LIVING ON THE EDGE

POPULATION GENETICS OF FINNO-UGRIC-SPEAKING HUMANS IN NORTH EURASIA

Ville Nikolai Pimenoff

Department of Forensic Medicine
University of Helsinki
Helsinki, Finland

Departament de Ciències de la Salut i de la Vida
Unitat de Biologia Evolutiva
Universitat Pompeu Fabra
Barcelona, Spain

Academic dissertation

To be publicly presented with the permission of the Medical Faculty of the University of Helsinki, in the lecture hall of the Department of Forensic Medicine on October 31st 2008 at 12 o'clock noon.

Helsinki 2008

Supervisors

Professor Antti Sajantila
University of Helsinki
Helsinki, Finland

Professor David Comas
Universitat Pompeu Fabra
Barcelona, Spain

Reviewers

Professor Ulf Gyllensten
University of Uppsala
Uppsala, Sweden

Professor Pekka Pamilo
University of Helsinki
Helsinki, Finland

Opponent

Doctor Lluís Quintana-Murci
Institut Pasteur
Paris, France

ISBN 978-952-92-4331-0 (paperback)
ISBN 978-952-10-4913-2 (pdf)
http://ethesis.helsinki.fi

Yliopistopaino
Helsinki 2008

"We'll never deal with the devils in the details unless we see the big picture."

Paul R. Ehrlich
Human Natures—Genes, Cultures,
and the Human Prospect

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles, which are referred to in the text by their Roman numerals. Study III has also been included in Enattah NS (2005) Molecular Genetics of Lactase Persistence, PhD thesis. University of Helsinki, Finland.

I. 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 Science International 142(1):37–43.

II. Pimenoff VN, Comas D, Palo JU, Vershubsky G, Kozlov A and Sajantila A (2008) Northwest Siberian Khanty and Mansi populations in the junction of West and East Eurasian gene pools as revealed by uniparental markers. European Journal of Human Genetics advance online publication 28 May 2008 (DOI 10.1038/ejhg.2008.101).

III. Enattah NS, Trudeau A, Pimenoff V, Maiuri L, Auricchio S, Greco L, Rossi M, Lentze M, Seo JK, Rahgozar S, Khalil I, Alifrangis M, Natah S, Groop L, Shaat N, Kozlov A, Verschubskaya G, Comas D, Bulayeva K, Mehdi SQ, Terwilliger JD, Sahi T, Savilahti E, Perola M, Sajantila A, Jarvela I, Peltonen L (2007) Evidence of still-ongoing convergence evolution of the lactase persistence T_{-13910} alleles in humans.
American Journal of Human Genetics 81(3):615–25.

IV. Pimenoff VN, Lavall G, Comas D, Palo JU, Gut I, Cann H, Excoffier L and Sajantila A. Fine-scale recombination and linkage disequilibrium in the CYP2C and CYP2D cytochrome P450 gene subfamily regions in European populations and implications for association studies of complex pharmacogenetic traits (submitted).

Additional unpublished data and supplementary material have also been included in this thesis.

ABBREVIATIONS

CEPH	Centre d'Étude du Polymorphisme Humain
cM	centimorgan
CYP	cytochrome P450
CYP2C19	cytochrome P450 2C19 gene
CYP2C9	cytochrome P450 2C9 gene
CYP2D6	cytochrome P450 2D6 gene
DMEs	drug-metabolizing enzymes
HGDP	Human Genome Diversity Project
HGP	Human Genome Project
HVS	Hypervariable segment of mtDNA
indel	insertion/deletion
LD	linkage disequilibrium
LNP	lactase non-persistence
LP	lactase persistence
LPH	lactase-phlorizin hydrolase
MAF	minor allele frequency
mtDNA	mitochondrial DNA
NCBI	National Center for Biotechnology Information
N_e	effective population size
NRY	non-recombining part of the Y chromosome
OMIM	Online Mendelian Inheritance in Man
SNP	single nucleotide polymorphism
STR	short tandem repeat polymorphism
tagSNP	tagging SNP

SUMMARY

In this thesis, I have explored the origins and distributions of genetic variation among the Finno-Ugric-speaking human populations living in remote areas of North Eurasia; it aims to disentangle the underlying molecular and population genetic factors which have shaped the genetic diversity of these human populations.

To determine the genetic variation within and between these human populations I have used mitochondrial, Y-chromosomal and autosomal genetic markers. In mitochondrial DNA analysis, we sequenced the HVS-I and HVS-II parts of the hypervariable control region along with phylogenetically informative SNP from the coding region of the mitochondrial genome. Multiple STR and SNP markers were also genotyped from the non-recombining part of the Y chromosome to assess the paternal variation among the particular North Eurasian populations. Moreover, multiple SNPs were genotyped across the LCT, CYP2C and CYP2D gene regions for the autosomal genetic diversity analysis of biomedical relevance. The obtained genotypes were further analyzed using various population genetic methods.

Our results revealed unique patterns of genetic diversity among the Finno-Ugric-speaking populations. Uniparental genetic diversity suggested that some of the Finno-Ugric-speaking populations in North Eurasia have resided in the contact zone of western and eastern Eurasian gene pools. This fact, along with the reduced uniparental and biparental genetic diversity found, emphasize the complex genetic background of these Finno-Ugric-speaking populations shaped by recurrent founder effects, admixture and genetic drift. Moreover, the high frequency of lactase persistence T_{-13910} allele among the Finno-Ugric-speaking populations and the haplotype background shaped by recent positive selection suggests a local adaptive response to a lactose rich diet in North Eurasia. The Finno-Ugric-speaking Saami show a significant difference in haplotype structure and LD within the cytochrome P450 CYP2C and CYP2D gene subfamily region mainly due to genetic drift, although the role of selection on these genes responsible for xenobiotic metabolism can not be excluded.

Based on our observations, the Finno-Ugric-speaking human populations show unique genetic features due to the complex background of genetic diversity shaped by molecular and population genetic processes and adaptation to remote areas of Boreal and Arctic North Eurasia.

1 REVIEW OF THE LITERATURE

1.1 INTRODUCTION

Since the discovery of numerous polymorphic markers in the human genome (Landsteiner 1931) and a well established population genetic theory pioneered by Haldane (1924), Fisher (1930) and Wright (1931), a great interest has focused on studies of genetic variation in natural human populations (Collins 2003, Jobling et al. 2003, Kidd et al. 2004). Similarly, the recent publication of the entire map of the human genome along with the vast number of available genetic markers and new computational tools have enabled the analysis of whole genomes (Lander et al. 2001, Venter et al. 2001, Collins et al. 2003, International Human Genome Sequencing Consortium 2004). The observed genomic diversity within and among human individuals, groups and closely related species has challenged us to understand the comprehensive heritable variation in humans (Carroll 2003, Collins 2003, Kidd et al. 2004). How much does the variation have any functional significance? How rare or common is a particular fraction of the variation? What is the distribution of the variation within humans and what molecular or population level factors caused the distribution of variation that we see today? In this thesis, I have examined the origin and distribution of genetic variation among the North Eurasian Finno-Ugric-speaking populations by using mitochondrial, Y-chromosomal and autosomal molecular markers.

1.1.1 Genetic characteristics of the Finno-Ugric-speaking populations

It has been suggested that some of the North Eurasian Finno-Ugric-speaking populations (*e.g.* the Finns, Saami, Khanty and Mansi) may hold genetic traces of early Upper Paleolithic people, who first colonized the North Eurasian regions c. 12,000 BP (Cavalli-Sforza et al. 1994, Derbeneva et al. 2002, Norio 2003b, Ross et al. 2006). The Finno-Ugric speakers represent populations, which can be clearly classified into linguistic groups (Figure 1A) and which inhabit relatively large and geographically remote areas in North Eurasia (Figure 1B).

It has been shown that the anatomically modern human (*i.e.* Homo sapiens) colonized the ice-free Eurasia including some areas north of the Arctic Circle already during the Upper Paleolithic (< 40,000 BP) (Pavlov et al. 2001, Vasil'ev et al. 2002, Goebel et al. 1993). The Last Glacial Maximum (23,000–14,000 BP) unquestionably limited the northward spread of modern humans, until a rapid global warming around 12,000 BP initiated the melting of the continental ice sheet revealing novel areas for colonization (Hewitt 1999). Indeed, the permanent arrival of modern humans into North Eurasia (Nordqvist 2000, Vasil'ev et al. 2002, Bergman et al. 2004) is dated to the early Boreal period (< 12,000 BP), although the geographical origin of these settlements is not entirely clear (Nordqvist 2000, Dolukhanov et al. 2002, Kuzmin and Keates 2004).

The Finnish population is one of the genetically well-studied Finno-Ugric-speaking human population (Nevanlinna 1972, Kere 2001, Norio 2003a; 2003b; 2003c). A particular reason for this has been the Finnish Disease Heritage, a highly specific spectrum of more than 30 inherited, mostly recessive diseases with high prevalence in the Finnish population but rare or absent in

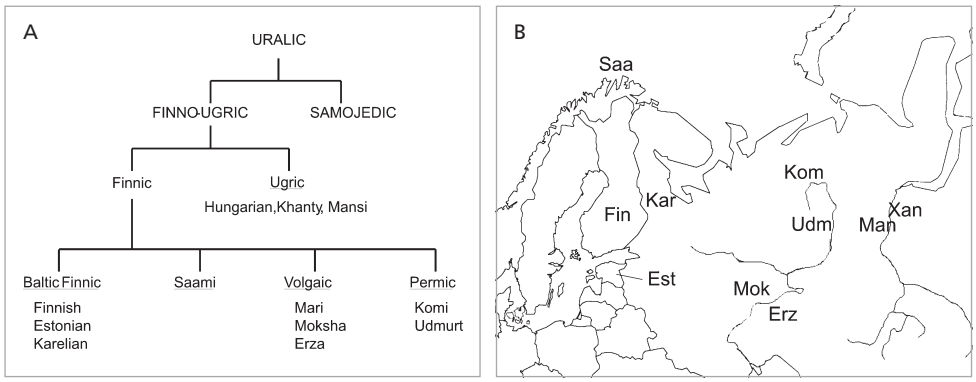

Figure 1. A) Human populations speaking Finno-Ugric languages belong to a specific branch of the Uralic language family which is distinct from the Samoyed-speaking branch also within the Uralic group (Greenberg 2000). The Finno-Ugric language group is further divided into four subclusters within the Finnic group, i) Baltic languages of Finnish, Estonian and Karelian, ii) Saami languages, iii) Volgaic languages of Erza, Moksha and Mari, iv) Permic languages of Komi and Udmurt, while the Ugric group consists of the Khanty, Mansi and Hungarian languages (Abondolo 1998). B) A map showing the geographic locations of the North Eurasian Finno-Ugric-speaking populations used in this study (I–IV).

other populations (Perheentupa 1995, Norio 2003c). Finns are also considered an ethnically more homogenous than several other European populations (Nevanlinna 1972, Kere 2001). However, already based on classical protein markers the Finnish population was shown to share genetic roots not only with the West European populations but also with more eastern populations (Nevanlinna 1972; 1984, Guglielmino et al. 1990). These observations attracted studies of mitochondrial (mtDNA) and Y-chromosomal markers to characterize the maternal and paternal Finnish gene pool, respectively (Vilkki et al. 1988, Pult et al. 1994, Sajantila et al. 1994; 1995; 1996, Lahermo et al. 1996; 1999, Zerjal et al. 1997; 2001, Kittles et al. 1998; 1999, Finnilä et al. 2001, Meinilä et al. 2001, Raitio et al. 2001, Hedman et al. 2007, Lappalainen et al. 2006, Palo et al. 2007). Mitochondrial studies have shown a clear western origin and diversity of the Finnish gene pool, but also minor traces (<5%) of eastern gene flow have been observed (eg. haplogroup Z, U4 and U7; Sajantila et al. 1995, Meinilä et al. 2001, Hedman et al. 2007). The Y-chromosome variation has revealed local reduction in the genetic diversity (Sajantila et al. 1996) and significant genetic differences between Western and Eastern Finland (Kittles et al. 1998; 1999, Lahermo et al. 1999, Zerjal et al. 2001, Lappalainen et al. 2006, Palo et al. 2007). A clear eastern component (haplogroup N3; Rootsi et al. 2007) in the Finnish Y-chromosome gene pool (>50%) has been observed (Zerjal et al. 1997, Lappalainen et al. 2006). Recent accumulation of the autosomal genetic data has clarified Finns as part of the western cluster of the Eurasian genetic landscape, although Finns are outliers among the general European populations (Cavalli-Sforza et al. 1994, Kidd et al. 2004, Lao et al. 2008).

The Saami is another relatively well-studied European Finno-Ugric speaking ethnic group populating the northernmost parts of Norway, Sweden, Finland and Kola Peninsula of Russia (Ross et al. 2006). Sev-

eral studies have shown the Western European genetic affinity of the Saami people, although their origin is still controversial (Cavalli-Sforza et al. 1994, Sajantila et al. 1995, Tambets et al. 2004, Ross et al. 2006, Ingman and Gyllensten. 2007, Johansson et al. 2008). However, it has been demonstrated that Saami are genetically extreme outliers within Europe (Cavalli-Sforza et a. 1994), with strikingly low mtDNA diversity (Sajantila et al. 1995; Lahermo et al. 1996, Tambets et al. 2004). The Saami mtDNA lineages are mainly of Western Eurasian origin while two East Eurasian lineages (*i.e.* haplogroup D5 and Z) indicate minor (≤ 6%) Asian contribution as well (Torroni et al. 1998, Meinilä et al. 2001, Tambets et al. 2004, Ingman and Gyllensten 2007). Similarly, the Y-chromosome variation separates the Saami from other North Eurasian populations, and shows low genetic diversity (Sajantila et al. 1996, Lahermo et al. 1999, Tambets et al. 2004). Two West Eurasian lineages (*i.e.* haplogroup I and R1a) together and one East Eurasian N3 lineage account for ~40% of the Saami Y chromosomes, respectively (Wells et al. 2001, Tambets et al. 2004). Uniparental and autosomal marker studies suggest that the origin of the Saami gene pool is an admixture of Western and Eastern Eurasian genetic components (Cavalli-Sforza et al. 1994, Tambets et al. 2004, Ross et al. 2006, Ingman and Gyllensten 2007, Johansson et al. 2008). Based on the unique genetic diversity of Saami it is interpreted that the Saami population has remained in small and constant size since its origin (Sajantila et al. 1995, Laan and Pääbo 1997, Ross et al. 2006).

The Ugric-speaking Khanty and Mansi ethnic groups originate from a common Ob-Ugric population on the western side of the Ural mountains (Kolga et al. 2001, Derbeneva et al. 2002). Currently, they populate mainly the Ob-river valley region in East Eurasia (Kolga et al. 2001, Derbeneva et al. 2002, Karafet et al. 2002). The Mansi mtDNA variation has shown a high frequency of Western Eurasian lineages, which has been interpreted as a genetic continuum of the early Upper Paleolithic populations expanding from Near East/Southeast Europe to North Eurasia (Derbeneva et al. 2002). However, a clear East Eurasian derived mtDNA component (~38%) is present in the Mansi (Derbeneva et al. 2002). Therefore, it is proposed that the present day Mansi may also contain genetic traces of the early Upper Paleolithic people originating from Central Asia/South Siberia (Wells et al. 2001, Derbeneva et al. 2002, Karafet et al. 2002, Derenko et al. 2007). Interestingly, the Y-chromosome variation has shown some specificity (*i.e.* N2 lineage) among the Siberian populations speaking Uralic languages, including the Khanty (Karafet et al. 2002). However, these Uralic-speaking populations as a whole are not characterized by a discrete set of founder Y-chromosome lineages (Karafet et al. 2002). The uniparental genetic composition of the Khanty and Mansi has been interpreted to represent a recurrent amalgamation of small Eurasian population groups along with the spread of humans into North Eurasia (Derbeneva et al. 2002, Karafet et al. 2002).

The genetic diversity of the other North Eurasian Finno-Ugric-speaking populations (*i.e.* Estonian, Karelian, Erza, Moksha, Mari, Komi and Udmurt) has been studied less comprehensively. The main interpretation among these Finno-Ugric-speaking populations is that they have closer genetic affinities with each other than with the non-Finno-Ugric speakers (except the Latvians and Lithuanians) living in North Eurasia (Zerjal et al. 1997; 2001, Khusnutdinova et al. 1999, Rosser et al. 2000, Raitio

et al. 2001, Derbeneva et al. 2002, Karafet et al. 2002, Laitinen et al. 2002, Kutuev et al. 2006, Tambets et al. 2004, Rootsi et al. 2007).

1.1.2 Sampling in human genetic studies and the concept of a population

The sampling of individuals and the criteria for defining a population are fundamental issues in genetic studies of natural populations (Waples and Gaggiotti 2006). In biological terms a population is defined as a group of organisms of the same species that interbreed and occupy a particular space at a particular time (Krebs 1994). In practice, however, population boundaries are often notoriously difficult to define, and humans make no exception. It is thus apparent that there is no single consensus to define a population but instead it depends on the context and objectives of the study (reviewed by Waples and Gaggiotti 2006). In practical terms, one definition of a population in humans is often a group of individuals that can be clustered according to some shared social or physical characteristic; *e.g.* geography, ethnicity, linguistic affiliation, culture, subsistence pattern and self-identity are often taken as proxies to define a population (Jobling et al. 2003). Consequently, several important ethical issues associated with genetic variation, ethnicity and race have been brought up (Greely 2001b, Tishkoff and Kidd 2004). Especially the indigenous peoples have been concerned about the consequences as genetic studies of human populations are of potential social and legal impact (Greely 2001ab, Collins et al. 2003, Jobling et al. 2003). It is important to note that in humans racial definitions are not based on biology. Up to 95% of the genetic variation in humans resides within populations and only 5–13% between populations (Lewontin 1972, Barbujani et al. 1997, Jorde et al. 2000, Romualdi et al. 2002, Rosenberg et al. 2002). These results clearly show that human races do not have any biological basis and should rather be considered within a complex historical and social context (Lewontin 1972, Collins et al. 2003).

Due to the sensitive issues concerning human population genetic studies, ethical guidelines have been enforced since the reports of World Health Organization (1964), World Medical Association (1964) and North American Regional Committee of the Human Genome Diversity Project (1997) (reviewed by Greely 2001b). In general, these guidelines aim to inform individuals and groups of people who want to participate actively in such population genetic studies, having access to the genetic data created, and also having a possibility to influence the concept of the study. Currently, the most important ethical requirement for human population genetic studies is an informed consent from each individual sampled and, if possible, a group consent obtained from the appropriate cultural authorities of the particular population or ethnic group (Greely 2001b). In addition, guidelines for giving scientific feedback and for explaining the obtained results to the volunteers of the particular study have been proposed (Greely 2001b).

1.2 ORGANIZATION OF THE HUMAN GENOME

1.2.1 General structure

A single haploid human genome is estimated to contain about 3.2 billion nucleotides with an average of 22,000 genes (Lander et al. 2001, International Human Sequencing Consortium 2004, Jobling et al. 2003). Less than 2% of the human genome is as-

sumed to encode proteins (Lander et al. 2001, International Human Sequencing Consortium 2004). This means that the number of genes in humans is greatly less than earlier estimates of 100,000 and even less than identified in the nematode worm *Caenorhabditis elegans* (23,000), the fruit fly *Drosophila melanogaster* (26,000) and the rice *Oryza sativa* (45,000). Moreover, the non-coding repeat elements showed to comprise more than 50% of the human genome, outnumbering that seen in the *Caenorhabditis elegans* (7%) and in the *Drosophila melanogaster* (3%) (Lander et al. 2001). Human and chimpanzee genomes only diverge by 1.2%, and human and Neanderthal genomes are estimated to diverge only by 0.5% (Chimpanzee sequencing and analysis consortium 2005, Green et al. 2006). Moreover, compared to the great apes, humans as a species show low genetic diversity and low genetic structure, which indicates a demographic bottleneck in the early history of humans (Kaessmann and Pääbo 2002). Consequently, human genomes are between 99.5–99.9% similar to each other (Kidd et al. 2004, Goldstein and Cavalleri 2005, International HapMap Consortium 2005; 2007). However, even 0.1% difference between two human genomes denotes over 3 million differing bases, and around 12 million possible variants (Kruglyak and Nickerson 2001, Kidd et al. 2004). This variation is enough to ensure a unique genome for every human individual (Kidd et al. 2004).

1.2.2 Variation in the Human Genome

Genome variation occurs in various forms, such as single-base substitutions (single nucleotide polymorpisms, SNPs), insertions or deletions of tandem (short tandem repeats, STRs; variable number of tandem repeats, VNTRs) or dispersed elements (short interspersed elements, SINEs; long interspersed elements, LINEs), rearrangements (copy number variants, CNVs) and other more complex variables (Figure 2; Denoeud et al. 2003, Watkins et al. 2003, Jobling et al. 2003, Kidd et al. 2004, Redon et al. 2006).

These genomic variations can also be categorized into short (< 10bp), interme-

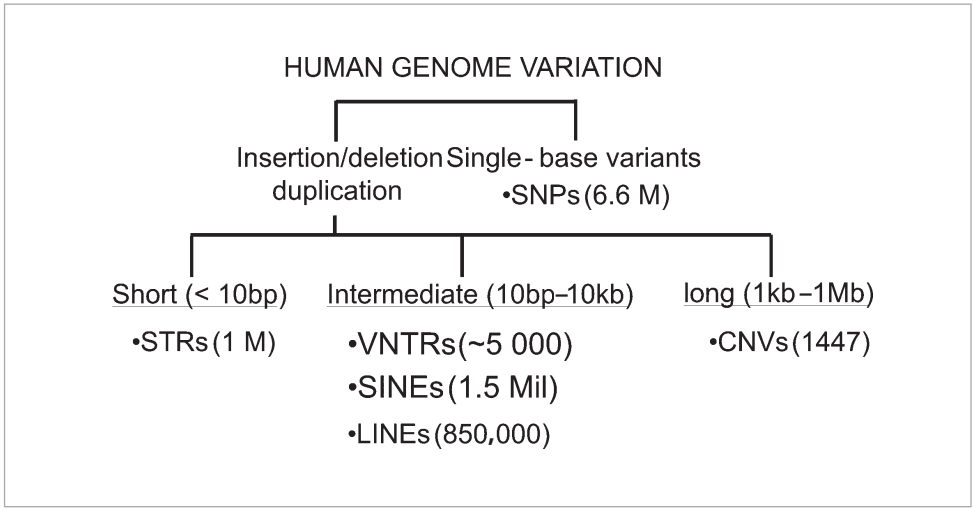

Figure 2. Classes of typical human genome variants.

diate (10bp–10kb) or long size (1kb–1Mb) genomic variants (Figure 2; Jobling et al. 2003). For practical purposes, a variant in a DNA sequence is defined as a polymorphism when at least two alleles are present in a population, and the frequency of the minor allele is MAF ≥0.01. The polymorphisms may be within coding, non-coding or regulatory regions. Within a coding region a polymorphism may alter the amino-acid composition or terminate the translation of the corresponding protein, while variation within a regulatory region may change the level of expression of the particular protein. As the majority of the human genome is non-coding, most of the variants do not affect the amino acids. However, there is a fraction of polymorphisms that do affect the amino-acid composition (0.7% of SNPs, dbSNP128) or gene expression (1.5% of SNPs, dbSNP128). The variants created by mutation can also be rearranged by recombination, which further increases diversity. Both non-homologous recombination between chromosomes and homologous recombination within a chromosome are fundamental genomic processes exchanging genetic material and increasing the genetic variation initially caused by mutations (Nachman 2001). The rate of recombination is typically estimated in centimorgans (cM), which describe the genetic distance in units of recombination frequency (1cM = 1% recombination). The whole genome average recombination rate is 1cM/Mb. However, the recombination rate is shown to vary extensively along the genome from 0.1cM/Mb to more than 3cM/Mb (Kong et al. 2002). More importantly, Jeffreys et al. (2001) estimated that recombination rates are organized into sharp local peaks and valleys (*i.e.* hotspots and coldspots, respectively) along the genome. These recombination hotspots are defined as small fractions of genome between 1–2kb in which the recombination rate is ten or more times higher than in surrounding regions, although regions up to several Mb have also been observed with high recombination rates (Arnheim et al. 2003). The overall existence of the hotspots has been confirmed by larger analysis estimating between 30,000 and 50,000 hotspots along the entire genome (Myers et al. 2006). However, very little is known about the underlying biological mechanisms creating hotspots (reviewed by Arnheim et al. 2003). Based on allele-specific hotspots, it is suggested that the main factor causing hotspots is the distribution of recombination initiation sites along the genome (Jeffreys and Neumann 2002, Arnheim et al. 2003). However, most hotspots have been shown to lack these motifs, indicating multiple causes for the recombination hotspots in the human genome (Myers et al. 2005).

1.2.3 Population processes shaping genetic variation

Genetic diversity created by molecular mechanisms (*i.e.* mutation and recombination) is further modified by population-level processes of genetic drift, migration and natural selection (Jobling et al. 2003, Kidd et al. 2004). These evolutionary forces may affect the variation differently within and among the populations or genomic regions. Hence, although humans are genetically ~99.9% identical, evolutionary mechanisms have created a substantial amount of genetic variation, observed 95% within and 5–13% among human populations (Kidd et al. 2004, McVean et al. 2005).

In a given population, each generation represents a finite sample from the previous generation. This random sampling of gametes known as genetic drift changes allele and haplotype frequencies between generations until the variant becomes ei-

ther fixed or lost (Wright 1931). Under neutral evolution, the number of new alleles generated by mutation is mainly shaped by random genetic drift (Kimura 1968, Ohta 2002). Drift is modeled through the ideal population model and measured as the effective population size (N_e). In this context, N_e is the size of an idealized Wright-Fisher population, *i.e.* population with infinite size, equal sex ratio, non-overlapping generations and random mating that experiences the same amount of genetic drift as the one under study (Wright 1931). The smaller the N_e the greater the genetic drift and vice versa. Therefore, reductions in population size (*i.e.* bottleneck/ founder effect) increasing genetic drift can dramatically change the allele frequencies in populations. An example of genetic drift in humans is the reduced genetic diversity of modern human populations whose ancestors migrated out of Africa and experienced a bottleneck, and lower effective population sizes compared to most sub-Saharan populations (Cann et al. 1987, Armour et al. 1996, Reich et al. 2001).

Migration and subsequent gene flow homogenizes allele frequencies between populations, which reduces the effect of random genetic drift (Jobling et al. 2003). The extreme scenario of a gene flow is an admixture of two populations into an admixed population. The extent of admixture is often inferred using a predefined set of parental populations from which the admixed population is assumed to have derived (Bertorelle and Excoffier 1998). Indeed, estimates of the global admixture of human populations inferred purely from the genetic structure have also been reported (Pritchard et al. 2000, Rosenberg et al. 2002).

Natural selection, originally set by Darwin (1859) and later refined by Fisher (1930), defines the differential survival of phenotypes in succeeding generations. In other words, individuals with allele combinations better adapted to the prevailing conditions are more likely to have higher chances to survive and reproduce. Alleles that reduce the survival are subject to negative (*i.e.* purifying) selection and reduce in frequency, while variants that increase survival undergo positive selection and increase in frequency. Moreover, other selective forces, such as balancing selection, may prefer heterozygote loci or maintain alleles at low frequency, creating high genetic diversity, as observed at the HLA loci responsible for immune response (Beck and Trowsdale 2000, Jobling et al. 2003). Recently, several studies have used molecular data to estimate the departures of allele frequency distributions from neutral expectations and thus to detect natural selection in humans (Akey et al. 2002, Bustamante et al. 2005, Sabeti et al. 2006; 2007, Wang et al. 2006, Williamson et al. 2007). Based on current observations, it is evident that selection has a strong role in shaping human genetic variation, although the relative contribution of the positive, negative and balancing selection to human genetic variation is still unclear (Kelley et al. 2006, Kryukov et al. 2007, Nielsen et al. 2007). It is estimated that the majority of the natural selection acting on genomes is of negative selection removing new deleterious mutations (Nielsen et al. 2007). But most of the genetic surveys have focused on detecting positive selection to disentangle the molecular-level traces of evolutionary adaptations and subsequent factors relating humans to their environment (Nielsen et al. 2007).

Even a weak selective benefit might cause substantial changes in allele frequencies over generations. However, in natural populations different evolutionary forces overlap with one another. Consequent-

ly, genetic drift affects the same alleles subjected to natural selection, and thus it may be difficult to identify loci subject to natural selection (Kelley et al. 2006). It is known that in populations of small N_e stronger selection is needed to influence allele frequencies, whereas allele frequencies of larger populations might be shaped by weaker selective forces (Nielsen et al. 2007).

1.2.4 Visualizing genomic variation

1.2.4.1 Molecular markers used in human genetic studies

DNA sequencing is the ultimate tool for detecting all different genetic variants in any particular genomic region (Sanger et al. 1977), but the method is often technically limited and more expensive to conduct than genotyping individual SNP or STR loci (Mir and Southern 2000, Syvänen 2001). Moreover, the vast number of identified SNP and STR loci in humans and new high-throughput methods enable efficient and simultaneous genotyping of these markers (Mir and Southern 2000, Syvänen 2001, Collins et al. 2003).

It is estimated that SNP markers account for most of all human genetic variation, and are also likely to play a crucial role in how humans respond to exogenous pathogens, chemicals, drugs and other therapies (Collins et al. 2004, Kidd et al. 2004, International Human Genome Sequencing Consortium 2004, International HapMap Consortium 2007, McVean et al. 2005). SNPs are mostly biallelic with an estimated average mutation rate of 2.3×10^{-8} per site per generation (Nachman and Crowell 2000). On average, there is one SNP within every 200bp along the human genome, as currently there are more than 18 million SNPs listed in the human genome, from which more than 6.6 million are validated and 7.5 million are found within genes (Build 129, April 2008). Therefore, SNPs are well-suited for several genetic and evolutionary studies including candidate gene or causal variant mapping, for assessing genetic diversity and divergence, and recently for disentangling whole genome associations of complex traits and risk factors such as cancer, diabetes and vascular diseases (Collins et al. 2004, Kidd et al. 2004, McVean et al. 2005).

STRs (*i.e.* microsatellites) are tandem repeat markers of 1–6 nucleotides (*e.g.* … CACACA… dinucleotide repeats) which are the most variable types of DNA sequences in humans (Weber 1990). The STR loci have typically a high number of different alleles per locus, each with different number of repeat motifs (reviewed by Ellegren 2004). Moreover, STRs are found in all chromosomes, and with a high repeat number variability between individuals. Most of the known STRs are most likely neutral, although some are involved in human diseases. In the disease causing microsatellites, the causal factor is often the increase of the repeat number over some threshold level (*e.g.* >30 and up to 2000 repeats in myotonic dystrophy; Mahadevan et al. 1992). Estimations using human pedigree and sperm sample analysis have shown that the STR mutation rate is between 10^{-3}–10^{-4} per locus per generation (Heyer et al. 1997, Brinkmann et al. 1998, Sajantila et al. 1999, Kayser et al. 2000, Xu et al. 2000). The STR mutation is mainly modeled by the stepwise mutation model (SMM), which postulates that the mutations, *i.e.* gain or loss of one repeat, occur at fixed rate independent of repeat length (Ohta and Kimura 1973). However, this is not totally true as the rate and direction of the STR mutations are shown to be length-dependent (reviewed by El-

legren 2004). Microsatellites, due to their informativeness, Mendelian inheritance and relative ease of genotyping, have proven extremely powerful for linkage analysis of Mendelian disorders (Jorde et al. 1997, Zhivotovsky et al. 2003) and for studies of evolution and population genetic structure (Rosenberg et al. 2002, Ellegren et al. 2004) as well as for genetic identification and paternity testing relevant in forensic medicine (Jobling et al. 1997).

Due to the high number of SNP and STR markers identified in the human genome, some of the markers in the same chromosome are close to each other. Often alleles of markers in close physical proximity are passed on from parents to offspring together. These allele combinations, called haplotypes, either directly from haploid mtDNA and Y-chromosome loci or from diploid chromosomes can be considered as single alleles from a gamete. Haplotypes, compared to single markers, provide a greater statistical power for analysis and reduce the sample size needed for analysis of significant association (Clark 2004). However, the deduction of haplotypes from diploid marker data is not always straightforward. To illustrate the problem we may assume three different cases (Figure 3), where two SNPs in a same chromosome a) are homozygous, b) only one SNP is heterozygous and c) both SNPs are heterozygous. Homozygous SNPs will produce two identical haplotypes, whereas two different haplotypes are observed when only one site is heterozygous. However, in the case of two heterozygous SNPs, the allele combinations are often shuffled by recombination making it more difficult to determine the true haplotypes. To overcome these limitations, genealogical, molecular and statistical methods have been used. In principle, using pedigree data from parents and grandparents often enables accurate haplotype estimations. Molecular methods include amplification of a single cell genome (Ruano et al. 1990), allele specific amplification (Michalatos-Beloin et al. 1996) or construction of mouse-human hybrid cells with haploid human genome (Patil et al. 2001). Statistical approaches are based on the assumption of a common ancestor homozygous at all sites. The steps from the observed variation to this common ancestor within a population are then estimated (Clark 1990). These statistical methods have shown to be powerful for accurate estimation of the haplotypes from diploid genotypes (Excoffier and Slatkin 1995, Stephens et al. 2001).

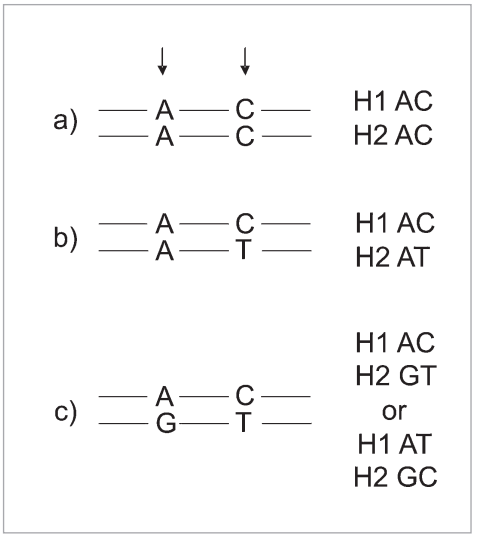

Figure 3. Haplotype phase estimation between two diploid loci of a) homozygotes, b) homozygote and heterozygote, and c) heterozygotes. Arrows denote the two diploid loci and H1/H2 denote the deduced haplotypes.

1.2.4.2 Special characteristics of the uniparental markers

In human cells, there are hundreds to thousands copies of cytoplasmic organelles called mitochondria. Each mitochondria

contain at least a single copy of mitochondrial DNA (mtDNA), organized in a small (~16.6kb) circular double-stranded DNA molecule, which is transmitted without recombination only through the mother. The mtDNA genome contains 37 genes and a non-coding control region including three known hypervariable segments (HVS-I, HVS-II and HVS-III) (Anderson et al. 1981, Andrews et al. 1999, Bandelt et al. 2006). The mtDNA has a much greater average mutation rate ($3.4 \times 10^{-7} – 3.6 \times 10^{-6}$) than nuclear genome ($2.5 \times 10^{-8}$) (Ingman et al. 2000, Nachman and Crowell 2000, Richards et al. 2000). As a haploid non-recombining molecule the variation within mtDNA results only from the accumulation of mutations. The mtDNA haplotypes are often further clustered into mtDNA lineages or haplogroups (Torroni et al. 2006), which possess a molecular record of the maternal genealogical history. In humans, analysis of mitochondrial genomic diversity and phylogeography, an analysis of geographical distribution of the variation, were initially studied merely with the HVS-I (360bp) and HVS-II (268bp) regions and more recently using complete mtDNA genomes. The mtDNA has proven powerful for assessing both the micro-geographic female population histories and reconstructing broader prehistoric human dispersal (Cann et al. 1987, Ingman et al. 2000, Torroni et al. 2006). In addition, the high copy number of this small circular genome has enabled several successful ancient DNA analyses (Pääbo 1989, Cooper and Poinar 2000, Pimenoff and Korpisaari 2004).

Y chromosome, the sex-determining male specific locus of the human genome and a uniparental haploid counterpart of the mtDNA, consists mainly of non-recombining DNA (NRY, 57Mb–60Mb), which is transmitted only from father to male offspring (Jobling and Tyler-Smith 2003). Only two pseudoautosomal telomeric segments recombine with the X chromosome, but these amount to less than 5% of the total length of the chromosome (Jobling and Tyler-Smith 2003). The NRY part of the Y-chromosome is extremely gene poor, coding for only 27 proteins but enriched with many types of DNA repeats and variants. To date, more than 200 binary polymorphisms (*i.e.* SNPs), over 200 microsatellites and several other repeat polymorphisms within the NRY have been characterized (Jobling and Tyler-Smith 2003, Kayser et al. 2004). These two marker classes have differing mutation rates; the slow evolving SNP markers allow construction of common Y-chromosome clusters (*i.e.* haplogroups) and their phylogeny, whereas STRs within these haplogroups enable a more detailed haplotype resolution (Knijff 2000, YCC 2002). The differing resolution obtained with these markers has allowed a detailed phylogeographic analysis of the human male populations (Underhill et al. 2000, Wells et al. 2001, also reviewed by Jobling and Tyler-Smith 2003).

Both mtDNA and Y chromosome loci possess only one quarter of an effective population size compared to the nuclear DNA. Therefore the genetic diversity and phylogenetic structure of uniparental loci are more sensitive to changes in the demography (*e.g.* bottlenecks) and generally show greater genetic differences between different groups or populations. All these evolutionary characteristics combined with the well-defined phylogeny and unified nomenclature system (YCC 2002, Torroni et al. 2006), make uniparental markers ideal tools for investigating the recent human evolution, with additional important applications in medical and forensic genetics (Jobling et al. 1997, Howell et al. 2003, Jobling and Tyler-Smith 2003).

1.2.5 Linkage Disequilibrium

1.2.5.1 Processes shaping LD

Linkage disequilibrium (LD) defined as a non-random association of alleles at linked loci may be broken by recombination during meiosis. Alleles at loci lying in close proximity in a chromatid recombine less frequently than those far apart and are more likely to be in LD. Thus, a new mutation arising in the genome is initially in complete LD with the adjacent marker alleles, which is indicated by only three of the four possible haplotypes between the two loci within a population (Figure 4, reviewed by Ardlie et al. 2002).

The most commonly used measures of pairwise linkage disequilibrium are |D'| (Lewontin 1964) and r^2 (Hill and Weir 1994), which both vary between complete (1.0) and no (0.0) association between loci. Moreover, a recently developed Bayesian method to estimate the population recombination parameter $\rho = 4N_e r$ from genotypes has proven to be an efficient way to quantify LD differences between populations (Li and Stephens 2003, Crawford et al. 2004, Evans and Cardon 2005). However, there is no consensus on what is the best statistic for LD as LD measures are known for several stochastic limitations *e.g.* differing sensibility to sampling and population genetic processes, although the ρ estimate seems to be more robust to fluctuations of ascertainment bias and marker density than the pairwise LD estimates (Lewontin 1988, Weiss and Clark 2002, Phillips et a. 2003, Evans and Cardon 2005).

The strength of LD between a pair of markers depends on both molecular and population genetic factors, which have shown to generate varying patterns of LD across the human genome and populations (Ardlie et al. 2002, Wang et al. 2002, Bertranpetit et al. 2003, Tishkoff and Verelli 2003, Wall and Pritchard 2003, Slatkin 2008). Mutations generally create LD, but locus with a high mutation rate or a high number of alleles (*e.g.* STRs) tend to erode LD (Ardlie et al. 2002), although recombination and gene conversion are the main molecular factors reducing LD. The population genetic factors (drift, migration and selection) have more diverse effects on LD. Consequently, the increased drift of small populations tends to increase LD as haplotypes are lost from the population (Terwilliger et al. 1998). Thus, small isolate populations have been shown to possess higher extended LD compared to populations of a larger size (reviewed in Ardlie et al. 2002). It is noteworthy that inbreeding as a phenomenon inseparable from drift can produce similar results of reduced heterozygosity as random genetic drift in small populations (Slatkin et al. 2008). Therefore, the combined effect of drift and inbreeding in some human population have been proposed to have caused the extended tracts of genomic homozygosity (Gibson et al. 2006). Interestingly, gene flow between populations also enhances the LD. Immediately after two populations have admixed, LD is proportional to the allele frequency differences between the parental populations and not related to the distances between markers. In the following genera-

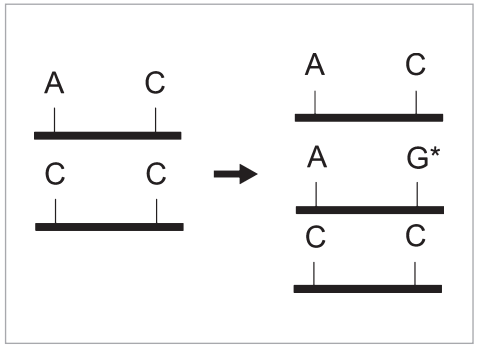

Figure 4. A new mutation G* associated with A allele at nearby locus.

tions, this artificial LD between unlinked markers fades, while LD between nearby markers is often slowly dissipated by recombination. Strong positive selection increases the frequency of an advantageous allele but also alleles closely linked to it, creating unusually strong LD between the causal and neutral alleles. This phenomenon is called genetic hitch-hiking (Smith and Haigh 1974), in which an entire segment of DNA (*i.e.* haplotype) flanking an advantageous variant can rapidly rise to high frequency or even fixation. The selective sweeps have shown significantly elevated LD and reduce heterozygosity among the closely linked neutral markers within particular regions (Smith and Haigh 1974, Kim and Stephan 2002, Nielsen et al. 2005). Similarly, although the overall effect is generally not so strong, negative selection against a deleterious variant may increase LD as the deleterious haplotypes are deleted from the population.

1.2.5.2 Block structured genome, tagSNPs and LD mapping

Based on simulation studies it was assumed that genomic LD rarely extends over 3kb (Kruglyak 1999). However, recent studies have shown that a great fraction of LD in the human genome is organized into discrete sets of loci of low haplotype diversity and high LD between markers (*i.e.* haplotype or LD blocks) separated by short regions (1–2kb) of intense hotspots of recombination (Jeffreys et al. 2000, Jeffreys et al. 2001, Daly et al. 2001, Gabriel et al. 2002, Goldstein 2001, Patil et al. 2001, May et al. 2002). This led to the hypothesis that most of the human genome has a block-like structure with an average LD block between few kb and 100kb (Wall and Pritchard 2001). Hence, it was proposed that only few SNPs at each block would be successful for mapping most of the common genomic variation (Carlson et al. 2004). The structure and distribution of LD blocks along the genome has been shown to be shared by diverse human populations and would indicate a common feature in the human genome (Daly et al. 2001, Gabriel et al. 2002). But the quest for common haplotypes of the human genome has shown to be more difficult a task than expected with no clear current consensus (Daly et al. 20001, Gabriel et al. 2002, Zhang et al. 2002, Phillips et al. 2003, Stumpf and Goldstein 2003, Ding et al. 2005, Zeggini et al. 2005, International HapMap Consortium 2007). However, some common features can be deduced in agreement with most LD studies. The sub-Saharan African populations tend to have shorter LD blocks compared to non-African populations. This is explained by the interplay of more recent recombination and a bottleneck leading to genetic drift experienced by modern humans since the expansion out of Africa as opposed to the present-day sub-Saharan Africans (Tishkoff et al. 1996, Jorde 2000, Gabriel et al. 2002, Wall and Pritchard 2003, Conrad et al. 2006). Moreover, in a whole genome analysis Hinds et al. (2005) estimated that non-African and African-American populations have around 95,000 and 236,000 LD blocks with an average block size of 23.0kb and 8.8kb, respectively. Therefore, it was proposed that further studies with larger sets of human populations are needed to establish more reliable definitions of the block boundaries along the human genome.

Regardless of the block criteria, certain SNPs along the genome show complete LD with each other even with longer distances (>5kb) (Johnson et al. 2001). These tightly correlating SNPs are often called haplotype tagging SNPs (tagSNPs) as it is shown

that typing a few such tagSNPs allows to predict most other variants within the same LD block (Johnson et al. 2001). Currently there are several approaches with congruent results to identify tagSNPs (Chi et al. 2006). These include: i) the identification of LD blocks within the genomic region of interest, ii) the estimation of pairwise LD values within the LD block, and iii) the selection of a few SNPs that capture most of the variation within the LD block (Carlson et al. 2004). However, alternative methods to define tagSNPs without LD block criteria are also currently used (Halldorsson et al. 2004). More importantly, it has been shown that tagSNPs are often well-transferable across populations at least within continental regions (Gonzalez-Neira et al. 2006, Mueller et al. 2005).

In practise, if a marker (*e.g.* tagSNP) is in LD with a disease-causing allele, the strength of LD between the marker and the disease variants can be used to predict the causal allele (Johnson et al. 2001). This population-based LD mapping rests on the assumption that the disease causing mutation stays linked with markers in its physical vicinity for a certain amount of time due to the slower decay of LD with tightly linked markers (Lewontin and Kojima 1960, reviewed by Slatkin 2008). Moreover, recent observations have led to the hypothesis that populations of small and constant size are ideal for LD mapping due to the drift-enhanced disease and allele frequency differences within a population between the case and control samples (Terwilliger et al. 1998). Similarly, admixture may offer another efficient approach for LD-mapping using hybrid populations compared to non-admixed populations (Chakraborty and Weiss 1988). However, the success of the admixture mapping depends heavily on the time since the admixture and the frequency differences of the disease and associated alleles in parental populations (Chakraborty and Weiss 1988). Based on these observations and unique demographic histories of the Finns and Saami, these populations have often shown markedly higher levels of extended LD compared to other European populations (Varilo et al. 1996; 2000; 2003, Laan et al. 1997; 2005, Kaessmann et al. 2002, May et al. 2002, Kauppi 2003, Johansson et al. 2005; 2007, Service et al. 2006). In this context, LD has been successfully used for mapping monogenic diseases prevalent in the Finnish population (Hästbacka et al. 1992, de la Chapelle and Wright 1998, Peltonen et al. 2000). The Saami have also been proposed as a promising target population for LD drift mapping of complex traits (Terwilliger et al. 1998, Kaessmann et al. 2002, Ross et al. 2006).

1.3. HUMAN GENOME DIVERSITY AND THE HAPMAP PROJECT

Shortly after the announcement of the Human Genome Project (HGP), Cavalli-Sforza et al. (1991) proposed for a worldwide survey of the human genome variation known as the Human Genome Diversity Project (HGDP). The aim of this project was to disentangle the structure and distribution of the genetic diversity in humans. Despite the difficulties in ethical issues and criticism from scientists and indigenous people (Greely 2001a), the HGDP successfully collected and announced a worldwide sample set of 1064 individuals representing 52 populations from all continents (HGDP CEPH cell line panel, Cann et al. 2002). These samples have since been used in a number of population genetic studies and the results are continuously collected into a publicly available database (Cavalli-Sforza 2005).

The discovery of the punctuate LD along the human genome (Ardlie et al. 2002, Gabriel et al. 2002) combined with the previous hypothesis of common disease/ common variant (Lander 1996, Reich and Lander 2001) and the available high-throughput genotyping methods boosted the foundation of the International HapMap Project (The International HapMap Consortium 2003). The primary aims of the HapMap were i) to discover new ascertained SNPs across human genome, ii) to characterize a genome-wide set of SNPs validated in four human populations and iii) to produce a common haplotype map of the entire human genome using 269 DNA samples from four ethnic human groups (*i.e.* of African, European, Japanese and Han Chinese origin). The primary use of the common haplotype map is in whole genome association studies of complex traits (The International HapMap Consortium 2003). So far, as a phase I result the HapMap has characterized more than 4 million SNPs along the human genome, and recently completed phase II has identified additional 6 million SNPs (The International HapMap Consortium 2005; 2007). Currently the project has been improved by the addition of more populations (HapMap phase 3 data, www.hapmap.org). Moreover, numerous fine-scale genomic analysis and genome-wide association studies have benefitted from HapMap data (Deloukas and Bentley 2004, McVean et al. 2005). Despite recent criticism (Terwilliger and Hiekkalinna 2006), the HapMap project has already strongly contributed to our quest for understanding the significance of the heritable genetic variation in modern humans and to disentangle the genetic variants relevant in complex traits of human health and disease (Deloukas and Bentley 2004, McVean et al. 2005, The International HapMap Consortium 2007).

2 AIMS OF THE PRESENT STUDY

In this thesis and the articles within I have explored the underlying molecular and population genetic factors and processes shaping genetic variation. The main focus of this thesis has been the Finno-Ugric-speaking populations living in remote and relatively extreme geographic locations in North Eurasia.

Specifically I have focused on the following themes:

1) To study the genetic history and diversity of the Finno-Ugric-speaking populations by using uniparental markers (I, II).

2) To determine the prevalence and haplotype background of lactase persistence variant C/T_{-13910} in North Eurasian populations (III)

3) To assess the recombination rate variation, haplotype structure and LD pattern within clinically significant cytochrome P450 CYP2C and CYP2D gene subfamily regions in European populations including the North Eurasian Finno-Ugric-speaking Saami and Finns (IV)

3 MATERIALS AND METHODS

3.1 SAMPLES

DNA samples consisted in total of 3119 healthy unrelated individuals of 53 human populations with informed consent. Moreover, a total of 5697 reference samples of 42 Eurasian populations were obtained from the literature. All these samples were used in the analysis but with differing sets as described in the original publications (I–IV). It is noteworthy that our main interest concentrates on the North Eurasian Finno-Ugric-speaking population shown in detail in Table 1 and also described in Pimenoff and Sajantila (2002).

3.2 MOLECULAR DATA

To study the maternal neutral genetic diversity and evolutionary relationships of different North Eurasian human populations, we assessed the mtDNA HVS-I and HVS-II region sequences between positions 16024–16383 and 72–340, respectively. In addition, we analyzed seven mtDNA coding region SNP markers to confirm the observed mtDNA control region lineages (II). To assess the paternal neutral genetic diversity and dispersal among the North Eurasian populations, we used 17 Y-chromosome-specific SNP markers describing

Table 1. Finno-Ugric-speaking populations used in each study (I-IV)

Population	n[a]	Linguistic affiliation	Geographic affiliation	Subsistence	Population size[b]	References within
Finns	400	Finnic (Finno-Ugric)	Northeast Europe	Agriculture	5,000,000	I, II,III,IV
Saami	114	Finnic (Finno-Ugric)	Northeast Europe	Reindeer breeding	80,000	II,III,IV
Estonians	28	Finnic (Finno-Ugric)	Northeast Europe	Agriculture	1,300,000	II
Karelians	83	Finnic (Finno-Ugric)	Northeast Europe	Agriculture	140,000	II
Moksha	30	Volgaic (Finno-Ugric)	Northeast Europe	Agriculture	380,000	II,III
Erza	30	Volgaic (Finno-Ugric)	Northeast Europe	Agriculture	760,000	II,III
Udmurt	30	Permic (Finno-Ugric)	Northeast Europe	Agriculture	640,000	II,III
Komi	28	Permic (Finno-Ugric)	Northeast Europe	Agriculture	340,000	II,III
Khanty	106	Ugric (Finno-Ugric)	Northwest Siberia	Reindeer breeding	21,000	II,III
Mansi	161	Ugric (Finno-Ugric)	Northwest Siberia	Reindeer breeding	8,000	II,III

a Total amout of unrelated DNA samples used in this study b Laakso 1991, Kolga et al. 2001, Karafet et al. 2002

the paternal haplogroup distribution along with 12 Y-chromosome specific microsatellite markers, with four additional Y STRs analysed in the Finnish population (I, II). For the haplotype analysis of lactase persistence T_{-13910} allele among populations, eight SNPs and one indel polymorphism with minor allele frequencies MAF > 0.07 distributed across a 30kb region of LCT gene was used with additional sequences (~ 700kb) flanking the whole LCT gene region in particular individuals (III). To disentangle the allele and haplotype distribution of clinically significant cytochrome P450 CYP2C and CYP2D gene subfamily regions we used 55 and 97 SNP markers with MAF> 0.05 in dbSNP with a mean spacing of 7.8kb and 7.6kb, respectively (IV). All the genotyping methods are described in detail in the original publications (I-IV).

3.3 DATA ANALYSIS

Population diversity indices, allele frequencies, Hardy-Weinberg (HW) equilibrium, and population pairwise F_{ST}- or R_{ST}-values along with the exact test of population differentiation and the analysis of molecular variance (AMOVA) were estimated using Arlequin software v3.0 (Excoffier at al. 2005) (I–IV). Phylogenetic median-joining networks were constructed using program package Network 4.5.0.0 (www.fluxus-techology.com) and when required locus weights described by Bandelt et al (2002) or Bosch et al (2006) were used (II, IV). To estimate the coalescence age of specific lineages within a uniparental network, the ρ-statistic along with mutation rates from Forster et al. (1996) and Saillard et al. (2000) were implemented (II). To define and test the uniparental phylogeographic structures both spatial analysis of molecular variance (SAMOVA; Dupanloup et al. 2002) and autocorrelation indices for DNA analysis (AIDA; Bertorelle and Barbujani 1995) were performed (II). Importantly, correlations between mtDNA and Y chromosome distance matrices (II) as well as between F_{ST} and population recombination rate delta distances (IV) were estimated using the Mantel test (Excoffier et al. 2005). Allele frequencies and uniparental lineages were also geographically visualized using the MapView 6.0 program (StatSoft™) (II). Moreover, pairwise F_{ST} and R_{ST} values were visualized using multidimensional-scaling (MDS) procedure implemented in the STATISTICA software package (StatSoft™) (II, IV). For each population, autosomal haplotypes were inferred separately either using the Arlequin software (Excoffier et al. 2005, III) or PHASE v.2.1 software package with 1000 iterations (Stephens et al. 2001, III–IV). Moreover, recombination rate parameter ρ was inferred separately for each population and genomic region using software PHASE v.2.1 with 1000 iterations (Stephens et al. 2001, IV). Non-parametric Spearman correlations between population recombination estimates and Wilcoxon test for adjacent SNP r^2-values between populations were performed with SPSS 7.0 (IV). In addition, most of the autosomal genotype data modifications were performed prior analysis with Perl scripts and Perl 5.8.7 (IV).

4 RESULTS AND DISCUSSION

4.1 UNIPARENTAL GENETIC LANDSCAPE IN NORTH EURASIA (I, II)

Mitochondrial and Y-chromosome studies suggest that not only the Southwestern Europe (Semino et al. 2000, Torroni et al. 2001) but also Central Asia (Wells et al. 2001, Zerjal et al. 2002, Comas et al. 2004, Quintana-Murci et al. 2004) and South Siberia (Derenko et al. 2003; 2007ab) have had an important role in the early settlement of the modern humans into North Eurasia. However, the genetic roots and dispersals of the North Eurasian Finno-Ugric-speaking populations are not entirely clear (Cavalli-Sforza et al. 1994, Derbeneva et al. 2002, Karafet et al. 2002, Norio 2003b, Ross et al. 2006).

To explore uniparental neutral variation among the North Eurasian Finno-Ugric-speaking populations and situate them into the North Eurasian genetic landscape, 42 ad 33 Eurasian mtDNA and Y chromosome population samples were analyzed, respectively. In addition, Y chromosome STR haplotypes from 15 Eurasian populations were used in further comparisons (Pimenoff et al. unpublished data of 85 Finno-Ugric-speaking Erza, Moksha and Udmurt individuals were also included).

In our analysis, geographically associated uniparental haplotypes showed statistically significant frequency trends along the East-West axis of North Eurasia (study II, Figure 1AB, 2). This is congruent with the current view of the clinal distribution of West and East Eurasian uniparental lineages (Richards et al. 2000, Semino et al. 2000, Underhill et al. 2000, Wells et al. 2001, Kivisild et al. 2002, Metspalu et al. 2004, Rootsi et al. 2007). Correspondence analysis also revealed east-west patterns of North Eurasian maternal lineages (study II, Figure 4A), where Finno-Ugric-speaking populations form distinct clusters at an edge of the mtDNA haplogroup distribution. However, the geographical pattern is not so clear within the Y-chromosome, except the clustering of Finno-Ugric and Samojedic populations together along with the Yakut population (study II, Figure 4B). Similarly (study II, Figure 3AB), mtDNA pairwise F_{ST} distances identify Finno-Ugric-speakers as distinct clusters between Northeast Europe and South Siberia/Central Asia, while the Y-chromosome R_{ST} distances appeared less structured. Even, when the Erza, Moksha and Udmurt Y-chromosomes (Pimenoff et al. unpublished) are added, the R_{ST} distances show the Finno-Ugric population totally dispersed with no clear structure. A mantel test between mtDNA and Y-chromosome pairwise distances showed nonsignificant correlation.

Indeed, most of the Finno-Ugric-speaking populations showed to possess both West and East Eurasian associated uniparental lineages (Figure 5AB, see also study II, Figure 1AB). This unique amalgamation of West and East Eurasian gene pools may indicate either mixed origin of these populations from genetically distinguishable Eastern and Western Eurasia or that North Eurasia was initially colonized by humans carrying both West and East Eurasian lineages. Previous studies of the Saami and Finns support the idea of mixed origin in these populations (Norio et al. 2003b, Ross et al. 2006, Johansson et al. 2006, Ingman and Gyllensten 2007). However, the Central Asian (Karafet et al. 2002, Comas et al. 2004), Southwest Asian (Quintana-Murci et al. 2004) and South Siberian (Kara-

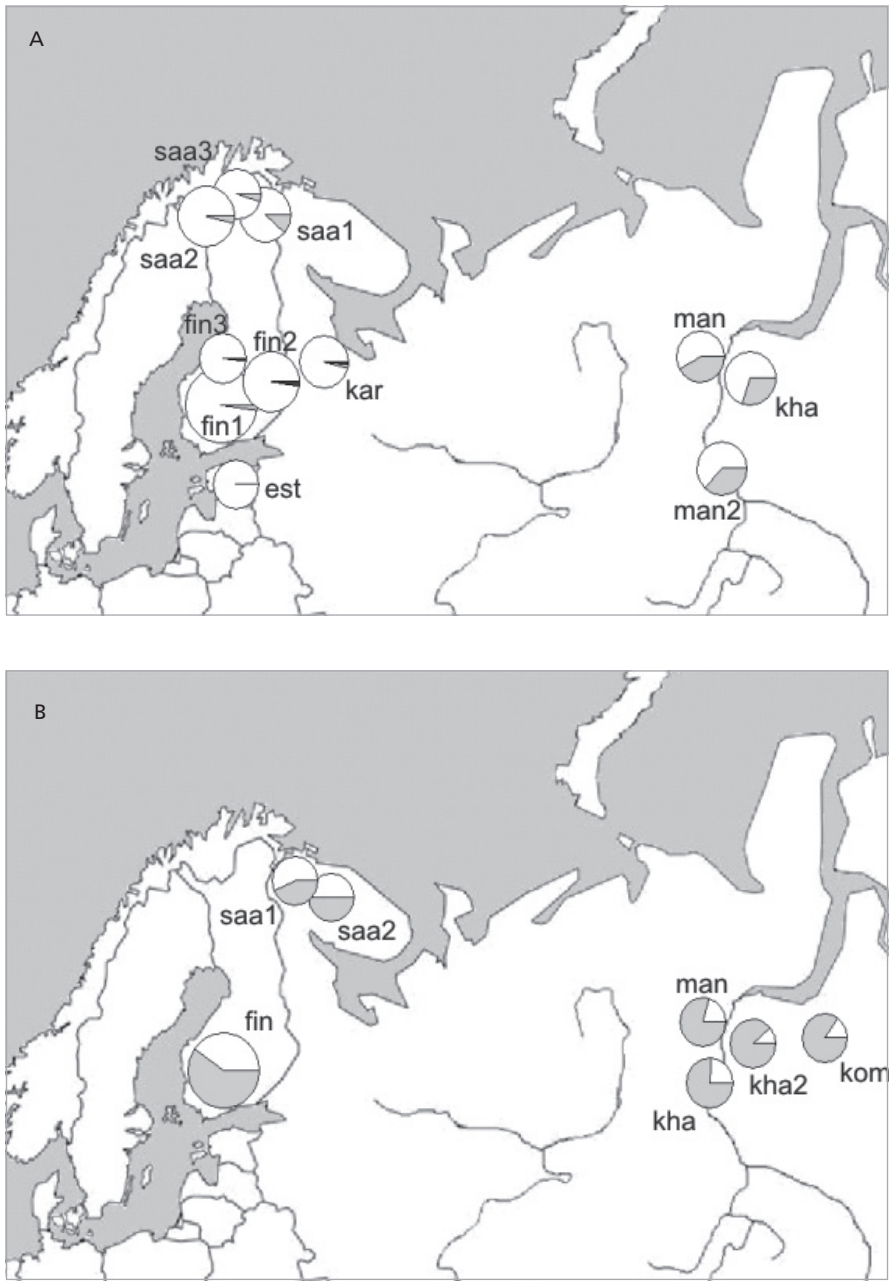

Figure 5. Distribution of the geographically associated A) mtDNA and B) Y-chromosome lineages among the Finno-Ugric-speaking Finns (fin), Khanty (kha), Komi (kom), Mansi (man) and Saami (saa) populations. Colors for the associated haplogroups are the following: white, West Eurasian; gray, East Eurasian; black, South Asian (geographical classification based on study II). Population abbreviations refer to the same samples and abbreviations used in study II (Figure 1AB), except in 5B (saa2; Pimenoff et al. unpublished data).

fet et al. 2002, Derenko et al. 2003; 2006; 2007b) populations have shown an admixture of West and East Eurasian lineages with higher overall genetic diversity compared to Finno-Ugric-speakers. This supports the idea that the edge of North Eurasia was colonized through Central Asia/ South Siberia by human groups already carrying West and East Eurasian lineages. Moreover, the observed mitochondrial U7 and Y-chromosome N2 sublineages indicate a more recent gene flow probably from Central Asia to North Eurasia (Rootsi et al. 2007). In this context it could be explained that the geographical region from Central Asia to Northeast Europe and Northwest Siberia has been a contact zone of genetically distinguishable Western and Eastern Eurasian lineages formed by recurring migrations and admixture of distinct population groups. This unique admixture is currently seen in North Eurasian Finno-Ugric-speaking populations

When the number of observed Y-chromosome STR haplotypes was divided by the number of sampled individuals within a population, a lower fraction of heterogeneity was observed in the Finns (43%), Khanty (43%), Mansi (52%) and Saami (52%) compared to Finno-Ugric Erza (81%), Komi (69%), Moksha (86%) and Udmurt (67%) or most other more southern populations. The difference in heterogeneity could be explained by the at least four times smaller population size (Table 1) and thus the greater sensitivity to genetic drift in the Khanty, Mansi and Saami compared to the Erza, Moksha, Komi and Udmurt populations. However, in the Finns it is not the population size but a probable population bottleneck in the founding of the Finnish population (Sajantila et al. 1996, Lahermo et al. 1999), which explains the reduced Y-chromosome heterogeneity (Figure 6). Moreover, when the samples were divided into six geographical subpopulations a clear local reduced heterogeneity in Eastern Finland and a significant genetic difference between Western and Eastern Finland was observed (Figure 6; see also study I, Table 3, Lappalainen et al. 2006, Palo et al. 2007).

Apart from the unique genetic amalgamation and clinal distribution of western and eastern uniparental lineages, the North Eurasian Finno-Ugric-speaking populations are genetically a heterogeneous group, mostly showing lower haplotype diversities among and within uniparental haplogroups when compared to more southern populations (see also Karafet et al. 2002, Derenko et al. 2003; 2006; 2007b, Comas et al. 2004). In a broader perspective, the ensuing loss of genetic diversity in populations living in Arctic and Boreal regions compared to more southern areas has been clearly demonstrated in a range of other species as well (Hewitt 2001). It is explained by the presence of southern refuges through the Last Glacial Maximum (c. 13,000 years ago) and subsequent founder effects, gene flow and genetic drift shaping the genetic diversity of small population groups migrating to the North Eurasia.

4.2 DISTRIBUTION OF LACTASE PERSISTENCE ALLELE IN NORTH EURASIA (III)

Most humans cannot digest lactose, *i.e.* the main milk carbohydrate, after weaning due to the natural reduced activity of the lactase-phlorizin hydrolase (LPH) enzyme in intestinal cells (Sahi et al. 1973, reviewed by Swallow 2003). These individuals are considered as lactase non-persistent (LNP) [MIM223100]. However, some people maintain the LPH activity throughout life, *i.e.* are lactase persistent

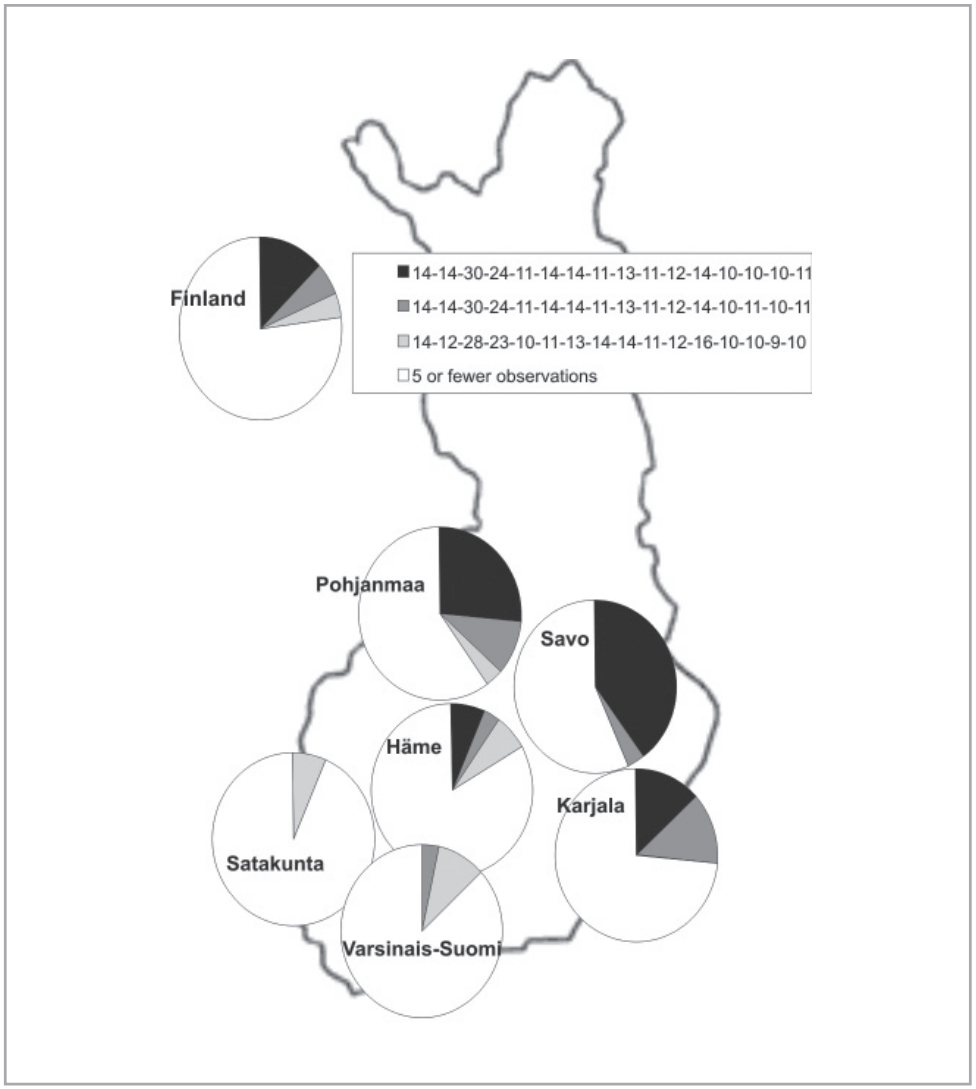

Figure 6. Distribution of the most common Y-chromosome 16-loci STR haplotypes within six subpopulations of Finland (study I, Figure 3B).

(LP) [MIM223100], and thus can use milk products without metabolic difficulty. Interestingly, among the North Eurasian and some sub-Saharan African populations, often with high dairy product consumption, the lactase persistence is relatively common (>80%), whereas the lactase non-persistence is predominant among the rest of populations worldwide (Swallow 2003). Previously a single SNP C/T_{-13910} was shown to correlate completely with the LP/LNP phenotype among the North Europeans (Enattah et al. 2002). This T_{-13910} variant correlating with LP is located 14kb upstream of the LCT gene, which encodes for LPH (Enattah et al. 2002). Previous haplotype analysis showed that all LP alleles among Finns originated from one common ancestor indicating a single introduction of lactase persistence allele into North

Europe (Enattah et al. 2002). It was also confirmed that the LNP is the ancestral state in humans like in most mammals and mutations causing LP have arisen probably due to a recent positive selection as an adaptation to energy need in lactose rich diet coincident with the animal domestication (Hollox et al. 2001, Enattah et al. 2002, Beja-Pereira et al. 2003, Bersaglieri et al. 2004, Myles et al. 2005, Tishkoff et al. 2007).

To investigate the allelic background of LP variant T_{-13910} in North Eurasia, we genotyped eight SNPs and one indel polymorphism, including C/T_{-13910} variant, covering the ~30kb of the LCT region in 37 worldwide populations (study III, Figure 1). Our results showed a high frequency of LP T_{-13910} allele especially among the Finno-Ugric-speaking Finns (58%), Udmurt (33%), Moksha (28%) and Erza (27%) (study III, Table 3). Interestingly, the reindeer breeding Khanty (Ob-Ugric, 3%), Mansi (Ob-Ugric, 3%) and Saami (17%) exhibit the lowest LP T_{-13910} allele frequencies among the Finno-Ugrics compared to the agriculturalists with the exception of the Komi (15%; N=10) (Table 1).

Furthermore, we identified nine different haplotypes carrying the T_{-13910} LP variant and 14 haplotypes with alleles carrying the C_{-13910} LNP variant, each with a frequency of >4% in at least one of the populations (study III, Table 5). One of the nine LP haplotypes dominate in LP alleles in most populations including the Finno-Ugrics, in which the highest frequencies are among the agriculturalists. Six other LP haplotypes were also observed at the reasonably frequency (between 2% and 11%) in the Finno-Ugric-speaking populations mostly among the agriculturalists.

A median-joining network constructed from these common haplotypes (MAF >4%) revealed two distinct clusters of LP haplotypes carrying the T_{-13910} allele (Figure 7). The first cluster was observed only among the Finno-Ugric Udmurts (15%), Mokshas (11%) and Erzas (5%) along with Iranians (5%), while the second cluster of LP haplotypes including the dominant H98 LP haplotype was observed in almost all populations (Figure 7; study III, Table 5). This observation of multiple LP haplotype clusters among the Finno-Ugric-speaking populations indicate two independent origins of the LP T_{-13910} allele in North Eurasia. We also indentified a probable single LNP haplotype responsible for the background on which the most common major LP haplotype was derived. This LNP background haplotype shows the highest frequency among the Finno-Ugric-speaking populations (between 33 and 35%) along with Han Chinese (36%), which might indicate an East Eurasian origin of the particular ancestral haplotype. However, the molecular and demographic factors may bias interpretation based purely on population frequencies. The age estimates showed that the common LP T_{-13910} haplotype cluster (12,000–5,000 BP) is older than the LP T_{-13910} haplotype cluster (3,000–1,400 BP) mainly observed among the North Eurasian Finno-Ugric-speaking agriculturalist populations. This LP T_{-13910} allele and haplotype frequency distribution in Finno-Ugric-speaking populations along with previous reports among the sub-Saharan populations (Tishkoff et al. 2007, Ingram et al. 2007) strongly imply that the LP T_{-13910} variant has been introduced independently more than once into North Eurasia. Moreover, the observed results support the role of still-ongoing convergent evolution of the lactase persistence among the Finno-Ugrics in response to adult milk consumption coincident with the change in subsistence at the edge of North Eurasia (Table 1).

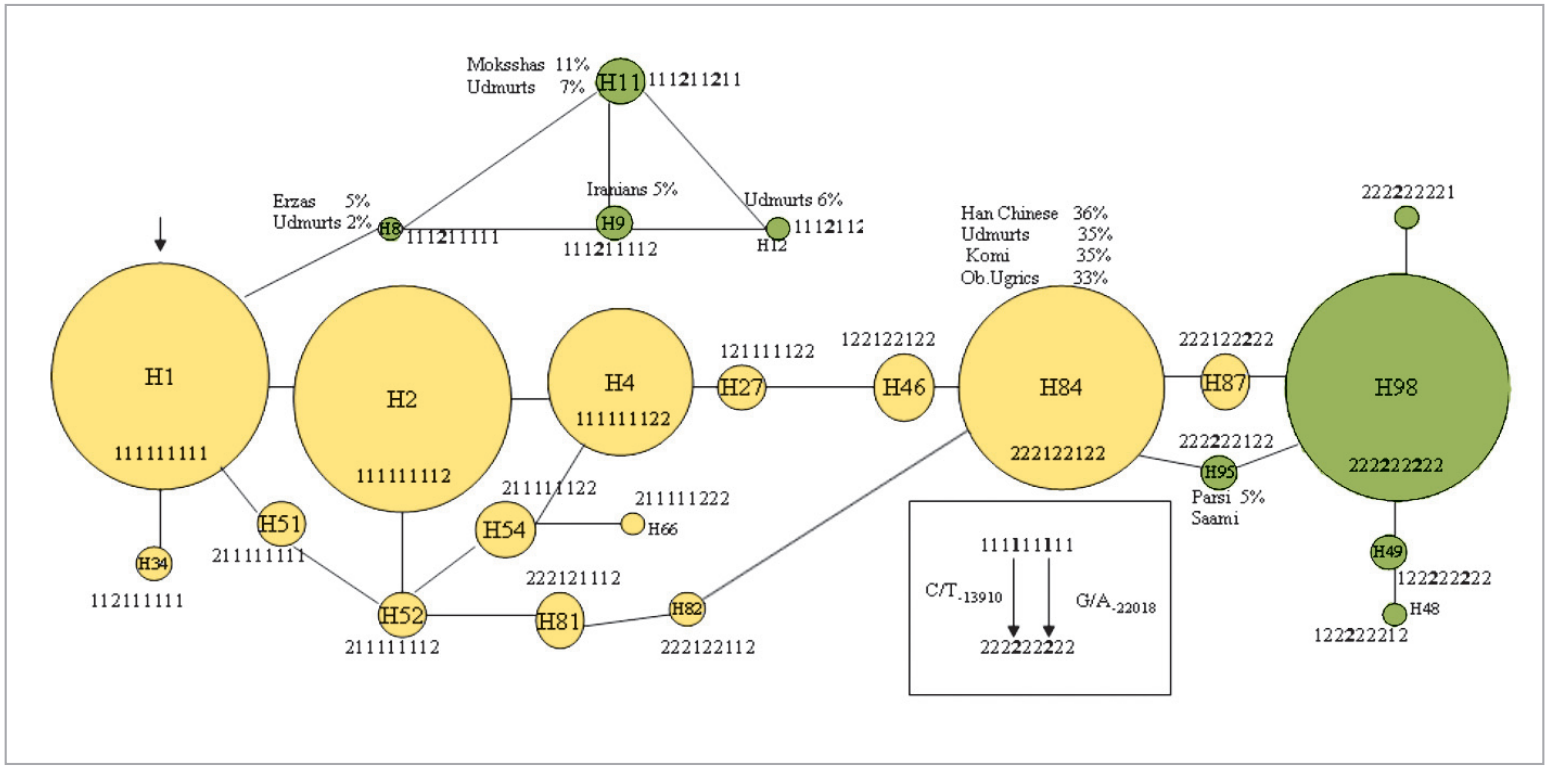

Figure 7. MJ-network of common (MAF>4%) LNP/LP haplotypes constructed from eight SNPs and one indel marker across the 30kb LCT gene region among 37 worldwide populations (study IV, Figure 5). Arrow denotes the root of the network. LNP haplotypes are shown with yellow and LP haplotypes with green color. The size of the circles are proportional to the estimated haplotype frequencies. Haplotype frequencies for the Finno-Ugric-speaking populations discussed in the text are shown. The positions of the C/T_{-13910} and G/A_{-22018} alleles within the haplotype are shown in the box. The SNPs have been coded for each site as 1 for the ancestral SNP and 2 for the derived SNP.

4.3 PATTERNS OF LD IN CYP2C AND CYP2D GENE SUBFAMILY REGIONS IN EUROPE (IV)

The cytochrome P450 oxidase gene family comprises a set of evolutionary-related genes that code for xenobiotic metabolism enzymes (Ingelman-Sundberg 2004). In humans, genes within the CYP2C and CYP2D regions of the cytochrome P450 gene subfamily code for CYP2C8, CYP2C9, CYP2C18, CYP2C19 and CYP2D6 drug-metabolizing enzymes (DMEs) (Wilkinson 2005). These genes are highly polymorphic with several known genetic variants associated to variable drug reactions of significant clinical relevance (Lewis 2004).

To characterize the recombination rate variation, LD distribution and haplotype structure in the CYP2C and CYP2D regions we genotyped 144 SNPs across these two regions in Finno-Ugric-speaking Saami and Finns along with nine other European and one African population (study IV). A further aim was to disentangle the past molecular and population genetic processes responsible for the observed LD distribution, inferring from known differences in demographic history of Saami and Finns compared to other European populations. In agreement with results obtained from other marker analysis (Cavalli-Sforza et al. 1994, Ross et al. 2006, Lao et al. 2008), the Finno-Ugric-speaking Saami and Finns showed significantly different CYP2C and

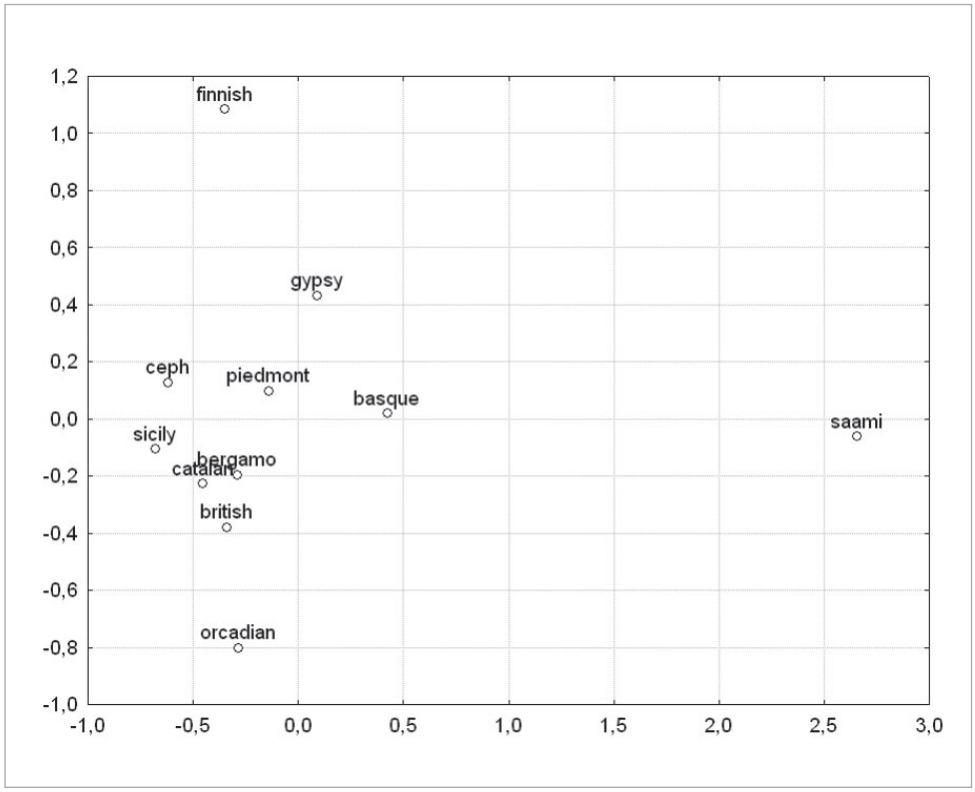

Figure 8. Multidimensional scaling (MDS) of population pairwise F_{ST} distances between 11 European populations across CYP2C and CYP2D gene regions (stress value = 0.073).

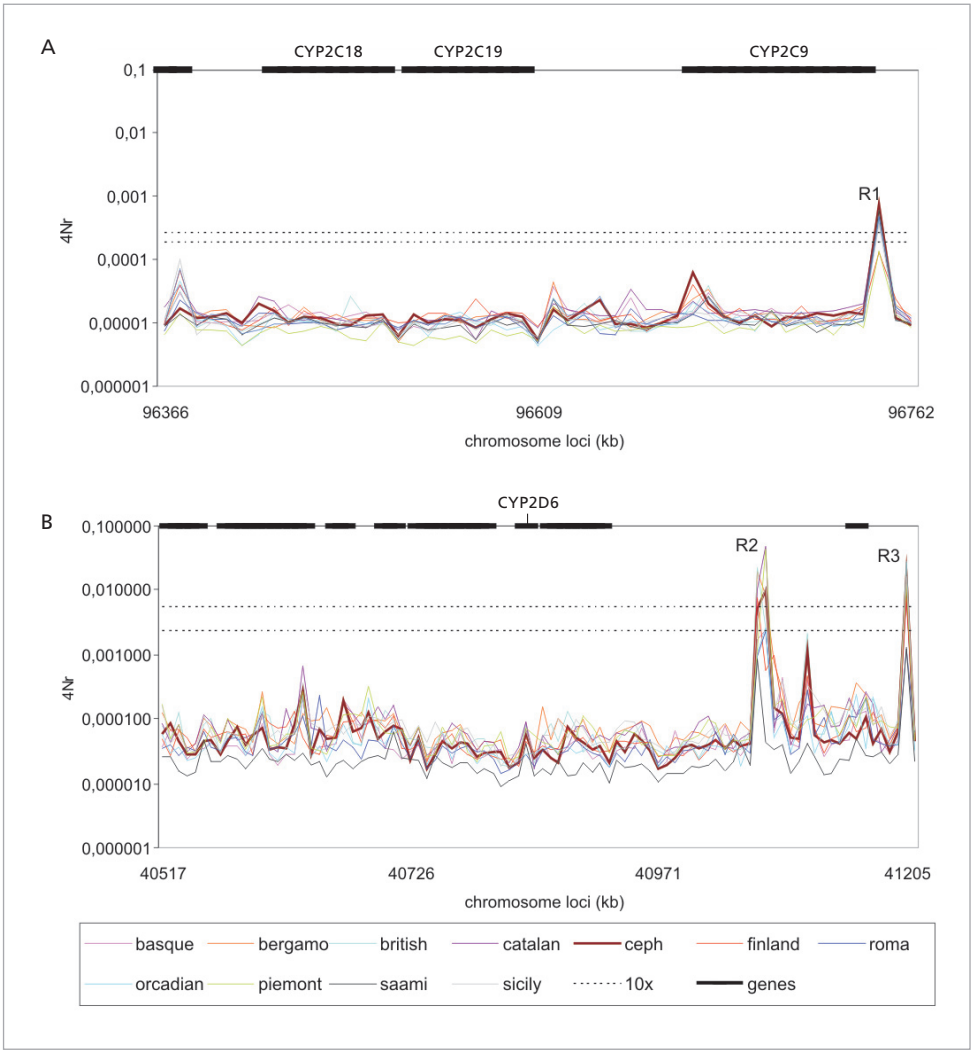

Figure 9. Recombination rate estimates across the A) CYP2C and B) CYP2D regions. Y axis is expressed in log scaled units of recombination rate (4Nr). The dash lines represent the upper and lower 95% confidential intervals of the 10-fold average recombination rate among the European populations at either region. The position of genes is shown as horizontal black bars on top of each graph.

CYP2D allele frequencies from other European populations (Figure 8). For the rest of the European populations including the CEPH sample representing the general European population as such in the HapMap project, the observed locus-specific and population pairwise F_{ST}-values indicate low degree of allele frequency differentiation for the two cytochrome P450 regions.

The estimated patterns of recombination rate variation revealed significant but lower correlation among European populations compared to correlations observed between continental groups (Evans and Cardon 2005). Regardless of the allele frequency differences and recombination rate heterogeneity among the studied populations, the location and magnitude of de-

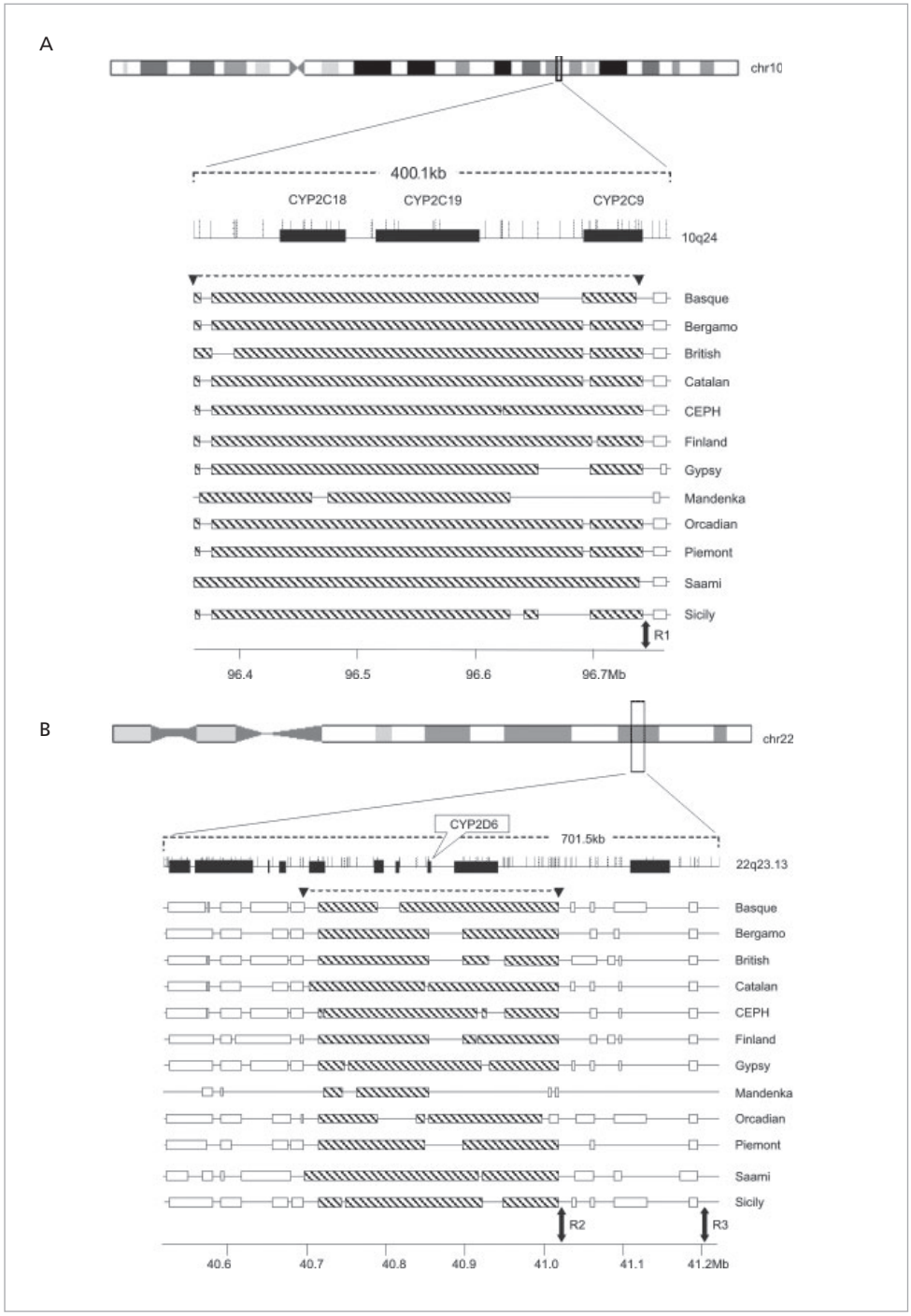

Figure 10. The haplotype blocks identified at CYP2C (A) and CYP2D (B) loci based on Gabriel et al. (2002) are shown in bars. Bars containing diagonal lines are those identified within the extended LD region at both loci. Empty bars are LD blocks characterized outside the extended LD region. The position of genes is shown as horizontal black bars (only CYP2 genes identified) below the depicted chromosome. Vertical arrows denote estimated recombination hotspots of R1-R3.

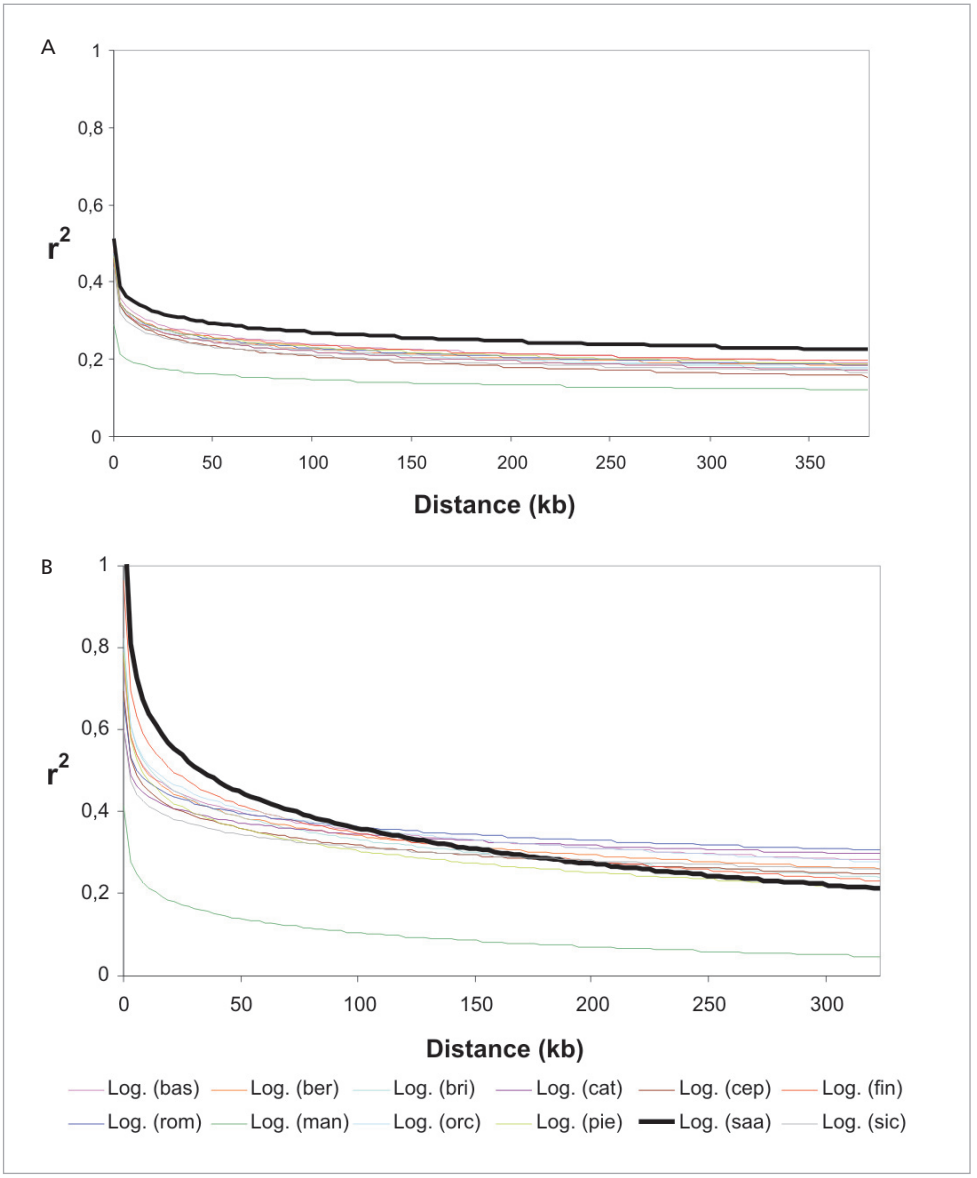

Figure 11. Decay of r^2 against the distance (kb) between marker pairs within the extended LD region defined as logarithmic best-fit curves along A) CYP2C and B) CYP2D regions. Population abbreviations are as reported in study IV (Table 1).

tected recombination hotspots R1–R3 are conserved in all 11 European populations (Figure 9) and the African Mandenka population. Interestingly, the CEPH European reference sample shows a very similar recombination profile with other European populations suggesting that a fine-scale recombination map inferred from the HapMap CEPH data would be applicable for other European populations. However, the loci with lower recombination rate exhibit more variation in the rates between populations indicating differences either in recombination histories or in past demography. In

more detail, the Saami shows the lowest recombination rate profile for CYP2D but not for CYP2C region, suggesting that recombination rate estimates are also shaped by other evolutionary factors than population history alone (Figure 9).

In previous studies the extended LD within the CYP2C region has shown to be conserved across populations, although the particular LD block structure is still under debate (Ahmadi et al. 2005, Walton et al. 2005, Vormfelde et al. 2007). Similarly, our data show a clear extended LD across the CYP2C region (Figure 10A) but with a varying pattern of adjacent marker r^2 values across the populations. Moreover, the LD block structure is consistent across populations within CYP2C region with few exceptions (Figure 10A). Firstly, the sub-Saharan Mandenka population shows significantly the shortest LD blocks (Figure 10 A) with lowest proportion of high LD (study IV, Table 1), which is due to the longer period of recombination compared to European populations who migrated out of Africa and experienced a demographic bottleneck. Secondly, the CEPH sample shows a different pattern of LD block distribution (Figure 10A) and the lowest proportion of high LD among Europeans (study IV, Table 1), while the Saami exhibit the highest proportion of high LD, significantly different adjacent marker r^2 values and the longest single LD block among all populations (Figure 10A). Moreover, the decay of high LD within the CYP2C extended LD region is slowest in the Saami and fastest in the Mandenka sample (or CEPH sample within Europe) (Figure 11A).

Within the CYP2D region also a clear extended LD region is observed (Figure 10B), although the adjacent marker r^2 values and the LD block distribution are more heterogenous compared to CYP2C region. The sub-Saharan Mandenka show significantly the shortest LD blocks and the lowest proportion of high LD while the Saami show again the highest proportion of high LD (study IV, Table 1), significantly different pattern of adjacent marker r^2 values and the longest single LD block (Figure 10B). The decay of LD within CYP2D region shows the slowest decay in the Saami, but this is only seen in distances below 50kb as thereafter the decay is much faster in the Saami and Finns compared to other population. It is also noteworthy that the Saami possess also the lowest number of CYP2C and CYP2D haplotypes but higher CYP2D haplotype diversity compared to most other Europeans (study IV, Table 1).

The Saami population has remained small and constant in size throughout its history and is considered to have an admixed West and East Eurasian genetic origin (Ross et al. 2006). The high and extended LD with fluctuating haplotype diversity in the Finno-Ugric-speaking Saami could be linked to the small and constant effective population size with admixture and/or subsequent long-term genetic drift shaping the genetic diversity and LD compared to other European populations (Ross et al. 2006). Both admixture and drift are shown to generate enhanced but random patterns of genetic diversity and LD (Terwilliger et al. 1998, Ardlie et al. 2002).

Moreover, a CYP2C19*2 A_{-19154} mutation causing reduced activity for the CYP2C19 DME shows significantly higher allele frequency in Saami (95% CI: 28.1–47.4%) compared to other European populations (95% CI: 13.4–16.0%; Xie et al. 1999). Interestingly, similar high frequencies (≥ 0.30) of the altered activity allele are observed in Central and East Asia (Xiao et al. 1997, Kimura et al. 1998, Wedlund et al. 2000, Niu et al. 2004). This, assuming an East Eurasian origin of the CYP2C19*2 A_{-19154} mutation indicates significant Asia

contribution to the Saami gene pool similarly as observed in HLA markers (Johansson et al. 2008), although the combined effect of selection and drift may have enhanced the level of A_{-19154} allele frequency.

The Finno-Ugric-speaking Saami, based on the lower level recombination and higher extended LD within the CYP2C and CYP2D regions exhibit an optimal genetic structure to tag haplotypes more efficiently with larger genomic coverage compared to other populations analyzed. This strategy of LD-based drift mapping originally proposed by Terwilliger et al (1998) may offer a great advantage for further identification of alleles associated to common complex pharmacogenetic traits.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

The major aim of this thesis was to examine the origins and distribution of uniparental and autosomal genetic variation among the Finno-Ugric-speaking human populations living in Boreal and Arctic regions of North Eurasia. In more detail, I aimed to disentangle the underlying molecular and population genetic factors which have produced the patterns of uniparental and autosomal genetic diversity in these populations. Among Finno-Ugrics the genetic amalgamation and clinal distribution of West and East Eurasian gene pools were observed within uniparental markers. This admixture indicates that North Eurasia was colonized through Central Asia/ South Siberia by human groups already carrying both West and East Eurasian lineages. The complex combination of founder effects, gene flow and genetic drift underlying the genetic diversity of the Finno-Ugric-speaking populations were emphasized by low haplotype diversity within and among uniparental and biparental markers. A high prevalence of lactase persistence allele among the North Eurasian Finno-Ugric agriculturalist populations was also shown indicating a local adaptation to subsistence change with lactose rich diet. Moreover, the haplotype background of lactase persistence allele among the Finno-Ugric-speakers strongly suggested that the lactase persistence T_{-13910} mutation was introduced independently more than once to the North Eurasian gene pool. A significant difference in genetic diversity, haplotype structure and LD distribution within the cytochrome P450 CYP2C and CYP2D regions revealed the unique gene pool of the Finno-Ugric Saami created mainly by population genetic processes compared to other Europeans and sub-Saharan Mandenka population. From all studied populations the Saami showed also significantly the highest allele frequency of a CYP2C19 gene mutation causing variable drug reactions. The diversity patterns observed within CYP2C and CYP2D regions emphasize the strong effect of demographic history shaping genetic diversity and LD especially among such small and constant size populations as the Finno-Ugric-speaking Saami. Moreover, the increased LD in Saami due to genetic drift and/or admixture was shown to offer an advantage for further attempts to identify alleles associated to common complex pharmacogenetic traits.

A challenge in future studies of human genome variation is to understand the molecular basis of common complex diseases, and variable sensitivity to drugs, pathogens and other environmental factors when recent developments in genotyping and genomic resequencing have enabled the high-throughput genome-wide studies such as the HapMap (International HapMap Consortium 2007) and 1000 Genomes (Kaiser 2008). These studies aim primarily at validating genetic variation without ascertainment bias, and secondarily to explore the evolutionary factor shaping genetic diversity. Both large scale genotyping projects and fine-scale resequencing studies of restricted genome regions assessed in different populations are needed for further refinements of the recombination hotspot and LD block structures within the haplotype map of the human genome. A solid understanding of the human genomic variation and haplotype structure within will enable further determination of our evolutionary past and enhance the identification of genetic variants underlying common complex traits and diseases.

6 ACKNOWLEDGEMENTS

This study was carried out between 2002–2008 in the Department of Forensic Medicine at the University of Helsinki and as a visiting scientist in the Evolutionary Biology Unit at the University of Pompeu Fabra along with a one and a half year vacation fulfilling the National Military Service.

The study was financially supported by The Finnish Cultural Foundation, the Federation of European Biochemical Societies and the Finnish Graduate School in Population Genetics along with grants from the EU and the University of Helsinki.

I wish to thank the former head of the Forensic Department, Professor Antti Penttilä, and the new director Professor Erkki Vuori, for providing the excellent research facilities with great academic atmosphere. I also wish to thank Professor Jaume Bertranpetit for inviting me to visit the inspiring Evolutionary Biology Unit at the University of Pompeu Fabra.

The deepest gratitude I wish to express to my supervisor Professor Antti Sajantila at the Department of Forensic Medicine. When I finally managed to meet up with the world famous Finno-Ugric professor I really got excited. Your personal enthusiasm and deep knowledge concerning North Eurasian Finno-Ugric populations and human population genetics in general hooked me. Since then I have learned a lot from you about how to conduct scientific work. Your broad understanding and interest in science, genetics and life itself have constantly carried on and stimulated my own sometimes restless and narrow mind. Your patience with me has been unbelievable, especially when things have gone wrong. I have been very lucky to have had the opportunity to work under your supervision.

I am also greatly indebted to my second supervisor Professor David Comas for a chance to work with him in ever enjoyable atmosphere. Your valuable advice and broad knowledge in human population genetics have kept me on a right track. Moreover, your never ending good humour and social skills to survive with the whining PhD student have always impressed me.

Professor Ulf Gyllensten and Professor Pekka Pamilo deserve enormous compliments for carefully reviewing my thesis during their summer holiday season and for their valuable comments. I am also indepted to Professor Kimmo Kontula for important advice to conduct my final years in PhD studies.

It has always been inspiring and challenging to work in the OLL-BIO laboratory at the Department of Forensic Medicine. I have learned everything about genotyping and forensic laboratory work during these years in Kytösuontie 11. Especially I want to thank my colleague Minttu Hedman, with whom I have not only shared an office and scientific papers but also moments of scientific joy while filling bureaucratic applications or glasses of wine. Jukka Palo is also greatly thanked for revising most of my scientific writings and restlessly explaining to me the basics of population genetics. I also want to express my sincere gratitude to the rest of the former or current oll-bio members: Antti L, Hanna, Silvia, Johanna, Anna, Yukiko, Eve, Teija, Kirsti, Pia, Helmuth, Katarina, Hannu and Mikko. Thank you for all the great moments and good coffee breaks.

I also had a great pleasure to take part in the LD-EUROPE project and work with great scientists: Laurent Excoffier, Gillaume Lavall, Howard Cann, Sir Walter Bodmer, Susan Tonks, Irina Evseeva, Al-

berto Piazza, Francesca Crobu, Silvana Santachiara-Benerecetti, Ornella Semino, Anna Gonzalez-Neira and Claudia de Toma. Thank you for your collaboration.

I also wish to thank all the bioevo-people in Pompeu Fabra with whom I had the opportunity to enjoy scientifically inspiring moments around the coffee machine—and obviously all the GFs in Bitacora: Monica, Roger, Stephanie, Ferran, Ixa, Urko, Elodie, Hafid, Araceli, Chiara, Andrés, Karla, Michelle, Gemma, Carles, Josep, Ricardo, Marta, Belen, Anna F, Anna R, Olga, Rui, Begoña, Oscar, Elena, Fracesc and Arcadi. Thank you and hope to see you soon.

Special BCN thanks go to Martin and Renia for being superkind hosts for homeless Finns and always being eager to share common weekend adventures. Bruno and Jordi. It's always a joy even just to talk with you and enjoy life while watching operación triumfo.

Numerous friends in Capoeira Forca Natural are greatly acknowledged for various relaxing moments and aching muscles outside the scientific world. Antti Korpisaari and Jussi Korhonen, thank you for your archaeological inspiration and intelligent company while digging treasures in eastern Finland. Risto and Pyry, you are greatly acknowledged for keeping an old guy sane in the Finnish forest. Juha and Aki, the two Hakunila musketeers, thank you for sharing fantastic journeys and private discussions. Gerald Matei (1974–2007), your never ending energy to get people to smile and enjoy life are greatly missed. And for Sampsa and Petri – thank you for your friendship. I also wish to thank all relatives. Especially Konsta, your visual eye and hilarious humour is more than greatly thanked during the printing of this thesis.

A warmest gratitude and sincere thanks are due to my family. Dear Heli and Georg. Thank you for all your mental and physical care. And do not worry, as your prodigal son will return. Vilma and Wander. Thank you for all the late dinners and great moments, which have kept me going.

Agnes "Oma", dear grandmother (1910–2008). I wish you would be here to share a cup of tea with lemon.

7 REFERENCES

Abondolo D (1998) The Uralic Languages. London and New York

Ahmadi KR, Weale ME, Xue ZY, Soranzo N, Yarnall DP, Briley JD, Maruyama Y, Kobayashi M, Wood NW, Spurr NK, Burns DK, Roses AD, Saunders AM, Goldstein DB (2005) A single-nucleotide polymorphism tagging set for human drug metabolism and transport. Nat Genet 37:84–89

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

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309

Armour JA, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, Bertranpetit J, Paabo S, Jeffreys AJ (1996) Minisatellite diversity supports a recent african origin for modern humans. Nat Genet 13:154–160

Arnheim N, Calabrese P, Nordborg M (2003) Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. Am J Hum Genet 73:5–16

Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150–1160

Bandelt HJ, Richards M, Macaulay V (eds) (2006) Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer, Berlin Heidelberg New York

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

Beck S, Trowsdale J (2000) The human major histocompatability complex: Lessons from the DNA sequence. Annu Rev Genomics Hum Genet 1:117–137

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, Erhardt G (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet 35:311–313

Bergman I (2004) Deglaciation and colonization: Pioneer settlements in Northern Fennoscandia. Journal of World Prehistory:155–177

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (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

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298–1311

Bertranpetit J, Calafell F, Comas D, Gonzalez-Neira A, Navarro A (2003) Structure of linkage disequilibrium in humans: Genome factors and population stratification. Cold Spring Harb Symp Quant Biol 68:79–88

Bosch E, Calafell F, Gonzalez-Neira A, Flaiz C, Mateu E, Scheil HG, Huckenbeck W, Efremovska L, Mikerezi I, Xirotiris N, Grasa C, Schmidt H, Comas D (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann Hum Genet 70:459–487

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408–1415

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J et al. (2002) A human genome diversity cell line panel. Science 296:261–262

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

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452

Carroll SB (2003) Genetics and the making of Homo sapiens. Nature 422:849–857

Cavalli-Sforza LL (2005) The human genome diversity project: Past, present and future. Nat Rev Genet 6:333–340

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

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King MC (1991) Call for a worldwide survey of human genetic diversity: A vanishing opportunity for the human genome project. Genomics 11:490–491

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U.S.A. 85:9119–9123

Chi PB, Duggal P, Kao WH, Mathias RA, Grant AV, Stockton ML, Garcia JG, Ingersoll RG, Scott AF, Beaty TH, Barnes KC, Fallin MD (2006) Comparison of SNP tagging methods using empirical data: Association study of 713 SNPs on chromosome 12q14.3–12q24.21 for asthma and total serum IgE in an African Caribbean population. Genet Epidemiol 30:609–619

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Clark AG (2004) The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333

Collins FS, Green ED, Guttmacher AE, Guyer MS, US National Human Genome Research Institute (2003) A vision for the future of genomics research. Nature 422:835–847

Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, Calafell F, Bertranpetit J (2004) Admixture, migrations, and dispersals in central asia: Evidence from maternal DNA lineages. Eur J Hum Genet 12:495–504

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

Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at all. Science 289:1139

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706

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

Darwin CR (1859) The Origin of Species by Means of Natural Selection. John Murray, London.

de Knijff P (2000) Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet 67:1055–1061

de la Chapelle A, Wright FA (1998) Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. Proc Natl Acad Sci U.S.A.95:12416–12423

Deloukas P, Bentley D (2004) The HapMap project and its application to genetic studies of drug response. Pharmacogenomics J 4:88–90

Denoeud F, Vergnaud G, Benson G (2003) Predicting human minisatellite polymorphism. Genome Res 13:856–867

Derbeneva OA, Starikovskaya EB, Wallace DC, Sukernik RI (2002) Traces of early Eurasians in the Mansi of Northwest Siberia revealed by mitochondrial DNA analysis. Am J Hum Genet 70:1009–1014

Derenko M, Malyarchuk B, Denisova GA, Wozniak M, Dambueva I, Dorzhu C, Luzina F, Miscicka-Sliwka D, Zakharov I (2006) Contrasting patterns of Y-chromosome variation in South Siberian populations from Baikal and Altai-Sayan regions. Hum Genet 118(5):591–604

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

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee HK, Vanecek T, Villems R, Zakharov I (2007b) Phylogeographic analysis of mitochondrial DNA in Northern Asian populations. Am J Hum Genet 81:1025–1041

Derenko MV, Grzybowski T, Malyarchuk BA, Dambueva IK, Denisova GA, Czarny J, Dorzhu CM, Kakpakov VT, Miscicka-Sliwka D, Wozniak M, Zakharov IA (2003) Diversity of mitochondrial DNA lineages in South Siberia. Ann Hum Genet 67:391–411

Ding K, Zhou K, Zhang J, Knight J, Zhang X, Shen Y (2005) The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. Mol Biol Evol 22:148–159

Dolukhanov PM, Shukurov AM, Tarasov PE, Zaitseva GI (2002) Colonization of Northern Eurasia by modern humans: Radiocarbon chronology and environment. Journal of Archaeological Science 29:593–606

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11:2571–2581

Ellegren H (2004) Microsatellites: Simple sequences with complex evolution. Nat Rev Genet 5:435–445

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237

Evans DM, Cardon LR (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet 76:681–687

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47–50

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

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

Fisher RA (1930) The Genetical Theory of Natural Selection. Clarendon Press, Oxford, Oxford University Press, Oxford

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

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789–795

Goebel T, Derevianko AP, Petrin VT (1993) Dating the middle-to-upper-paleolithic transition at Kara-Bom. Curr Anthropol 34:452–458

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111

Goldstein DB, Cavalleri GL (2005) Genomics: Understanding human diversity. Nature 437:1241–1242

Gonzalez-Neira A, Ke X, Lao O, Calafell F, Navarro A, Comas D, Cann H, Bumpstead S, Ghori J, Hunt S, Deloukas P, Dunham I, Cardon LR, Bertranpetit J (2006) The portability of tagSNPs across populations: A worldwide survey. Genome Res 16:323–330

Greely HT (2001a) Human genome diversity: What about the other human genome project? Nat Rev Genet 2:222–227

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

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444:330–336

Greenberg JH (2000) Indo-European and its Closest Relatives; the Eurasiatic Language Family. Stanford, Stanford University Press Volume 1, Grammar

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

Haldane JBS (1924) A mathematical theory of natural and artificial selection. Transactions of the Cambridge philosophical society Part I:19–41

Halldorsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S (2004) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res 14:1633–1640

Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. Nat Genet 2:204–211

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

Hewitt GM (1999) Post-glacial recolonization of European biota. Biological Journal of the Linnean Society:87–112

Hewitt GM (2001) Speciation, hybrid zones and phylogeography – or seeing genes in space and time. Mol Ecol 10:537–549

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54:705–714

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM (2001) Lactase haplotype diversity in the old world. Am J Hum Genet 68:160–172

Howell N, Oostra RJ, Bolhuis PA, Spruijt L, Clarke LA, Mackey DA, Preston G, Herrnstadt C (2003) Sequence analysis of the mitochondrial genomes from Dutch pedigrees with Leber hereditary optic neuropathy. Am J Hum Genet 72:1460–1469

Ingelman-Sundberg M (2004) Human drug metabolising cytochrome P450 enzymes: Properties and polymorphisms. Naunyn Schmiedebergs Arch Pharmacol 369:89–104

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

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, Swallow DM (2007) A novel polymorphism associated with lactose tolerance in Africa: Multiple causes for lactase persistence? Hum Genet 120:779–788

International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796

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

International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222

Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet 31:267–271

Jeffreys AJ, Ritchie A, Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum Mol Genet 9:725–733

Jobling MA, Hurles M, Tyler-Smith C (2003) Human evolutionary genetics: Origins, peoples and disease. Garland Science

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110:118–124

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: An evolutionary marker comes of age. Nat Rev Genet 4:598–612

Johansson A, Ingman M, Mack SJ, Erlich H, Gyllensten U (2008) Genetic origin of the Swedish Sami inferred from HLA class I and class II allele frequencies. Eur J Hum Genet (in press)

Johansson A, Vavruch-Nilsson V, Cox DR, Frazer KA, Gyllensten U (2007) Evaluation of the SNP tagging approach in an independent population sample--array-based SNP discovery in Sami. Hum Genet 122:141–150

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

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435–1444

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci U S A 94:3100–3103

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

Kaessmann H, Paabo S (2002) The genetical history of humans and the great apes. J Intern Med 251:1–18

Kaessmann H, Zollner S, Gustafsson AC, Wiebe V, Laan M, Lundeberg J, Uhlen M, Paabo S (2002) Extensive linkage disequilibrium in small human populations in Eurasia. Am J Hum Genet 70:673–685

Kaiser J (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. Science 319:395

Karafet TM, Osipova LP, Gubina MA, Posukh OL, Zegura SL, Hammer MF (2002) High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum Biol 74:761–789

Kauppi L, Sajantila A, Jeffreys AJ (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Hum Mol Genet 12:33–40

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA, Sajantila A, Tyler-Smith C (2004) A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet 74:1183–1197

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580–1588

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

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

Khusnutdinova EK, Viktorova TV, Fatkhlislamova RI, Khidiatova IM (1999) Restriction polymorphism of the major non-decoding region of mitochondrial DNA in human populations from the Volga-Ural region. Genetika 35:695–702

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

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M, Ieiri I, Mamiya K, Urae A, Higuchi S (1998) Genetic polymorphism of cytochrome P450s, CYP2C19, and CYP2C9 in a Japanese population. Ther Drug Monit 20:243–247

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, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171–1179

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19:1737–1751

Kolga M, Tõnurist I, Vaba L, Viikberg J (2001) The Red Book of the Peoples of the Russian Empire. Tallinn, NGO Red Book

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Krebs CJ (1994) Ecology: The experimental analysis of distribution and abundance.

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

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. Am J Hum Genet 80:727–739

Kutuev I, Khusainova R, Karunas A, Yunusbayev B, Fedorova S, Lebedev Y, Hunsmann G, Khusnutdinova E (2006) From east to west: Patterns of genetic diversity of populations living in four Eurasian regions. Hum Hered 61:1–9

Kuzmin YV, Keates SG: Comment on "Colonization of Northern Eurasia by Modern Humans: Radiocarbon Chronology and Environment" by P.M. Dolukhanov, A.M. Shukurov, P.E. Tarasov and G.I. Zaitseva. Journal of Archaeological Science 29, 593–606 (2002). Journal of Archaeological Science, 2004; 31: 141–143

Laakso J (1992) Uralilaiset kansat. Tietoa suomen sukukielistä ja niiden puhujista. WSOY, Juva

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

Laan M, Wiebe V, Khusnutdinova E, Remm M, Paabo S (2005) X-chromosome as a marker for population history: Linkage disequilibrium and haplotype study in Eurasian populations. Eur J Hum Genet 13:452–462

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

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

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

Landsteiner K (1931) Individual differences in human blood. Science 73:403–409

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe.Curr Biol. 18:1241–1248

Lappalainen T, Koivumaki S, Salmela E, Huoponen K, Sistonen P, Savontaus ML, Lahermo P (2006) Regional differences among the Finns: A Y-chromosomal perspective. Gene 376:207–215

Lewis DF (2004) 57 varieties: The human cytochromes P450. Pharmacogenomics 5:305–318

Lewontin RC (1964) The interaction of selection and linkage. I. general considerations; heterotic models. Genetics 49:49–67

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, Narang M, Barcelo J, O'Hoy K (1992) Myotonic dystrophy mutation: An unstable CTG repeat in the 3' untranslated region of the gene. Science 255:1253–1255

May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ (2002) Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nat Genet 31:272–275

McVean G, Spencer CC, Chaix R (2005) Perspectives on human genetic variation from the HapMap project. PLoS Genet 1:e54

Meinila M, Finnila S, Majamaa K (2001) Evidence for mtDNA admixture between the finns and the saami. Hum Hered 52:160–170

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R (2004) Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 5:26

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841–4843

Mir KU, Southern EM (2000) Sequence variation in genes and genomic DNA: Methods for large-scale analysis. Annu Rev Genomics Hum Genet 1:329–360

Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76:387–398

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324

Myles S, Bouzekri N, Haverfield E, Cherkaoui M, Dugoujon JM, Ward R (2005) Genetic evidence in support of a shared Eurasian-North African dairying origin. Hum Genet 117:34–42

Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. Trends Genet 17:481–485

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304

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

Nevanlinna HR (1984) The roots of the Finns in the light of genetic research into distinguishing genetic characteristics. Soc Sci Fennica:157–174

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

Niu CY, Luo JY, Hao ZM (2004) Genetic polymorphism analysis of cytochrome P4502C19 in Chinese Uigur and Han populations. Chin J Dig Dis 5:76–80

Nordqvist B (2000) Coastal adaptations in the mesolitic. A study of costal sites with organic remains from the Boreal and Atlantic periods in Western Sweden. GOTARC Series B 13

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

Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A 99:16134–16137

Ota T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201–204

Paabo S (1989) Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci U S A 86:1939–1943

Palo JU, Hedman M, Ulmanen I, Lukka M, Sajantila A (2007) High degree of Y-chromosomal divergence within Finland – forensic aspects. Forensic Sci Int Genet 1:120–124

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Pavlov P, Svendsen JI, Indrelid S (2001) Human presence in the European Arctic nearly 40,000 years ago. Nature 413:64–67

Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. Nat Rev Genet 1:182–190

Perheentupa J (1995) The Finnish disease heritage: A personal look. Acta Paediatr 84:1094–1099

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387

Pimenoff V, Korpisaari A (2004) A preliminary report on the genetic analysis of the osteological remains of Tiwanaku tombs in Tiraska and Aymara chullpa C17 in kewaya. Report submitted to DINAAR (National Directorate of Archaeology and Anthropology, Bolivia).

Pimenoff V, Sajantila A (2002) Genetic Tools in Molecular Anthropology and Archaeology. The Roots of Peoples and Languages of the Northern Eurasia IV Ed. K. Julku, Gummerus OY. Jyväskylä, Finland.

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

Pult I, Sajantila A, Simanainen J, Georgiev O, Schaffner W, Paabo S (1994) Mitochondrial DNA sequences from switzerland reveal striking homogeneity of european populations. Biol Chem Hoppe Seyler 375:837–840

Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K (2004) Where west meets east: The complex mtDNA landscape of the southwest and central asian corridor. Am J Hum Genet 74:827–845

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

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H et al. (2006) Global variation in copy number in the human genome. Nature 444:444–454

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

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

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C et al. (2000) Tracing european founder lineages in the near eastern mtDNA pool. Am J Hum Genet 67:1251–1276

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602–612

Rootsi S, Zhivotovsky LA, Baldovic M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, Ilumae AM, Khusnutdinova EK, Voevoda MI, Osipova LP, Stoneking M, Lin AA, Ferak V, Parik J, Kivisild T, Underhill PA, Villems R (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from southeast asia towards europe. Eur J Hum Genet 15:204–211

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

Ross AB, Johansson A, Ingman M, Gyllensten U (2006) Lifestyle, genetics, and disease in sami. Croat Med J 47:553–565

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W 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

Ruano G, Kidd KK, Stephens JC (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Proc Natl Acad Sci U S A 87:6296–6300

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312:1614–1620

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918

Sahi T, Isokoski M, Jussila J, Launiala K, Pyorala K (1973) Recessive inheritance of adult-type lactose malabsorption. Lancet 2:823–826

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S (2000) mtDNA variation among greenland eskimos: The edge of the beringian expansion. Am J Hum Genet 67:718–726

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, Beckman L, Tranebjaerg L, Gedde-Dahl T, Issel-Tarver L, DiRienzo A, Paabo S (1995) Genes and languages in europe: An analysis of mitochondrial lineages. Genome Res 5:42–52

Sajantila A, Lukka M, Syvanen AC (1999) Experimentally observed germline mutations at human micro- and minisatellite loci. Eur J Hum Genet 7:263–266

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

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of paleolithic homo sapiens sapiens in extant europeans: A Y chromosome perspective. Science 290:1155–1159

Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38:556–560

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

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity 76:377–383

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stumpf MP, Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13:1–8

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. Annu Rev Genet 37:197–219

Syvanen AC (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942

Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogvali EL, Tolk HV 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

Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the "fundamental theorem of the HapMap". Eur J Hum Genet 14:426–437

Terwilliger JD, Zollner S, Laan M, Paabo S (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum Hered 48:138–154

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36:S21–7

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in africa and europe. Nat Genet 39:31–40

Tishkoff SA, Verrelli BC (2003) Role of evolutionary history on haplotype block structure in the human genome: Implications for disease mapping. Curr Opin Genet Dev 13:569–575

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, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major late paleolithic population expansion from southwestern to northeastern europe. Am J Hum Genet 62:1137–1152

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V et al. (2001) A signal, from human mtDNA, of postglacial recolonization in europe. Am J Hum Genet 69:844–852

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L (2000) Linkage disequilibrium in isolated populations: Finland and a young sub-population of kuusamo. Eur J Hum Genet 8:604–612

Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of finnish populations with different histories. Hum Mol Genet 12:51–59

Varilo T, Savukoski M, Norio R, Santavuori P, Peltonen L, Jarvela I (1996) The age of human mutation: Genealogical and linkage disequilibrium analysis of the CLN5 mutation in the finnish population. Am J Hum Genet 58:506–512

Vasil'ev SA, Kuzmin YV, Orlova LA, Dementiev VN (2002) Radiocarbon-based chronology of the paleolithic of siberia and its relevance to the peopling of the new world. Radiocarbon 44(2):503–530

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO et al. (2001) The sequence of the human genome. Science 291:1304–1351

Vilkki J, Savontaus ML, Nikoskelainen EK (1988) Human mitochondrial DNA types in finland. Hum Genet 80:317–321

Vormfelde SV, Schirmer M, Toliat MR, Meineke I, Kirchheiner J, Nurnberg P, Brockmoller J (2007) Genetic variation at the CYP2C locus and its association with torsemide biotransformation. Pharmacogenomics J 7:200–211

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597

Walton R, Kimber M, Rockett K, Trafford C, Kwiatkowski D, Sirugo G (2005) Haplotype block structure of the cytochrome P450 CYP2C gene cluster on chromosome 10. Nat Genet 37:915–6; author reply 916

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

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234

Waples RS, Gaggiotti O (2006) What is a population? an empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol 15:1419–1439

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB (2003) Genetic variation among world populations: Inferences from 100 alu insertion polymorphisms. Genome Res 13:1607–1618

Weber JL (1990) Human DNA polymorphisms and methods of analysis. Curr Opin Biotechnol 1:166–171

Wedlund PJ (2000) The CYP2C19 enzyme polymorphism. Pharmacology 61:174–183

Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19–24

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L et al. (2001) The eurasian heartland: A continental perspective on Y-chromosome diversity. Proc Natl Acad Sci U S A 98:10244–10249

Wilkinson GR (2005) Drug metabolism and variability among patients in drug response. N Engl J Med 352:2211–2221

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

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Xiao ZS, Goldstein JA, Xie HG, Blaisdell J, Wang W, Jiang CH, Yan FX, He N, Huang SL, Xu ZH, Zhou HH (1997) Differences in the incidence of the CYP2C19 polymorphism affecting the S-mephenytoin phenotype in chinese han and bai populations and identification of a new rare CYP2C19 mutant allele. J Pharmacol Exp Ther 281:604–609

Xie HG, Stein CM, Kim RB, Wilkinson GR, Flockhart DA, Wood AJ (1999) Allelic, genotypic and phenotypic distributions of S-mephenytoin 4'-hydroxylase (CYP2C19) in healthy caucasian populations of european descent throughout the world. Pharmacogenetics 9:539–549

Xu X, Peng M, Fang Z (2000) The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24:396–399

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

Zeggini E, Barton A, Eyre S, Ward D, Ollier W, Worthington J, John S (2005) Characterisation of the genomic architecture of human chromosome 17q and evaluation of different methods for haplotype block definition. BMC Genet 6:21

Zerjal T, Beckman L, Beckman G, Mikelsaar AV, Krumina A, Kucinskas V, Hurles ME, Tyler-Smith C (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, Kayser M, Roewer L, Santos FR, Schiefenhovel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of asians and northern europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into central asia. Am J Hum Genet 71:466–482

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE (2002) Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci U S A 99:17004–17007

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186