THE LINKAGE DISEQULIBRIUM AND THE SELECTION OF GENETIC MARKERS FOR ASSOCIATION STUDIES IN EUROPEAN POPULATIONS

REEDIK MÄGI

Table of contents

List of original publications

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.; Lohmmussaar, E.; Zernant, J.; Tonisson, N.; Remm, M.; **Mägi, R.**; Puurand, T.; Vilo, J.; Kurg, A.; Rice, K.; Deloukas, P.; Mott, R.; Metspalu, A.; Bentley, D.R.; Cardon, L.R.; Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature, 418(6897), 544 - 548.

Mueller, J. C.; Lõhmussaar, E.; **Mägi, R.**; Remm, M.; Bettecken, T.; Lichtner, P.; Huber, S.; Illig, T; Luedemann, J.; Schreiber, S.; Wichmann, H. E.; Pramstaller, P.; Romeo, G.; Testa, A.; Metspalu, A.; Meitinger, T. (2005). Linkage Disequilibrium Patterns and tagSNP Transferability among European Populations. American Journal of Human Genetics, 76(3), 387 - 398.

Montpetit, A.; Nelis, M.; Laflamme, P.; **Mägi, R.**; Ke, X.Y.; Remm, M.; Cardon, L.; Hudson, T.J.; Metspalu, A. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population . PLoS Genetics, 2(3), 282 - 290.

Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M. (2007). Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. BMC Genomics, 8(1):159.

Author's contributions:

Ref I: performed marker validation and LD calculations of Estonian samples

Ref II: participated in creating new methodology, analysis of data and preparation of the manuscript

Ref III: participated in analysis of data and preparation of the manuscript

Ref IV: participated in planning, experimental phase, analysis and in the writing of the manuscript

List of abbreviations

APEX	arrayed primer extension
bp	base pair
CD40	CD40 molecule, TNF receptor superfamily member 5
CEPH	Centre d'Etude du Polymorphisme Humain
CHB	Han Chinese in Beijing, China
CNV	copy number variation
cSNP	coding SNP
DNA	deoxyribonucleic acid
ENCODE	Encyclopedia of DNA elements
FKBP5	FK506 binding protein 5
G6PD	glucose-6-phosphate dehydrogenase
GWAS	genome-wide association study
HLA	human leukocyte antigen
htSNP	haplotype-tagging SNP
HWE	Hardy-Weinberg equilibrium
JPT	Japanese in Tokyo, Japan
Kb	kilobase, 1,000 base pairs
KORA	Cooperative Health research in the Region of Augsburg
LD	Linkage disequilibrium
LMNA	lamin A/C
MAF	minor allele frequency
MALDI-TOF	matrix-assisted laser desoption/ionization time-of-flight
Mb	megabase, 1,000,000 base pairs
PCR	polymerase chain reaction

PLAU	plasminogen activator, urokinase
POPGEN	population sample collected in Schleswig-Holstein, Germany
RFLP	restriction fragment length polymorphism
SHIP	Study of Health in Pomerania
SNCA	synuclein, alpha
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
STR	short tandem repeat
tagSNP	tagging SNP
VNTR	variable number of tandem repeats
YRI	Yoruba population from Ibadan, Nigeria

Introduction

According to a recent study by Levy et al, the interchromosomal difference between the genome sequences of two human beings is five times larger than previously estimated – only 99.5% similarity exists between the two chromosomal copies (Levy et al. 2007). Only a small proportion of those genomic variations are located in regions that can be identified to encode a protein. Until recently, most of the association studies have focused on genes and the importance of the non-coding genomic landscape has not been emphasized. The results of the ENCODE (Encyclopedia of DNA elements) project have revealed that a large proportion of non-coding DNA is involved in gene regulation and that there are non-protein-coding transcripts practically everywhere in the human genome (Birney et al. 2007). These findings indicate the importance of genome-wide association studies (GWAS) as a tool for detecting genetic causes of complex diseases.

Association analyses are based on the linkage disequilibrium between loci. In case of a tight linkage between the marker and the disease causing locus, it is possible to localize the disease gene by genotyping neighboring markers. The difference in allele or genotype frequencies of a marker can be observed between the case and control samples if the marker locus itself is causing the disease or is in linkage with the susceptibility locus (Risch and Merikangas 1996).

As the disease causing locus can be investigated through neighboring markers, it is possible to describe most of the common variants with a reduced number of tagging SNPs. One of the crucial prerequisites for a successful association study is estimating the power to detect association (Purcell et al. 2003). Additionally to the sample size, the frequencies of causal and marker alleles and the disease penetrance, the power is determined by the correlation between causal allele and observed marker. Therefore, it is important to estimate the coverage of selected marker sets to evaluate how successfully they can capture the genetic variation of an observed genomic region.

Due to the large variance of allelic association in different regions of the human genome and also between populations, it is necessary to create an empirical LD map of the population under observation. The detailed structure of the LD patterns gives us an opportunity to create optimal sets of tagging SNPs for association studies. In addition, it is important to estimate how similar the LD patterns are between populations. In the case of close populations, the tagging SNP sets could be transferrable, creating an opportunity for fixed SNP panels.

The first part of this thesis provides an overview of genetic markers, the concept of linkage disequilibrium and its variability in the human genome. This is followed by an introduction to association studies and a discussion of marker selection strategies for effectively finding the disease causing genes and alleles. The research part of this dissertation is focused on: 1) Describing the variability of LD and comparing the patterns of LD among European populations 2) Evaluating the performance of the tagSNPs and their transferability among European populations 3) Evaluating the coverage of whole-genome marker sets for genome-wide association studies.

1. Review of literature

1.1. Linkage disequilibrium in human genome

1.1.1. Genetic markers

Rapid development of technology and progress in molecular biology has boosted the number and size of association studies in the last decade. It is now possible to make an association study with thousands of cases and controls genotyped for hundreds of thousands of markers across the human genome.

The first genetic markers which were sufficiently numerous and adequately spaced across the human genome were restriction fragment length polymorphisms (RFLPs). They were introduced as a genetic mapping tool in 1975 (Grodzicker et al. 1975) and in 1980 Botstein et al. describes the construction of the human genetic linkage map using restriction fragment length polymorphisms (Botstein et al. 1980). RFLPs are generally based on single-nucleotide changes and therefore have heterozygosities below 50% (Ott 1999).

Many more polymorphic markers were described by Nakamura et al. in 1987 (Nakamura et al. 1987). These are called variable number of tandem repeats (VNTRs). These polymorphic sequences contain 20-50 copies of 6-100 bp repeats. There are approximately 150,000 minisatellites across the human genome, of which approximately 20% are polymorphic i.e. VNTRs (Feuk et al. 2006). As minisatellites were not believed to be as easy to clone and characterize as microsatellites and also are concentrated mostly in telomeric regions of chromosomes, they were replaced by microsatellite markers (Weber 1990). Still there is some renewed interest in VNTRs as they might have important functional roles (Nakamura et al. 1998) and according to more recent studies they are abundant and widespread across the human genome (Näslund et al. 2005).

Microsatellites , also referred to as short tandem repeats (STRs) or simple sequence repeats (SSRs) are repeats of 1-6 bp units totaling < 200 bp in length (Toth et al. 2000). With the advent of PCR in the late 1980s, the genotyping of STRs became straightforward and they became the markers of choice for genome mapping (Ellegren 2004). They are mostly $(CA)_n$ repeats, but also any other short motifs, accounting for approximately 3% of the human genome sequence (Lander et al. 2001).

Single nucleotide polymorphisms (SNPs) are the most abundant form of genomic variation. According to their definition, SNPs are single base pair positions in genomic DNA at which different sequence alternatives exist wherein the least frequent allele have an abundance of 1% or greater (Brookes 1999). More than 10 million SNPs have been identified in the human genome (Hinds et al. 2005). SNPs are usually bi-allelic, but might theoretically have up to four alleles. In the current dbSNP build 127, there are thousands of validated SNPs with more than two alleles, but they still contribute approximately 0.2% of all SNPs. Therefore, a single SNP is a lot less informative than a mini- or microsatellite, but their large number and possibilities of automation make them a very attractive tool for mapping the human genome (Wang et al. 1998). Current genotyping platforms enable the

genotyping of hundreds of thousands of SNPs quickly and affordably (Barrett and Cardon 2006; Pe'er et al. 2006). New SNP panels should also be able to detect other types of variability of genomes – copy number variations (CNVs), inversions, insertions, deletions and other complex rearrangements (Feuk et al. 2006).

1.1.2. Linkage disequilibrium

The concept of linkage disequilibrium (LD), also known as allelic association, dates back to 1917, when Jennings published his work on the linkage between factor pairs (Jennings 1917). In his study, he compared the frequencies of gametes derived from independent and linked pairs of phenotypic traits. LD occurs when two alleles at adjacent loci tend to co-occur more frequently than expected by their allele frequencies.

The LD between a mutant allele in a disease locus and marker alleles at flanking loci is complete when the mutation occurs. If the evolutionary factors are ignored, the LD will decay due to recombination events. The decay in LD is related to the recombination fraction between loci. Recombinations between tightly linked loci are rare and therefore LD remains strong for many generations. The LD between loosely linked loci will decay quickly with generations until the frequencies of four possible haplotypes between bi-allelic loci are determined by their allele frequencies. Recurrent mutation (in either mutant or marker locus) might also decrease the association between these loci, but it would be a very rare case and there is no evidence that it contributes significantly to the erosion of LD between SNPs (Ardlie et al. 2002). Recombination events and mutations are not the only factors that affect the LD – demographic, molecular and evolutionary forces also play major role in shaping the LD structure (De La Vega et al. 2005; Jorde 2000; Pritchard and Przeworski 2001).

Long range LD can be caused by an extreme founder effect or a bottleneck: a period, when the population size is so small that a few ancestors gave rise to most of the haplotypes that exist today (Reich et al. 2001). Also, the genetic drift is particularly important in shaping LD, if the population has gone through bottleneck of founder effect (Risch et al. 2003). It is also shown, that in principle, long-range LD might be generated by a recent population mixture or by migration (Chakraborty and Weiss 1988; Pfaff et al. 2001; Pritchard and Przeworski 2001; Zhu et al. 2005).

Various aspects of population structure may affect the pattern of LD. For example, population subdivision might be an important factor in establishing the patterns of LD. In the study of *Arabidopsis* it was shown that inbreeding can cause high levels of LD (Nordborg et al. 2002).

Positive selection causes a rapid rise of allele frequency, occurring so quickly that recombinations can not break the haplotype on which the selected mutation occurs. Therefore, positive selection causes an allele having unusually long-range LD (Przeworski 2002; Sabeti et al. 2002). For example, positive selection has created high LD around the human dopamine receptor D4 gene locus which has been found to be associated with hyperactivity disorder and also the personal trait of novelty seeking (Ding et al. 2001). Another well-known example is the positive selection around genes G6PD and CD40, which is related to resistance to malaria (Sabeti et al. 2002; Saunders et al. 2005).

Gene conversion has been shown to be an important mechanism in the breakdown of LD over short distances (Przeworski and Wall 2001; Wiuf and Hein 2000) and also contributing to the decay of LD in human population level (Frisse et al. 2001). In the case of gene conversion, a short stretch of one copy of chromosome is transferred to another during meiosis - the effect is the same as that of two very close recombinations (Ardlie et al. 2002; Chen et al. 2007).

It is well-known that the processes of generating LD are highly stochastic (Pritchard and Przeworski 2001), therefore the variation in LD levels across different genomic regions is expected to occur by chance (Frisse et al. 2001).

1.1.3. Measures of LD

The most common LD measure, *D*, quantifies the disequilibrium as the difference between the observed frequency of the two-locus haplotype and the frequency, it would be expected to have if those two loci would segregate randomly. The formula was developed by Lewontin more than 40 years ago (Lewontin 1964):

$$D = P_{AB} - P_A \times P_B$$

Where P_{AB} is the observed frequency of two locus haplotype with allele A in the first locus and B in the second locus, P_A is the frequency of allele A in first locus and P_B is the frequency of allele B in second locus (Table 1). Although the measure *D* is good for explaining LD, its numerical value is not a very good tool for measuring the strength of and comparing levels of LD due to its dependence on allele frequencies. To make the LD measure less dependent on allele frequencies, another measure *D'* was proposed by Lewontin in the same paper (Lewontin 1964):

$$D' = \begin{cases} \dfrac{D}{\min(P_A \times P_b, P_a \times P_B)} & D > 0 \\ \dfrac{D}{\min(P_A \times P_B, P_a \times P_b)} & D < 0 \end{cases}$$

Where the quantity in the denominator is the absolute maximum *D* that could be achieved with the given allele frequencies at the two loci. *D'* = 1 between two loci only if they haven't been separated by recombination. In this case, at most three out of four possible two-locus haplotypes are observed in a sample (Figure 1). Values of *D'* < 1 show that complete LD has been disrupted. The problem with the *D'* measure is that it is strongly inflated in small samples, even with common alleles, but especially for SNPs with rare alleles. Because the magnitude of *D'* is dependent on sample size, it should not be used for comparisons of the strength of LD between studies (Ardlie et al. 2002).

One of the most popular LD measure is r^2 *(also denoted by* Δ^2*),* which is currently the measure of choice for quantifying and comparing the LD (Hill and Robertson 1968). This common scaling of *D* is calculated:

$$r = \frac{D}{\sqrt{(P_A \times P_a \times P_B \times P_b)}}$$

It is equal to $\sqrt{\chi^2/N}$, where the χ^2 statistic can be obtained from a 2x2 table of haplotype frequencies and *N* is the total number of haplotypes in the sample. This provides a means of testing the statistical significance of *r*. In the paper by Pritchard and Przeworski, they show that r^2 arises naturally in the context of association mapping (Pritchard and Przeworski 2001). It can be shown that in order to achieve the same power of association as you would have genotyping disease causing allele, you need to increase the sample size by $1/r^2$, where r^2 is the coefficient of LD between marker and disease mutation (Kruglyak 1999).

Another common two-locus disequilibrium statistic δ can be calculated as:

$$\delta = \frac{D}{P_{B\times}P_{ab}}$$

Where P_B is the population frequency in the disease allele. Several studies have shown that δ and D' give more reliable estimates of physical distance between the disease locus and the marker than D and r^2 (Devlin and Risch 1995; Guo 1997). Devlin and Risch have showed that δ is directly proportional to the recombination fraction and thereby a desirable measure for genetic distance (Devlin and Risch 1995), but the measure becomes unpredictable, when the marker is very close to the disease locus (Guo 1997). This might be one reason why popular marker selection methods are based on D' and r^2 statistics and δ has not been used for this purpose.

1.1.4. The extent of LD in human genome

As the recombination rates vary across the human genome, the length and strength of LD between loci is not solely defined by the physical distances between loci. Many studies have shown that the LD structure is highly variable in different regions of the human genome – close markers are not always in strong LD (Ardlie et al. 2001; Clark et al. 1998; Moffatt et al. 2000) and sometimes LD has been reported between quite distant markers (Abecasis et al. 2001; Daly et al. 2001; Gabriel et al. 2002; Patil et al. 2001; Reich et al. 2001). At the extreme, variations in several orders of magnitude are observed between meiotic recombination frequencies per unit of physical distance (Lichten and Goldman 1995). It has been proposed that the human genome might consist of non-recombining blocks separated by regions of high recombination, so called recombination "hot-spots" (Ardlie et al. 2001; Jeffreys et al. 2001; Lichten and Goldman 1995). This theory was also supported by the study of Jeffreys et al. where they estimated the recombination frequencies in sperms and found three clusters of recombinational hotspots accounting for 94% of the observed recombinations of HLA class II region (Jeffreys et al. 2001). A possible cause for this has been shown in that the in human genome recombination rates tend to be higher in regions with higher gene density (Fullerton et al. 2001). There is also some evidence to indicate that mammalian recombination hot-spots might be associated with GC rich repetitive DNA sequences (Petes 2001). It has even been shown that the recombination hotspot reduces the effect of strong positive selection - the rapid decay in LD upstream of the HbC allele demonstrates the large effect the ß-globin hotspot has in decreasing the effects of positive selection on linked variation (Wood et al. 2005).

Another important fact is that the LD varies remarkably across populations - the pattern of LD might give information about population history and migrations (Plagnol and Wall 2006; Reich et al. 2001).

The highest levels of LD are usually observed in isolated populations due to genetic drift (Peltonen et al. 2000). These populations can successfully be used for studying rare diseases – they have greater phenotypic homogeneity and similar environmental conditions, reduced genetic diversity and higher prevalence of some diseases. Due to high LD, sparsely spaced markers can describe most of the genetic variations. Valuable isolates are for example Finns, Amish, Sardinians and Bedouins, which show high frequencies of certain Mendelian diseases (Peltonen et al. 2000). The lowest values of LD are described in African populations (Reich et al. 2001). They are possibly caused by the differences in demographic history as the biological determinants of LD are expected to be constant across populations (Gabriel et al. 2002). If a population has been stable, and of a large enough size sufficient amount of time, there can be enough recombinations to reduce LD between distant markers to unmeasurable levels (Goldstein 2001). In European and Asian populations, clear LD block structure is reported, hinting that large blocks of LD can be described by a few markers (Gabriel et al. 2002; Goldstein 2001).

Understanding the nature of LD is essential for planning the association studies. LD enables scientists to map the locations of mutations that cause genetic diseases by genotyping other marker SNPs (Hey 2004). Knowing the possible causes and locations of recombination hot spots and cold spots could make it possible to represent much of the genomic variations using only a small number of SNPs found within the region.

1.2. The selection and evaluation of genetic markers for association studies

1.2.1. The principles of association studies

Common inherited disorders are difficult to study because they are the result of a combination of various genes and environmental factors (Gambano et al. 2000). These disorders tend to be inherited in families, but they don't follow the typical mendelian inheritance patterns. Therefore, they are usually referred to as non-mendelian disorders, or complex disorders. Pedigree based linkage analysis is a successful method for studying mendelian disorders, but has not been as successful in the studies of complex genetic traits. This indicates a need for identifying different approaches (Goddard et al. 2000). The possibility of using dozens or even hundreds of past generations of recombination to achieve the fine-scale gene localization is one of the advantages of association studies (Jorde 2000). The linkage disequilibrium mapping has proved itself as a valuable tool for finding disease related genes (Feder et al. 1996; Fujita et al. 1990; Kerem et al. 1989; Petrukhin et al. 1993). Due to its success in localizing Mendelian disease genes, it is hoped that this approach is also useful for localizing genes of complex diseases, even in whole-genome association studies (Risch and Merikangas 1996).

Association analyses are based on linkage disequilibrium. With tight linkage between the disease and marker loci, the possibility of recombination is small and the allelic association can be used to localize the disease gene. The allele frequency of marker locus is different in case and control samples if the marker locus itself is causing the disease or is in tight linkage with the susceptibility locus (Cardon and Bell 2001; Gambano et al. 2000; Risch and Merikangas 1996).

There are two complementary approaches for selecting markers. The first approach exploits the LD between loci in many parts of the genome. If a subset of SNPs is genotyped, they capture information about adjacent loci and regional haplotypes. There are several strategies for selecting informative subsets of SNPs (also referred to as tagging SNPs or tagSNPs), which will be discussed later (2.2.2.1). The second approach tries to access the variation of the genome with likely functional effect. Obvious targets are non-synonymous coding variants, as they alter the aminoacid sequence in the gene product (Hattersley and McCarthy 2005). Their advantage is that the number of common coding SNPs (cSNPs) is several orders of magnitude smaller than the number of common SNPs overall (Carlson et al. 2004a). Recently, the structural variants and non-protein-coding sequence has gained attention with their possible influence to phenotype (Birney et al. 2007; Feuk et al. 2006). Therefore, genotyping only non-synonymous coding variants might lead to false negative results.

The association study design can also be classified according to its magnitude. Due to the cost of genotyping and multiple testing corrections, the candidate gene approach has been widely used. In thit case, one or only a few genes are studied. These genes are selected according to the results of previous studies, or on the basis of other evidence that the gene might be associated to the studied disease.

There are numerous association studies that cannot be replicated, which has led to skepticism about that approach. Therefore, the importance of a good study design is crucial for a successful association study (Cardon and Bell 2001). Current genome-wide association study (GWAS) contains multiple steps:

- **The Genome-wide SNP Genotyping**: Current SNP panels contain hundreds of thousands or even millions of SNPs, which can be quickly and automatically genotyped for each individual. To reduce the cost of large-scale association study, it is possible to use DNA pooling (Docherty et al. 2007; Sham et al. 2002).
- **Validation with subset of SNPs**: SNPs showing the strongest association with a disease are selected to the next step. Independent case and control samples are tested for association with custom SNP panels. In case of the genome-wide marker set, the number of SNPs analyzed is rather high and the multiple-testing problem arises. Analysis of each single SNP can be treated as an independent test and multiple testing correction must be used for declaring a significant association in GWAS (Risch and Merikangas 1996). To overcome this problem, the multi-stage screening approach can be used, assuming a smaller number of susceptibility markers are genotyped in the second stage. The significance level for declaring association doesn't have to be as strict as for all markers in the whole-genome marker panel (Hirschhorn and Daly 2005). It has been shown that the joint analyzing the data from both stages almost always has more power to detect genetic association then the replication based analysis (Skol et al. 2006). Therefore, the joint analysis for all two-stage genome-wide analysis should be used.
- **Independent replication:** Independent replication from different population to validate the results.

In the case of a complex disease, the number of cases and controls should be relatively high. This is due to the modest genetical effects of single causal alleles (Cardon and Bell 2001; Hattersley and McCarthy 2005). Current association studies use thousands or even tens of thousands of individuals to analyze the effect of disease causing alleles (Herbert et al. 2006; Sladek et al. 2007; Smyth et al. 2006).

1.2.2. The selection of genetic markers for association studies

In 1998, Chapman and Wijsman claimed that single-marker LD testing with bi-allelic markers was feasible only for rapidly growing genetic isolates because the mapping resolution was not high enough (Chapman and Wijsman 1998). Since then, the number of SNPs in the public databases has been rising and currently more than 10 million validated SNPs are available (Hinds et al. 2005).

Alleles of closely located SNPs are often correlated, resulting in a reduced genetic variability and the defining of a limited number of „haplotype blocks" (Patil et al. 2001). Some of these haplotypes might extend only for a few kilobases (kb) and others may extend more than 100 kb (Abecasis et al. 2001; Clark et al. 1998; Reich et al. 2001). The extent of LD is formed by mutation, recombination, selection, population history and stochastic events. This suggests that a comprehensive description of the haplotype structure of the human genome is possible only by empirical studies with dense SNP sets.

The first large-scale publicly available SNP set was generated by Perlegen Sciences, Inc. in 2001 (Patil et al. 2001). It contained 20 independent copies of chromosome 21, representing the African, Asian and European chromosomes. In this study, 35,989 SNPs were identified and 24,047 of them had minor allele represented more than once in their dataset.

In October 2002 the International HapMap Project was launched with the goal of determining the common variants of the DNA sequence in the human genome and then determining their frequencies and identifying correlations between them (Gibbs et al. 2003). A total of 269 DNA samples were studied:

1. 90 CEPH samples from US Utah population with Northern and Western Europe ancestry. These samples were collected by Centre d'Etude du Polymorphisme Humain and contained info of 30 trios of two parents and one adult child.
2. 90 Yoruba people in Ibadan, Nigeria (also 30 trios).
3. 44 unrelated Japanese in Tokyo, Japan.
4. 45 unrelated Han Chinese in Beijing, China.

The sample sizes were found to be sufficient to identify 99% of the haplotypes with an allele frequency of 5% and higher (Gibbs et al. 2003). In the first genotyping phase, the goal was to genotype 600,000 markers with an average spacing of 5kb. This was completed in 2005, when a resource consisting over a million accurate SNP genotypes in 269 individuals was released (Altshuler et al. 2005). Two years later, in 2007, the second genotyping phase was finished, adding over 2.1 million validated SNPs to the database.

The genome-wide genotype information of the HapMap project can be used for locating recombination hotspots, LD blocks, and regions with low haplotype diversity. This enables improved selection of tagging SNPs (tagSNPs), thereby increasing the power of an association analysis. In the next phase of the HapMap project, seven additional populations from the USA, Kenya and Italy have been selected for genotyping. The main goals for the next phase are the comparison of genome-wide patterns of variation and assessing the transferability of tagSNPs between populations.

1.2.2.1. TagSNP selection methods

In the case of an „indirect" association study, the marker set will be selected to describe most of the common variations in the genomic region under observation. There are many algorithms available to select the most informative set of common single-nucleotide polymorphisms (tagSNPs).

The tagSNP selection methods can be based on haplotype blocks (haploblocks) – regions with a low recombination rate. In 2001, Patil et al described a greedy tagSNP selection algorithm, which was based on minimizing the number of tagSNPs required to describe most of the common haplotypes in each block for the entire chromosome (Patil et al. 2001). As a greedy algorithm gives an approximate solution, Zhang et al. developed a dynamic programming algorithm for haplotype block partitioning as an optimal solution to the problem (Zhang et al. 2002). In order to describe 80% of the most common haplotypes on the entire 21 chromosome, with 24,047 common SNPs genotyped, 3,582 SNPs in the 2,575 haplotype blocks were required. The main disadvantage of this solution is the high variability of haplotype block borders and selected SNPs. The marker and haploblock selection is

highly dependent on the underlying samples and marker density (Cardon and Abecasis 2003; Carlson et al. 2004b).

To make the haplotype block selection less dependent on marker density, several haploblock selection methods were proposed based on LD measures. Block borders, according to the four-gamete test (Hudson and Kaplan 1985), can reflect the LD regions with no evidence of recombination (Wang et al. 2002). According to this method, the haplotype block is extended as long as four possible haplotypes of a pair of bi-allelic loci are not present (Figure 1). The addition of the fourth haplotype indicates a recombination event and therefore identifies the block border.

Another method is based on the solid spine of LD (Barrett et al. 2005). This method searches for a „spine" of LD, where the first and last marker of the haplotype block are in strong LD with all of the intermediate markers, but does not require that the intermediate markers are necessarily in LD with each other.

The third and most popular haploblock selection method is based on the confidence intervals of *D'*. It is a normalized measure of allelic association which reflects the history of recombinations between the SNP pairs. As the point estimate of *D'* tends to be biased upward, if a small number of samples or rare alleles are analyzed, Gabriel et al. defines the 95% confidence interval for each *D'* estimate (Gabriel et al. 2002). The authors of this method claim that haplotype blocks defined by this method are robust to study-specific differences, like the frequencies of SNPs and sample sizes.

However, the haploblock borders tend to fluctuate in different populations or even in the same population if different set of samples is selected. Cardon and Abecasis claimed in their paper in 2003 that the confidence intervals method shares the same problems as all methods that are based on LD measures: it is unclear how haplotype ancestry is reflected in a matrix of pairwise LD coefficients that are not independent and therefore the defined blocks are subjective and arbitrary (Cardon and Abecasis 2003).

For the marker selection from the haplotype blocks, the haplotype tagging SNP method (htSNPs) has been proposed (Johnson et al. 2001). HtSNPs are defined as a minimum number of informative SNPs, which can be used to distinguish between all common haplotype variants in a block. A single SNP can distinguish between two haplotypes; two SNPs could distinguish between up to four haplotypes, etc. The disadvantage of this tagSNP selection method is that the relationship between the tagSNPs selected to describe haplotypes and the power to detect disease risk associated with the existing polymorphism are poorly addressed (Carlson et al. 2004b).

The htSNPs are not the most efficient way of describing common haplotypes for association studies, due to the correlation or overlap between neighboring haploblocks (Daly et al. 2001). This indicates that tagSNP selection methods, which allow for interblock disequilibrium or ignore block boundaries altogether, could be more effective than those who treat chromosomes as a series of discrete blocks (Cardon and Abecasis 2003).

For a haploblock independent tagSNP selection, an algorithm based on the r^2 LD statistic was proposed by Carlson et al., because r^2 is directly related to the statistical power to detect disease associations with markers (Carlson et al. 2004b). The r^2-bin method is based on a greedy algorithm. The SNP exceeding the r^2 threshold, with a maximum number of other SNPs, is identified. This

maximally informative SNP and all the associated markers are grouped as a first r^2-bin. One tagSNP is selected from each bin and can be selected for assay on the basis of genomic context, ease of assay design, or other user specified criteria. The binning process is iterated, analyzing not-binned markers as long as all markers are in bins. If an SNP does not exceed the r^2 threshold with any other SNP in the region, it is placed in a singleton bin. As expected, r^2-bin based tagSNPs are more powerful than an equivalent number of either haplotype-selected htSNPs or randomly selected SNPs for detecting the association between a tagSNP and disease phenotype (Carlson et al. 2004b).

It should be pointed out, that most of the tagSNP selection methods are focused on describing of the common variants. Rare mutations and SNPs with low minor allele frequency may not be detected by reduced number of marker SNPs.

1.2.2.2. Commercial SNP sets

Currently, two major companies are providing fixed SNP panels for genome-wide SNP genotyping – Illumina Inc. and Affymetrix Inc. Both of these companies are offering very high throughput and accuracy with low cost per SNP analysis. There are obvious advantages of having fixed SNP panels, including the possibility of combining datasets across laboratories and designing statistical methods for commonly used panels (Barrett and Cardon 2006). These SNP panels could be used for DNA pooling (Macgregor 2007; Steer et al. 2007) and for detecting copy number variations in the human genome regions (Komura et al. 2006; Peiffer et al. 2006).

Mapping 10K Array Sets from Affymetrix GeneChip series were introduced in 2003, when a work about rapid genotyping of 14,548 SNPs in three different human populations was published (Kennedy et al. 2003). This was followed by Mapping 100K (Di et al. 2005) and Mapping 500K Array sets (Nicolae et al. 2006). The HumanHap 300 and later the HumanHap 550 Array Sets from the Illumina Infinium series were developed after the end of the Hapmap's first genotyping phase, providing 318,000 and 555,000 tagSNPs accordingly. While the SNPs included on the Affymetrix SNP panels are selected on the basis of their technical quality and are evenly distributed across human genome, SNPs on the Illumina platforms are selected using the r^2-bin method from HapMap genotype data (Pe'er et al. 2006). The HumanHap 300 is based on CEPH data, therefore the most efficient option for describing the genome variation of Northern and Western European populations. The HumanHap 500 has been altered by adding tagSNPs to describe the variation of other HapMap populations as well (JPT+CHB, YRI). To provide even more comprehensive coverage in African and African-American populations, the new Illumina HumanHap 650Y has been developed, with over 100,000 Yoruba-specific tagSNPs.

To capture other types of genetic differences, such as copy number variations, even larger SNP panels have been designed. The first the Affymetrix SNP Array 5.0, was introduced with the Mapping 500K Array Set combined with 420,000 additional non-polymorphic probes that can measure other genetic differences. The Illumina Inc. responded by developing the Human1M BeadChip, featuring over one million markers to interrogate human genetic variation, using SNPs and CNV probes. The latest Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers, including more than 906,600 SNPs and 946,000 CNV probes.

1.2.3. The evaluation of the performance of genetic markers

It is important to estimate the power of detecting association with the existing marker set in any study design and disease scenario. In the case of sparser marker set, a larger number of samples must be collected in order to achieve sufficient power to detect the hypothesized effect (Purcell et al. 2003). A a simple way to evaluate the performance of markers is by using the square of the correlation coefficient (r^2) between a marker and a putative causal allele (Pritchard and Przeworski 2001). This pairwise measure has become a standard for evaluating the performance of marker sets (Barrett and Cardon 2006; Pe'er et al. 2006). This coverage measure can be biased upwards if tagSNPs themselves are part of the coverage calculation or part of the reference set was used to select tag SNPs. Therefore, the coverage measure must be corrected considering the number of markers and the count of tagSNPs in the reference set and the estimate for the total number of SNPs in a described region (Barrett and Cardon 2006).

2. Results and discussion

2.1. Aims of the present study

1. To define the extent and LD patterns of European populations. This information is essential for the marker selection of association studies. The extent of LD is necessary for estimating the number of SNPs necessary to describe certain genomic regions. The comparison of LD patterns across European populations provides information as to whether SNP panels can be universally used for association studies in different populations (Ref I, II).
2. To evaluate the performance of tagSNPs and calculate how well the tagSNPs of one population sample can be used to describe the variability of another population. Different haploblock and tagSNP selection methods have been proposed for marker selection. It is necessary to evaluate whether the selected tagSNP set can describe the full heterogeneity of the genomic regions. The genotype information of four HapMap population samples is used for marker selection of association studies from different populations. Therefore, it is necessary to know how much information might get lost by transferring the tagSNPs (Ref II, III).
3. To evaluate how well can new whole-genome SNP panels describe the variability of the Estonian population. The currently available whole-genome SNP panels are the cheapest and most convenient way of performing genotyping studies. As the SNP selection strategies differ for these panels, it is necessary to compare different marker sets to choose the best one for the Estonian population samples (Ref IV).

2.2. Population samples

Three different sets of population samples were used in this study (Table 2). Also, publically available HapMap population samples were used for comparisons(Gibbs et al. 2003).

In the chromosome 22 study, three populations were used – 77 CEPH (Coriell Cell Repositories) family DNAs, 90 unrelated DNAs from UK and 51 unrelated Estonian DNAs. CEPH and UK population markers were genotyped using the Third Wave Technologies Invader assay (Mein et al. 2000). The Estonian population samples were genotyped using APEX technology (Kurg et al. 2000). In the CEPH DNA panel 1,504 were successfully genotyped, in the UK panel 1,285 and in the Estonian panel 908 SNPs were successfully genotyped. Monomorphic markers and markers with segregation error, Hardy-Weinberg equilibrium deviations and other quality issues were not used in this analysis.

In study of four gene regions across nine European populations, 90 – 170 samples were genotyped. In addition to CEPH and Estonian samples, German populations from Pomerania (SHIP) , Schleswig-Holstein (POPGEN) and urban regions of the southern part of Germany (KORA); two alpine populations from Vinschgau (VIN) and from Grödnertal and Gadertal (LAD); an Italian population from the Emilia-Romagna region (BRISI) and Calabria (CALA). SNPs were selected from four different genomic regions, which all contain candidate genes for complex diseases. For each region, SNPs were

selected evenly, covering the gene and 76-174 kb of flanking regions in both sides of the gene. SNPs were genotyped by the primer extension reaction of multiplex PCR products, with the detection of allele-specific extension products by matrix-assisted laser desoption/ionization time-of-flight (MALDI-TOF) mass spectroscopy.

In the third and fourth study of two ENCODE regions on chromosome 2 (ENr112 on 2p16.3 and ENr131 on 2p37.1), 1,090 ethnical Estonian population samples were genotyped. These regions differ in their average recombination rates and content of known genes. From two regions, 1420 SNPs were randomly selected and genotyped using the Illumina GoldenGate® Assay. As a marker validation, genotyping success >= 95%, p-level for HWE >=0.001 and two cut-off levels of MAF were used.

Additionally, the publically available SNP genotype data from the HapMap Consortium was used. Nearly 4 million non-redundant validated SNPs of 60 founders of CEPH (Utah residents with ancestry from northern and western Europe), 60 founders of Yoruba (Ibadan, Nigeria) and 90 unrelated individuals from mixed Asian (Han Chinese in Beijing and Japanese in Tokyo) population samples have been genotyped in the International HapMap project. In the marker validation process, each marker was tested for Mendelian inconsistencies (< 2), genotyping success (>80%), p-level of HWE >=0.0001 and duplicate discrepancies (<2). Markers had to be polymorphic in at least one of the studied populations.

2.3. Linkage disequilibrium patterns in European populations (Ref I, II)

Each copy of a chromosome in a given population might be a unique mosaic of ancestral chromosomes. Still, those haplotypes do not occur in the population at frequencies expected by random recombination events in the past. This is caused by linkage disequilibrium (LD) – consecutive marker alleles tend to be correlated with each other. Knowing the pattern of LD is critical for marker selection strategies in association studies. To study the pairwise LD in European populations, it was measured along the complete sequence of human chromosome 22 (Table 2 Ref I). To describe the pairwise LD between markers, the most commonly used D' and r^2 statistics were calculated. According to the results, the LD decays with distance, but shows very large variability (Figure 1 in ref. I). The analysis of the LD with a sliding window reveals that the LD pattern is highly variable across the chromosome (Figure 2 in ref. I). Regions with nearly complete LD up to 804kb were interspersed with regions of little or undetectable LD. When the LD patterns of different European population samples were compared, a strong correlation between high and low recombination spots was revealed. This study demonstrates that the recombination rates are similar across different European populations and it is possible to create genome-wide maps of LD.

To study the LD patterns of nine distinct European populations, another genotyping study was performed with markers from four genomic regions (Table 2 Ref II). Pairwise LD plots were drawn for each population and each genomic region. Standard LD plots of pairwise LD revealed similar patterns across population samples (Figure A2 [online only] in ref. II). To compare the LD structure across populations in a robust probability-based assessment, haplotype blocks were calculated with block

overlaps allowed. The bootstrap frequencies of specific boundary positions were evaluated. The calculations were based on a sample size of 100 individuals per population, to exclude variation in sample size as a potential cofounder. The general patterns of LD were found to be similar across populations (Figure 2 in ref. II). The LMNA gene region shows the most conserved block starts and ends across the populations studied. Only one, the largest haploblock varied across the populations being shifted in the range of 7-15 kb. The overall variability of block structure among the populations was explored using multidimensional scaling of all four gene regions. Most extreme block structures were indicated for EST, LAD, VIN, BRISI, and CALA population samples (Figure 3 in ref. II). The German populations, SHIP, POPGEN, KORA, and the reference population, CEPH, appeared in the center, indicating an intermediate block structure. In general, a conservation of LD patterns across European population samples was observed. Still, the bootstrapping procedure revealed shifts in the positions of LD block boundaries between population samples.

2.4. Evaluation of the performance of the tagSNPs and their transferability among populations (Ref II, III)

One of the main purposes of the HapMap project is defining the most common haplotypes of the human genome and then making this information freely available in the public domain (Gibbs et al. 2003). Tagging SNPs, calculated according to this information, should be able to adequately describe other similar populations, i.e. Yoruba population information to describe other African populations, CEPH – Northern and Western Europeans and Chinese and Japanese mixed samples – Asian populations. Therefore, it is important to evaluate how well tagSNPs can be transferred among different populations.

TagSNP transferability was measured in two different studies (Table 2 Ref II, III). The first study evaluates the performance of tagSNPs in 9 different European populations, giving an overview of how well each CEPH tagSNPs covered different European populations. The second study focused on the Estonian population and addressed the questions regarding the effect of sample size, SNP density, and the minor allele frequency of SNPs to the tagging performance. In both studies the tagSNPs were calculated according to the r^2-bin method(Carlson et al. 2004b). According to this method, tagging SNPs are selected from marker groups with a high correlation coefficient between the markers in group (r^2>0.8).

The performance of tagSNPs in the CEPH population sample was compared with local population samples across Europe. For each population, the relative proportion of SNPs was calculated, which was tagged by CEPH tagSNPs. For comparison, the tagSNPs were also calculated according to a reduced number of local samples. The selected tagSNP sets were then tested on the full SNP set in all populations.

The tagSNPs selected according to the CEPH information performed adequately well for three genes out of four: SNCA, FKBP5 and LMNA (Figure 5 in ref. II). More than 70% of the typed SNPs had a r^2 value higher than 0.8 with best tagSNP. In LMNA and SNCA genes tagSNPs which were calculated according to CEPH population samples performed better than local samples with 20 individuals. It takes 40 or 60 individuals from local samples to get comparable results to the CEPH samples. In FKBP5, CEPH samples performed better than 20 local samples in all populations, except VIN and CALA. The CEPH tagSNPs did not describe populations in the PLAU gene region – six populations

showed a ratio of tagged SNPs < 70% (CALA 53%). Data of 20 random individuals of local populations performed better then CEPH trios. Through CEPH tagSNPs outperformed 20 local samples of LAD and POPGEN population samples.

Additionally, the performance of tagSNPs calculated according to HapMap populations in two ENCODE regions was measured. The transferability of tagSNPs was tested between different HapMap populations and Estonian population data. CEPH tagSNPs performed equally well in both of these regions for the Estonian population samples. More than 90% of the SNPs were correlated with an r^2>0.8 for all SNPS with MAF 5% or more (Figure 5 in ref. III). The CHB+JPT tagSNPs didn't perform as well as the CEPH ones. They captured less than 80% of SNPs in most cases. In the ENCODE 1 region, tagSNPs were selected to have a minimum MAF of 10% showed the best performance at capturing SNPs in any population. In the ENCODE 2 region, MAF 5% showed the best performance. THE YRI tagSNP set worked well in all population samples, but at the expense of using 2-3 times more tagSNPs. TagSNP sets which were generated as a combination of CEPH and JPT+CHB population samples improved tagging performance about 2-5%, but needed 20-30% more tagSNPs.

The effect of minor allele frequency of SNPs was also measured for tagSNP transferability (Figure 6 in ref. III). Markers with higher MAF in the Estonian population samples tended to be better correlated with CEPH tagSNPs. An important aspect is that markers with very low MAF (less than 5%) were poorly described by CEPH tagSNPs. For these SNPs, 17% of the ENCODE 1 region and 23% of the ENCODE 2 region markers had a r^2 lower than 0.5 with best tagSNP. On the other hand, all SNPs with a MAF>20% had correlation to best tagSNP r^2>0.7.

The effect of sample size used to derive the tagSNPs was calculated using random sets of 10 to 1,000 Estonian samples. For each dataset, an average of 100 tests was used to evaluate the performance of tags relative to all polymorphic SNPs on the CEU sample (Figure 7 in ref. III). Sample size mainly had an effect for less frequent SNPs (MAF < 5%). For markers with higher MAF, the optimal tagging was obtained with 90-100 independent samples. However, the difference in tagging performance using the sample size of 60 was non-significant for these SNPs.

The effect of marker density for tagSNP selection was assessed using six different datasets. The 500-kbp ENCODE regions were divided into equal-sized windows and one polymorphic SNP in each CEPH population was selected randomly from each window. The tagSNPs were selected from these sparser CEPH datasets and their performance was measured on Estonian population samples. The tagging performance is worse in the case of each decreased density that was studied (Figure 8 in ref. III). The effect is larger in the low LD ENCODE 2 region.

According to the results of different European populations, CEPH performed well as a reference for tagSNP design in the two gene regions studied – SNCA and LMNA. For PLAU and FKBP5, CEPH is not as reliable of a reference. TagSNPs calculated according to a small number of local samples will give better results. However, the increase of marker density in reference population gives better tagSNP efficiency and transferability. In the second study with Estonian samples, where denser SNP data was available for the CEPH population, the results were better.

2.5. Evaluation of the performance of whole-genome marker sets for genome-wide association studies (IV)

We compared four commercial SNP panels: HapMap 300 and HapMap 550 Array Sets from the Illumina Infinium series and the Mapping 100K and Mapping 500K Array sets from the Affymetrix GeneChip series. Tagging performance of these panels was evaluated among HapMap CEPH, mixed Asian (JPT+CHB) and Yoruba (YRI) population samples and also on the Estonian population sample (Table 2 Ref IV). To evaluate the performance of commercial panels, for each marker present in the HapMap data, we calculated the best tagging SNP from each commercial panel. The percentage of SNPs covered with $r^2>0.8$ and the mean r^2 between each marker and their best tagging SNP for the investigated population was calculated. This was done with two MAF cut-offs: 1% and 5%. According to the results, all SNP panels had poor coverage on the Yoruba population (Figure 1 A-B in ref. IV), but the coverage of CEPH and JPT+CHB populations reaches up to 80-90% on HumanHap 550. The previously unpublished HumanHap 550 had whole-genome coverage estimations as follows: 86% CEPH, 83% JPT+CHB and 48% for YRI population sample. It showed the best performance among the technologies analyzed, although the increase over HumanHap 300 is not large on the CEPH population.

Since fewer SNPs were genotyped for the Estonian population, the mean r^2 and coverage could not be directly compared with the results of HapMap populations. Many tagSNPs were not genotyped in the Estonian samples and therefore their pairwise LD could not be calculated. As a solution, we reduced the marker counts of other populations so that only markers present in the Estonian dataset were used for pairwise LD calculation. The relative performance of the commercial SNP sets were calculated on the reduced SNP set and the results were expressed as the fractions of the coverage of the CEPH sample (Figure 2A-D in ref. IV). The results show that the coverage of the commercial SNP panels have the same efficiency in Estonian, JPT+CHB and CEPH population samples and have lower tagging performance in the Yoruba population sample. Some studies, including our own, have already shown that CEPH population data from HapMap samples can successfully used to tag other European populations(Conrad et al. 2006; Gonzalez-Neira et al. 2006; Willer et al. 2006). Also, it is known that most of the common SNPs are captured by first generation whole genome SNP panels(Barrett and Cardon 2006; Pe'er et al. 2006). Our study supports the combination of these results with this new conclusion: commercial SNP panels can capture most of the common SNPs from non-reference European population samples.

Current whole-genome SNP panels still don't cover ~14% of markers ($r^2<0.8$) which is a quite large number of markers – if we assume that we would like to cover circa 7.5 million markers, approximately one million of them are poorly covered. Unfortunately, any of these might be the disease-causing SNP that we are looking for in the association study. Our hope is that new, larger commercial platforms will be able to cover most of the currently uncovered SNPs by adding tagSNPs.

To investigate how universal the SNP sets of commercial platforms are for studying different populations we counted the tagSNPs used for describing only one population and those that could identify SNPs from multiple populations. We determined whether each commercial SNP was the best describer of SNPs in one, two or all three populations. Thus, it was possible to compare the universality of coverage of the different commercial platforms in different populations (Figure 3 in ref. IV). Strong bias towards CEPH specific markers in the HumanHap 300 panel was observed, which could be explained by the SNP selection strategy used for this SNP panel – SNPs were picked

according to the HapMap CEPH population data using the r^2 based method (Carlson et al. 2004b). In contrast, GeneChip panels describe population-specific markers from all three populations fairly equally.

Our results show that universal markers constitute 63-82% of all SNPs and these numbers are similar in all the commercial platforms studied. Approximately 10% of the SNPs in commercial panels describe SNPs from only a single population sample. The markers that are able to tag different populations are expected to be useful in many populations. This information is important for planning association studies in non-HapMap populations.

Conclusions

The LD pattern is highly variable along the 22nd chromosome. Long regions of nearly complete LD are separated by regions with almost undetectable LD. The LD is not only dependent on the physical distance between markers but also differs along genomic regions. At the same time, results indicate that in general, the LD patterns are conserved across European populations, which makes it possible to create a general LD map for all European populations. Still, some shifts between the boundaries of high-LD regions could be observed between populations.

CEPH samples of the International HapMap Project are performing well for tagSNP design in most of the analyzed European populations. As the tagSNP transferability is depending on the marker density of the reference population, the sparser coverage of the HapMap phase I data did not provide a reliable marker set. TagSNPs calculated according to the denser phase II data could be transferred between populations without significant power loss. TagSNP coverage was also found to be dependent on the SNP minor allele frequency. Markers with higher MAF in the Estonian population samples tended to be better correlated with the CEPH tagSNPs.

The studies have shown that HapMap populations could efficiently be described by commercial genome-wide SNP panels (Barrett and Cardon 2006; Pe'er et al. 2006). Several studies have also been performed to evaluate how well other European population samples can be described by tagSNPs calculated from HapMap CEPH data (Gonzalez-Neira et al. 2006; Willer et al. 2006). In addition to this information, it is also useful to know how well current genome-wide genotyping arrays can capture genetic variation of a non-HapMap population. According to the results, commercial SNP panels provide similar levels of coverage for non-reference European population (Estonian) as to those in the HapMap SNP sample.

References

Abecasis, G., E. Noguchi, A. Heinzmann, J. Traherne, S. Bhattacharyya, N. Leaves, G. Anderson, Y. Zhang, N. Lench, and A. Carey. 2001. Extent and Distribution of Linkage Disequilibrium in Three Genomic Regions. *The American Journal of Human Genetics* **68:** 191-197.

Altshuler, D., L. Brooks, A. Chakravarti, F. Collins, M. Daly, and P. Donnelly. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

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

Ardlie, K., S. Liu-Cordero, M. Eberle, M. Daly, J. Barrett, E. Winchester, E. Lander, and L. Kruglyak. 2001. Lower-Than-Expected Linkage Disequilibrium between Tightly Linked Markers in Humans Suggests a Role for Gene Conversion. *The American Journal of Human Genetics* **69:** 582-589.

Barrett, J. and L. Cardon. 2006. Evaluating coverage of genome-wide association studies. *Nature Genetics* **38:** 659-662.

Barrett, J., B. Fry, J. Maller, and M. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263-265.

Birney, E. J.A. Stamatoyannopoulos A. Dutta R. Guigo T.R. Gingeras E.H. Margulies Z. Weng M. Snyder E.T. Dermitzakis R.E. Thurman M.S. Kuehn C.M. Taylor S. Neph C.M. Koch S. Asthana A. Malhotra I. Adzhubei J.A. Greenbaum R.M. Andrews P. Flicek P.J. Boyle H. Cao N.P. Carter G.K. Clelland S. Davis N. Day P. Dhami S.C. Dillon M.O. Dorschner H. Fiegler P.G. Giresi J. Goldy M. Hawrylycz A. Haydock R. Humbert K.D. James B.E. Johnson E.M. Johnson T.T. Frum E.R. Rosenzweig N. Karnani K. Lee G.C. Lefebvre P.A. Navas F. Neri S.C. Parker P.J. Sabo R. Sandstrom A. Shafer D. Vetrie M. Weaver S. Wilcox M. Yu F.S. Collins J. Dekker J.D. Lieb T.D. Tullius G.E. Crawford S. Sunyaev W.S. Noble I. Dunham F. Denoeud A. Reymond P. Kapranov J. Rozowsky D. Zheng R. Castelo A. Frankish J. Harrow S. Ghosh A. Sandelin I.L. Hofacker R. Baertsch D. Keefe S. Dike J. Cheng H.A. Hirsch E.A. Sekinger J. Lagarde J.F. Abril A. Shahab C. Flamm C. Fried J. Hackermuller J. Hertel M. Lindemeyer K. Missal A. Tanzer S. Washietl J. Korbel O. Emanuelsson J.S. Pedersen N. Holroyd R. Taylor D. Swarbreck N. Matthews M.C. Dickson D.J. Thomas M.T. Weirauch J. Gilbert J. Drenkow I. Bell X. Zhao K.G. Srinivasan W.K. Sung H.S. Ooi K.P. Chiu S. Foissac T. Alioto M. Brent L. Pachter M.L. Tress A. Valencia S.W. Choo C.Y. Choo C. Ucla C. Manzano C. Wyss E. Cheung T.G. Clark J.B. Brown M. Ganesh S. Patel H. Tammana J. Chrast C.N. Henrichsen C. Kai J. Kawai U. Nagalakshmi J. Wu Z. Lian J. Lian P. Newburger X. Zhang P. Bickel J.S. Mattick P. Carninci Y. Hayashizaki S. Weissman T. Hubbard R.M. Myers J. Rogers P.F. Stadler T.M. Lowe C.L. Wei Y. Ruan K. Struhl M. Gerstein S.E. Antonarakis Y. Fu E.D. Green U. Karaoz A. Siepel J. Taylor L.A. Liefer K.A. Wetterstrand P.J. Good E.A. Feingold M.S. Guyer G.M. Cooper G. Asimenos C.N. Dewey M. Hou S. Nikolaev J.I. Montoya-Burgos A. Loytynoja S. Whelan F. Pardi T. Massingham H. Huang N.R. Zhang I. Holmes J.C. Mullikin A. Ureta-Vidal B. Paten M. Seringhaus D. Church K. Rosenbloom W.J. Kent E.A. Stone S. Batzoglou N. Goldman R.C. Hardison D. Haussler W. Miller A. Sidow N.D. Trinklein Z.D. Zhang L. Barrera R. Stuart D.C. King A. Ameur S. Enroth M.C. Bieda J. Kim A.A. Bhinge N. Jiang J. Liu F. Yao V.B. Vega C.W. Lee P. Ng A. Yang Z. Moqtaderi Z. Zhu X. Xu S. Squazzo M.J. Oberley D. Inman M.A. Singer T.A. Richmond K.J. Munn A. Rada-Iglesias O. Wallerman J. Komorowski J.C. Fowler P. Couttet A.W. Bruce O.M. Dovey P.D. Ellis C.F. Langford D.A. Nix G. Euskirchen S. Hartman A.E. Urban P. Kraus S. Van Calcar N. Heintzman T.H. Kim K. Wang C. Qu G. Hon R. Luna C.K. Glass M.G. Rosenfeld S.F. Aldred S.J. Cooper A. Halees J.M. Lin H.P. Shulha M. Xu J.N. Haidar Y. Yu V.R. Iyer R.D. Green C. Wadelius P.J.

Farnham B. Ren R.A. Harte A.S. Hinrichs H. Trumbower H. Clawson J. Hillman-Jackson A.S. Zweig K. Smith A. Thakkapallayil G. Barber R.M. Kuhn D. Karolchik L. Armengol C.P. Bird P.I. de Bakker A.D. Kern N. Lopez-Bigas J.D. Martin B.E. Stranger A. Woodroffe E. Davydov A. Dimas E. Eyras I.B. Hallgrimsdottir J. Huppert M.C. Zody G.R. Abecasis X. Estivill G.G. Bouffard X. Guan N.F. Hansen J.R. Idol V.V. Maduro B. Maskeri J.C. McDowell M. Park P.J. Thomas A.C. Young R.W. Blakesley D.M. Muzny E. Sodergren D.A. Wheeler K.C. Worley H. Jiang G.M. Weinstock R.A. Gibbs T. Graves R. Fulton E.R. Mardis R.K. Wilson M. Clamp J. Cuff S. Gnerre D.B. Jaffe J.L. Chang K. Lindblad-Toh E.S. Lander M. Koriabine M. Nefedov K. Osoegawa Y. Yoshinaga B. Zhu and P.J. de Jong. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799-816.

Botstein, D., R. White, M. Skolnick, and R. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32:** 314-331.

Brookes, A. 1999. The essence of SNPs. *Gene* **234:** 177-186.

Cardon, L. and G. Abecasis. 2003. Using haplotype blocks to map human complex trait loci. *Trends Genet* **19:** 135-140.

Cardon, L. and J. Bell. 2001. Association study designs for complex diseases. *Nat Rev Genet* **2:** 91-99.

Carlson, C., M. Eberle, L. Kruglyak, and D. Nickerson. 2004a. Mapping complex disease loci in whole-genome association studies. *Nature* **429:** 446-452.

Carlson, C., M. Eberle, M. Rieder, Q. Yi, L. Kruglyak, and D. Nickerson. 2004b. Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *The American Journal of Human Genetics* **74:** 106-120.

Chakraborty, R. and K. Weiss. 1988. Admixture as a Tool for Finding Linked Genes and Detecting that Difference from Allelic Association between Loci. *Proceedings of the National Academy of Sciences* **85:** 9119-9123.

Chapman, N. and E. Wijsman. 1998. Genome Screens Using Linkage Disequilibrium Tests: Optimal Marker Characteristics and Feasibility. *The American Journal of Human Genetics* **63:** 1872-1885.

Chen, J.M., D.N. Cooper, N. Chuzhanova, C. Ferec, and G.P. Patrinos. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*.

Clark, A., K. Weiss, D. Nickerson, S. Taylor, A. Buchanan, J. Stengaard, V. Salomaa, E. Vartiainen, M. Perola, and E. Boerwinkle. 1998. Haplotype Structure and Population Genetic Inferences from Nucleotide-Sequence Variation in Human Lipoprotein Lipase. *The American Journal of Human Genetics* **63:** 595-612.

Conrad, D., M. Jakobsson, G. Coop, X. Wen, J. Wall, N. Rosenberg, and J. Pritchard. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38:** 1251-1260.

Daly, M., J. Rioux, S. Schaffner, T. Hudson, and E. Lander. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* **29:** 229-232.

De La Vega, F., H. Isaac, A. Collins, C. Scafe, B. Halldorsson, X. Su, R. Lippert, Y. Wang, M. Laig-Webster, and R. Koehler. 2005. The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Research* **15:** 454-462.

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

Di, X., H. Matsuzaki, T. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, and G. Yang. 2005. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* **21:** 1958-1963.

Ding, Y., H. Chi, D. Grady, A. Morishima, J. Kidd, K. Kidd, P. Flodman, M. Spence, S. Schuck, and J. Swanson. 2001. Evidence of positive selection acting at the human dopamine receptor D4 gene locus. *Proceedings of the National Academy of Sciences***:** 12464099.

Docherty, S., L. Butcher, L. Schalkwyk, and R. Plomin. 2007. Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics* **8:** 214.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5:** 435-445.

Feder, J., A. Gnirke, W. Thomas, Z. Tsuchihashi, D. Ruddy, A. Basava, F. Dormishian, R. Domingo, M. Ellis, and A. Fullan. 1996. A novel MHC class I–like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics* **13:** 399-408.

Feuk, L., C. Marshall, R. Wintle, and S. Scherer. 2006. Structural variants: changing the landscape of chromosomes and design of disease studies. *Human Molecular Genetics* **15:** R57.

Frisse, L., R. Hudson, A. Bartoszewicz, J. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene Conversion and Different Population Histories May Explain the Contrast between Polymorphism and Linkage Disequilibrium Levels. *The American Journal of Human Genetics* **69:** 831-843.

Fujita, R., A. Hanauer, G. Sirugo, R. Heilig, and J. Mandel. 1990. Additional Polymorphisms at Marker Loci D9S5 and D9S15 Generate Extended Haplotypes in Linkage Disequilibrium with Friedreich Ataxia. *Proceedings of the National Academy of Sciences* **87:** 1796-1800.

Fullerton, S., A. Carvalho, A. Clark, and O. Journals. 2001. Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Molecular Biology and Evolution* **18:** 1139-1142.

Gabriel, S., S. Schaffner, H. Nguyen, J. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, and M. Faggart. 2002. The Structure of Haplotype Blocks in the Human Genome. In *Science*, pp. 2225-2229.

Gambano, G., F. Anglani, and A. D'Angelo. 2000. Association studies of genetic polymorphisms and complex disease. *The Lancet* **355:** 308-311.

Gibbs, R., J. Belmont, P. Hardenbol, T. Willis, F. Yu, H. Yang, L. Ch'ang, W. Huang, B. Liu, and Y. Shen. 2003. The International HapMap Project. *Nature* **426:** 789-796.

Goddard, K., P. Hopkins, J. Hall, and J. Witte. 2000. Linkage Disequilibrium and Allele-Frequency Distributions for 114 Single-Nucleotide Polymorphisms in Five Populations. *The American Journal of Human Genetics* **66:** 216-234.

Goldstein, D. 2001. Islands of linkage disequilibrium. *Nat Genet* **29:** 109-111.

Gonzalez-Neira, A., X. Ke, O. Lao, F. Calafell, A. Navarro, D. Comas, H. Cann, S. Bumpstead, J. Ghori, and S. Hunt. 2006. The portability of tagSNPs across populations: A worldwide survey. *Genome Research* **16:** 323-330.

Grodzicker, T., J. Williams, P. Sharp, and J. Sambrook. 1975. Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harb Symp Quant Biol* **39:** 439-446.

Guo, S. 1997. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Hum Hered* **47:** 301-314.

Hattersley, A. and M. McCarthy. 2005. What makes a good genetic association study? *The Lancet* **366:** 1315-1323.

Herbert, A., N. Gerry, M. McQueen, I. Heid, A. Pfeufer, T. Illig, H. Wichmann, T. Meitinger, D. Hunter, and F. Hu. 2006. A Common Genetic Variant Is Associated with Adult and Childhood Obesity. In *Science*, pp. 279-283. American Association for the Advancement of Science.

Hey, J. 2004. What's So Hot about Recombination Hotspots? *PLoS Biology* **2:** 0730–0733.

Hill, W. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics* **38:** 226-231.

Hinds, D., L. Stuve, G. Nilsen, E. Halperin, E. Eskin, D. Ballinger, K. Frazer, and D. Cox. 2005. Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science* **307:** 1072-1079.

Hirschhorn, J. and M. Daly. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6:** 95-108.

Hudson, R. and N. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111:** 147-164.

Jeffreys, A., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29:** 217-222.

Jennings, H.S. 1917. The Numerical Results of Diverse Systems of Breeding, with Respect to Two Pairs of Characters, Linked or Independent, with Special Relation to the Effects of Linkage. *Genetics* **2:** 97-154.

Johnson, G., L. Esposito, B. Barratt, A. Smith, J. Heward, G. Di Genova, H. Ueda, H. Cordell, I. Eaves, and F. Dudbridge. 2001. Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29:** 233-237.

Jorde, L. 2000. Linkage Disequilibrium and the Search for Complex Disease Genes. *Genome Research* **10:** 1435-1444.

Kaplan, N. and B. Weir. 1992. Expected behavior of conditional linkage disequilibrium. *Am J Hum Genet* **51:** 333-343.

Kennedy, G., H. Matsuzaki, S. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, and J. Zhang. 2003. Large-scale genotyping of complex DNA. *Nature Biotechnology* **21:** 1233-1237.

Kerem, B., J. Rommens, J. Buchanan, D. Markiewicz, T. Cox, A. Chakravarti, M. Buchwald, and L. Tsui. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245:** 1073.

Komura, D., F. Shen, S. Ishikawa, K. Fitch, W. Chen, J. Zhang, G. Liu, S. Ihara, H. Nakamura, and M. Hurles. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research* **16:** 1575.

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

Kurg, A., N. Tonisson, I. Georgiou, J. Shumaker, J. Tollett, and A. Metspalu. 2000. Arrayed Primer Extension: Solid-Phase Four-Color DNA Resequencing and Mutation Detection Technology. *Genetic Testing* **4:** 1-7.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G.

Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lewontin, R. 1988. On Measures of Gametic Disequilibrium. *Genetics* **120:** 849-852.

Lewontin, R.C. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49:** 49-67.

Levy, S., G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov, Y. Lin, J.R. Macdonald, A.W. Pang, M. Shago, T.B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S.A. Kravitz, D.A. Busam, K.Y. Beeson, T.C. McIntosh, K.A. Remington, J.F. Abril, J. Gill, J. Borman, Y.H. Rogers, M.E. Frazier, S.W. Scherer, R.L. Strausberg, and J.C. Venter. 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* **5:** e254.

Lichten, M. and A. Goldman. 1995. Meiotic recombination hotspots. *Annu Rev Genet* **29:** 423-444.

Macgregor, S. 2007. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *European Journal of Human Genetics* **15:** 501-504.

Mein, C., B. Barratt, M. Dunn, T. Siegmund, A. Smith, L. Esposito, S. Nutland, H. Stevens, A. Wilson, and M. Phillips. 2000. Evaluation of Single Nucleotide Polymorphism Typing with Invader on PCR Amplicons and Its Automation. In *Genome Research*, pp. 330-343. Cold Spring Harbor Lab.

Moffatt, M., J. Traherne, G. Abecasis, and W. Cookson. 2000. Single nucleotide polymorphism and linkage disequilibrium within the TCR a/d locus. *Human Molecular Genetics* **9:** 1011-1019.

Nakamura, Y., K. Koyama, and M. Matsushima. 1998. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *Journal of Human Genetics* **43:** 149-152.

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, and E. Kumlin. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235:** 1616.

Nicolae, D., X. Wu, K. Miyake, and N. Cox. 2006. GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics* **22:** 1942.

Nordborg, M., J. Borevitz, J. Bergelson, C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. Maloof, T. Noyes, and P. Oefner. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics* **30:** 190-193.

Näslund, K., P. Saetre, J. von Salomé, T. Bergström, N. Jareborg, and E. Jazin. 2005. Genome-wide prediction of human VNTRs. *Genomics* **85:** 24-35.

Ott, J. 1999. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press.

Patil, N., A. Berno, D. Hinds, W. Barrett, J. Doshi, C. Hacker, C. Kautzer, D. Lee, C. Marjoribanks, and D. McDonough. 2001. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. In *Science*, pp. 1719-1723.

Pe'er, I., P. de Bakker, J. Maller, R. Yelensky, D. Altshuler, and M. Daly. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics* **38:** 663-667.

Peiffer, D., J. Le, F. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. Shaw, and J. Belmont. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research* **16:** 1136.

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

Petes, T. 2001. Meiotic recombination hot spots and cold spots. *Nat Rev Genet* **2:** 360-369.

Petrukhin, K., S. Fischer, M. Pirastu, R. Tanzi, I. Chernov, M. Devoto, L. Brzustowicz, E. Cayanis, E. Vitale, and J. Russo. 1993. Mapping, cloning and genetic characterization of the region containing the Wilson disease gene. *Nature Genetics* **5:** 338-343.

Pfaff, C., E. Parra, C. Bonilla, K. Hiester, P. McKeigue, M. Kamboh, R. Hutchinson, R. Ferrell, E. Boerwinkle, and M. Shriver. 2001. Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. *The American Journal of Human Genetics* **68:** 198-207.

Plagnol, V. and J. Wall. 2006. Possible Ancestral Structure in Human Populations. *PLoS Genet* **2:** 972–979.

Pritchard, J. and M. Przeworski. 2001. Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics* **69:** 1-14.

Przeworski, M. 2002. The Signature of Positive Selection at Randomly Chosen Loci. *Genetics* **160:** 1179-1189.

Przeworski, M. and J. Wall. 2001. Why is there so little intragenic linkage disequilibrium in humans? *Genet Res* **77:** 143-151.

Purcell, S., S. Cherny, and P. Sham. 2003. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19:** 149-150.

Reich, D., M. Cargill, S. Bolk, J. Ireland, P. Sabeti, D. Richter, T. Lavery, R. Kouyoumjian, S. Farhadian, and R. Ward. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Risch, N. and K. Merikangas. 1996. The Future of Genetic Studies of Complex Human Diseases. *Science* **273:** 1516.

Risch, N., H. Tang, H. Katzenstein, and J. Ekstein. 2003. Geographic Distribution of Disease Mutations in the Ashkenazi Jewish Population Supports Genetic Drift over Selection. *The American Journal of Human Genetics* **72:** 812-822.

Sabeti, P., D. Reich, J. Higgins, H. Levine, D. Richter, S. Schaffner, S. Gabriel, J. Platko, N. Patterson, and G. McDonald. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832-837.

Saunders, M., M. Slatkin, C. Garner, M. Hammer, and M. Nachman. 2005. The Extent of Linkage Disequilibrium Caused by Selection on G6PD in Humans. *Genetics* **171:** 1219-1229.

Sham, P., J. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA Pooling: a tool for large-scale association studies. *Nature Reviews Genetics* **3:** 862-871.

Skol, A., L. Scott, G. Abecasis, and M. Boehnke. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38:** 209-213.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, and S. Hadjadj. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445:** 881-885.

Smyth, D., J. Cooper, R. Bailey, S. Field, O. Burren, L. Smink, C. Guja, C. Ionescu-Tirgoviste, B. Widmer, and D. Dunger. 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature genetics* **38:** 617-619.

Steer, S., V. Abkevich, A. Gutin, H. Cordell, K. Gendall, M. Merriman, R. Rodger, K. Rowley, P. Chapman, and P. Gow. 2007. Genomic DNA pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis. *Genes and Immunity* **8:** 57-68.

Zhang, K., M. Deng, T. Chen, M. Waterman, and F. Sun. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences* **99:** 7335.

Zhu, X., A. Luke, R. Cooper, T. Quertermous, C. Hanis, T. Mosley, C. Charles Gu, H. Tang, D. Rao, and N. Risch. 2005. Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics* **37:** 177-181.

Toth, G., Z. Gaspari, and J. Jurka. 2000. Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. In *Genome Research*, pp. 967-981. Cold Spring Harbor Lab.

Wang, D., J. Fan, C. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, and J. Spencer. 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280:** 1077-1082.

Wang, N., J. Akey, K. Zhang, R. Chakraborty, and L. Jin. 2002. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. *The American Journal of Human Genetics* **71:** 1227-1234.

Weber, J. 1990. Human DNA polymorphisms and methods of analysis. *Curr Opin Biotechnol* **1:** 166-171.

Willer, C., L. Scott, L. Bonnycastle, A. Jackson, P. Chines, R. Pruim, C. Bark, Y. Tsai, E. Pugh, and K. Doheny. 2006. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol* **30:** 180–190.

Wiuf, C. and J. Hein. 2000. The Coalescent With Gene Conversion. *Genetics* **155:** 451-462.

Wood, E., D. Stover, M. Slatkin, M. Nachman, and M. Hammer. 2005. The ß-Globin Recombinational Hotspot Reduces the Effects of Strong Selection around HbC, a Recently Arisen Mutation Providing Resistance to Malaria. *Am J Hum Genet* **77:** 637-642.

Summary in Estonian

Geneetiliste haiguste kaardistamiseks kasutatav assotsiastioonianalüüs põhineb kromosoomil lähestikku paiknevate lookuste vahelisel aheldatusel. Haigust põhjustavat lookust on võimalik leida genotüpiseerides seda ümbritsevaid markerlookusi. Haigete indiviidide ja kontrollgrupi markeri alleelisageduste erinevus viitab sellele, et markerlookus ise põhjustab haigust või on markerlookus tugevalt korreleeritud haigust põhjustava lookusega.

Markerlookuste valikul tuleb silmas pidada seda, et võimalikult suur hulk potensiaalselt haigusega seotud genoomi varieeruvusi saaks nendega kirjeldatud. Kuivõrd lookuste vaheline aheldatuse ulatus ja tugevus on inimese genoomis piirkonniti erinev, on optimaalseks markerite valikuks vaja uuritava populatsiooni markerite vahelise aheldatuse kaardistamine.

Antud töö teoreetiline osa tutvustab olemasolevaid geneetilisi markereid, nende vahelist aheldatust (LD) ning aheldatusel põhinevat assotsiatsioonianalüüsi. Ülevaade on antud ka populaarsematest marker valiku metoodikatest ning olemasolevatest ülegenoomsetest markeripaneelidest.

Töö praktilises osas on uuritud alleelse aheldatuse (LD) varieeruvust inimese 22. kromosoomi andmete põhjal, tagSNPde valiku metoodikaid ning ülegenoomsete SNP markeripaneelide sobivust eestlaste populatsiooni varieeruvuse kirjeldamiseks.

Uurides 22. kromosoomi alleelset aheldatust mitmes Euroopa päritolu populatsioonis (Eesti, Suurbritannia, CEPH) leidsime, et LD ulatus ei sõltu vaid füüsilisest lookuste vahelisest distantsist vaid on piirkonniti väga erinev. Leidus pikki kromosoomi lõike, mis olid omavahel tugevalt assotseerunud. Sellised lõigud olid üksteisest eraldatud piirkondadega, kus aheldatus oli väga nõrk, andes tunnistust sellest, et seal toimuvad väga tihedalt rekombinatsioonid. Samal ajal Euroopa populatsioone võrreldes leidsime, et LD muster on erinevates populatsioonides küllaltki sarnane. See teeb võimalikuks luua üldiseid LD kaarte mis kirjeldaks kõiki Eurropa populatsioone. Teatavaid nihkeid tugeva aheldatusega regioonide piirides siiski esineb.

Rahvusvahelise HapMap projekti raames kogutud CEPHi populatsiooni valimi põhjal arvutatud markerid (tagSNPd) on hästi kirjeldanud sagedasemaid genoomi varieeruvusi paljudes analüüsitud Eurropa populatsioonides. TagSNPe, mis on arvutatud vastavalt HapMap-i teise faasi andmetele, saab kasutatda teiste populatsioonide kirjeldamiseks kaotamata oluliselt hilisema uuringu statistilises võimsuses. Oma töös leidsime, et tagSNPd kirjeldavad paremini suurema minoorse alleelisagedusega markereid. Haruldasemate polümorfismide kirjeldamiseks ei pruugi tagSNPdel põhined lähenemine olla edukas.

Senised uuringud on näidanud, et tänapäevased ülegenoomi SNP paneelid kirjeldavad edukalt HapMap-i projekti populatsioone. Ka on uuritud seda, kui sarnase alleelse aheldatusega on HapMap projekti populatsioonid teiste sama regiooni populatsioonidega. Oma töös uurisime lisaks, et kas ülegenoomsed SNP paneelid kirjeldavad sarnaselt HapMap populatsioonidele hästi ka eestlaste populatsiooni. Leidsime, et kommerstiaalsed SNP paneelid kirjeldavad Eesti populatsiooni sama efektiivselt kui HapMap projekti raames genotüpiseeritud CEPH populatsiooni valimit.

Acknowledgements

I would like to thank all the co-authors of the papers which this dissertation is based on. I am also thankful to all the people who have helped and taught me during my studies.

I am most grateful to my supervisor, Maido Remm, for providing an interesting topic of study, teaching me bioinformatics, and for being supportive and encouraging during my PhD studies.

I would like to thank my first supervisor, Dr. Kristjan Zobel, for teaching me statistics and study design in the Department of Plant Ecology.

I am thankful to Prof. Thomas F. Wienker for teaching me so much about genetics and for providing me with the great opportunity of working in the Institute for Medical Biometry, Informatics and Epidemiology during my visit at the Rheinische Friedrich-Wilhelms-University in Bonn.

It has been a great pleasure to work with Dr. Elin Org from the Department of Biotechnology during all those years. Thank you for all of the collaboration and support.

I would like to thank Jody Novakoski for the language correction on so many occasions and also for teaching me some basics of English.

I wish to thank all of my present and former colleagues from the Department of Bioinformatics, Asper Biotech, Biodata, and the Department of Plant Ecology for providing your help and a great work atmosphere.

Last but not least I would like to thank my parents and all of my family members and friends for their love, patience and the educational discussions we have had. Especially Merike.

Publications

Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M. (2007). Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. BMC Genomics, 8(1):159.

Kuningas, M.; **Mägi, R.**; Westendorp, R.G.; Slagboom, P.E.; Remm, M.; van Heemst, D. (2007). Haplotypes in the human Foxo1a and Foxo3a genes; impact on disease and mortality at old age. European Journal of Human Genetics, 15(3), 294 - 301.

Kaminski, S.; Brym, P.; Rusc, A.; Wojcik, E.; Ahman, A.; **Mägi, R**. (2006). Associations between milk performance traits in Holstein cows and 16 candidate SNPs identified by arrayed primer extension (APEX) microarray. Animal Biotechnology, 17(1), 1 - 11.

Montpetit, A.; Nelis, M.; Laflamme, P.; **Magi, R.**; Ke, X.Y.; Remm, M.; Cardon, L.; Hudson, T.J.; Metspalu, A. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population . PLoS Genetics, 2(3), 282 - 290.

Mägi, R.; Kaplinski, L.; Remm, M. (2006). The Whole Genome TagSNP Selection and Transferability Among HapMap Populations. *In: Pacific Symposium on Biocomputing 2006: 2006.* World Scientific Publ Co Pte Ltd, 2006, 535 - 543.

Mueller, J. C.; Lõhmussaar, E.; **Mägi, R.**; Remm, M.; Bettecken, T.; Lichtner, P.; Huber, S.; Illig, T; Luedemann, J.; Schreiber, S.; Wichmann, H. E.; Pramstaller, P.; Romeo, G.; Testa, A.; Metspalu, A.; Meitinger, T. (2005). Linkage Disequilibrium Patterns and tagSNP Transferability among European Populations. American Journal of Human Genetics, 76(3), 387 - 398.

Liira, J.; Zobel, K.; **Magi, R.**; Molenberghs, G. (2002). Vertical structure of herbaceous canopies: the importance of plant growth-form and species-specific traits. Plant Ecology, 163(1), 123 - 134.

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.; Lohmmussaar, E.; Zernant, J.; Tonisson, N.; Remm, M.; **Magi, R.**; Puurand, T.; Vilo, J.; Kurg, A.; Rice, K.; Deloukas, P.; Mott, R.; Metspalu, A.; Bentley, D.R.; Cardon, L.R.; Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature, 418(6897), 544 - 548.

Tables

Table 1. Sample allele frequencies in a 2x2 table for two loci. Locus 1 has alleles A and a, locus 2 has alleles B and b.

Locus 2 \ Locus 1	A	a	
B	P_{AB}	P_{aB}	P_B
b	P_{Ab}	P_{ab}	P_b
	P_A	P_a	1

Table 2. General information of used datasets.

study ID	region	region size	Studied populations and sample size	# of studied SNPs
Ref I	chromosome 22q	33.4 Mb	Estonia (51), CEPH (77), UK (90)	594
Ref II	FKBP5, SNCA, LMNA and PLAU gene regions	749 kb	CEPH (30 trios), Estonia (170), POPGEN (160), SHIP (100), KORA (170), VIN (170), LAD (160), BRISI (98), CALA (100)	169
Ref III, IV	encode regions ENr112 and ENr131	1 Mb	CEPH (30 trios), Estonia (1,090)	1420
HapMap	Whole human genome	~3 Gb	CEPH (30 trios), YRI (30 trios), JPT (44), CHB (45)	~3.8 million

FKBP5 – FK-506 binding protein 5; SNCA – synuclein α, LMNA – lamin A/C, PLAU – plasminogen activator, urinary; CEPH – Centre d'Etude du Polymorphisme Humain, UK – United Kingdom, POPGEN – population samples collected from Schleswig-Holstein; SHIP – Study of Health in Pomerania, KORA – Cooperative health research in the region of Augsburg; VIN – Vinchgau; LAD – Ladinia; BRISI – Brisighella; CALA – Calabria; YRI – Yoruba in Ibadan, Nigeria; JPT - Japanese in Tokyo, Japan; CHB – Han Chinese in Beijing, China.

Figure captions

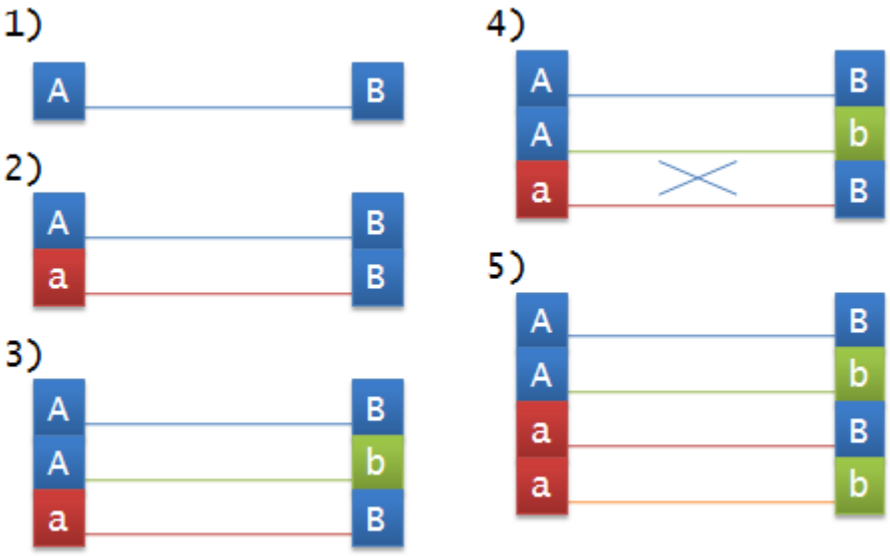

Figure 1. The formation of new haplotypes through mutation and the erosion of LD through recombination. 1) We have two monomorphic loci, both with single allele: A and B accordingly. Between these two loci only one haplotype (AB) exists. 2) In first locus mutation occurs creating a new allele a. Together with allele B in second locus, new haplotype evolves (aB). 3) In second locus also a mutation occurs (b) on initial AB haplotype creating a new, third haplotype (Ab). 4) Recombination event between haplotypes Ab and aB occurs 5) This event creates fourth possible haplotype (ab).