DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
115

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
115

THE COMPARATIVE PATTERNS OF LINKAGE DISEQUILIBRIUM IN EUROPEAN POPULATIONS AND ITS IMPLICATION FOR GENETIC ASSOCIATION STUDIES

ELIN LÕHMUSSAAR

TARTU UNIVERSITY
PRESS

Institute of Molecular and Cell Biology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in molecular biomedicine) on 07.12.2005, by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Opponent: Aarno Palotie, Prof., PhD, Helsingi Ülikool, Soome

Commencement: Room No 217, Riia 23, Tartu, on January 19th, at 13.00

The publication of this dissertation is granted by the University of Tartu

ISBN 9949–11–221–4 (trükis)
ISBN 9949–11–222–2 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 610. 2005

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications referred to in the text by their Roman numbers:

I. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, **Lõhmussaar E,** Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544–8.

II. Mueller JC. *, **Lõhmussaar E*,** Mägi 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 and Meitinger T. (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. *American Journal of Human Genetics* 76, 387–98.
* authors contributed equally to this work

III. Binder EB, Salyakina D, Lichtner P, Wochnik GM, Ising M, Putz B, Papiol S, Seaman S, Lucae S, Kohli MA, Nickel T, Kunzel HE, Fuchs B, Majer M, Pfennig A, Kern N, Brunner J, Modell S, Baghai T, Deiml T, Zill P, Bondy B, Rupprecht R, Messer T, Kohnlein O, Dabitz H, Bruckl T, Muller N, Pfister H, Lieb R, Mueller JC, **Lõhmussaar E,** Strom TM, Bettecken T, Meitinger T, Uhr M, Rein T, Holsboer F, Muller-Myhsok B. (2004) Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment *Nature Genetics* 36, 1319–25.

IV. **Lõhmussaar E,** Gschwendtner A, Mueller JC., Org T, Wichmann E, Hamann G, Meitinger T and Dichgans M. (2005) The *ALOX5AP* gene and the *PDE4D* gene in a Central-European population of stroke patients. *Stroke* 36, 731–6.

My contribution to the articles referred in the current thesis is as follows:

Ref. I designed and performed the experiments for Estonian samples, participated in analysis of data and preparation of the manuscript

Ref. II designed and performed the experiments, analysed the experimental data and participated in writing of the paper

Ref. III designed and performed the experiments of the LD analysis part of the study, performed the analysis of LD data

Ref. IV designed and performed the experiments, analysed the experimental data and writing of the paper

LIST OF ABBREVIATIONS

ALOX5AP	arachidonate 5-lipoxygenase activating protein
APEX	Arrayed Primer Extension
ApoE	apolipoprotein E
bp/kb/Mb	base pair (s)/kilo base pairs/mega base pairs
cAMP	cyclic adenosine monophosphate
cSNP	coding Single Nucleotide Polymorphism
CARD15	caspase recruitment domain-containing protein 15
CD/CV	Common Disease/Common Variant
CEPH	Centre d'Etude du Polymorphisme Humain
CLPS	colipase
ENCODE	Encyclopedia of DNA elements
FKBP5	FK-506 binding protein 5
FLAP	5-lipoxygenase activating protein
G6PD	glucose 6-phosphate dehydrogenase
GR	glucocorticoid receptor
GWA	genome-wide association
HPA	hypothalamic-pituitary-adrenal
htSNP	haplotype tagging SNP
KORA	Cooperative Health Research in the Region of Augsburg
LD	linkage disequilibrium
LMNA	lamin isoforms A and C
LTA4	leucotriene A4
LTB4	leucotriene B4
LTC4	leucotriene C4
MAF	minor allele frequency
MI	myocardial infarction
mRNA	messenger ribonuclei acid
PCR	polymerase chain reaction
PDE4D	phosphodiesterase 4D
PLAU	plasminogen activator, urinary
POPGEN	population sample collected in Schleswig-Holstein, Germany
PPARγ	peroxisome proliferator-activated receptor γ
SHIP	Study on Health in Pomerania
SNCA	synuclein, alpha
SNP	Single Nucleotide Polymorphism
STRK1	stroke susceptibility to, 1
STRP	Short Tandem Repetitive Polymorphism
tagSNP	tagging SNP
TDT	transmission/disequilibrium test

INTRODUCTION

The human genome contains a large amount of individual differences in DNA sequence, which largely determine the functional and phenotypic variability between individuals. Knowledge about the genetic basis of human variability provides the opportunity to identify the causes of human diseases. Recent progresses, such as vast improvement of genotyping technologies and finishing of the HapMap Project (Altshuler et al. 2005), have made the task of finding the genes responsible for common complex human diseases a realistic undertaking in coming decade. Identifying the functional genes and causal variants underlying the pathogenesis of disease would be the first step towards improving prevention, diagnosis and treatment of disease.

Risch and Merikangas proposed a decade ago that population-based genetic association studies would be the most effective strategy for dissecting the genetic basis of complex diseases (Risch and Merikangas 1996). The principal genetic targets for association studies are SNPs, the most common form of variation in the human genome. The association between disease and causal variant can be investigated through correlation patterns among nearby variants (known as linkage disequilibrium or LD). The pattern of LD shows high variability across the genome, but is relatively conserved among different human populations. These patterns are reflecting a complex interplay between recombination and the population's demographic and evolutionary history. A large proportion of the human genome is organized in regions of high LD and low haplotype diversity. Understanding the detailed structure of LD patterns have given us a great opportunity to select the optimal set of SNPs and design whole genome based genetic association studies. Although theoretical and empirical studies have improved our knowledge about genetic association studies, we are still in the beginning of understanding the causal links between genetic factors and disease risk in patients. However, only now we have the proper tools (advanced information of LD, population based samples and technology) in order to be successful.

The first part of the present thesis gives an overview of LD and haplotype structure in the human genome and aspects about designing, performing and analyzing the genetic association studies of complex diseases. The research part of this dissertation entails the following areas; (i) characterizing the LD and haplotype structure in different regions of the human genome, (ii) investigating the LD and haplotype variability among European populations, (iii) evaluating the performance and transferability of selected tagSNPs among populations, and (iiii) evaluation of two candidate genes involved in stroke development by a case-control based association study.

1. REVIEW OF LITERATURE

1.1. Linkage disequilibrium (LD) and haplotype structure in the human genome

1.1.1. Variations in the human genome

Most of the human genome sequence is identical between any two individuals and variations in it contribute to phenotypic differences, including susceptibility to or protection against diseases. As the general mutation rate of the mammalian genome is low (on average 2 x 10^{-9} per base pair per year) the majority of inter-individual genetic variability is inherited (Kumar and Subramanian 2002). Several types of genetic variations exist in the human genome, ranging from a single base pair to thousands of base-pairs in size: single nucleotide polymorphisms (SNPs), repeat polymorphisms (minisatellites and microsatellites), small insertions or deletions (indels) and copy number polymorphisms. More than two decades ago it was recognized that different variations (or polymorphisms) in human DNA could be effectively used as genetic markers in the search for genetic factors underlying human diseases (Botstein et al. 1980). Microsatellies or short tandem repetitive polymorphisms (STRPs) and single base changes in DNA or SNPs are the most commonly used markers for gene mapping because of their abundance.

Microsattellites are fast evolving markers, having a moderate to high mutation rate (usually 10^{-5}–10^{-2} per generation) and a high degree of heterozygosity (Weber and Wong 1993; Chakraborty et al. 1997). Microsatellites are thought to mutate via the "stepwise" gain or loss of single-repeat units, although larger "jumps" in repeat size occasionally do occur (Valdes et al. 1993; Ellegren 2004). Their high information content due to the high number of alleles make them ideal markers for pedigree-based linkage analysis and have lead to the identification of genes involved in many monogenic diseases and some polygenic diseases (Ellegren 2004).

The most common form of DNA variation in the human genome is SNPs, making up about 90% of all human genetic variations. In the human genome SNPs occur on average once per 300 bp, but the density varies up to ten fold between different regions of the genome (Kruglyak and Nickerson 2001; Sachidanandam et al. 2001). The interest in SNPs has been increased by the progress made with the sequencing of the human genome. This, together with the rapid improvement of genotyping technologies (Syvanen 2005), enabled the identification of a large number of SNP sites (Sachidanandam et al. 2001; Venter et al. 2001). Currently, the public SNP database (dbSNP) includes over 10 million human reference SNPs (http://www.ncbi.nlm.nih.gov/SNP build 125, 2005). Since the genetic variability in the human genome is relatively limited, most genes have only a handful of common variants in their coding regions and

the vast majority of alleles are exceedingly rare (Lander 1996). It has been estimated that out of 11–15 million existing SNPs about 7 million are common around the world, with a minor allele frequency of at least 5% (Kruglyak and Nickerson 2001; Salisbury et al. 2003; Miller et al. 2005). Although the frequency of any allele may vary considerably between populations, the most common SNPs are found in most major populations (Romualdi et al. 2002; Hinds et al. 2005). However, the number of SNPs is greater and more population specific SNPs are found in African populations as compared to European populations, which clearly indicates the evolutionary history of human populations (Crawford et al. 2005; Hinds et al. 2005).

Two processes, the misincorporation of nucleotides during replication or chemical and physical mutagenesis, can give rise to base substitutions in a DNA sequence. In principle, SNP could be bi-, tri- or tetra- allelic variations, but tri- and tetra-allelic SNPs are very rare (Brookes 1999). In humans, all combinations of substitution polymorphisms are observed, with A/G substitution SNPs (including reverse complement T/C) being the most prevalent (Taillon-Miller et al. 1999; Miller et al. 2001). This is related to 5-methylcytosine deamination reactions that are known to occur frequently, particularly at CpG dinucleotides. Compared to repeat polymorphisms, SNPs are associated with low mutation rates (mutation rate about 10^{-8} per generation) and presumably arose only once in human history.

SNPs can have a different role in medical genetic studies. On one hand they can have functional impact and directly contribute to the disease phenotype. On the other hand they can serve as genetic markers for gene mapping studies. SNPs that occur within regions of functional significance, such as coding regions of the gene (cSNPs), splice junctions, and promotor regions, are of particular interest, because changes in these genomic regions can have direct impact at the phenotypic level (Peltonen and McKusick 2001; Belanger et al. 2005). SNPs in coding regions may be synonymous or non-synonymous, and both can potentially alter the structure or function of the protein (Drysdale et al. 2000). It has been observed that roughly half of the cSNPs change the encoded amino acid and more conservative amino acid changes are more common than radical changes (Salisbury et al. 2003; Hinds et al. 2005). This likely reflects purifying selection acting against deleterious alleles during human evolution.

Many attributes, such as their high density, slow mutation rate and automated detection, have made SNPs a marker of choice for gene mapping studies. Although the SNPs are individually less informative than microsatellites, the comparable level of heterozygosity with multi-allelic markers can be achieved by assembling multiple SNPs together as haplotypes. Construction of dense SNP maps over the human genome allowed the identification of regions that are ancestrally conserved. Within these regions the neighbouring SNPs show associations between alleles in a population and it has been suggested that this kind of structure plays a fundamental role in gene mapping studies. The ability to analyse the high density of SNPs across the genome has lead to remarkable

progress in recent years in characterizing and understanding the patterns of association between adjacent markers.

1.1.2. The nature of LD and factors shaping patterns of LD

LD is defined as the nonrandom gametic association of alleles at different loci in a population. Synonymous terms are “allelic association” or “gametic phase disequilibrium”. LD is said to occur when two alleles are found together on the same chromosome more often than expected by random segregation. This type of association is generated when a new mutation occurs on a chromosome that carries a particular allele at a nearby locus or this allele entered a particular population through migration (Figure 1) (Ardlie et al. 2002).

LD patterns observed in current human populations are the result of a complex interplay between biological factors and the population's demographic and evolutionary history (Figure 2) (Jorde 2000; Pritchard and Przeworski 2001; Ardlie et al. 2002; Reich et al. 2002; De La Vega et al. 2005; Smith et al. 2005). Each new mutation arises on a particular haplotype background (Figure 1). Haplotypes may gain high frequency by random genetic drift, and may subsequently be cleaved into segments by recombination. LD decays with increasing physical distance and this gradual decay of LD is dependent on both the time when the ancestral mutation event occurred, and on the local recombination rate (Abecasis et al. 2001; Clark et al. 2003; Bhangale et al. 2005).

Human populations have a history of both size reduction and expansion, which can largely influence the magnitude and pattern of population variation. The effects of genetic drift on variation is a function of the effective population size, whereby large populations can maintain higher levels of variation and small effective populations are more subject to random fluctuations in allele frequencies. For example, simulation studies have demonstrated that population expansion tends to decrease the extent of LD, especially if this takes place for a long period of time (Pritchard and Przeworski 2001). The effect of genetic drift is particularly severe in populations that have undergone a bottleneck event or a founding event (Finland, Ashkenazi Jewish) (Ober et al. 2001; Arcos-Burgos and Muenke 2002; Risch et al. 2003). The true demographic history of a human population is very complex, with populations in different parts of the globe experiencing varying degrees of isolation, admixture, migration, bottlenecks and expansion.

One factor that may inflate LD in the human genome directly and strongly is selection. Extended LD can be a signature of positive selection where allele frequency rapidly increases without allowing recombination to erode the ancestral haplotypes where these alleles originated. Also selection against deleterious variants can inflate LD, as the deleterious haplotypes are swept from the population by recombination. A number of studies have inferred the action of positive selection on LD structure (Bersaglieri et al. 2004; Stefansson et al.

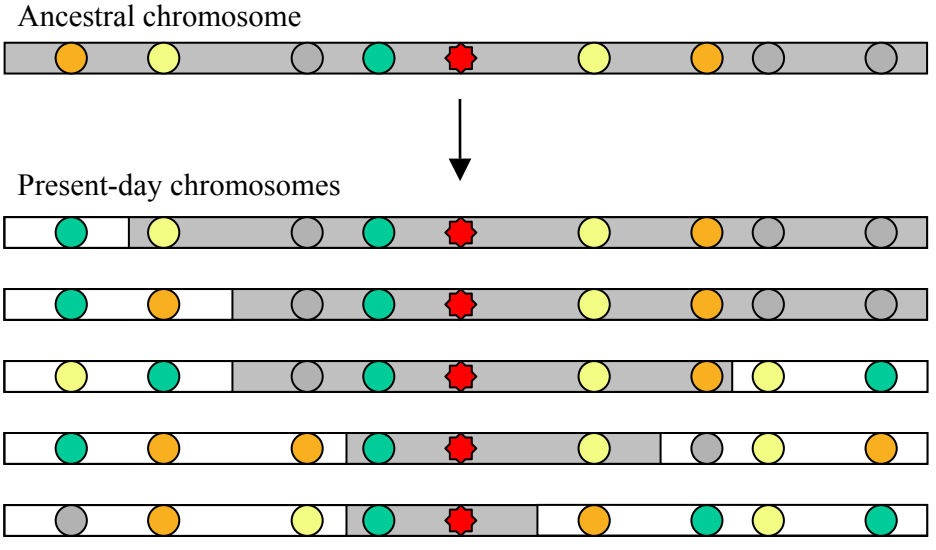

Figure 1. Linkage disequilibrium around the ancestral mutation. In the ancestral chromosome the mutant allele (indicated by the red star) is in strong LD with all other markers (indicated by circles with different colours). With time, the LD between adjacent marker alleles will be broken down by recombination. Present-day carriers of this mutant allele will also carry a small surrounding stretch of the ancestral chromosome (indicated by gray colour).

2005). Well-known illustrative example of recent positive selection is the alleles at the G6PD and CD40 loci, which confer resistance to malaria (Sabeti et al. 2002; Saunders et al. 2005). The genome-wide screen for the impact of past selection indicates that the evidence for positive selection in human genome is widespread (Fay et al. 2001; Akey et al. 2002). Epistatic selection can also lead to the association of particular alleles at different loci (Ardlie et al. 2002). It is still unclear exactly how our past demography has interacted with recombination in shaping the patterns of LD. It has been argued that the demographic history of a population influences the pattern of variation across the entire genome, whereas natural selection, mutation, and recombination influence the patterns of variation at particular genetic loci (Tishkoff and Verrelli 2003; Sawyer et al. 2005).

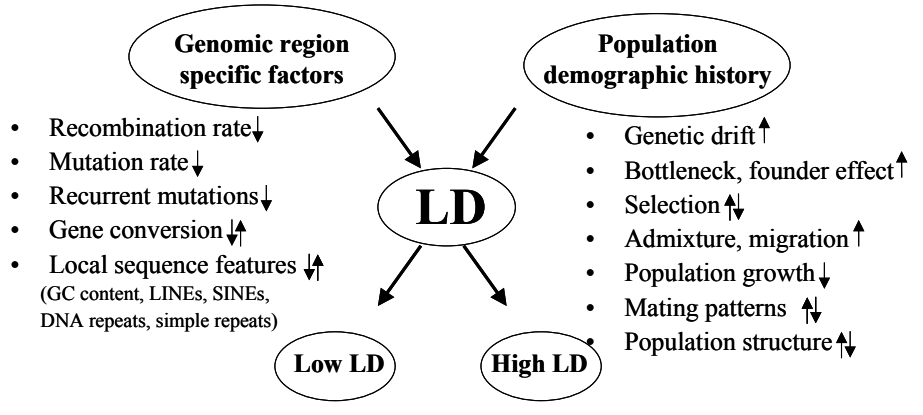

Figure 2. Factors shaping the patterns of linkage disequilibium. The level and pattern of LD is influenced by genomic region specific and population specific factors, both of which can lead to either low or high LD (indicated by arrows).

LD analysis has a wide range of applications. Understanding the patterns of LD within the genome is a key step towards the identification of susceptibility alleles to common complex diseases by genome-wide association studies (Tabor et al. 2002; Ke et al. 2004b; Wang et al. 2005). LD has also been used extensively to describe demographic and evolutionary processes in human populations (such as admixture or migration between populations) (Tishkoff et al. 1996; Wall 2001).

1.1.3. Measuring LD

The degree of LD between alleles at two loci can be evaluated by a variety of statistics (Devlin and Risch 1995; Jorde 2000; Hudson 2001), but in practice mainly two, termed *D'* and r^2, are widely used (Table 1). The basic concept of both pairwise disequilibrium measures is the difference (*D*) between the observed frequency of a two-locus haplotype and the frequency expected if the alleles were associated at random (Lewontin 1964). The statistical significance of LD can be tested by standard contingency table tests. If *D* is significantly different from zero, LD is said to exist and whether it is positive or negative depends on the arbitrary labeling of the alleles. The maximum value that *D* can have depends strongly on allele frequencies. To overcome this dependence *D'* can be used (Lewontin 1964). *D'*=1 denotes complete LD, and historical recombination results in the decay of *D'* toward zero. However, *D'* values are known to fluctuate upwards when small number of samples or rare alleles are examined. It is therefore suggested to rely on confidence intervalls of *D'* rather than point estimates (Gabriel et al. 2002).

The second common pairwise measure of LD is statistic r^2 (Δ^2), the square of the correlation coefficient between the two loci. r^2 =1 only when the marker loci have identical allele frequencies and every occurrence of an allele at each of the markers perfectly predicts the allele at the other locus. By contrast, *D'* can reach a value of 1.0 when the allele frequencies vary widely, as it reflects the correlation only since the most recent mutation occurred (Weiss and Clark 2002). r^2 is typically lower than *D'* for any chromosomal distance.

D' and r^2, have very different properties and may be applied for different purposes (Table1). *D'* is useful to assess the probability for historical recombination in a given population, whereas r^2 is useful in the context of association studies, because its magnitude can be translated directly to the sample size that is required for an association study (Sham et al. 2000; Pritchard and Przeworski 2001). To achieve the same power to detect association at the marker locus as we would have at the causal locus, sample size needs to be increased by a factor of $1/r^2$ (Pritchard and Przeworski 2001). Moreover, r^2 is the most appropriate in selecting tagSNPs, because it measures how well one SNP can act as a surrogate (proxy) for another (Carlson et al. 2004).

One relevant question is what amount of LD is useful for association studies. Calculations for that depend on the statistical properties of the different LD measures, which are mostly related to allele frequency and sample size dependencies. Studies on average *D'* levels have used values of *D'*=0.5 or "*D*' half-length" (Abecasis et al. 2001; Reich et al. 2001a) to describe the extent of LD along chromosome segments. For r^2 statistics it has been suggested to use a value of r^2=0.10 for describing the "useful LD", which would require 10 times the sample size of the best outcome (Kruglyak 1999).

Table 1. Summary of widely used LD measures

Measure	Formula	Advantages	Disadvantages
D	$D = P_{AB}\text{-}P_A x P_B$	Theory well understood	Strongly influenced by allele frequency
D'	$D' = D/D_{max}$	Useful to assess the probability of historical recombination	Depends on sample size; influenced by low allele frequency
r^2 (Δ^2)	$r^2 = D^2/P_A P_a P_B P_b$	Sample size estimation for association studies; tagSNP selection	Influenced by allele frequency and sample size

A,a; B,b – alleles of two loci (A and B); P_A;P_B;P_a;P_b – frequencies of the alleles at two loci

A very important step for LD studies is to determine haplotype frequencies. The traditional method to determine haplotypes is analyzing family members to acquire phase-information. The true haplotype can be optimally determined by direct molecular haplotyping. Phased genotype data could be obtained experimentally either using allele-specific long-range PCR (Michalatos-Beloin et al. 1996), somatic-cell hybrid method (diploid to haploid conversion) (Douglas et al. 2001) or by analyzing individual DNA molecules directly (Kwok and Xiao 2004). For example a straight-forward cloning approach (cell hybrids) has been used for haplotyping in a high-resolution scan of chromosome 21 (Patil et al. 2001). Molecular haplotyping is technologically difficult and cost-prohibitive, which makes it difficult to use especially for large-scale studies. Developing new cost-effective and high throughput methods for direct molecular haplotyping is probably the largest challenge of the future genotyping technology (Kwok and Xiao 2004). Therefore, most often haplotypes are determined from genotype data by statistical methods (Clark 1990; Excoffier and Slatkin 1995; Stephens et al. 2001; Niu et al. 2002). The most frequently used algorithm for estimating accurately the frequencies of common haplotypes is the expectation-maximization (EM) algorithm (Fallin and Schork 2000; Tishkoff et al. 2000). This algorithm works especially well when significant disequilibrium exists and for common haplotypes (Zhang et al. 2001).

1.1.4. Structure and extent of LD in the human genome

The early studies that characterized the extent and range of LD used low-density microsatellite markers, but the recent studies have mostly foucused on SNPs. The remarkably different characteristics of microsatellites and SNPs influence the measured LD patterns. Recurrent mutation in microsatellites can explain the lower levels of LD for tightly linked markers, and the more recent

origin of microsatellite alleles can explain the slower observed decay of LD with physical distance (Varilo et al. 2000; Abecasis et al. 2001). However, most of our understanding of how LD is shaped in human genome and populations came from research on recent studies with high density SNPs across the genome. SNPs are numerous and their low mutation rate allows the retention of LD signature of historical demographic events longer.

The first genome-wide estimation of the average extent of LD in the human genome showed that LD extended over much longer distances than would be expected by standard population genetic models and assumptions (Reich et al. 2002). The patterns of LD across the human genome show a high degree of variability and are unpredictable. Genetic markers that are immediately adjacent on a chromosome might be statistically independent, whereas those that are far away from each other might be highly correlated (Abecasis et al. 2001; Ardlie et al. 2001; Stephens et al. 2001; Reich et al. 2001a). Therefore LD is not a simple function of the distance between markers, but has been observed as a complicated pattern of regions of extensive LD separated by regions of low LD across the genome (Patil et al. 2001; Reich et al. 2001a; Phillips et al. 2003). Abecasis et al (2001) estimated in their study that physical distance could account for less than 50% of the variation in LD. They proposed that the remaining variation was probably due to variable rates of mutation, recombination as well as genetic drift, demographic factors and selection.

Empirical studies have shown that in current human populations LD extends for relatively short distances in most genomic regions (on average 60–200 kb) (Jorde, 2000; McVean et al. 2004), but in a few genomic regions LD may extent for longer distances (>500 kb) (Abecasis et al. 2001; Reich et al. 2001a). Recent fine-scale study across the genome showed remarkable extent of LD especially in the centromeric regions, as well as several regions in chromosome X (Altshuler et al. 2005). Extremely long stretches of LD are usually observed in studies where populations with small effective size (genetic isolates) (Laan and Pääbo 1997; Varilo et al. 2000), or populations that have undergone recent admixture, have been used (Wright et al. 1999; Zhu et al. 2005).

1.1.5. Haplotype blocks and tagSNPs

Initial studies about LD were focused just on average LD levels in the genome, but further investigations with a high density of markers indicated specific patterns of LD throughout the genome. The genome has been portrayed as stretches of consistently high LD ("blocks") interspersed with short intervals of rapid LD breakdown. Such blocks of high pairwise LD exhibit limited haplotype diversity, so that a small number of distinct haplotypes account for most of the chromosomes in a population (Figure 3) (Daly et al. 2001; Goldstein 2001; Johnson et al. 2001; Patil et al. 2001; Gabriel et al. 2002).

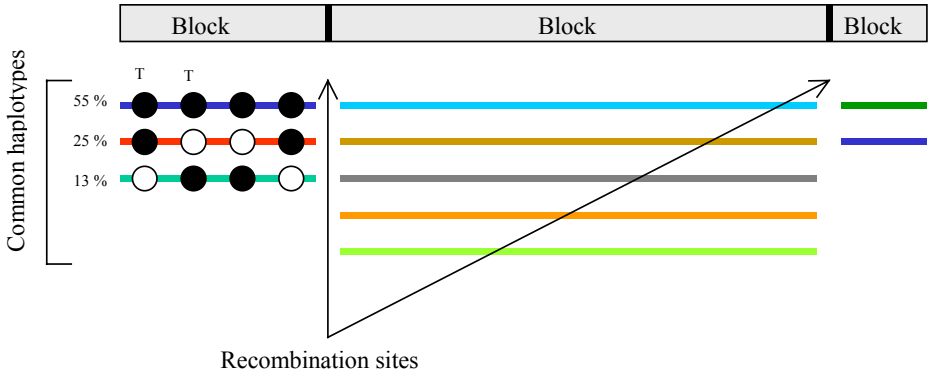

Figure 3. Representation of the block-like structure of LD. This figure illustrates a stretch of DNA sequence with three haplotype blocks, interrupted with regions of recombination. Within blocks typically two to five common haplotypes (at frequencies above 5%, indicated as a bar with different colours) account for most of the variation in human population (typically 90–95%). For example block 1 is defined by 4 common SNPs (indicated by circles) and the three most common haplotypes account for 93% of the sample. The black and white circles indicate the allele patterns of four SNPs and genotyping only two htSNPs (labelled T) is sufficient to distinguish the three common haplotypes in this block.

It has been demonstrated that blocks are regions of low recombination bounded by precisely localized recombination hotspots (Daly et al. 2001). The direct estimates of local recombination rates in humans have been done using sperm typing experiments. Data are limited to selected genomic regions due to laborious experiments (Jeffreys et al. 2001; Kauppi et al. 2003). Recent genome-wide estimations by computational methods indicate that recombination is highly variable and very often clusters in locally high intensity regions (or "hotspots") (Kauppi et al. 2004; McVean et al. 2004; Ptak et al. 2004; Altshuler et al. 2005; De La Vega et al. 2005). Moreover, comparison of fine-scale recombination rate in human and chimpanzee shows little concordance in the location of hotspots (Winckler et al. 2005). Humans have about twice as much recombination as mouse and rat, and in many mammalian species, including humans, the recombination rate in females is higher than in males (Kong et al. 2002; Jensen-Seaman et al. 2004). Instead of precise hotspots, block boundaries can be also shaped by stochastic recombination events, where random genetic drift in finite populations can generate high LD and limited haplotype diversity in regions of uniform recombination (Wang et al. 2002; Phillips et al. 2003; Zhang et al. 2003). If the extent of haplotypes is determined by random recombination, then all haplotypes encompassing a given point in the genome would not have the same length and we could see a few high values of LD extending outside the main blocks of LD.

For several reasons it is unlikely that blocks are discrete entities with clear-cut boundaries. First, the recombination hotspots are a region rather than a single base pair boundary (Templeton et al. 2000; Jeffreys et al. 2001) and secondly, a wide variety of demographic and population genetic phenomena can counteract the effect of recombination (Bamshad and Wooding 2003). Therefore, blocks should be viewed as a model that tries to capture the main features of LD patterns. A recombination pattern in human populations can also reflect the pattern from the "out of Africa" time, before large population expansions occurred, since 100 000 years is probably too short a time to destroy this pattern (Tishkoff and Williams 2002).

It is clear, that if such block structures exist in the human genome, the simplicity of the block structure within high LD regions could enhance association studies of complex diseases (Cardon and Abecasis 2003). Regions of high LD in fine-scale maps can often be indicative of SNPs that are in perfect LD (r^2=1) or have a high correlation (r^2>0.8) with one another (Carlson et al. 2003; Ke et al. 2004b; Lawrence et al. 2005). Knowledge of the haplotype structure would therefore enable the identification of a minimum number of SNPs, needed to uniquely tag all common haplotypes in high LD regions and therefore lead to a considerable reduction in genotyping effort (Figure 3) (Daly et al. 2001; Johnson et al. 2001). Based on applied methods, these minimum sets of SNPs are called either haplotype tagging SNPs (htSNPs) or tagging SNPs (tagSNPs).

Detecting haplotype blocks in the genome is sensitive to choice of method and parameters. The structure of blocks is dependent on many factors including SNP density, allele frequency, choice of markers and the studied population's demographic and evolutionary history (Wang et al. 2002; Stumpf 2004; Ke et al. 2004b; Nothnagel and Rohde 2005). Different studies investigating the structure of haplotype blocks have used distinct definitions based on various subjective criteria. The consensus finding is that denser marker maps, larger sample sizes and the use of common variants lead to higher sequence coverage of haplotype blocks, whereas the average size of haplotype blocks decreases (Cardon and Abecasis 2003; Wall and Pritchard 2003; Ke et al. 2004b). Moreover, the average block density is higher and blocks are longer in non-African compared to African populations (Wall and Pritchard 2003; Costas et al. 2005). Recent whole genome analysis indicated that in regions where all common SNPs were analysed a high proportion of the human genome (over 80% in non-African populations) seemed to fit the haplotype-block concept quite well (Altshuler et al. 2005). However, it has also been demonstrated that in these high SNP density regions (marker spacing < 2 kb) the block structure is dependent on the chosen SNP set and this may lead to different haplotype block patterns, as well as different haplotype frequencies (Nothnagel and Rohde 2005). Since the block structure is influenced by the complex interplay between factors that shape LD, the block-like chromosomal patterns are very complicated and obscure. This has often raised the question in the literature, how useful and reliable the concept of haplotype block is. We are not able to describe the complete and final picture of blocks and haplotypes before we have information of all genetic variations in the region. Finally, despite that the latest data claims that most of genomic regions show limited haplotype diversity, there are also regions in which the structure of LD is more complicated than a simple block description (Phillips et al. 2003; Wall and Pritchard 2003; Yalcin et al. 2004; De La Vega et al. 2005).

1.1.6. Methods for defining haplotype blocks and selecting tagSNPs

Because it is currently infeasible to genotype every available SNP in genetic-association studies, key questions are how many and which SNPs should be chosen for an association study so that it would have sufficient power to detect an association with a disease-causing variant (Kruglyak 1999; Cardon and Abecasis 2003; Wang et al. 2005). Estimates of the numbers of tagging SNPs required to cover the human genome have varied widely, ranging from 100,000 to 1,000,000 (Cardon and Abecasis 2003; Wang and Todd 2003; Ke et al. 2004b; Wang et al. 2005). The number of tagSNPs required to explain variation across large genomic regions fluctuates and depends on the extent of LD and allele frequencies (Weale et al. 2003; Miretti et al. 2005).

Several methods have been developed for haplotype block partitioning and tagSNP selection based on haplotype or genotype data. The most important factors on haplotype block partitioning and tagSNPs selection are the density of SNPs, the allele frequency of SNPs, genotyping error rate and missing data (Carlson et al. 2004; Schulze et al. 2004; Zhang et al. 2004; Ahmadi et al. 2005; Ke et al. 2005). Available methods can be classified into three categories. In the first category, haplotype blocks are first obtained based on a pairwise LD pattern (Gabriel et al. 2002) or a four-gamete test (Wang et al., 2002). When pairwise measures are used, a block is defined whenever all pairwise coefficients within a region exceed some pre-defined threshold. The most commonly used Gabriel *et al.* method defines the blocks using confidence limits on the pairwise coefficients and imposing constraints on marker number and spacing. Values of *D'* are divided into three categories: (i) strong LD (*D'* near 1); (ii) weak LD (*D'* significantly < 1); and (iii) intrmediate/unknown LD (pairs of SNPs with intermediate values of *D'* and with wide confidence intervals). Two or more SNPs can be grouped together into a block if the outermost pairs of SNPs are in strong LD and if the number of pairs in strong LD is at least 19-fold greater than the number of pairs in weak LD (Gabriel et al. 2002). Minimum numbers of SNPs (called htSNPs) are then selected as a follow-up study in each resulting block.

The second group is based on the concept of "chromosome coverage", with a haplotype block containing htSNPs that account for the majority of common haplotypes (Patil et al. 2001) or a reduced level of haplotype diversity (Daly et al. 2001). For example, Patil *et al*. required that in haplotype blocks, at least 80% of the observed haplotypes should be observed two or more times. Here the objective is to minimize the total number of htSNPs over a region of interest or the whole genome (Patil et al. 2001; Zhang et al. 2002; Stram 2004).

The last category involves methods, which were implemented after large numbers of SNPs became available and are currently most widely used in practise. It contains programs that are inherently block free in their approach towards selection of tagSNPs and is based on pairwise measures of LD. Ignoring defined block boundaries allows the use of long-range LD to efficiently represent genetic variations and to select tagSNPs simply on the basis of their pairwise r^2 values with tagged SNPs (Carlson et al. 2003) or by using a multiple-marker criterion (haplotype r^2) (Goldstein et al. 2003; Weale et al. 2003). A recent study has demonstrated that the haplotype-based tagging method increases tagging efficiency compared to pairwise tagging methods (de Bakker et al. 2005). Importantly, selecting tagSNPs using the pairwise measure r^2 optimizes the power for association tests (Pritchard and Przeworski 2001; Wang et al. 2005). One possibility is to look at genealogical relationships among haplotypes and choose tagSNPs based on phylogeny (Altshuler et al. 2005). When the distribution of tags and causal variants correlate perfectly to the phylogeny, then it is likely that the haplotype is tagged by a relatively small subset of markers and there is chance to detect small genetic effects (Cordell and Clayton 2005).

Simulation and empirical studies have indicated that optimal tagSNPs would have the same allele frequency as the ungenotyped SNPs they are meant to tag (Weale et al. 2003; Zondervan and Cardon 2004). The level of the tagging threshold has important implications for the efficiency. The general consensus is that an r^2 of 0.8 or greater is sufficient to obtain a good coverage of ungenotyped SNPs (Wang et al. 2005). This threshold allows the genotyping of a lower number of SNPs with relatively small losses in power. For a good choice of markers we have to understand LD across the markers and this was one of the reasons why the International HapMap project was created (for more details see pp 23).

1.1.7. The variability of LD and haplotype patterns in human populations

One crucial attribute of an LD map is that we have a good understanding of its utility in different human populations. Local differences in LD will likely necessitate selection of some population-specific SNPs for an optimal LD map, but it would be very useful to have a core set of SNPs that are informative in many populations. Therefore, an understanding of how conserved or variable LD and haplotype patterns are across populations is of key importance for the efforts to identify disease genes by association with marker loci.

Considerable effort has been devoted to characterizing the extent of genomic variation in modern human populations. It has been estimated in several studies with different types of markers that within-population differences among individuals account for most of genetic variation (85–90%), and only small proportion of genetic differences (up to 15%) occur among different populations. Well-known studies from Lewontin *et al.* (1972) and Barbujani *et al.* (1997) claimed that the division of populations into discrete groups is not possible according to genetic data. However, these early calculations ignored the fact that different loci are not inherited independently. When applying a model-based clustering algorithm that identifies subgroups with distinct allele frequencies, then analyzis of a global human sample showed that individuals cluster discretely according to their continents of origin (like Africa, Europe/ Middle East, Asia, Oceania, New World) (Cavalli-Sforza et al. 1991; Tishkoff et al. 1996; Rosenberg et al. 2002; Gonzalez-Neira et al. 2004; Shriver et al. 2005). Continental structure has been also observed recently by analyzing LD and haplotype patterns using globally diverse population panels (Gonzalez-Neira et al. 2004; Sawyer et al. 2005; Shriver et al. 2005). It has been known for a long time that allele frequencies of individual markers may vary widely across populations (Cavalli-Sforza et al. 1994) and instead of clear boundaries between continents and populations, gradients of allele frequencies exist (Barbujani et al. 1997; Wilson et al. 2001; Rosenberg et al. 2002; Serre and Pääbo 2004; Shriver et al. 2005).

There are still only limited empirical data available where geographic variability in LD and haplotype structure within large a number of human populations at the genome-wide level has been compared. Most recent studies either compare the variability at the whole chromosome or genome-wide level using only samples that represent the main ethnic groups (Patil et al. 2001; De La Vega et al. 2005; Hinds et al. 2005) or using a large number of populations for restricted genomic regions (Reich et al. 2001a; Gonzalez-Neira et al. 2004; Tsunoda et al. 2004; Sawyer et al. 2005). However, some general conclusions can be drawn from available datasets. It has been shown that broad views of LD tend to be stable across populations and most common haplotypes exist in worldwide populations, presumably reflecting common localization of recombination hot spots (Tsunoda et al. 2004; Ke et al. 2004a; De La Vega et al. 2005; Hinds et al. 2005).

It has been observed consistently that the decay of LD with increasing physical distance tends to be faster and hapoltype diversity higher in samples from African populations in comparison with non-African samples (Tishkoff et al. 1996; Tishkoff et al. 1998; Kidd et al. 2000; Stephens et al. 2001; Ke et al. 2004a). Moreover, African populations show the largest number of population-specific alleles and a broad range of haplotypes, whereas populations in other continents harbor only a subset of the genetic diversity present in Africa (Tishkoff et al. 1998; Tishkoff and Williams 2002; Shifman et al. 2003; Wall and Pritchard 2003; De La Vega et al. 2005; Hinds et al. 2005). All these findings indicate that African populations are the most variable, ancestral and are the origin of all modern humans. The migration of anatomically modern humans out of Africa is the most significant event that has influenced the patterns of genetic variation in current non-African populations (Marth et al. 2003; Tishkoff and Verrelli 2003). Knowledge of the African genetic diversity is critical for reconstructing human evolutionary history, for understanding the genetic basis of phenotypic variation, and for mapping genetic diseases prevalent in people of recent African origin (Tishkoff and Williams 2002). Figure 4 represents a model of human demographic history and explains the common theory why less LD is found in populations inside of Africa compared to populations outside of Africa.

The patterns of LD and haplotypes in the human genome are more complex than simply differences between African and non-African populations because subsequent demographic events have influenced the specific LD and haplotype patterns in different populations. The fine-scale LD analysis has shown both similarities in haplotype composition between populations of close ancestry (Bonnen et al. 2000; Nejentsev et al. 2004; Ke et al. 2004b) and also remarkable variation even between populations from similar geographic regions (for example inside of Europe) (Clark et al. 2003; Crawford et al. 2004; Beaty et al. 2005; Evans and Cardon 2005; Sawyer et al. 2005). However, even if different haplotypes or haplotype frequencies to some extent exist, it seems that both common and rare haplotypes are often shared across ethnically similar

populations. Detected differences are typically the result of rare, population specific SNPs or haplotypes, which are relatively young and reflect a recent demographic history of the population. In addition, the observed population differences may be due to the effect of strong positive selection (Enattah et al. 2002; Swallow 2003; Bersaglieri et al. 2004).

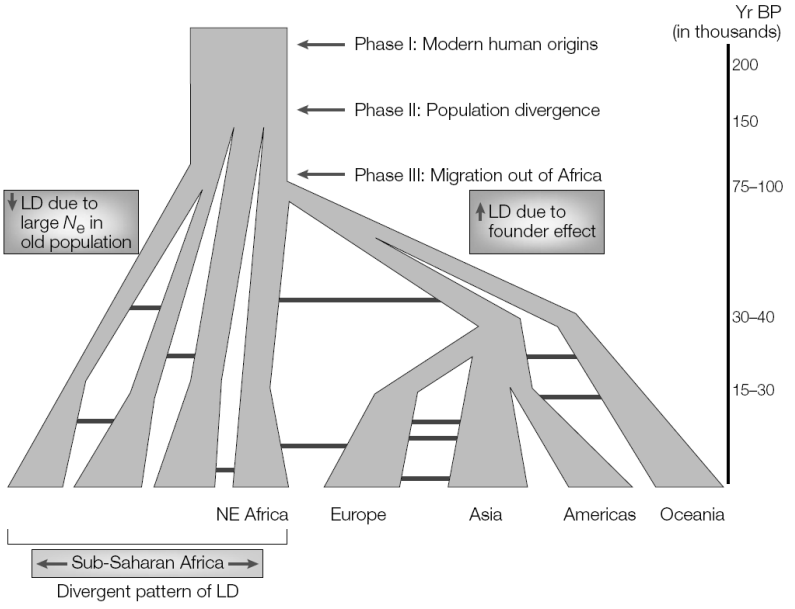

Figure 4. Model of human demographic history. A small subset of the population migrated out of Africa 100,000 years ago and rapidly expanded throughout the rest of the globe. Ancestral African populations have less LD compared with non-African populations, which is due to a large effective population size and a subdivided population structure throughout their evolutionary history. The bottleneck event that is associated with the founding of non-African populations and rapid population expansions resulted in reduced genetic variation, greater LD and less haplotype variation. N_e, effective population size; YrBP, years before present (figure taken from Tishkoff and Williams, 2002).

Comparison of Asian and European populations usually shows quite similar patterns of LD structure. However examples where major European haplotypes are absent in Asia and visa versa exist, illustrating the impact of demographic history to the local fine-scale LD and haplotype patterns (Ng et al. 2004; Laan et al. 2005).Usually populations that show strong divergent patterns are rather small with particular demographic histories (for example the Basques, Sardinians, Icelanders, Lapps, Finns) (Barbujani and Goldstein 2004). Moreover, populations distributed in geographic borderlines of Europe and Asia exhibit a more complex pattern of haplotye distribution, where clear admixture

of genetically different migrants can be seen in haplotype composition (Laan et al. 2005; Sawyer et al. 2005). Population-specific LD and haplotype patterns may have great practical relevance in disease mapping studies, starting with marker selection for association studies and ending with interpretation of the results (Pritchard et al. 2000; Reich and Goldstein 2001).

1.1.8. HapMap project

The early LD and haplotype block studies indicated that the construction of the haplotype maps of the human genome would facilitate the genetic association studies of human diseases. Based on this assumption, the International Haplotype Map Project ("HapMap") was started in 2001. This large project involves 9 research groups in 6 countries and has a goal of determining the common patterns of human genetic variation and to make this information freely available in the HapMap homepage (http://www.hapmap.org) (The International HapMap Consortium 2003). The HapMap dataset provides researchers with the opportunity to use fine scale LD data, and to estimate the approximate number of tagSNPs necessary to cover the candidate gene regions or the whole human genome (The International HapMap Consortium 2003). This information offers a framework for researchers for designing and interpreting genome-wide association studies (Evans and Cardon 2005). In addition, the knowledge about the patterns of recombination, correlation of LD with sequence elements and features of natural selection and how they affect LD patterns will be improved.

The available dataset from HapMap project contains information over 3 million SNPs (Release 19, October 2005) genotyped in 269 samples (Table 2). The first phase of the HapMap project provided information on the patterns of variations and LD using a SNP density on average of 1 SNP per 5 kb (Interational HapMap Concortium, 2003). One main feature of tagSNPs – performance is influenced by marker density. High marker density allows better selection of tagSNPs and extends genome coverage. For that reason the HapMap Consortium planned phase II of the project to increase marker density up to 1 SNP per 600 bp. Recently, the Phase II dataset was reported, which contains information from 10 genomic regions with higher SNP density (about 1 SNP per 1–2 kb) (Altshuler et al. 2005). These 500 kb long regions are part of the Encyclopedia of DNA elements (ENCODE) project, which was created in order to recognize the need to annotate the list of functional SNPs of the human genome (Collins et al. 2003). The ENCODE project involves the resequencing of 96 chromosomes to ascertain all common variants and the genotyping of all SNPs that are either in the dbSNP database or that were identified by resequencing (http://www.hapmap.org).

The available HapMap dataset gives us a fine-scale overview about LD and haplotype structure across the whole genome, as well as important knowledge for tagSNPs selection. For example, when the marker density is high like in the ENCODE regions, then indeed most of the genome spanned by blocks with high LD and showed limited haplotype diversity (Table 2). Moreover, the efficiency of tagging is also encouraging, indicating that SNPs with an average density of 1–5 kb capture most common genetic variation in regions of extended LD (Altshuler et al. 2005). These results are consistent with the previously published Perlegen (http://genome.perlegen.com) dataset (Hinds et al. 2005).

Recent calculations indicate that about 10–20% of genomic regions need a SNP density higher than 1 SNP per kb (Wang et al. 2005; Carlson et al., 2004; Ke et al. 2004). Regions with low LD require high marker density to allow comprehensive coverage by tagSNPs (Carlson et al. 2003; Wang et al. 2005). Finally, to fully capture all common sequence variants SNP data need to be integrated with duplication, copy number and inversion polymorphisms (Iafrate et al. 2004; Sebat et al. 2004). For example, segmental duplications, which often carry functional genes (Bailey et al. 2002), vary considerably in copy number and in sequence content (Fredman et al. 2004) and could also have a direct influence on individual differences in risk of common diseases (Brookes and Prince 2005).

One of the crucial aspects of marker selection in populations is the transferability of chosen tagSNPs sets to other populations. The evaluation of the transferability of HapMap tagSNPs among populations will determine if they can serve as a universal reference for the selection of tagSNPs in other populations in the future. The efficiency of transferring the same set of tagSNPs across populations in similar geographic regions or even between continental regions depends first on similar haplotype content among populations and also on the number of tagSNPs that form a particular tagSNP set (Ke et al. 2004b; Liu et al. 2005; Miretti et al. 2005). Currently, only four different populations have been analyzed during the HapMap project and presumably this sample selection does not represent the whole existing variability of the world human populations. Therfore, one further aim is to collect samples from several other populations, which allows the testing of tagging efficiency among different populations (http://www.hapmap.org).

Table 2. Phase I datasets from the International HapMap project

Parameters	Studied Populations		
	YRI	CEU	JPT and CHB
Sample Size	90 (30 trios)	90 (30 trios)	44 (JPT); 45 (CHB) unrelated
Unique SNPs	1,076,392	1,104,980	1,087,305
Monomorphic	156,290 (15%)	234,482 (21%)	268,325 (25%)
Polymorphic	920,102 (85%)	870,498 (79%)	818,980 (75%)
Fraction of genome spanned by blocks (%)*	67	87	81
Average length per block (kb)*	7.3	16.3	13.2
Average number of haplotypes (MAF≥0.05) per block*	5.57	4.66	4.01
Common SNPs captured (%)[#]			
Simulated Phase I[a]	45	74	72
Simulated Phase II (ENCODE)[s]	81	94	94

CEU – from Centre d'Etude du Polymorphisme Humain (CEPH), Utah, USA; YRI – Yoruba in Ibadan, Nigeria; CHB – Han Chinese in Beijing, China; JPT – Japanese in Tokyo, Japan; * – haplotype block calculated in ENCODE regions (calculated by *D'* method), [#] – common SNPs are predicted to have a proxy with maximum $r^2 \geq 0.8$; [a] average density of 1 SNP per 5 kb in Phase I HapMap; [s] overall density of 1 SNP per 1 kb in Phase II HapMap (Data derived from Altshuler et al. 2005).

Recent observations indicate that the number of tagSNPs required is similar in non-African populations from both Europe and Asia (Gabriel et al. 2002; Ke et al. 2004a; Ahmadi et al. 2005). It has been also shown that tagSNPs selected in the CEPH families of European origin in the HapMap project can be successfully applied to other Europeans with only a moderate (or no) loss of power (Nejentsev et al. 2004; Ke et al. 2004a; Miretti et al. 2005). The largest differences in terms of tagSNP selection are between African and non-African populations, where for example less than half of the tagSNPs are required for capturing most of the diversity in European populations as compared to African samples (Nejentsev et al. 2004; Ke et al. 2004a).

1.2. Applications of LD: Association studies to identify disease susceptibility alleles for complex diseases

The main interest of human genetic studies is to find genetic factors responsible for disease phenotypes. For years, the main study design for investigating the genetic basis of inherited disease has been linkage analysis in families. This study design has been especially powerful for identifying rare high-risk disease

alleles with high penetrance (Kerem et al. 1989; Hastbacka et al. 1992). Dissecting the genetic architecture of human complex disease is the next big challenge of human genetics for years to come. Existing studies have indicated features that are typical to complex diseases, like: multiple genes with individually small or moderate effects, underlying variants are frequently in non-coding and regulatory regions, genetic effects may have strong interactions with other genes (epistasis) and can vary in different environments and lifestyles. This complex nature is a reason why only limited success has been achieved using classical linkage analysis for mapping susceptibility loci responsible for complex diseases (Altmuller et al. 2001; Freimer and Sabatti 2004). Currently, the most promising approach for complex disease gene mapping is association studies (Risch and Merikangas 1996).

1.2.1. Genetic association studies

Association studies compare the allele frequency of a polymorphic marker, or a set of markers (haplotype), in unrelated patients (cases) and healthy controls drawn from a general population to identify marker with frequencies that differ significantly between the two groups (Risch and Merikangas 1996; Risch 2000; Cardon and Bell 2001; Carlson et al. 2004; Cordell and Clayton 2005). Genetic associations arise only because human populations share common ancestry. Since associations operate usually over shorter distances in the genome, a large number of markers are required to detect associations. The success of association studies relies highly on the patterns of LD in the human genome. To find a positive association, a tested marker must either be the causal allele or in LD with the causal allele (Kruglyak 1999).

Compared with traditional linkage studies, association studies based on LD have two major advantages. First, in a population-based study the region around a marker that is shared identically by descent in unrelated affected individuals will be much smaller because of a much higher number of generations from the most recent common ancestor has passed in comparison with related individuals in pedigrees. Second, the use of unrelated individuals makes it feasible to obtain sample sizes large enough to capture modest genetic effects between genotype and phenotype. One limitation of early association studies was the modest number of polymorphisms available. Currently it is possible to analyze a large number of genetic variations simultaneously which allows researchers to design genome-wide association experiments and increase the probability of finding genetic factors underlying complex diseases (Carlson et al. 2004; Hirschhorn and Daly 2005; Wang et al. 2005).

1.2.2. Strategies for genetic association studies

Two main approaches for mapping the genes that underlie complex diseases and quantitative traits exist: candidate gene studies and genome-wide studies. These strategies, together with illustrative examples, are summarized in Table 3. Until recently, most association studies were based on candidate polymorphisms in which only a few SNPs within a gene of interest were studied for association with a phenotype. Most of the candidate genes were selected for further study either based on the results of previous linkage studies, or on the basis of other evidence that they might affect a disease risk (Tabor et al. 2002). Since our current knowledge about the function of the genes in disease processes is insufficient, the ability to predict functional candidate genes and variants is limited.

Table 3. Different strategies with illustrative examples for mapping genes underlying complex disease

Approaches	Potential advantages	Gene example	Disease	References
Candidate-gene:	Analysis of small genomic regions;		Lipid	Cohen et al.
(i) Resequencing	possible to confirm potential candidacy	ABCA1	metabolism	2004
(ii) Association	and biological pathways; requires only a few SNPs	PPAR-γ	Type II diabet	Altshuler et al. 2000
Genome-wide association:	No prior hypothesis			
(i) Functional SNPs	required; power to detect common alleles	Lymphotoxin-α	MI	Ozaki et al. 2002
(ii) GWA	with modest effect	CFH	AMD	Klein et al. 2005

ABCA1 the adenosine triphosphate binding cassette (ABC) transporter A1; PPAR-γ peroxisome proliferator-activated receptor γ: AMD-age-related macular degeneration; *CFH* complement factor H; MI myocardial infarction

In candidate-gene based studies the genotyped variations typically had putative phenotypic consequences, like cSNPs that alter or terminate amino acid sequence, disrupt splice sites, or occur in promoter regions (Lander 1996; Kruglyak and Nickerson 2001; Botstein and Risch 2003; Crawford et al. 2005). These kind of studies are referd as a “direct” approach in the literature, since putative functional variants are tested directly. Several recent gene-based resequencing studies show that only 4% of all SNPs are within coding regions of genes and 2.2% of all SNPs are classified as nonsynonymous (Stephens et al.

2001; Haga et al. 2002; Crawford et al. 2005). On the basis of these studies, the estimated total number of such functional SNPs in the human genome would be between 50,000 and 100,000 (Botstein and Risch 2003). Therefore one option is to perform whole-genome association studies based on functional variants only (Table 3). However, this approach would be currently impractical to use because there is only limited data available about cSNPs. Furthermore, at the moment we are far from a complete understanding of the functionally important elements that predispose to disease development. Empirical studies have already shown that both non-coding as well as variants from regulatory regions can be reliably associated with complex diseases (Bennett and Todd 1996; Stefansson et al. 2002; Abelson et al. 2005). For example, such variants may cause differential splicing or variation in gene regulation and expression. Moreover, data from a recent comparative genomic study demonstrate that the level of evolutionary conserved non-coding sequences is comparable to the amount of evolutionary conserved exonic sequences (Schwartz et al. 2003).

Therefore, instead of searching genetic effects underlying complex diseases by using only 5% of the genome (the estimated proportion of genomic sequence that contains genes), it would be more successful to use whole-genome association studies. This approach aims to test all common SNPs for function by assaying a subset of tagSNPs, such that all unassayed SNPs would be detected through LD with tagSNPs (known also as an "indirect" strategy).

1.2.3. Genome-wide association studies

The advantage of a genome-wide association (GWA) approach is that no prior hypothesis or identification of a specific candidate gene is required and therefore new candidate genes or regions can be identified and tested. Instead, the approach uses the whole genome to localize the underlying genes and also gives the opportunity to investigate different gene-gene interactions.

The GWA analysis includes several different steps (Figure 5) and it is important to understand the conditions under which a study leads to successful results. In the first step, a full set of SNPs is genotyped in a fraction of samples. Available HapMap and Perlegen datasets give an opportunity to identify a set of genome-wide tagSNPs, based on the information of genome-wide LD patterns. Currently, the most marker-dense available map for genome-wide analysis contains 500,000 SNPs (www.affymetrix.com), which were chosen in regard to their LD patterns with neighbouring SNPs using a dataset of 1.54 million SNPs (Hinds et al. 2005). Another biotechnology company Illumina will soon offer a whole-genome BeadChip array, where SNP selection is based on HapMap datasets, containing 250,000 tagSNPs (http://www.illumina.com). The two GWA assays available have used different SNP selection criterions, which probably have an effect on the coverage of the genome.

When true positive signals from new candidate loci are detected then in order to confirm the candidacy the pattern of LD around the associated variants should be assessed to determine association interval and exclude other nearby loci (Goldstein 2003). Next, to get conclusive evidence, an independent replication and validation of candidate loci should be done using samples from different populations. In order to obtain a reasonable power to detect moderate genetic effects, it is important to use a relatively high number of samples for the replication study (Hattersley and McCarthy 2005). Finally, to understand molecular as well as biological function of candidate gene(s), the functional effects of causal variants should be assessed. Moreover, it would be very helpful to consider also the information from expression profiles and comparative genomic analysis is assessing new sets of genes or biological pathways.

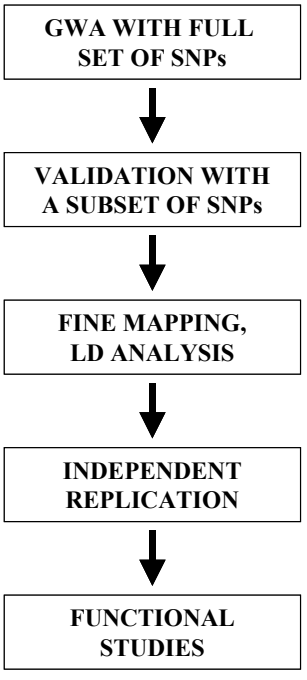

Figure 5. Overview of the genome-wide analysis (GWA) approach for gene discovery in complex disease studies.

During the analysis we can consider whether to analyze each genetic marker alone for an association with a disease or to analyze multi-marker haplotypes. Haplotype-based analysis allows testing simultaneously for an association with multiple potentially causal variants. Haplotypes may also be a proxy for untyped causal markers and capture them more efficiently than single markers

(de Bakker et al. 2005; Newton-Cheh and Hirschhorn 2005). For example, it was recently shown in the study with the *CARD15* gene (associated with Crohn's disease) that haplotypes of common variants would be sufficient to capture disease alleles with modest frequencies (1–5%) (Vermeire et al. 2002).

One of the problems in GWA studies arises with statistical analysis, because all recorded *p*-values must be corrected for the large number of hypotheses tested. Therefore, a conservative significance threshold in the order of $P<10^{-6}$ has been proposed for declaring a significant association in a genome-wide study and to distinguish false positives from real associations (Risch and Merikangas 1996; Dahlman et al. 2002; Freimer and Sabatti 2004; Neale and Sham 2004). To overcome this problem a multi-stage approach has been suggested, where by during the initial GWA analysis a liberal *p*-value threshold is used to identify a subset of SNPs with putative associations (Hirschhorn and Daly 2005). In the next step the aim is to re-test the SNPs identified by using populations that are larger or similarly sized. The results of this scan can then be used to distinguish the few true-positive associations identified in the GWA analysis from the many false-positive results that occur by chance.

The first successful genome-wide association studies have already been carried out (Carrasquillo et al. 2002; Ozaki et al. 2002; Mitra et al. 2004; Hirschhorn and Daly 2005; Klein et al. 2005). These first studies used a relatively small number of markers (maximally 100,000 SNPs) and therefore have covered probably close to 50% of the genome. The results of GWA analysis with higher number of markers should determine the efficacy of this approach and show whether it will become truly practical for finding genetic risk factors of complex diseases.

1.2.4. Complications of mapping the genetic components of complex diseases by association studies

Currently the genetic complexity underlying common diseases is largely unknown, but both theoretical models and practical experiments have demonstrated that typically no single factor is either necessary or sufficient for complex disease but rather multiple loci, each with a range of effects, are involved (Wright et al. 1999; Pritchard and Cox 2002; Terwilliger and Weiss 2003). Several confounding factors (Table 4) make the investigation of the genetic background of complex diseases especially difficult (Lander and Schork 1994; Risch 2000; Cardon and Bell 2001). For example, features like different allelic variants or loci causeing a similar phenotype (genetic heterogeneity), a single gene (or variant) affecting many characteristics of the phenotype (pleiotropy), environmental factors induce a particular phenotype that resembles the phenotype produced by a mutation (phenocopies), may all contribute to complex phenotype (Jorde 2000; Glazier et al. 2002; Zondervan and Cardon 2004; Salanti et al. 2005). Moreover, complicated interactions between modifier

and susceptibility genes and gene-environment interactions also modulate the phenotype of individuals with disease (Nadeau 2003). If these confounding factors are abundant then the power to detect a genuine causative or predisposing effect can fail even in very large cohort. All these mentioned factors can also be potential sources of variable findings among different studies.

Table 4.

Factors complicating analysis of complex diseases
• Oligo/polygenic inheritance
• Quantitative phenotypes
• Incomplete penetrance
• Unknown mode of inheritance
• Genetic and environmental heterogeneity
• Late onset of the disease
• Phenocopies
• Pleiotropy
• Epistasis
• Limited statistical power
• Multiple testing
• Publication bias

The statistical power of an association study depends on many parameters, including (i) LD among nearby variants and how efficiently selected tagSNPs capture untagged markers; (ii) allele frequencies of both causal variants and marker alleles; and (iii) the strength of phenotypic effects (Zondervan and Cardon 2004; Clark et al. 2005; Wang et al. 2005).

1.2.4.1. Allelic spectrum of human disease genes

One critical factor is the "allelic architecture" of the genes underling complex diseases, because this will determine which study designs will lead to successful results (Wright et al. 1999; Pritchard and Cox 2002; Smith and Lusis 2002; Weiss and Clark 2002). The allelic architecture of a disease refers to the number

of genetic variants that exist, their frequencies and the risks that they confer (Reich and Lander 2001b; Pritchard and Cox 2002).

The typical frequencies of variants that underlie common disease are largely unknown. There are two contrasting models of the allelic diversity underlying complex diseases. The common disease/common variant (CD/CV) hypothesis proposes that most of the genetic risk for common, complex diseases is due to disease loci with one common variant or with relatively small pool of common polymorphic disease-associated alleles (with frequency >1%) (Lander 1996; Reich and Lander 2001b). If this is the case then association mapping using current HapMap resource should be powerful to capture most common variants through strong correlations (Altshuler et al. 2005). Alternatively, the multiple rare-variant hypothesis (or disease heterogeneity model) proposes that there are multiple low frequency risk alleles (MAF less than 0.01) at a large number of loci with varying effects on the disease risk (Smith and Lusis 2002). Several examples have been reported where common variants influence a disease susceptibility (Altshuler et al. 2000; Rioux et al. 2001; Stefansson et al. 2002; Lohmueller et al. 2003; Klein et al. 2005) thus implying that the common disease-common variant hypothesis may be valid for complex diseases (Brookes and Prince 2005). Two well-known exsamples of disease alleles with high population frequency are the *APOE* locus (ε4 allele, increases risk to Alzheimer disease and heart disease) (Fullerton et al. 2000) and the PPARγ locus (the Pro12Ala polymorphism, implicated in type 2 diabetes (Altshuler et al. 2000). However, these few variants that have already been found may not be representative of complex disease variants in general. The current genome-wide studies focus on common variants, but they will not entirely explain the genetic component of common disease risk. For example, the allelic architecture at the *NOD2* locus is far more complicated, where the disease haplotype frequency for *NOD2* represents the combined frequency of three rare SNPs, all which predispose to the development of Crohn's disease (Hugot et al. 2001; Ogura et al. 2001). The success for identifying this locus was due to the large effect size which produces a high power of detection, even with the rare-allele models (Hugot et al. 2001; Ogura et al. 2001; Zondervan and Cardon 2004).

Different evolutionary processes, including positive and purifying selection can result in a different allelic spectrum. Shift towards common variants might be a result of positive selection. A classical example is lactase persistence (or lactose intolerance), where alleles have reached higher population frequencies due to recent positive selection (Enattah et al. 2002; Swallow 2003; Bersaglieri et al. 2004). Genes that have evolved under recent positive selection reflect on one hand the local adaption to different environments, or on the other hand may cause differences in susceptibility to modern diseases in current human populations (Altshuler and Clark 2005). However, a disease with no common alleles may indicate a purifying selection, where mutations that occurred long time ago have been lost, leaving only the more recent, rare mutations in the population.

It seems that to fully understand and dissect the allelic spectrum underlying complex diseases requires genotyping both, common as well as rare SNPs (Pritchard and Cox 2002). If the modest-risk variants associated with complex disease are very rare then they will go undetected by LD mapping. Moreover, when the genetic effect is weak or when gene-gene interactions and different environmental factors occur then also common alleles may lack the necessary power to be detected (Altshuler and Clark 2005). Therefore denser SNP maps and probably the development of new approaches which allow capturing both variants (common and rare) and moderating genetic effects will be required in the near future.

1.2.4.2. Replication of association studies

The best way to convincingly demonstrate a true association between a genetic marker and a disease is a replication with independent samples. Independent replication has become especially important, but unfortunately difficult to obtain in the genetic studies of complex diseases (Ioannidis et al. 2001; Hirschhorn et al. 2002; Colhoun et al. 2003; Lohmueller et al. 2003). Only a small fraction of associations have been replicated by others, which leads to the assumption, that many of the results are false positive (Hirschhorn et al. 2002). When the genetic effect is large and the predisposing allele is common, then most independent researchers can readily obtain similar results with strong levels of statistical significance. An illustrative example is the *CARD 15* gene, where the causal variants associated with Chron's disease phenotype have been replicated by many groups (Schreiber et al. 2005). However, when genetic effects are weak, and possibly context-dependent (e.g., they may vary by sex, ethnicity, or different subtypes of a disease) replication may be particularly difficult (Colhoun et al. 2003).

In principle, there are three possible explanations for inconsistencies between reports: false-positive results, false negative results, and true variability in association among different populations (Figure 6) (Lohmueller et al. 2003; Newton-Cheh and Hirschhorn 2005). The etiological complexity of a disease, allelic and locus heterogeneity, population structure, sample selection, statistical fluctuations that arise by chance, and technical difficulties can all influence non-replication of initial findings (Cardon and Bell 2001; Clayton and McKeigue 2001; Colhoun et al. 2003). A major cause for both false-positive and false-negative results is inadequate sample collection and sample size (Ioannidis et al. 2001; Hattersley and McCarthy 2005). The problems can arise when selected samples suffer hidden population stratification and admixture (Cardon and Palmer 2003; Freedman et al. 2004; Marchini et al. 2004; Campbell et al. 2005; Helgason et al. 2005). Population stratification can arise, when the population of interest is not homogenous but consists of subgroups that have different demographic histories and also different allele frequencies and disease

prevalence. One option to reduce the effect of population stratification is to select controls from the families of affected probands and to apply a transmission disequilibrium test (TDT) (Spielman et al. 1993). The other possibility is to check population stratification with the use of genomic controls (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Marchini et al. 2004).

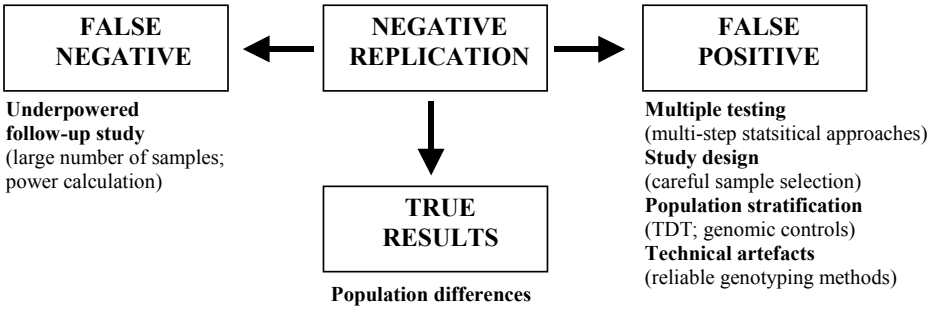

Figure 6. The possible causes for the non-replication of association studies. Possible solutions to overcome negative replication are indicated within brackets.

However, non-replication does not necessarily imply lack of causality but might point to the need for additional studies in certain populations or more detailed studies of the function of the gene. To avoid spurious association results a gene-based approach has been proposed in which all common variations within putative candidate gene are considered jointly (Neale and Sham 2004). Different populations may have a unique profile of disease alleles, including different frequency of disease-causing alleles, the pattern of association between disease-causing alleles and single markers or haplotypes and finally, interacting genetic and environmental factors. Usually genetically isolated populations are different in comparison with outbred populations, as a result of genetic drift and founder effects. Population isolates (like Sardinians, Icelanders, Finns) usually harbor reduced genetic variation and consequently reduced genetic complexity of the disease trait (Wright et al. 1999; Dunning et al. 2000; Peltonen et al. 2000; Service et al. 2001; Bonnen et al. 2002; Pardo et al. 2005). Therefore, one suggestion to increase the efficiency of association studies has been to use founder populations, which reduce the number of markers that need to be genotyped. Founder populations might be most useful in detecting an initial association region, while "older" or general populations could be more useful for fine-scale mapping (Varilo et al. 2000; Jorde et al. 2001). However, the empirical data are fairly mixed, some studies have shown only little difference in LD patterns and allelic diversity in population isolates (Dunning et al. 2000; Eaves et al. 2000) and therefore the usefulness of isolates in gene mapping studies has to be clarified by further empirical studies (Jorde 2000).

1.3. An example of an association study: search for a genetic component of stroke

Stroke is a common and serious disease, being the third leading cause of death and the main cause of long-term disability in the western societies (American Heart Association, 2004). The clinical phenotype of stroke is complex and there are different classification systems for categorizing stroke subtypes (Meschia 2002). The most widely used systems are the Trial of Org 10172 in Acute Stroke Treatment (TOAST) (Adams et al. 1993) and Oxfordshire Community Stroke Project (OSCP) system (Bamford et al. 1991). Stroke can be divided broadly into two major varieties: ischaemic and hemorrhagic. The most common form of stroke is ischaemic (accounting for 80-90%), which occurs when the blood supply to the brain is interrupted, usually by a blood clot. These clots may be caused by arteriosclerosis in the carotid arteries, which feed the head and brain with oxygen-rich blood. Ischaemic stroke can be further subdivided into: (i) large vessel occlusive disease or carotid stroke (usually due to atherosclerosis and plaque formation within common and internal carotid arteries); (ii) small-vessel occlusive disease (due to the involvement of small perforating end-arteries within the brain); and (iii) cardiogenic stroke (secondary to blood clots arising from the heart) (Adams et al, 1993 Stroke). The second type of stroke is hemorrhagic (10–20%), which occurs when there is bleeding into or around the brain.

Ischaemic and hemorrhagic stroke are believed to result from both shared and different determinants, although the genetic variants that influence these clinical endpoints are probably different (Humphries and Morgan 2004). A large number of epidemiological studies have shown that all forms of stroke share risk factors, such as hypertension, diabetes, hyperlipidemia, obesity and smoking (Hassan and Markus 2000; Humphries and Morgan 2004). However, there is a substantial portion of patients who do not have any of these risk factors, leading to a speculation that there are other risk factors that have not been identified yet. Furthermore, many suffer stroke despite the aggressive treatment of known risk factors.

During the past years there have been a number of significant discoveries in the field that have increased our knowledge and understanding of the importance of genetic factors involvement in stroke. Reports from family histories, twin studies and animal models support the evidence for a considerable genetic component in stroke development**,** but the extent of any genetic predisposition remains uncertain (Hassan and Markus 2000; Flossmann et al. 2004; Rubattu et al. 2004). An epidemiological study of environmental and genetic risk factors estimated that two-thirds of the population-attributable risk of stroke was due to genetic factors (Jamrozik et al. 1994; Hajat et al. 2004). As common stroke is pathologically an extremely heterogeneous disorder, there are probably many loci that act through different pathways, each providing only a minor amount of additional risk.

A large number of candidate-gene based association studies have been attempted to identify genes implicated in stroke, but unfortunately most of them have not been replicated. Several mutations are known in specific genes that cause rare Mendelian forms of stroke (Palsdottir et al. 1988; Joutel et al. 1996) but most of these are probably not involved in common forms of stroke related to arteriosclerosis (Hassan and Markus 2000; Dong et al. 2003; Zee et al. 2004). Due to the complex nature of stroke, several methodological considerations have been purposed for studying stroke genetics (Hajat et al. 2004; Dichgans and Markus 2005). For example, instead of looking at the endpoint of stroke itself, several groups have started focusing on the intermediate phenotypes of stroke (Rubattu et al. 2004; Dichgans and Markus 2005; Flossmann et al. 2005). Several candidate-gene case-control association studies have found only suggestive associations of modest effects (Rubattu et al. 2004; Rubattu et al. 2004; Zee et al. 2004). Most of these candidate-gene based studies have considered genes related to inflammation, thrombosis and lipid metabolism. However, recent meta-analysis indicate that out of thirty-two genes only four showed statistically significant associations with ischaemic stroke (Casas et al. 2004). Lack of replication can be due to relatively small sample-size or the examination of only the coding variants. Focusing only on coding variants might lead to a severely underpowered study, limited by the lack of haplotype diversity over the candidate gene of interest.

Recently, the phosphodiesterase 4D gene (*PDE4D*) and the 5-lipoxygenase activating protein gene (*ALOX5AP*) were reported to confer risk of stroke independently of conventional risk factors (Table 5) (Gretarsdottir et al. 2003; Helgadottir et al. 2004). The candidacy of these genes was identified through genome-wide linkage scans and subsequent case-control association studies in individuals from Iceland. *PDE4D* encodes phosphodiesterase 4D, which is a member of a large super-family of cyclic nucleotide phosphodiesterases. The phosphodiesterase 4D enzyme is involved in the selective degradation of second messenger cAMP, and proposed to control the level of smooth muscle proliferation and immune function in vessels thereby leading to increased or decreased atherosclerosis and hence ischaemic stroke risk (Palmer et al. 1998; Netherton and Maurice 2005).

The *ALOX5AP* (or *FLAP*) gene encodes for 5-lipoxygenase-activating protein, which is required for the cells that cluster at the injured sites in blood vessels and are implicated in the progression of atherosclerosis (Samuelsson 1983; Mehrabian et al. 2002; Spanbroek et al. 2003). In this biosynthetic pathway, unesterified arachidonic acid is converted to leucotriene A4 (LTA4) by the action of 5-lipoxygenase (5-LO) and its activating protein *FLAP*. The unstable epoxide LTA4 is further metabolized to LTB4 or LTC4 by LTA4 hydrolase and LTC4 synthase, respectively (Dixon et al. 1990). In addition, LTA4 can be exported to neighbouring cells that are devoid of 5-LO activity and become subject to transcellular leucotriene biosynthesis (Sala et al. 1996). The role of an upregulation of the leucotriene pathway in atherosclerosis is

further supported by studies on human atheromas that have shown an abundant expression of members of the 5-LO pathway in the lesions, and the number of 5-lipoxygenase-positive cells (macrophages, dendritic cells and neutrophils) is markedly increased in advanced lesions (Spanbroek et al. 2003). Furthermore, the *ALOX5AP* gene is involved in the development of atherosclerosis in mice, because the loss of only one *ALOX5* allele confers protection against atherosclerosis in LDL-receptor $^{-/-}$ mice (Mehrabian et al. 2002). On the basis of recent work it has been suggested that the elevated levels of LTB4 might contribute to atherogenesis and/or plaque instability by increasing inflammation of the atherosclerotic plaques (Gulcher et al. 2005).

Table 5. Two candidate genes, *PDE4D* and *ALOX5AP*, in stroke susceptibility

Description	***PDE4D***	***ALOX5AP (FLAP)***
Chromosome Region	5q12	13q12-13
Gene Mapping method	Linkage; association	Linkage; association
Biological function	PDE4D enzyme degrades cAMP and controls the level of smooth muscle proliferation and immune function	FLAP protein is a regulator of a pathway in the genesis of leucotriene inflammatory mediators, which are implicated in atherosclerosis
Association	Single markers, haplotypes	Haplotypes
RR*	1.98	1.67
Stroke subtype	Carotid and cardiogenic	Ischemic
Functional data	At-risk haplotype carriers had lower expression of two PDE4D isoforms.	At-risk haplotype carriers had greater production of leukotriene-B4 (LTB4)
Potential pathway in the pathogenesis of stroke	Atherosclerosis	Atherosclerosis
Studied populations	Iceland, Germany, England, US	Iceland, Scotland, Germany, England, US

* RR risk ratio for carrying the most significant haplotype; cAMP cyclic adenosine monophosphate

Both genes are good candidates for stroke susceptibility, because they encode enzymes (PDE4D and FLAP), which probably play an important role in development of atherosclerosis, one of the background processes for cardio-vascular diseases (Gulcher et al. 2005). The evaluation of both genes in stroke development has been recently studied in many populations outside of Iceland (Bevan et al. 2005; Helgadottir et al. 2005; Meschia et al. 2005). However, the replication results are fairly mixed and therefore, the involvement of both genes in stroke development still need stronger evidence.

2. PRESENT INVESTIGATIONS AND DISCUSSION

2.1. Aims of the present study

The main goal of the present study was to investigate LD and haplotype structure in the human genome and to study the variation patterns in Europe in order to facilitate further association studies.

The specific aims of the current thesis were following:

1. to describe the LD and haplotype structure in different genomic regions (Ref. I, II, III)
2. to investigate the genetic variability of LD and haplotype structure among European populations (Ref. II)
3. to provide a framework for association mapping by studying tagSNP performance and transferability among European populations (Ref. II)
4. to study the contribution of two candidate genes in the development of stroke in continental Europe population (Ref. IV)

2.2. Characterization of general LD and haplotype structure in different genomic regions (Ref. I, II, III)

As described in the literature overview, the design and feasibility of whole-genome association studies is critically dependent on the extent of LD between markers. An extensive knowledge about the patterns of LD and hapotypes in the human genome is required. To investigate the structure and the patterns of LD in the human genome we analysed different regions of the genome (Table 6).

Table 6. General information about studied genome regions

Region/Gene	Region Size	No. of Common SNPs*	Average Spacing (kb)	Studied Populations and Sample Size (n)	Study Purpose	Ref.
Chromosome 22q	33.4 Mb	679	15	Estonia (51), CEPH (77), UK (90)	Extent of LD; patterns of LD	I
FKBP5 6p21.13	289 kb	37	6.7	CEPH (30 trios); Estonia (170);	Extent of LD; patterns of LD and haplotypes; block boundaries; tagSNP performance and transferability	II, III
SNCA 4q21	188 kb	73	2.1.	POPGEN (160); SHIP (100); KORA (170);		II
LMNA 1q21.2	177 kb	27	4.4	Vinchgau (170); Ladinia (160);		II
PLAU 10q22.2	95 kb	32	2.2	Brisighella (98); Calabria (100)		II

FKBP5 -FK-506 binding protein 5; *SNCA* – synuclein, alpha; *LMNA* – lamin A/C; *PLAU* – plasminogen activator, urinary; CEPH – Centre d'Etude du Polymorphisme Humain; UK – United Kingdom; POPGEN – population samples collected in Schleswig-Holstein, Germany; SHIP – Study on Health in Pomerania; KORA- Cooperative health research in the Region of Augsburg; * MAF > 20% (for Chr. 22) and >5% for other regions.

2.2.1. First-generation LD map of chromosome 22 (Ref. I)

Chromosome 22 is the second smallest chromosome in humans, comprising 1.6-1.8% of the whole genomic DNA (Morton 1991; Dunham et al. 1999). Due to the relatively small size, chromosome 22 was the first fully-sequenced human chromosome (Dunham et al. 1999). It is one of the five human acrocentric chromosomes, each of which shares substantial sequence similarity in the short arm, which encode tandemly repeated ribosomal RNA genes and contain a series of other tandem repeat arrays. There is no evidence to indicate the presence of any protein coding genes on the short arm of chromosome 22 (22p). The long arm of chromosome 22 (22q) is 33.4 Mb long and is rich in genes (at least 545 genes and 134 pseudogenes) compared to other chromosomes (Deloukas et al. 1998). The full sequence and a SNP map of chromosome 22 (Mullikin et al. 2000; Dawson et al. 2001) provided an excellent opportunity to characterize the LD patterns across the whole human chromosome.

All genotyped markers were selected from chromosome 22q (33.4 Mb) in the Sanger Center. Markers were obtained from publicly available SNPs and small insertions/deletions (indels) by walking through the chromosome from the centromere in 15 kb steps and choosing the nearest variant that was suitable for the Invader genotyping assay. The genotyping in the Sanger Center was done using the CEPH sample collection (77 samples in seven three-generation pedigrees) and 90 unrelated UK Caucasian samples. Here, in Estonia, we aimed to genotype all previously selected SNPs using the Arrayed Primer EXtension (APEX) method. This method is based on a allele specific primer extension in microarray format (Kurg et al. 2000). We first designed the assays for studying 1638 markers simultaneously on one chip. If a marker was not amenable to use in the APEX assay then it was rejected from the chip design. We found that only 1279 SNPs had good parameters to design assays for the APEX method. The assays failed mostly because it was not possible to design suitable primers for the APEX reaction. The smaller number of genotyped SNPs in Estonia is partly caused by the fact that SNPs were originally selected for genotyping with the Invader assay. Moreover, the ongoing SNP discovery project in the Sanger Center allowed researchers to continuously add markers to the assay. We were able to develop a specific chip for the simultaneous analysis of 1279 SNPs and genotyped 51 individuals from the Estonian population. After genotyping, the data was cleaned to remove non-polymorphic markers, SNPs with low number of calls (quality issues due to primer variants or different experimental failures) and SNPs with Hardy-Weinberg Equilibrium deviations. After the cleaning procedure, 908 SNPs remained. In order to avoid inconsistencies caused by genotyping errors using different technologies, we genotyped the same set of CEPH samples as the Sanger Center. The final marker set had a median SNP spacing of 34.72 kb (average 61.42 kb) and included 594 SNPs in common with the initial CEPH SNP set. The final set of markers in CEPH family panel consisted of 1504 polymorphic SNPs and 1286 SNPs in the UK samples.

We examined the extent of LD among 661 biallelic markers and for each pair of markers the two most commonly used pairwise disequilibrium measures *D'* and r^2 were calculated. As the precision of individual LD estimates suffers at low allele frequencies we used SNPs with minor allele frequencies >0.2. Plotting the moving average of *D'* and r^2 (comprising 200 marker bins, 100 marker overlaps) demonstrated that the LD between neighboring SNPs is strong but decays rapidly with increasing distance. The dataset from three different samples (CEPH, UK Caucasians and Estonians) were compared and very similar results were obtained (fig. 1c, d in ref. I). Both LD measures show similar decay profiles although they differ in scale. On average, our results show that significant disequilibrium can be detected at least up to 50 kb distances, which is greater than predicted by simulations (Kruglyak 1999). However, the decay of disequilibrium varied considerably between marker pairs, where maximal *D'* values extend up to distances over 400 kb, contrasting with occurrences of no detectable LD between markers less than 5 kb apart (fig. 1a and b; ref. I). Our findings on the extent and variability of disequilibrium are consistent with other studies (Clark et al. 1998; Eaves et al. 2000; Templeton et al. 2000; Abecasis et al. 2001).

We assessed the pattern of LD along the chromosome by calculating average *D'* and r^2 for all marker pairs separated by 50 to 500 kb in a sliding window manner within contiguous 1.7 Mb stretches of DNA (1.6 Mb overlap). The results demonstrated that LD is not continuous and there are islands with high LD separated by low LD. Areas with very high levels of LD were detected at positions 11-16 Mb and 21-27 Mb of the reference sequence (fig.2a in ref. I). Unfortunately the marker density in the Estonian dataset was too coarse (median spacing 34.72 kb) for the formal delineation of specific regions. In the CEPH and UK samples, average *D'* levels in the regions of high LD were 2–5 times greater than background levels, presenting obvious distinctions of high and low LD tracts, whereas in the Estonian data, the highest LD region wes less than twice the background level (fig 2a in ref. I). This indicates that the first-generation LD map requires a median marker density greater than one marker per 35 kb. The SNP densities in current maps are on average 1 SNP per 2 to 5 kb, thus allowing a more precise estimation of LD properties (Altshuler et al. 2005; Hinds et al. 2005).

In summary, this first-scale LD structure of chromosome 22 yielded important knowledge for further fine-scale studies in this chromosome. The accumulation of a high density of markers will allow the systematic investigation of the patterns and causes of LD variability. Several complex diseases, including many psychiatric disorders, involve some genetic component on chromosome 22 and therefore the characterization of an LD map for chromosome 22 could facilitate the identification of genetic variants for common complex diseases.

2.2.2. Fine-scale LD structure across selected genomic regions (Ref. II, III)

The chromosome 22 LD study indicated that higher marker density is required for a detailed description of LD and hapotype patterns. Here, in this study we analyzed LD structure and haplotype diversities in four genomic regions which all contain candidate genes for different complex diseases. The SNPs were chosen to be evenly distributed throughout the regions with an average spacing of ~2-4 kb (Table 6; table 1 in ref. II). The size of the analyzed region varied according to the underlying candidate gene size, between 95 kb in the *PLAU* gene region to 288 kb in the *FKBP5* gene region. For every selected region the assays were designed to cover the gene and downstream and upstream flanking regions. Genotyping was done by detection of the allele-specific extension products using matrix-assisted laser desorption/ionization time of light (MALDI-TOF; Sequenom) mass spectroscopy. The results showed that each genomic region had a unique and the unpredictable pattern of LD (Figure 7; fig. A2 in ref. II). For example, in the *SNCA* region a clear LD break is visible within the gene, whereas around the *PLAU* gene the whole region showed strong allelic association between markers. High marker density allowed us to define blocks in each region. In the CEPH samples using the standard algorithm of Gabriel et al. (2002) we detected four blocks in *SNCA*, six blocks in *LMNA*, six blocks in *FKBP5* and two blocks in PLAU (fig2 in ref. II). Consistent with previous studies, each detected block in all regions consisted of only a few common haplotypes (fig.A3 in ref. II).

2.2.3. LD and block structure in the FKBP5 gene, associated with rapid response to antidepressant treatment (Ref. III)

During the association studies researchers may detect many genetic variants that show positive correlation with a disease phenotype. In order to confirm these positive findings the detailed characterization of LD structure and extent of LD across the investigated region is required. This knowledge may be very helpful for investigators in designing and interpretating results of association studies. Here, in this study several candidate gene polymorphisms were selected to investigate a possible association with susceptibility to depression and a response to antidepressants. The first part of the study found a significant association between three SNPs in the *FKBP5* gene and response to depression treatment. In order to draw conclusions about the role of the *FKBP5* gene in depression, we aimed to study in detail the whole region of the *FKBP5* gene. We genotyped additionally 27 SNPs covering 288 kb around the *FKBP5* gene to investigate whether the previous results could indeed be attributed to *FKBP5* and not to adjacent genes. The investigated region located in chromosome 6p21.31, included the 115 kb long *FKBP5* gene, the *TULP1* gene (3' of *FKBP5*), the hypothetical protein FLJ25390 and the *CLPS* gene (both 5' of

FKBP5). SNPs were selected to cover all genes in this region, with an average distance of 9.6 kb between SNPs. Based on *D'* we constructed an LD block map for this region.

We detected a single large LD block that encompasses most of the *FKBP5* gene plus a 5' region of the gene, with the block ending before FLJ25390 (fig. 1A in ref. III). New association tests with 27 SNPs of this candidate region revealed three SNPs with strong association with response to antidepressant treatment. The strongest associations were detected with SNPs within the largest LD block containing the *FKBP5* gene (fig. 1B in ref. III). One associated SNP located at the 5' end of the gene, the second in intron 2 and the third in the 3' untranslated region of *FKBP5*. Thus, the clustering of significant associations within the LD block containing *FKBP5* indicated that the observed associations can be most probably attributed to this gene. In summary, this part of the study indicates that determining the LD structure of a candidate gene region helps to confirm the gene candidacy, reduces the region size to be analyzed and finally, increases confidence of future observations. The overall study proposed that *FKBP5*, a glucocorticoid receptor (GR) regulating co-chaperon of hsp-90, plays a central role in regulating the hypothalamic-pituitary-adrenal (HPA) axis in causality of depression and the mechanism of action of antidepressant drugs.

2.2.4. Summary of LD structure based on studied regions

Based on the results and observations from different genomic regions, some general conclusions about LD patterns in the human genome can be drawn. First, results from all regions clearly indicate that LD decays rapidly with increasing physical distance. Secondly, we detected high heterogeneity in LD structure in different chromosomal regions. This is consistent with previous studies (Abecasis et al. 2001; Gabriel et al. 2002) and indicates that LD is unpredictable and has to be established for each gene region separately. For example the study with the *FKBP5* region demonstrates that detailed knowledge of the LD structure in a candidate-gene region allows researcher to improve the interpretation of association results.

Next, the regions of low LD were clearly separated from high LD regions that have limited haplotype diversity. These regions with limited haplotype diversity are of primary interest for practical implications. In the chromosome 22 study haplotypes were calculated only for the CEPH and UK samples. The collection of SNPs in the Estonian dataset was evenly distributed, but the mean distance between consecutive markers was only about 35 kb and therefore the haplotype structure was not studied. The distance between consecutive SNPs is critical for defining meaningful haplotype structure and the block length depends on the marker density used. Haplotype blocks, which can be as short as a few kilobases, may be unrecognized if the distance between consecutive SNPs is large relative to the size of the actual haplotype blocks. The longest haplotype

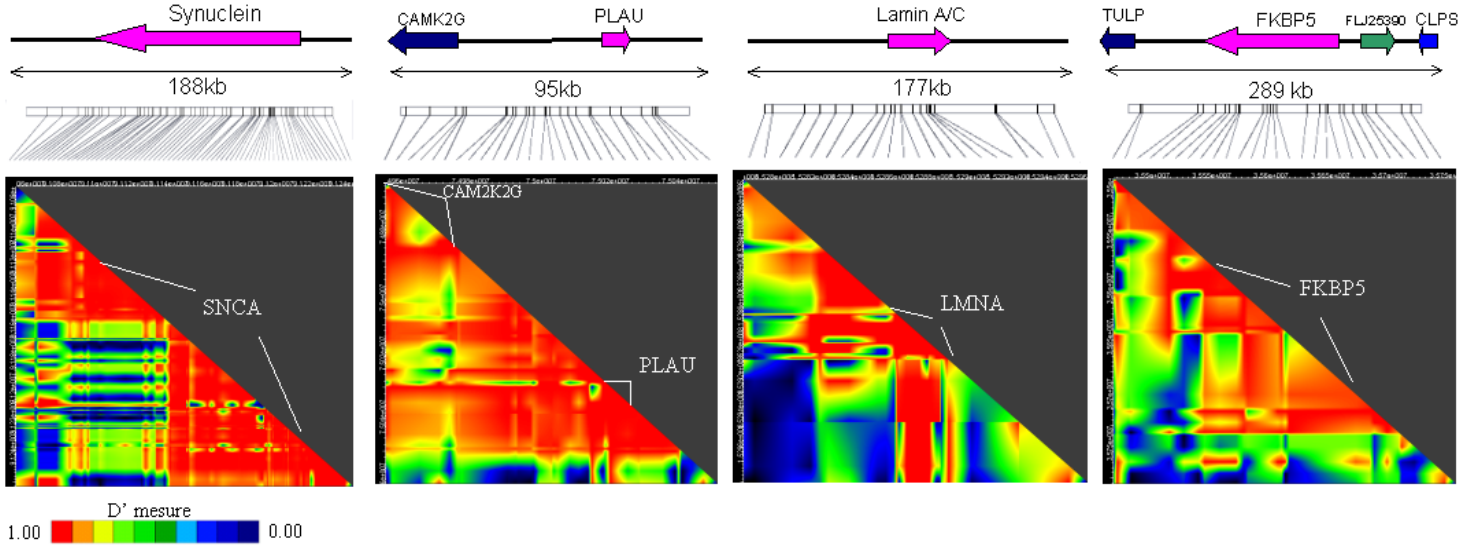

Figure 7. SNP distribution and fine-scale LD structure in four studied gene regions.

in chromosome 22 extended 804 kb (at 11.83 Mb, including 16 markers). The biological reasons for the distribution of LD and the existence of long haplotype blocks may be due to regional characteristics of DNA that induce differences in mutation or recombination rates. It has been shown in many studies that high LD regions correlate well with the general recombination rate of the investigated region (Jeffreys et al. 2001; Zhang et al. 2002). Previous studies have shown that compared to the genome average chromosome 22 has an elevated degree of recombination (Dunham et al. 1999; Yu et al. 2001). The detected high LD regions at 11-16 Mb and 21-27 Mb are located in regions of exceptionally low recombination rate relative to the chromosome average which explains the relatively long haplotypes and low haplotype diversity. On the other hand, many recent studies have stressed that marker density is the most important factor influencing haplotype block partitioning (Stumpf 2004; Ke et al. 2004b). The high marker density in four candidate gene regions allowed more detailed analysis of haplotype block structure and content. The average block length in observed regions falls into the same range as reported for non-African populations in other studies (Gabriel et al. 2002; Altshuler et al. 2005).

Finally, the comparison of general pairwise LD structure, as well as the decay of LD among different European population samples showed quite similar patterns across all studied regions (Figure 8; fig. 1c, d in ref. I; fig. A2 in ref. II). The chromosome 22 study involved samples from Estonia and UK. The candidate region study used eight different European populations selected along a line from north to south (fig 1. ref. II). Both studies analysed CEPH samples, which is an emigrant population of northern and western European origin. Since the biological determinants of LD (rate of recombination, mutation) are expected to be constant across populations, the similarty of LD among European populations presumably indicates the common evolutionary history. However, high SNP density allowed the detection of slight differences in the decay of LD across populations (Figure 8). For example in the distance range between 20 kb and 150 kb the Ladinians and the Italian populations from Calabria and Brisighella showed the highest r^2 values, whereas Estonia showed the lowest r^2 values. This sequence of populations stresses the infuence of population specific factors (like effective population size, relative isolation).

A low marker density, like in the chromosome 22 study, allowed us to investigate LD only in large-scale. However, chromosome 22 was one of the first studies where LD structure was described across the whole-chromosome and it became a model for further LD studies. During the last two years, our knowledge about fine-scale LD and haplotype structure has improved tremendously. At present, we have the possibility to use HapMap data to investigate the correlation of LD with sequence elements and recombination across the human genome with a marker density up to 1 kb.

2.3. The variability of LD and haplotype structure among European populations and implications for association studies (Ref. II)

Variable demographic factors are responsible for shaping patterns of LD in different population groups. Large-scale association studies are usually needed to detect small genetic effects on complex traits. The specific population demographic history may strongly influence the outcome of association studies. Therefore, the knowledge of genetic variability in human populations is important to obtain reliable results in association studies.

2.3.1. The European LD and haplotype variability

In order to investigate how much variability exists in LD structure and haplotype composition between European populations we analyzed the previously mentioned four autosomal regions in 8 European populations. To compare the LD structure across populations in a detailed and robust probability-based assessment, we developed a simple bootstrap approach based on the standard algorithm of Gabriel et al. (2002). We estimated the boundary frequencies for the start and end of blocks separately, which enabled us to track individual blocks. In addition, we allowed blocks to overlap each other, which gives a more natural framework. For each population equal numbers of individuals and exactly the same set of markers were used to evaluate the strength of block boundaries. The general patterns of block structure were similar across samples. However, we were also able to show clear examples of block boundary shifts and block fragmentation among European samples (fig2 in ref. II). For example in the *LMNA* gene, five of six blocks have nearly conserved block beginnings and ends across all studied populations. Only the end of the largest block varied between populations and presents a shift in the block extension between 7 kb and 15 kb (fig.2B in ref. II). Notably, the LD block boundaries revealed by the standard Gabriel et al. (2002) algorithm (without allowing overlapping blocks) did not always coincide with the position of the highest bootstrap frequency (for example the start of block 4 in CEPH in fig 2A in ref. II). Many methods have been proposed for haplotype block partitioning and it is well demonstrated that block structure strongly depends on marker density as well as on methods applied (Cardon and Abecasis 2003; Schwartz et al. 2003; Wall and Pritchard 2003; Ke et al. 2004b).

The overall variability in block structure among populations showed that the most different block structures for all gene regions were observed in Alpine populations (Vinschgau and Ladinia) and geographically peripheral populations (Estonia, Brisighella and Calabria) (fig. 3 in ref. II). The reference CEPH population and all studied German populations show a similar intermediate

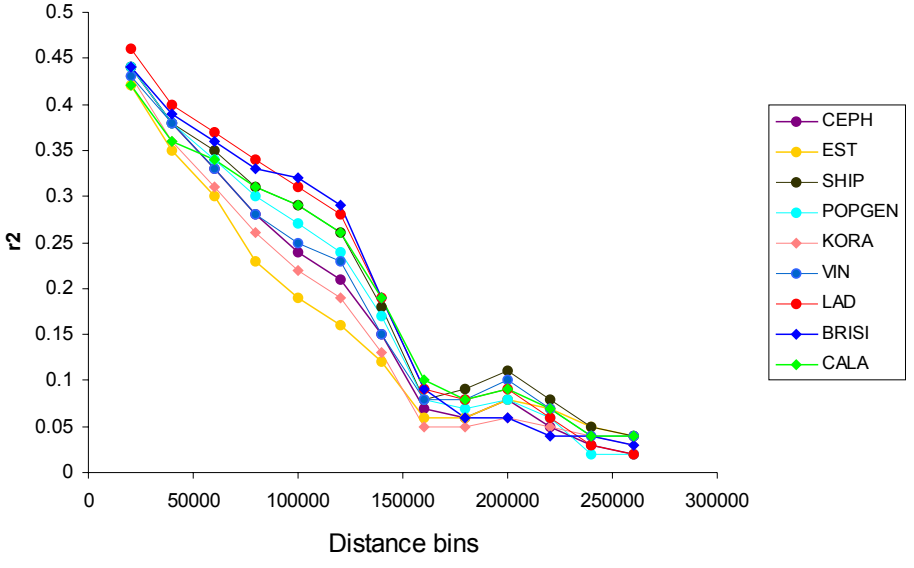

Figure 8. Long-range LD across four gene regions (*SNCA, FKBP5, LMNA, PLAU*) in European populations. Moving averages of r^2 versus physical distances, averages are calculated across 40 kb bins with 20 kb overlap (EST-Estonia, VIN-Vinschgau, LAD-Ladinia, BRISI-Brisighella, CALA-Calabria)

block structure. The values of similarity measure of block structure, which estimates the average probability that boundaries coincide, ranged from 0.72 among Ladinia and Brisighella to 0.87 among SHIP and POPGEN. The overall variation appeared in a pattern that was concordant with geography, where specific boundaries vary in a geographical sense and differentiate in accordance with the level of presumed genetic isolation of populations. Comparative analysis of haplotype block structure in other studies have revealed both a significant variation (Liu et al. 2004) and also a high degree of concordance among European populations (Nejentsev et al. 2004; Ng et al. 2004; Stenzel et al. 2004). This similar pattern between populations on one hand reflects a common variation pattern of recombination rate and on the other hand shared ancestry of human populations. However, it might also reflect the effect of uneven marker spacing in these studies. Importantly, for several reasons it is unlikely that blocks have a clear beginning and end. Rather blocks should be considered as models that try to capture the main features of LD patterns.

General LD and haplotype block profiles do not allow inference of the haplotypes or their frequencies. In contrast, underlying estimated haplotypes and their frequencies are the primary data from which the LD profiles can be generated. The relationships between LD and haplotype frequencies are complex. To study the haplotype composition we applied the standard Gabriel et al. (2002) algorithm to the reference CEPH population and applied exactly the same haplotype blocks to each studied population. The haplotype distribution in European populations exhibit strong similarity and all populations shared the same few common haplotypes, indicating the remarkable influence of population history. However, we also observed considerable differences in haplotype frequencies among populations. Three gene regions (*PLAU*, *FKBP5*, *SNCA*) had at least one block where significant common haplotype frequency differences among populations were found (fig. 4 in ref. II). These frequency patterns clearly indicate a geographical gradient among populations. We observed a frequency gradient for both, single marker allele frequencies as well as for haplotype frequencies. This gradient was most pronounced for the *PLAU* gene region, where in both blocks the most common haplotype frequency values between Estonia and Calabria differed by ca. 20% and a clear gradient among the remaining six populations was seen. The *PLAU* gene region showed overall remarkable differences in genetic variability among populations compared to the other three gene regions. This region showed the highest number of population-differentiating SNPs (79%). In the other three gene regions the number of population-differentiating markers varied between 6% and 23% (table1 in ref. II). *PLAU* is a putative susceptibility gene for Alzheimer disease and is located in chromosome 10q22.2. The underlying genetic variability may cause a different disease risk among populations. These differences emphasize that special care has to be taken to design and interpret the results of association studies in regions that show high genetic variability among populations.

As demonstrated above, the observed overall pattern of genetic differentiation reflects well the geographical location of the population. The maximum values for allele and haplotype frequency differences were observed between Estonia and Calabria, thus indicating a geographical gradient between the northern and southern populations. This finding corresponds relatively well to the European genetic variation described by previous studies (Barbujani and Sokal 1990; Cavalli-Sforza et al. 1994). An allele-frequency gradient between European populations is a repeatedly observed phenomenon and reflects the evolutionary and/or demographic history of Europe. Considerable frequency differences among populations within the same geographical areas have been observed in recent studies (Beaty et al. 2005; Sawyer et al. 2005). It seems that variability in haplotype frequencies among populations mostly appear in specific gene regions or in populations with specific demographic histories. For example, demographic events such as genetic drift, bottleneck and admixture are predicted to increase the LD level and have a strong impact on the extent and patterns of regional LD. In our population set the relatively strong genetic divergence of the Italian and the two Alpine populations from all other populations can be attributed to the isolation resulting from physical boundaries (Alps) or linguistic differences (Germanic-Romanic). Those populations showed significant differences from the CEPH samples and are therefore less well represented by this reference population. The knowledge of LD patterns assists the association mapping of complex diseases and the results of our study emphasize the importance of understanding the history, structure and variation of a study population. The observed population differences in haplotype frequencies and LD structure may affect the power to detect phenotype-genotype associations. Association signals with markers, which are correlated with a true causal variant, may appear at different positions in populations with an individual LD structure.

2.3.2. tagSNP performance and transferability among European populations

Since population differences can strongly influence the patterns of association, the selection process of markers to be genotyped is crucial. It is important to study how well would tagSNPs that are selected and account for the most common haplotypes in one population work in other populations. It is often stated that tagSNPs are population specific and should be newly assessed in each local population or geographical area in which an association study is planned (Thompson et al. 2003; Weale et al. 2003; Carlson et al. 2004). On the other hand, recent studies have demonstrated that HapMap data may be used to define tagSNPs for related populations (Ke et al. 2004a; Altshuler et al. 2005). The populations used in the HapMap Project are suggested as reference populations for the selection of tagging SNPs for association studies.

Alternatively a subset of local population samples can be used to create a haplotype map which can then be used for tagSNP selection. One debated question is whether the HapMap constructed map would be applicable for association mapping for all European populations or if more empirical data on different populations from Europe is needed. To address this question a simulation was used to test the transferability of tagSNP sets among European populations.

To compare the efficiency of tagSNPs we first used CEPH trios as a reference to select tagSNPs, which were defined to represent untagged SNPs with a high correlation coefficient (r^2 >0.8) (Carlson et al. 2004). TagSNPs were also defined for local population samples with different sample sizes. For this comparison the tagSNPs were selected from a reduced SNP set, which was comparable to the HapMap set (table 1 in ref. II). This reduced SNP set comprised about 40% of SNPs identical to the HapMap data and was then used to test against the full SNP set in all populations. Our results indicate that tag SNPs defined in the HapMap CEPH trios perform relatively well for two out of four candidate-gene regions, particularly in central European populations (fig 5 in ref. II). However, for two of the tested candidate genes (*PLAU* and *FKBP5*), CEPH is not a good reference. For example, in the PLAU region two populations showed a ratio of tagged SNPs of <70%, when the CEPH sample was used as a reference. For the same gene, data from 20 random individuals of most populations performed better as a reference than the CEPH trios. Alternatively, the haplotype-based SNP selection method (htSNP) (Zhang and Jin 2003) was used to test performance of CEPH trios as reference for SNP selection. This algorithm finds SNPs that represent common haplotypes within predefined blocks. When CEPH trios were used as a reference, the chromosomal coverage of tagged haplotypes was below the intended threshold (80% and 90% respectively) only for the *PLAU* gene (table A5 in ref. II), thus indicating similar performance with the tagSNP selection method, but differences between CEPH and local populations were weaker. We also tested how SNP density would influence the tagSNP performance. A substantial increase in tagSNP efficiency and transferability is achieved by increasing the density of genotyped SNPs in the reference sample (table A4 in ref. II). All our simulations indicated that by genotyping larger sample sizes (>20 individuals), the advantage of a local population reference will be stronger, but increasing sample size beyond 40 individuals is not very effective (fig5 in ref II). Finally, our results indicate that the majority of tagSNPs did not show strong population differences, underlining their universality (fig. 6 in ref. II).

In summary, two studied candidate-gene regions (*SNCA* and *LMNA*) revealed similar haplotype patterns and consistent blocks among European populations and were therefore well represened by tagSNPs identified in the CEPH population. By contrast, two other regions (*PLAU* and *FKBP5*) showed clear population differences in terms of haplotype frequencies and tagSNP transferability. Similar results have been shown in other studies, where tagSNPs identified in one population may not necessarily perform well in another (Weale

et al. 2003; Liu et al. 2004). For these regions special care has to be taken and it has been suggested that a preliminary study to identify tagSNPs and the later large-scale case-control study should be performed in the same population. Our results suggest that future HapMap releases with more genotype data will allow sufficient selection of tagSNPs and capture most variations or haplotypes in the majority of genes in the main European populations. This was recently shown by analysing ENCODE regions in the HapMap project. Moreover, it has been demonstrated that selecting a higher number of tagSNPs may work adequately in multiple human populations (so called "cosmopolitan" tagSNPs) even for gene regions where the patterns of LD are markedly different among populations (Ahmadi et al. 2005).

Only a limited number of studies have been published so far where tagSNP transferability among European populations has been evaluated. However, most of them are quite encouraging, showing that tagSNPs selected in the CEPH population (Ke et al. 2004a; Miretti et al. 2005) or in some local European population (Nejentsev et al. 2004) can be well transferred to another European population. Available studies also indicate that the CEPH sample in general may have been a good choice as a reference population in the HapMap project as this population seems to be suitable for characterizing a broad view of LD and haplotype structure for the general European population.

2.4. Replication of genetic association studies: the roles of *PDE4D* and *ALOX5AP* genes in stroke development (Ref. IV)

Association studies offer a potentially powerful approach to identify genetic variants that influence susceptibility to common disease, but unfortunately many of them are not consistently reproducible. Replication of the intitial findings in an independent group of subjects provides further evidence for the involvement of the gene in the trait. The aim of this study was to investigate the role of two previously reported candidate genes in susceptibility for stroke phenotype. Involvement of *PDE4D* and *ALOX5AP* in the susceptibility of stroke has been previously shown in deCODE studies (Gretarsdottir et al. 2003; Helgadottir et al. 2004). In our study a total of 639 stroke patients and 736 unrelated population-based controls from Germany were genotyped. In this study, we applied a 2-step replication approach. We first examined the most significant single markers and haplotypes that were associated with stroke in the Icelandic population. Analysis was then extended to additional markers and haplotypes.

In the original study from deCODE, a single *ALOX5AP* haplotype (termed HapA) was found to double the risk of MI and stroke in patients from Iceland (Helgadottir et al. 2004). Hap A (defined by four SNPs) is relatively common

and is carried by 27% of patients with stroke and conferred a 1.67-fold increased risk for stroke. In the current study, we analyzed 22 SNPs that had been used for haplotype analysis in the original study (Helgadottir et al. 2004). We were not able to replicate the association with the Icelandic at-risk haplotype (HapA) of *ALOX5AP*. However, we found a significant association with several SNP markers. The most significant association was found with one single SNP (SG13S114) constituting HapA haplotype (table 1 in ref. IV). Moreover, the association between SG13S114 and stroke was stronger in males than in females, which is similar to the findings from Iceland. All haplotypes that had shown significant association in Iceland contained either the T allele of SG13S114 or the A allele of SG13S100, both of which showed nominally significant P values in the current study.

Because we could not replicate the association with HapA, we expected that there might be other unidentified haplotypes in *ALOX5AP* that confer risk to stroke in the German population. This assumption was based on the fact that in the original work researchers found another at-risk haplotype (HapB) which was associated with MI in individuals from the United Kingdom (Helgadottir et al. 2004). We found several haplotypes that were significantly associated with stroke. However, after correction of multiple testing, associations were no longer significant (Table I in ref. IV). To check for possible differences in LD structure between German and the Icelandic samples, we determined the LD structure and haplotype diversity of *ALOX5AP* (Table I in ref. IV). Detected LD block structure was similar to that in the Icelandic population, where most studied SNPs were strongly correlated (fig. 1 in ref. IV).

In the *PDE4D* gene, the initial deCODE study found that specific single markers and haplotypes at *PDE4D* are associated specifically with carotid and cardiogenic stroke but not with other stroke subtypes. The most significant at-risk haplotype conferred a 1.98-fold increased risk for the combined group of carotid and cardiogenic stroke patients, whereas a protective haplotype gave a risk ratio of 0.68 (Gretarsdottir et al. 2003). Neither the risk nor the protective haplotype correlated with nearby missense or nonsense mutations, but the variants did correlate with the expression of *PDE4D*. The strongest association was found with markers in the alternative-promotor region of one of the eight isoforms of *PDE4D* (Gretarsdottir et al. 2003). For replication of the inital findings' we selected 2 SNPs and 1 microsatellite marker which had shown the most significant association with stroke in the original study. In addition, we analyzed an optimal set of haplotype tagging SNPs (htSNPs) that distinguished 95% of all chromosomes within two predefined LD blocks in the *PDE4D* gene region (Gretarsdottir et al. 2003). In the current study, no significant association was found between single markers or specific hapltypes of *PDE4D* and stroke (tables 3 and 4 in ref. IV). The LD structure analysis of the *PDE4D* gene revealed slight differences between the German and Iceland datasets, indicating less LD between markers in the German population (fig 2 in ref. IV). Since our analysis involved only htSNPs, it is possible that our set of htSNPs may have

covered too little of the genetic variability in *PDE4D* in the German population. Assuming a multiplicative model and the relative risk of single markers and haplotypes reported by deCODE, the power of our sample to detect the same associations should be >99%. Consistent with our findings, a recent replication study with an English stroke cohort reported also negative associations between *PDE4D* and the general stroke phenotype (Bevan et al. 2005). Those two reported independent replication studies, both with a high number of samples, indicate that the *PDE4D* gene is probably not a major risk factor for stroke in continental European populations.

Association studies in both stroke candidate genes indicate that different populations may have different patterns of association. In contrast with our study, the association with the HapA haplotype in *ALOX5AP* gene has been recently replicated with a Scottish ischaemic stroke cohort (Helgadottir et al. 2005). Historical and archaeological data suggest a Gaelic ancestry for both Icelanders and Scots (Helgason et al. 2000; Helgason et al. 2001). Given this common ancestry, it is possible that the two populations share the same disease-causing variant and this variant may reside on the same common haplotype background (HapA) (Helgadottir et al. 2005). The observed lack of association with HapA in our study may relate in part to population differences in allele and haplotype frequencies. In support of this, the frequency of HapA was much higher in our German samples of control individuals than in the Icelandic controls (table 2 in ref. IV). If we compare all three independent large-scale association studies in the *ALOX5AP* gene then we can see marked HapA frequency differences among the three studied populations (Table 7). Both German as well as Scottish samples have higher HapA frequency in controls in comparison with Iceland. This comparison also indicates a more than two times less significant association in the replication study with Scottish samples. Frequencies of the HapB haplotype seem to be similar among all studied populations (Table 7).

In the *PDE4D* gene, similar to the *ALOX5AP* gene, marked frequency differences in at-risk wild-type and protective haplotypes between Iceland and German controls were detected (table 4 in ref IV). It is well known that the Icelandic population is not typical, being a genetic isolate and may have strong founder effects. A founder effect in Iceland could explain the elevated level of LD in the *PDE4D* gene. Therefore it might be that outside of Iceland another undetected variant(s) or haplotypes in the *PDE4D* gene could confer risk to stroke. For example, a recent replication study with North-American stroke samples found an association with other PDE4D gene variants in comparison with the original work (Meschia et al. 2005). It has been suggested that differences between populations may occur due to different patterns of association of the high–risk allele with marker alleles and haplotypes (Botstein and Risch 2003; Clark 2003).

Despite detecting associations between both candidate-genes and ischaemic stroke, the deCODE group was unable to find specific mutations explaining this

increased risk. Therefore, they described a haplotype on the basis of a number of SNPs associated with increased risk. Focusing on the expansion of haplotype diversity over the LD blocks encompassing the candidate gene provides an opportunity to detect non-coding disease-associated variants. Our analyses involved upstream regions (putative promotor regions) of both candidate genes, but variants involved in complex disease can also be located in other regulatory regions of genes.

Table 7. Association of 'HapA' and 'HapB' (*ALOX5AP)* to stroke

Samples	HapA			HapB		
	% Cases	% Controls	P-value	% Cases	% Controls	P-value
Icelandic sample[1]	14.9	9.5	0.000095	6.7	7.3	ns
German sample[2]	14.5	15.2	ns	6.1	7.6	ns
Scottish sample[3]	18.4	14.2	0.007	5.8	6.8	ns

Haplotype frequencies determined by [1] Helgadottir et al., 2004; [2] Lõhmussaar et al., 2005; [3] Helgadottir et al., 2005

Both candidate genes have been suggested to be good biological candidates involved in the disease mechanism associated with arteriosclerosis, the pathological basis for most stroke and MI (Gulcher et al. 2005). Moreover, preliminary functional studies also support their candidacy (Gretarsdottir et al. 2003; Helgadottir et al. 2004). For example, functional studies demonstrated that *PDE4D* high-risk isoforms present in affected individuals were associated with lower gene expression. It has been argued that greater activity of one or a few splice variants alters the *PDE4D* enzymatic activity of the cell, decreasing the cAMP levels and thus altering the expression of cAMP-regulated isoforms. However, it is still unclear what the specific roles of both candidate genes in stroke development are and, therefore, additional well-designed association and functional studies with different population samples are required to clarify the role of both genes in involvement of the stroke phenotype.

CONCLUSIONS

The summarized results of the study:

1. The characterization of the LD structure of chromosome 22q reveals a high variability along the chromosome, in which extensive regions of nearly complete LD are interspersed with regions of little or no detectable LD. The LD is not dependent only on the distance between markers, but is also influenced by properties of the particular region in the human genome. Similar trends of the variability of LD were observed in different European populations in different genomic regions. This study demonstrates the feasibility of developing pan-European genome-wide maps of LD.

2. The LD structure of *FKBP5* gene region was established. The construction of a LD map in the chromosome 6p21.31 region allowed us to confirm the candidacy of *FKBP5* gene involvement in response to antidepressants and susceptibility to depression.

3. The comparison of LD patterns and haplotype variability in four complex trait candidate gene regions (*PLAU*, *SNCA*, *FKBP5*, *LMNA*) across European populations lead to several conclusions: (1) although there was significant conservation of LD patterns across European populations, several shifts in the position of boundaries of high LD regions was demonstrated among populations; (2) the observed population differences in allele and haplotype frequencies indicate a geographical gradient among the northern and southern populations; (3) the transferability of tagSNP sets was tested among populations and tagSNPs defined in the HapMap CEPH trios performed relatively well in the *SNCA* and *LMNA* gene regions. However, variation in two other gene regions (*PLAU* and *FKBP5*) predicts a restricted applicability of the CEPH derived tagging markers; (4) a higher SNP density substantially increases the HapMap applicability.

4. Evaluating of the candidacy of *ALOX5AP* and *PDE4D* as susceptibility genes for stroke suggests that sequence variants in the *ALOX5AP* gene are significantly associated with stroke, particularly in males in a central European population of stroke patients. Variants in the *PDE4D* gene are not a major risk factor for stroke. The observed population differences in allele and haplotype frequencies as well as LD structure may contribute to the observed differences between Icelandic and German populations.

REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191–197

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LSt, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310:317–320

Adams HP, Jr., Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, Marsh EE, 3rd (1993) Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 24:35–41

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

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936–950

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320

Altshuler D, Clark AG (2005) Genetics. Harvesting medical information from the human family tree. Science 307:1052–1053

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76–80

Arcos-Burgos M, Muenke M (2002) Genetics of population isolates. Clin Genet 61:233–247

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

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582–589

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M, Eichler EE (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet 70:83–100

Bamford J, Sandercock P, Dennis M, Burn J, Warlow C (1991) Classification and natural history of clinically identifiable subtypes of cerebral infarction. Lancet 337:1521–1526

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111
Barbujani G, Goldstein DB (2004) Africans and Asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119–150
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
Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci U S A 87:1816–1819
Beaty TH, Fallin MD, Hetmanski JB, McIntosh I, Chong SS, Ingersoll R, Sheng X, Chakraborty R, Scott AF (2005) Haplotype diversity in 11 candidate genes across 4 populations. Genetics
Belanger H, Beaulieu P, Moreau C, Labuda D, Hudson TJ, Sinnett D (2005) Functional promoter SNPs in cell cycle checkpoint genes. Hum Mol Genet
Bennett ST, Todd JA (1996) Human type 1 diabetes and the insulin gene: principles of mapping polygenes. Annu Rev Genet 30:343–370
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
Bevan S, Porteous L, Sitzer M, Markus HS (2005) Phosphodiesterase 4D gene, ischemic stroke, and asymptomatic carotid atherosclerosis. Stroke 36:949–953
Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14:59–69
Bonnen PE, Story MD, Ashorn CL, Buchholz TA, Weil MM, Nelson DL (2000) Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. Am J Hum Genet 67:1437–1451
Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res 12:1846–1853
Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl:228–237
Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331
Brookes AJ (1999) The essence of SNPs. Gene 234:177–186
Brookes AJ, Prince JA (2005) Genetic association analysis: lessons from the study of Alzheimers disease. Mutat Res 573:152–159
Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872
Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140
Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99
Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521
Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120
Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. Nat Genet 32:237–244
Casas JP, Hingorani AD, Bautista LE, Sharma P (2004) Meta-analysis of genetic studies in ischemic stroke: thirty-two genes involving approximately 18,000 cases and 58,000 controls. Arch Neurol 61:1652–1661
Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, N.J.
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, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A 94:1041–1046
Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122
Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev 13:296–302
Clark AG, Boerwinkle E, Hixson J, Sing CF (2005) Determinants of the success of whole-genome association testing. Genome Res 15:1463–1467
Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E (2003) Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet 73:285–300
Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612
Clayton D, McKeigue PM (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360
Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361:865–872
Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422:835–847
Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131
Costas J, Salas A, Phillips C, Carracedo A (2005) Human genome-wide screen of haplotype-like blocks of reduced diversity. Gene 349:219–225
Crawford DC, Akey DT, Nickerson DA (2005) The Patterns of Natural Variation in Human Genes. Annu Rev Genomics Hum Genet

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

Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, Cucca F, Guja C, Ionescu-Tirgoviste C, Stevens H, Carr P, Nutland S, McKinney P, Shield JP, Wang W, Cordell HJ, Walker N, Todd JA, Concannon P (2002) Parameters for reliable results in genetic association studies in common disease. Nat Genet 30:149–150

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

Dawson E, Chen Y, Hunt S, Smink LJ, Hunt A, Rice K, Livingston S, Bumpstead S, Bruskiewich R, Sham P, Ganske R, Adams M, Kawasaki K, Shimizu N, Minoshima S, Roe B, Bentley D, Dunham I (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Res 11:170–178

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nat Genet 37:1217–1223

De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, Su X, Lippert RA, et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res 15:454–462

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, et al. (1998) A physical map of 30,000 human genes. Science 282:744–746

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

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Dichgans M, Markus HS (2005) Genetic Association Studies in Stroke. Methodological Issues and Proposed Standard Criteria. Stroke

Dixon RA, Diehl RE, Opas E, Rands E, Vickers PJ, Evans JF, Gillard JW, Miller DK (1990) Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. Nature 343:282–284

Dong Y, Hassan A, Zhang Z, Huber D, Dalageorgou C, Markus HS (2003) Yield of screening for CADASIL mutations in lacunar stroke and leukoaraiosis. Stroke 34:203–205

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361–364

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci U S A 97:10483–10488

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, et al. (1999) The DNA sequence of human chromosome 22. Nature 402:489–495

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BA (2000) The extent

of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet 67:1544–1554
Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet 25:320–323
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, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927
Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959
Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–1234
Flossmann E, Schulz UG, Rothwell PM (2004) Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. Stroke 35:212–227
Flossmann E, Schulz UG, Rothwell PM (2005) Potential confounding by intermediate phenotypes in studies of the genetics of ischaemic stroke. Cerebrovasc Dis 19:1–10
Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861–866
Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388–393
Freimer N, Sabatti C (2004) The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet 36:1045–1051
Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881–900
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
Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. Science 298:2345–2349
Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111
Goldstein DB (2003) Pharmacogenetics in the laboratory and the clinic. N Engl J Med 348:553–556

Goldstein DB, Ahmadi KR, Weale ME, Wood NW (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. Trends Genet 19:615–622

Gonzalez-Neira A, Calafell F, Navarro A, Lao O, Cann H, Comas D, Bertranpetit J (2004) Geographic stratification of linkage disequilibrium: a worldwide population study in a region of chromosome 22. Hum Genomics 1:399–409

Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, Jonsdottir T, Gudmundsdottir T, et al. (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat Genet 35:131–138

Gulcher JR, Gretarsdottir S, Helgadottir A, Stefansson K (2005) Genes contributing to risk for common forms of stroke. Trends Mol Med 11:217–224

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. J Hum Genet 47:605–610

Hajat C, Tilling K, Stewart JA, Lemic-Stojcevic N, Wolfe CD (2004) Ethnic differences in risk factors for ischemic stroke: a European case-control study. Stroke 35:1562–1567

Hassan A, Markus HS (2000) Genetics and ischaemic stroke. Brain 123 (Pt 9):1784–1812

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

Hattersley AT, McCarthy MI (2005) What makes a good genetic association study? Lancet 366:1315–1323

Helgadottir A, Gretarsdottir S, St Clair D, Manolescu A, Cheung J, Thorleifsson G, Pasdar A, Grant SF, Whalley LJ, Hakonarson H, Thorsteinsdottir U, Kong A, Gulcher J, Stefansson K, MacLeod MJ (2005) Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population. Am J Hum Genet 76:505–509

Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, Samani NJ, et al. (2004) The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat Genet 36:233–239

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

Helgason A, Sigureth ardottir S, Nicholson J, Sykes B, Hill EW, Bradley DG, Bosnes V, Gulcher JR, Ward R, Stefansson K (2000) Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. Am J Hum Genet 67:697–717

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37:90–95

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

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Humphries SE, Morgan L (2004) Genetic risk factors for stroke and carotid atherosclerosis: insights into pathophysiology from candidate gene approaches. Lancet Neurol 3:227–235

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309

Jamrozik K, Broadhurst RJ, Anderson CS, Stewart-Wynne EG (1994) The role of lifestyle factors in the etiology of stroke. A population-based case-control study in Perth, Western Australia. Stroke 25:51–59

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

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538

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, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. Hum Mol Genet 10:2199–2207

Joutel A, Corpechot C, Ducros A, Vahedi K, Chabriat H, Mouton P, Alamowitch S, Domenga V, Cecillion M, Marechal E, Maciazek J, Vayssiere C, Cruaud C, Cabanis EA, Ruchoux MM, Weissenbach J, Bach JF, Bousser MG, Tournier-Lasserve E (1996) Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature 383:707–710

Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5:413–424

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

Ke X, Durrant C, Morris AP, Hunt S, Bentley DR, Deloukas P, Cardon LR (2004a) Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. Hum Mol Genet 13:2557–2565

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004b) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Ke X, Miretti MM, Broxholme J, Hunt S, Beck S, Bentley DR, Deloukas P, Cardon LR (2005) A Comparison of Tagging Methods and Their Tagging Space. Hum Mol Genet

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 66:1882–1899

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

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

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

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A 99:803–808

Kurg A, Tonisson N, Georgiou I, Shumaker J, Tollett J, Metspalu A (2000) Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. Genet Test 4:1–7

Kwok PY, Xiao M (2004) Single-molecule analysis for molecular haplotyping. Hum Mutat 23:442–446

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

Laan M, Wiebe V, Khusnutdinova E, Remm M, Pääbo 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

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

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lawrence R, Evans DM, Morris AP, Ke X, Hunt S, Paolucci M, Ragoussis J, Deloukas P, Bentley D, Cardon LR (2005) Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. Genome Res 15:1503–1510

Lewontin RC (1964) The Interaction Of Selection And Linkage. Ii. Optimum Models. Genetics 50:757–782

Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H (2004) Haplotype block structures show significant variation among populations. Genet Epidemiol 27:385–400

Liu PY, Zhang YY, Lu Y, Long JR, Shen H, Zhao LJ, Xu FH, Xiao P, Xiong DH, Liu YJ, Recker RR, Deng HW (2005) A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. J Med Genet 42:221–227

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182

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

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S, Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci U S A 100:376–381

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584

Mehrabian M, Allayee H, Wong J, Shi W, Wang XP, Shaposhnik Z, Funk CD, Lusis AJ (2002) Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. Circ Res 91:120–126

Meschia JF (2002) Addressing the heterogeneity of the ischemic stroke phenotype in human genetics research. Stroke 33:2770–2774

Meschia JF, Brott TG, Brown RD, Jr., Crook R, Worrall BB, Kissela B, Brown WM, Rich SS, Case LD, Evans EW, Hague S, Singleton A, Hardy J (2005) Phosphodiesterase 4D and 5-lipoxygenase activating protein in ischemic stroke. Ann Neurol 58:351–361

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

Miller RD, Phillips MS, Jo I, Donaldson MA, Studebaker JF, Addleman N, Alfisi SV, et al. (2005) High-density single-nucleotide polymorphism maps of the human genome. Genomics 86:117–126

Miller RD, Taillon-Miller P, Kwok PY (2001) Regions of low single-nucleotide polymorphism incidence in human and orangutan xq: deserts and recent coalescences. Genomics 71:78–88

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, Bentley DR, Rioux JD, Beck S, Deloukas P (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 76:634–646

Mitra N, Ye TZ, Smith A, Chuai S, Kirchhoff T, Peterlongo P, Nafa K, Phillips MS, Offit K, Ellis NA (2004) Localization of cancer susceptibility genes by genome-wide single-nucleotide polymorphism linkage-disequilibrium mapping. Cancer Res 64:8116–8125

Morton NE (1991) Parameters of the human genome. Proc Natl Acad Sci U S A 88:7474–7476

Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, Matthews LH, et al. (2000) An SNP map of human chromosome 22. Nature 407:516–520

Nadeau JH (2003) Modifier genes and protective alleles in humans and mice. Curr Opin Genet Dev 13:290–295

Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353–362

Nejentsev S, Godfrey L, Snook H, Rance H, Nutland S, Walker NM, Lam AC, Guja C, Ionescu-Tirgoviste C, Undlien DE, Ronningen KS, Tuomilehto-Wolf E, Tuomilehto J, Newport MJ, Clayton DG, Todd JA (2004) Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. Hum Mol Genet 13:1633–1639

Netherton SJ, Maurice DH (2005) Vascular endothelial cell cyclic nucleotide phosphodiesterases and regulated cell migration: implications in angiogenesis. Mol Pharmacol 67:263–272

Newton-Cheh C, Hirschhorn JN (2005) Genetic association studies of complex traits: design and analysis issues. Mutat Res 573:54–69

Ng MC, Wang Y, So WY, Cheng S, Visvikis S, Zee RY, Fernandez-Cruz A, Lindpaintner K, Chan JC (2004) Ethnic differences in the linkage disequilibrium and distribution of single-nucleotide polymorphisms in 35 candidate genes for cardiovascular diseases. Genomics 83:559–565

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Nothnagel MRK and Rohde K (2005) The Effect of Single-Nucletide Polymorphism Marker Selection on Patterns of Haplotype Blocks and Haplotype Frequency Estimates. American Jpurnal of Human Genetics 77:988–998

Ober C, Abney M, McPeek MS (2001) The genetic dissection of complex traits in a founder population. Am J Hum Genet 69:1068–1079

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Palmer D, Tsoi K, Maurice DH (1998) Synergistic inhibition of vascular smooth muscle cell migration by phosphodiesterase 3 and phosphodiesterase 4 inhibitors. Circ Res 82:852–861

Palsdottir A, Abrahamson M, Thorsteinsson L, Arnason A, Olafsson I, Grubb A, Jensson O (1988) Mutation in cystatin C gene causes hereditary brain haemorrhage. Lancet 2:603–604

Pardo LM, MacKay I, Oostra B, van Duijn CM, Aulchenko YS (2005) The effect of genetic drift in a young genetically isolated population. Ann Hum Genet 69:288–295

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

Peltonen L, McKusick VA (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291:1224–1229
Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. Nat Rev Genet 1:182–190
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
Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant.or not? Hum Mol Genet 11:2417–2423
Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14
Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181
Ptak SE, Voelpel K, Przeworski M (2004) Insights into recombination from patterns of linkage disequilibrium in humans. Genetics 167:387–397
Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001a) Linkage disequilibrium in the human genome. Nature 411:199–204
Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16
Reich DE, Lander ES (2001b) On the allelic spectrum of human disease. Trends Genet 17:502–510
Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135–142
Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al. (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228
Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
Risch N, Tang H, Katzenstein H, Ekstein J (2003) Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. Am J Hum Genet 72:812–822
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
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
Rubattu S, Di Angelantonio E, Stanzione R, Zanda B, Evangelista A, Pirisi A, De Paolis P, Cota L, Brunetti E, Volpe M (2004) Gene polymorphisms of the renin-angiotensin-aldosterone system and the risk of ischemic stroke: a role of the A1166C/AT1 gene variant. J Hypertens 22:2129–2134

Rubattu S, Gigante B, Stanzione R, De Paolis P, Tarasi D, Volpe M (2004) In the search for stroke genes: a long and winding road. Am J Hypertens 17:197–202
Rubattu S, Stanzione R, Di Angelantonio E, Zanda B, Evangelista A, Tarasi D, Gigante B, Pirisi A, Brunetti E, Volpe M (2004) Atrial natriuretic peptide gene polymorphisms and risk of ischemic stroke in humans. Stroke 35:814–818
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933
Sala A, Bolla M, Zarini S, Muller-Peddinghaus R, Folco G (1996) Release of leukotriene A4 versus leukotriene B4 from human polymorphonuclear leukocytes. J Biol Chem 271:17944–17948
Salanti G, Sanderson S, Higgins JP (2005) Obstacles and opportunities in meta-analysis of genetic association studies. Genet Med 7:13–20
Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC (2003) SNP and haplotype variation in the human genome. Mutat Res 526:53–61
Samuelsson B (1983) Leukotrienes: mediators of immediate hypersensitivity reactions and inflammation. Science 220:568–575
Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (2005) The span of linkage disequilibrium caused by selection on G6PD in humans. Genetics
Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK (2005) Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13:677–686
Schreiber S, Rosenstiel P, Albrecht M, Hampe J, Krawczak M (2005) Genetics of Crohn disease, an archetypal inflammatory barrier disease. Nat Rev Genet 6:376–388
Schulze TG, Zhang K, Chen YS, Akula N, Sun F, McMahon FJ (2004) Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. Hum Mol Genet 13:335–342
Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S (2003) Robustness of inference of haplotype block structure. J Comput Biol 10:13–19
Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Miller W (2003) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res 31:3518–3524
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
Serre D, Pääbo S (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Res 14:1679–1685
Service SK, Ophoff RA, Freimer NB (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate. Hum Mol Genet 10:545–551

Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Genet 66:1616–1630

Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet 12:771–776

Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, Jones KW (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2:81–89

Smith AV, Thomas DJ, Munro HM, Abecasis GR (2005) Sequence features in regions of weak and strong linkage disequilibrium. Genome Res 15:1519–1534

Smith DJ, Lusis AJ (2002) The allelic structure of common disease. Hum Mol Genet 11:2455–2461

Spanbroek R, Grabner R, Lotzer K, Hildner M, Urbach A, Ruhling K, Moos MP, Kaiser B, Cohnert TU, Wahlers T, Zieske A, Plenz G, Robenek H, Salbach P, Kuhn H, Radmark O, Samuelsson B, Habenicht AJ (2003) Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. Proc Natl Acad Sci U S A 100:1238–1243

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, et al. (2005) A common inversion under selection in Europeans. Nat Genet 37:129–137

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, et al. (2002) Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet 71:877–892

Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S (2004) Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. Hum Genet 114:377–385

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

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

Stram DO (2004) Tag SNP selection for association studies. Genet Epidemiol 27:365–374

Stumpf MP (2004) Haplotype diversity and SNP frequency dependence in the description of genetic variation. Eur J Hum Genet 12:469–477

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

Syvanen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37 Suppl:S5–10

Zee RY, Cook NR, Cheng S, Reynolds R, Erlich HA, Lindpaintner K, Ridker PM (2004) Polymorphism in the P-selectin and interleukin-4 genes as determinants of stroke: a population-based, prospective genetic analysis. Hum Mol Genet 13:389–396

Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. Hum Genet 113:51–59

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci U S A 99:7335–7339

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300–1301

Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908–916

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906–914

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

Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T, Gu CC, Tang H, Rao DC, Risch N, Weder A (2005) Admixture mapping for hypertension loci with genome-scan markers. Nat Genet 37:177–181

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89–100

Tabor HK, Risch NJ, Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391–397

Taillon-Miller P, Piernot EE, Kwok PY (1999) Efficient approach to unique single-nucleotide polymorphism discovery. Genome Res 9:499–505

Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am J Hum Genet 66:69–83

Terwilliger JD, Weiss KM (2003) Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. Ann Med 35:532–544

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

Thompson D, Stram D, Goldgar D, Witte JS (2003) Haplotype tagging single nucleotide polymorphisms and association studies. Hum Hered 56:48–55

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, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–522

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

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
Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3:611–621
Tsunoda T, Lathrop GM, Sekine A, Yamada R, Takahashi A, Ohnishi Y, Tanaka T, Nakamura Y (2004) Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. Hum Mol Genet 13:1623–1632
Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749
Wall JD (2001) Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr Opin Genet Dev 11:647–651
Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597
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
Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118
Wang WY, Todd JA (2003) The usefulness of different density SNP maps for disease association studies of common variants. Hum Mol Genet 12:3145–3149
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
Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551–565
Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128
Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19–24
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
Vermeire S, Wild G, Kocher K, Cousineau J, Dufresne L, Bitton A, Langelier D, Pare P, Lapointe G, Cohen A, Daly MJ, Rioux JD (2002) CARD15 genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. Am J Hum Genet 71:74–83
Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, Goldstein DB (2001) Genetic evidence for different male and female roles during cultural transitions in the British Isles. Proc Natl Acad Sci U S A 98:5078–5083
Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, Altshuler D (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111
Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. Nat Genet 23:397–404

Yalcin B, Fullerton J, Miller S, Keays DA, Brady S, Bhomra A, Jefferson A, Volpi E, Copley RR, Flint J, Mott R (2004) Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc Natl Acad Sci U S A 101:9734–9739

Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL (2001) Comparison of human genetic and sequence-based physical maps. Nature 409:951–953

SUMMARY IN ESTONIAN

Alleelse aheldatuse (LD) struktuuri uurimine Euroopa populatsioonides ja selle rakendused geneetilistes assotsiatsiooniuuringutes

Viimaste aastate jooksul tehtud teadusuuringud inimese genoomis esinevate varieeruvuste kirjeldamisel on lubanud püstitada mitmeid olulisi hüpoteese ja luua teooriaid, mis annavad võimaluse avastada uusi komplekshaigusi põhjustavaid või soodustavaid geneetilisi riskifaktoreid. Igal geenil on organismi elutegevuses oma kindel funktsioon ja seega pole olemas nn haigusgeene. Kui geeni järjestusse või selle ümbrusesse tekib muutus (mutatsioon), siis võib see viia genoomis valitseva tasakaalu rikkumisele ja geeni funktsiooni häirumisele, mille tulemusena võib tekkida haigus. Komplekshaiguste nagu infarkt, insult, teist tüüpi diabeet jpt., kujunemisel on olulised nii erinevad geneetilised - kui ka keskkonnafaktorid ning nende omavahelised kombinatsioonid. Veelgi enam, iga üksiku geeni effekt üldisesse haigestumise riski võib olla suhteliselt madal (suhteline risk 1.2–1.5) ning seetõttu pole õnnestunud mitmeid seni saadud uuringutulemusi korrata. Komplekshaiguste pärandumine ei allu klassikalisele Mendeliaalsel teel päranduvate haiguste mustritele, mistõttu riski andvate geneetiliste faktorite leidmiseks on vaja rakendada teistsuguseid lähenemisviise kui perekonna põhine aheldusuuring. Arvatakse, et ilmselt parimaks ja rohkem edu toovaks lähenemiseks on assotsiatsiooniuuringud, mille käigus võrreldakse omavahel samast populatsioonist pärit suurt hulka haigeid ja terveid inimesi.

Haigustega seotud geenide kaardistamisel on aegade jooksul kasutatud erinevaid geneetilisi markerid. Sagedasemad varieeruvused inimese genoomis on ühenukleotiidsed polümorfismid (SNPd). Lähedalasuvate markerite vahel võib esineda alleelne aheldatus ehk *linkage disequilibrium* (LD), mis vaheldub piirkondadega, kus selline assotsiatsioon on lõhutud. LD kujunemisel on olulised nii erinevad molekulaarsed (rekombinatsioon, mutatsiooni kiirus jne.) kui ka populatsiooni demograafilised ja evolutsioonilised faktorid (geenitriiv, migratsioon, selektsioon jne.). Mitmete faktorite koosmõju tulemusena on LD genoomi erinevates piirkondades väga varieeruv ja ettearvamatu ning võib suures ulatuses erineda erinevate populatsioonide vahel. Kõrge LD-ga genoomipiirkondadest on võimalik välja valida optimaalne arv geneetilisi markereid, mille analüüsimine annab infot ka lähedalolevate analüüsimata markerite kohta. Seega on LD täpne kirjeldamine olulise praktilise tähtsusega, võmaldades teostada üle kogu genoomi hõlmavaid suuremahulisi assotsiatsiooniuuringuid minimaalse arvu markeritega.

Käesolevas doktoritöö suurem osa eksperimentaalsest tööst kirjeldab LD ja haplotüübi mustreid erinevates genoomi piirkondades ning varieeruvust Euroopa populatsioonides seoses markerite valimise ja sobimisega hilisemateks assotsiatsiooniuuringutes. Kasutades mudelsüsteemina inimese kromosoom 22.

uuriti üldist LD esinemist ja ulatust ning erinevaid faktoreid, mis on seotud LD mustrite kujunemisega. Tulemustest selgus, et LD varieerub kogu kromosoomi ulatuses ning on seotud rekombinatsiooni mustritega. Tihedam markerite analüüs erinevates kompleskhaigustega seotud kandidaatgeenide piirkondades võimaldas täpsemalt kirjeldada esinenud LD ja haplotüüpide mustreid. Näiteks võimaldas *FKBP5* geeni ümbruse täpne LD analüüs kinnitada eelnevaid positiivseid assotsiatsioonianalüüsi tulemusi ning toetas antud kandidaatgeeni rolli depressiooni kujunemisel.

Üheks töös püstitatud eesmärgiks oli vaadata, kui palju esineb erinevates Euroopa populatsioonides LD ja haplotüüpide vahelist varieeruvust, mis lubab hinnata assotsiatsiooniuuringuteks kasutavate markerite sobivust populatsioonide vahel. Uuringutulemustest järeldus, et kõikides vaadeldud Euroopa populatsioonides esinevad samad sagedasemad haplotüübid ning enamus piirkondades olid populatsioonide vahelised erinevused markerite alleelisagedustes minimaalsed. Antud tulemus viitab Euroopa populatsioonide sarnasele demograafilisele ajaloole ning näitab, et assotsiatsiooniuuringuteks võib suure tõenäosusega Euroopas valida samad geneetilised markerid. Samas aga, ühes uuritud geenoomipiirkonnas täheldati ka populatsioonide vahelisi märkimisväärseid erinevusi alleeli- ja haplotüübisagedustes. Selles piirkonnas eristus selge haplotüübi sageduste gradient Põhja- ja Lõuna-Euroopa populatsioonide vahel. Genoomi regioonides, kus esineb populatsioonide vahel märkimisväärseid erinevusi tuleb markerite valikul edasistes assotsiatsiooniuuringuteks pöörata erilist tähelepanu, kuna erinevad populatsioonid võivad vajada erinevaid markereid, et tuvastada haigust põhjustavaid riskifaktoreid.

Käesoleva töö viimases osas vaadati kahe kandidaategeeni, *ALOX5AP* ja *PDE4D* seost insuldi tekkel. Teostati assotsiatsioonianalüüs Saksa populatsioonist pärit haigete ja kontrollindiviidega, kus vaatluse all olid varem Islandi populatsioonis leitud erinevate markerite ja haplotüüpide vahelised assotsiatsioonid. Antud töös leiti positiivne seos insuldi ja *ALOX5AP* geeni erinevate polümorfismide vahel, millest võib järeldada, et teatud *ALOX5AP* geeni variandid võivad soodustada insuldi tekkimist ka mandri-Euroopa populatsioonis. Leitud assotsiatsioon oli tugevam meestel. Positiivset seost aga insuldi ja ühegi *PDE4D* geeni variandi ega haplotüübi vahel polnud võimalik tuvastada. Tööst leitud populatsioonide vahelised erinvused mõlemas uuritud kandidaatgeenis annavad alust arvata, et erinevates populatsioonides võivad komplekshaiguste kujunemisel osaleda erinevad geneetilised riskifaktorid.

ACKNOWLEDGEMENTS

First of all I would like to give my acknowledge the co-autors of the experimental work presented in this dissertation. Thanks for all people who have helped and teached me during my studies.

I am grateful to my supervisor Prof. Andres Metspalu especially for giving me a possibility to work with so facinating topic and also for being supportive and encouraging throughout my studies.

I have a great pleasure to thank prof. Thomas Meitinger in Human Genetic Insitute in GSF, Munich. I am grateful for the opportunity to work his institute, for the great and educating discussions and for knowledges I got during my stay in Munich.

I am thankful to Jakob Mueller for very nice collaboration and for his patience and guidance in teaching and explaining me statistics and analysis we needed for our work.

I would like to thank Prof. Maido Remm and Reedik Mägi from Department of Bioinformatic for pleasant collaboration.

Thanks to Dr. Martin Dichgans and his colleques from Klinikum Grossharden in Munich for fruitful cooperation.

Tarmo Annilo and Prof. Maris Laan are also warmly thanked for reading and commenting the manuscript of this thesis. My sincere thanks goes to Maris for her understaning and very supportive attitude and finally, for giving me an opportunity to continue the work with the genetics of complex diseases.

I am thankful to Jack Favor for language correction and Krista Liiv for administrative assistance.

I wish to thank all my former and present collegues from Department of Biotechnology, Asper Biotech and Institute of Human Genetic, for help and especially for creating friendly athmosphere.

Many thanks to all my friends for making my life more colourful.

I owe my thanks to my parents. Their love and always optimistic and supportive attitude has been very important throughout the years of my studies. My dear Mom, I thank you for sending me many wounderful letters when I was alone abroad, their influence to me was invaluable.

Last but not least, my deepest gratitude belongs to Tõnis for many and varied kind of help. Special thanks for many fruitful discussions about my projects in any time I needed it, and also for critically reading all my manuscripts including this dissertation. Your help will never be forgotten.

PUBLICATIONS

CURRICULUM VITAE

Elin Lõhmussaar

Date and place of birth:	March 5, 1974 in Tartu, Estonia
Addrss:	Department of Biotechnology, Institute of Molecular and Cell Biology of Tartu University, 23 Riia Street, 51010 Tartu, Estonia
Phone:	+372 737 5006
E-mail:	elin@ebc.ee

Education and professional employment

1983–1992	Tartu Secondary School No. 5, silver medal
1992–1996	Heino Eller Music Scool
1992–1997	B.Sc University of Tartu, Institute of Molecular and Cell Biology
1997–2000	M.Sc University of Tartu, Institute of Molecular and Cell Biology, *cum laude*
From 2000	Ph.D. student in Department of Biotechnology
2001–2003	Asper Biotech, scientist
2003–2004	as a guest student in the Institute of Human Genetic, GSF, Munich
From 2004	University of Tartu, research scientist

Scientific work

I have been involved in project associated with development of microarray based genotyping platform for genetic analysis and molecular diagnostics. During the last years my research interest has been the association studies of complex diseases.

LIST OF PUBLICATIONS

1. Lamina C, Steffens M, Mueller J, **Lohmussaar E**, Meitinger T, Wichmann HE. Genetic diversity in german and European populations: looking for substructures and genetic patterns. Gesundheitswesen. 2005 Aug; 67 Suppl 1:S127–31.
2. **Lõhmussaar E**, Gschwendtner A, Mueller JC, Org T, Wichmann E, Hamann G, Meitinger T and Dichgans M. The *ALOX5AP* gene and the *PDE4D* gene in a Central-European population of stroke patients. Stroke 2005 Apr;36(4):731–6.
3. Mueller JC, **Lõhmussaar E,** Mägi 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 and Meitinger T. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 2005 Jan 6; 76 (3).
4. Binder EB, Salyakina D, Lichtner P, Wochnik GM, Ising M, Putz B, Papiol S, Seaman S, Lucae S, Kohli MA, Nickel T, Kunzel HE, Fuchs B, Majer M, Pfennig A, Kern N, Brunner J, Modell S, Baghai T, Deiml T, Zill P, Bondy B, Rupprecht R, Messer T, Kohnlein O, Dabitz H, Bruckl T, Muller N, Pfister H, Lieb R, Mueller JC, **Lohmussaar E,** Strom TM, Bettecken T, Meitinger T, Uhr M, Rein T, Holsboer F, Muller-Myhsok B. Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment. Nat Genet. 2004 Dec; 36(12): 1319–25. Epub 2004 Nov 21.
5. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, **Lohmussaar E**, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. A first-generation linkage disequilibrium map of human chromosome 22. *Nature.* 2002 Aug 1; 418(6897): 544–8.
6. Tõnisson, N., Kurg, A., **Lõhmussaar, E**., Metspalu, A. Arrayed primer extension on the DNA chip – method and applications. In "DNA Microarrayes: Biology and Technology" Ed by M. Schena (2000) BioTechniques Books ISBN 1-881299-37-6, 247–263.
7. Tõnisson N, Kurg A, Kaasik K, **Lõhmussaar E**, Metspalu A. Unravelling Genetic Data by Arrayed Primer Extension. *Clinical Chemistry and Laboratory Medicine* 2000, 38 (2), 165–177. By Walter de Gruyter, Berlin, New York.

ELULOOKIRJELDUS

Elin Lõhmussaar

Sünniaeg ja koht:	5. märts 1974, Tartu, Eesti
Aadress:	Molekulaar ja Rakubioloogia Instituut, Tartu Ülikool, Riia mnt. 23, 51010, Tartu, Eesti
Telefon:	+372 737 5006
E-mail:	elin@ebc.ee

Haridus ja erialane teenistuskäik

1983–1992	Tartu 5. keskkool, hõbemedal
1992–1996	Heino Elleri nim Tartu Muusikakool
1992–1997	Bakalaureus TÜ bioloogia-geograafia teaduskonna biotehnoloogia ja biomedistiini erialal
1997–2000	magister TÜ bioloogia-geograafia teaduskonna biotehnoloogia erialal, *cum laude*
2000–	TÜ MRI doktorant
2002–2003	Asper Biotech, teadur
2003–2004	GSF instituut Münhenis, Inimesegenetika labor; teadustöö külalistudengina
2004–	TÜ MRI teadur

Teadustegevus

Oma uurimistöö raames olen biotehnoloogia õppetoolis tegelenud projektiga, mille eesmärgiks oli uue genotüpiseerimismeetodi välja töötamine ja rakendamine meditsiinigeneetikas. Viimastel aastatel olen tegelenud uurimistööga, mis käsitleb LD struktuuri varieeruvuse analüüsi Euroopa populatsioonides ja selle rakendamist geneetilistes assotsiatsiooniuuringutes, et tuvastada uusi kompleks-haigusi soodustavaid geneetilisi riskifaktoreid.

PUBLIKATSIOONID

1. Lamina C, Steffens M, Mueller J, **Lohmussaar E**, Meitinger T, Wichmann HE. Genetic diversity in german and European populations: looking for substructures and genetic patterns. Gesundheitswesen. 2005 Aug; 67 Suppl 1: S127–31.
2. **Lõhmussaar E**, Gschwendtner A, Mueller JC, Org T, Wichmann E, Hamann G, Meitinger T and Dichgans M. The *ALOX5AP* gene and the *PDE4D* gene in a Central-European population of stroke patients. Stroke 2005 Apr;36(4):731–6.
3. Mueller JC, **Lõhmussaar E,** Mägi 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 and Meitinger T. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 2005 Jan 6; 76 (3).
4. Binder EB, Salyakina D, Lichtner P, Wochnik GM, Ising M, Putz B, Papiol S, Seaman S, Lucae S, Kohli MA, Nickel T, Kunzel HE, Fuchs B, Majer M, Pfennig A, Kern N, Brunner J, Modell S, Baghai T, Deiml T, Zill P, Bondy B, Rupprecht R, Messer T, Kohnlein O, Dabitz H, Bruckl T, Muller N, Pfister H, Lieb R, Mueller JC, **Lohmussaar E,** Strom TM, Bettecken T, Meitinger T, Uhr M, Rein T, Holsboer F, Muller-Myhsok B. Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment. Nat Genet. 2004 Dec; 36(12):1319–25. Epub 2004 Nov 21.
5. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, **Lohmussaar E**, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. A first-generation linkage disequilibrium map of human chromosome 22. *Nature.* 2002 Aug 1; 418(6897): 544–8.
6. Tõnisson, N., Kurg, A., **Lõhmussaar, E**., Metspalu, A. Arrayed primer extension on the DNA chip – method and applications. In "DNA Microarayes: Biology and Technology" Ed by M. Schena (2000) BioTechniques Books ISBN 1-881299-37-6, 247–263.
7. Tõnisson N, Kurg A, Kaasik K, **Lõhmussaar E**, Metspalu A. Unravelling Genetic Data by Arrayed Primer Extension. *Clinical Chemistry and Laboratory Medicine* 2000, 38 (2), 165–177. By Walter de Gruyter, Berlin, New York.