Tishkoff SA & Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu.Rev.Genomics Hum.Genet.* 4:293-340.

Torroni A, Achilli A, Macaulay V, Richards M & Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22:339-345.

Torroni A, Bandelt HJ, D'Urbano L *et al.* (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am.J.Hum.Genet.* 62:1137-1152.

Torroni A, Huoponen K, Francalacci P *et al.* (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics* 144:1835-1850.

Underhill PA & Kivisild T (2007) Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu.Rev.Genet.* 41:539-564.

Underhill PA, Passarino G, Lin AA *et al.* (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann.Hum.Genet.* 65:43-62.

Verrelli BC, Tishkoff SA, Stone AC & Touchman JW (2006) Contrasting histories of G6PD molecular evolution and malarial resistance in humans and chimpanzees. *Mol.Biol.Evol.* 23:1592-1601.

Virtaranta-Knowles K, Sistonen P & Nevanlinna HR (1991) A population genetic study in Finland: comparison of the Finnish- and Swedish-speaking populations. *Hum.Hered.* 41:248-264.

Voight BF, Kudaravalli S, Wen X & Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wallace DC (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu.Rev.Genet.* 39:359-407.

Wang ET, Kodama G, Baldi P & Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc.Natl.Acad.Sci.U.S.A.* 103:135-140.

Weir BS, Cardon LR, Anderson AD, Nielsen DM & Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15:1468-1476.

Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661-678.

Whittle A (1997) The First Farmers. In Cunliffe B: The Oxford Illustrated History of Prehistoric Europe. Oxford University Press, Oxford. pp. 136-167.

Wiik K (2002) Eurooppalaisten juuret. Atena, Jyväskylä.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R & Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc.Natl.Acad.Sci.U.S.A.* 102:7882-7887.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD & Nielsen R (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3:e90.

Wood B (2000) Investigating human evolutionary history. *J.Anat.* 197 (Pt 1):3-17.

Workman PL, Mielke JH & Nevanlinna HR (1976) The genetic structure of finland. *Am.J.Phys.Anthropol.* 44:341-367.

Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* 12:339-348.

Zerjal T, Beckman L, Beckman G *et al.* (2001) Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. *Mol.Biol.Evol.* 18:1077-1087.

Zerjal T, Dashnyam B, Pandya A *et al.* (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am.J.Hum.Genet.* 60:1174-1183.

Zhivotovsky LA, Underhill PA, Cinnioglu C *et al.* (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am.J.Hum.Genet.* 74:50-61.

Zhivotovsky LA, Underhill PA & Feldman MW (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol.Biol.Evol.* 23:2268-2270.