DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
176

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
176

MARI NELIS

Genetic structure of the Estonian population and genetic distance from other populations of European descent

TARTU UNIVERSITY
PRESS

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

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in molecular diagnostics) on 02.03.2010 by the Council of the Institute of Molecular and Cell Biology, University of Tartu

Supervisor: Prof. Andres Metspalu, M.D., Ph.D.
Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Estonia

Opponent: Dr. Markus Perola, M.D., Ph.D.
Public Health Genomics, National Institute for Health and Welfare, Finland;
Department of Medical Genetics, University of Helsinki, Finland

Commencement: Room No 217, 23 Riia Str., Tartu, on April 9th 2010, at 10.00.

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

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

ISSN 1024–6479
ISBN 978–9949–19–327–1 (trükis)
ISBN 978–9949–19–328–8 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 113

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

Ref.I Montpetit A*, **Nelis M***, Laflamme P, Mägi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A. 2006. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet 2(3):e27.

Ref.II 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 8:159.

Ref.III **Nelis M***, Esko T*, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskáčková T, Balaščák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann H-E, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kučinskas V, Kasnauskienė J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. 2009. Genetic Structure of Europeans: a view from the North-East. PLoS ONE 4(5):e5472.

* These authors contributed equally to this work.

Author's contributions:

Ref. I, III participated in study design, performed the experiments, analyzed the data, participated in preparation and writing of the paper

Ref. II performed the experiments, participated in study design and writing of the paper

LIST OF ABBREVIATIONS

bp	base pair
CHB	Han Chinese from Beijing
CEU	Utah residents with ancestry from Northern and Western Europe
ENCODE	encyclopedia of DNA elements
GWAS	genome-wide association study
JPT	Japanese from Tokyo
kb	kilobase
LD	linkage disequilibrium
MAF	minor allele frequency
OR	odds ratio
PC	principal component
PCA	principal component analysis
RFLP	restriction fragment length polymorphism
SNP	single nucleotide polymorphism
tagSNP	tagging SNP
YRI	the Yoruba people of Ibadan, Nigeria

INTRODUCTION

The availability of human DNA sequence and variations, together with advances in new technology, have enabled the detailed analysis of associations between genetic markers and phenotypes. More than a thousand genes have been identified that cause rare Mendelian disorders – in which a mutation has high penetrance, and a defect in a single gene is necessary and sufficient to cause disease. But most common diseases are caused by a combination of variations at many genetic loci that are not fully penetrant and are affected by environment and lifestyle. Phenotypic and genetic heterogeneity may also complicate the elucidation of the true cause of a disease. Understanding the genetic variation between different populations is a prerequisite to study the genetic component of common complex diseases, like type II diabetes, cardiovascular diseases, hypertension, cancer, etc., which is the topic of my thesis.

With the development of microarray technology and bioinformatics, the genome of thousands of individuals can be studied on a large scale, in a single study. Association studies were previously performed on a small scale, examining only markers from candidate genes. However, as whole-genome studies become feasible, hundreds of new single nucleotide polymorphisms (SNPs) associated with specific diseases, have been discovered. Current technology enables the genotyping of over one million markers from a single genome. As commercial chips become widely available, SNP panels must be evaluated so the correct arrays for particular studies and populations are selected.

An important preliminary task for association studies is the selection of cases and controls. Since the samples from one population may be insufficient for the required number of cases and controls, there is an opportunity to combine the studied individuals from various populations. However, the allele frequencies can vary widely throughout the world. In addition, confounding factors of association studies, such as heterogeneity in alleles, populations and phenotypes, may lead to false positives or negatives, and must be considered carefully before data analysis. Therefore, prior characterization of populations is required so that the most relevant ones can be combined without loss of study power.

This thesis gives an overview of some factors that should be considered in designing whole-genome genotype studies. The first part of the thesis provides an overview of the structure of the human genome, specifically the linkage disequilibrium patterns and tagging SNP (tagSNP) selection. Then the selection of an appropriate SNP panel and study subjects for genome-wide association studies is reviewed. Correlation between the genetic and geographic distances of human populations, as discussed in recent publications, is surveyed. Finally, the research section of this thesis concentrates on three issues: 1) transferability of tagSNPs selected from HapMap data to the Estonian population; 2) evaluation of the coverage of the Estonian population on commercial chip panels; and 3) characterization of the genetic structure of the Estonian population and comparison with other European populations.

1. REVIEW OF LITERATURE

1.1 Linkage disequilibrium in the human genome

1.1.1 Genetic markers

The first nucleotide sequence-based polymorphisms that achieved popularity for studying genetic diversity in the human genome were restriction fragment length polymorphisms (RFLPs), in the 1980s (Jeffreys, 1979; Chakravarti et al., 1984). The development of technologies such as polymerase chain reaction allowed the study of different types of genetic markers, like microsatellites (STRs – short tandem repeats), minisatellites (VNTRs – variable number of tandem repeats), short insertions and deletions, and finally, the most common polymorphic genetic markers – single nucleotide polymorphisms (SNPs). In addition, copy number variations (CNVs), which are structural variations greater than one kilobase in size, have increasingly been researched recently, since they account for most bases that vary among human genomes, and have been associated with many human diseases, schizophrenia (Walsh et al., 2008), Crohn's disease, psoriasis (Buchanan and Scherer, 2008; Gu et al., 2008; Conrad et al., 2009).

Currently, the most commonly used DNA variations in association studies are SNPs. Since SNPs are usually binary, they are well-suited to automated, high-throughput genotyping (Wang et al., 1998). Their frequency is approximately 1 per 300 base pairs, but their density varies up to ten-fold in different regions of the genome (Sachidanandam et al., 2001). In addition, the mutation rate of SNPs is lower than STRs, making them a good indicators for studying genetic variation in the human genome (Jorde, 2000).

The phenotypic effect of a SNP depends on its location in the gene. SNPs in the coding region may affect the protein structure and function, and tend to cause genetic diseases with dominant or recessive inheritance patterns. SNPs can alter the structure of proteins involved in drug metabolism, and are therefore often a direct target of pharmacogenetics (Roses, 2000). The majority of SNPs are located in the non-coding areas of genes, and probably have an impact on the regulation of gene expression. In addition, SNPs in introns and intergenic regions often appear in association studies of complex diseases (WTCCC, 2007). These SNPs may be in linkage disequilibrium (LD) with the DNA variations associated with the disease, or may influence expression of nearby genes.

The selection of SNPs for an association study usually depends on the nature of the study, and may follow one of two strategies:

1. Direct strategy – SNPs selected for the association study are putative causal variants. This type of study is easy to analyze, but the selection of candidate polymorphisms is difficult. Non-synonymous SNPs in the coding region of a gene lead to an amino acid change and are obvious candidates for causal

variants. Although most variants that alter gene regulation and expression are in non-coding regions in the genome and the selection of causal variant is more difficult. However, the direct strategy of SNP selection has the potential to discover the primary genetic cause of a disease.

2. Indirect strategy – selected SNPs are a surrogate for the causal locus. Here the disease-causing locus has been localized by linkage between two polymorphisms, so the occurrence of a particular SNP predicts the presence of the second (disease-causing)SNP (Kruglyak, 1999). There is a non-random, regular association between the alleles of the disease-causing mutation and the studied polymorphism i.e. linkage disequilibrium, or LD (Reich et al., 2001). The number of studied markers depends on the extent of LD observed in a particular population and genomic region. The extent of observed LD can vary widely throughout the genome, on scale of 1–100 kb, and occasionally extending up to hundreds of kilobases (Dawson et al., 2002; Gabriel et al., 2002). Thus, the number of markers needed for an association study ranges from 500,000 to 1 million, as this level of detection is now possible with current whole-genome genotyping platforms. However, this method is limited by variation in the extent of LD in the genome, and the fact that 0.5–1% of all high-frequency SNPs are untaggable, meaning that no other proxy SNPs occur within 100 kb (Frazer et al., 2007).

1.1.2 Linkage disequilibrium

The appearance of new mutations in a DNA sequence is infrequent, and mutations usually do not arise at the same site as previous mutations. Therefore, a SNP can be used as a marker to provide information about the presence of nearby variants. The non-random association of two polymorphisms at different loci is the basis of linkage disequilibrium. LD is affected by many factors, including new mutations, gene conversion or recombination events that lead to formation of new haplotypes (Weiss and Clark, 2002), which are particular combinations of alleles observed in a population. In principle, the number of haplotypes depends on the number of polymorphisms in the region, for example, three SNPs give rise to 2^3 different haplotypes, from which 3–5 haplotypes are usually found to be more frequent in the population than others.

1.1.2.1 Recombination

Recombination rates vary widely across the genome, and are a major determinant of LD. Regions with a high rate of recombination events are usually referred as recombination “hotspots”, and the regions with low recombination as “coldspots”. Typically, 80% of the recombination occurs within 10 to 20% of the genome. Recombination rates generally tend to be higher in telomeric

regions than in centromeric regions. The human genome has approximately 25,000–50,000 recombination hotspots, which is comparable to the total number of human genes. Recombination hotspots in human genomes preferentially occur within 50 kb of genes, but are usually located outside the transcribed domain (Myers et al., 2005).

1.1.2.2 Mutation rates

The average mutation rate per nucleotide site ranges from 1.3 x 10^{-8} and 2.7 x 10^{-8}, assuming a human generation time of 20 years. The human diploid genome contains 7 x 10^{9} bp (Marshall, 1999), which leads to an estimate of 175 new mutations per generation (range 91–238). The accuracy of the estimated mutation rate depends more on uncertainties in divergence time, ancestral population size, and generation time, than on estimates of molecular substitution rates, which have standard errors of approximately one-tenth of the mean values (Nachman and Crowell, 2000).

1.1.2.3 Measures of LD

The extent of LD can be characterized by two common pair-wise measures, D´ and r^2. D´ is defined as 1 in the absence of obligate recombination, and declines only because of recombination or recurrent mutation. In contrast, r^2 is the square of the correlation coefficient between two SNPs. If two SNPs are independent of each other, r^2=0. In the case of perfect LD, the allele frequencies of the two SNPs are the same, and r^2=1.

1.1.3 Haplotype blocks

In 2001, several groups described the block-like structure of human LD patterns (Daly et al., 2001; Patil et al., 2001; Jeffreys et al., 2004). Haplotype blocks are sizable regions which show little evidence of historical recombination, and which contain only a few common haplotypes (Gabriel et al., 2002). The idea raised that since these common haplotypes capture most of the genetic variation across a sizable region, these haplotypes and the undiscovered variants they contain, can be tested using a small number of SNPs, called tagSNPs. Furthermore, characterization of haplotype blocks could be the foundation for constructing a haplotype map of the human genome, which would facilitate comprehensive genetic association studies of human disease.

1.1.3.1 HapMap project

In 2002, a project called "The Haplotype Map of the Human Genome" (www.hapmap.org) was launched by the International HapMap Consortium. The main goal was to describe variation patterns in four populations in a way that would help researchers find complex disease genes by analyzing tagSNPs. The goal was to reduce the number of SNPs that needed to be tested in any given association study. A fine-scale genetic map of the human genome is a major requirement for designing and analyzing association mapping experiments.

The phase I HapMap showed variation patterns for four populations: 30 parent-offspring trios representing residents of Utah in the United States with ancestry from Northern and Western Europe (CEU), 30 trios from Yoruba in Ibadan, Nigeria (YRI), 45 Han Chinese in Beijing, China (CHB), and 44 Japanese in Tokyo, Japan (JPT). As the allele frequencies between CHB and JPT are generally similar, these populations are usually analyzed together. SNPs were selected at 5 kb intervals across the genome, with the requirement that the minor allele frequency (MAF) be ≥ 0.05, which defines a "common" SNP. Approximately 1.3 million SNPs were genotyped in phase I of the project (TheInternationalHapMapConsortium, 2005).

In phase II of the HapMap project, a further 2.1 million SNPs were genotyped for the same set of individuals. The resulting marker map had a SNP density of approximately one per kilobase. The phase II HapMap differs from the phase I HapMap not only in SNP density, but also in the frequency distribution of minor alleles, and in LD patterns. In phase II, the marker selection criteria did not include a requirement for only common SNPs, so this HapMap contains more low frequency SNPs, with a better representation of rare SNPs (Frazer et al., 2007).

The SNPs in the HapMap database were tested only in four populations, so questions have been raised about the transferability of its information to other populations. Several studies have addressed it in lower scale. Thus, the phase III HapMap has data for approximately 1.5 million genetic markers for 1115 individuals from 11 populations: the initial HapMap samples, and samples from seven additional populations (Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; people with African ancestry in the south-western U.S.; Gujarati Indians in Houston, Texas, U.S.; metropolitan Chinese in Denver, Colorado, U.S.; people of Mexican origin in Los Angeles, California, U.S.; and Tuscans in Italy).

Extending the catalog to include rare variants will require whole-genome sequencing of much larger samples. In 2008, the 1000 Genomes Project (www.1000genomes.org) was launched to sequence the genomes of approximately 1200 people from around the world. As of May 26, 2009, the first set of SNPs representing the preliminary analysis of four genome sequences is available for download.

1.1.3.2 ENCODE project

In September 2003, a public research consortium named ENCODE, for encyclopedia of DNA elements, (http://www.genome.gov/10005107) was launched by the U.S. National Human Genome Research Institute (NHGRI) to identify all functional elements in the human genome. The studied sequence corresponds to 30 megabases (Mb), or roughly percent of the total human genome (Birney et al., 2007). Ten of these regions were selected for HapMap project to compare the genome-wide resource with more complete database of common and rare variants. Each 500 kb region was sequenced for 48 individuals, and all SNPs discovered in these regions were genotyped in 269 HapMap samples. The regions were selected from different chromosomes and differed by gene density (0–5.9%), and recombination rate (0.5–2.6 cM/Mb) (TheInternationalHapMapConsortium, 2005).

1.1.4 Selection of tagSNPs

The HapMap project has created a significant resource for LD-based marker selection for genome-wide association studies (GWAS). Correlation among nearby variants (i.e. LD) enables the selection of informative tagSNPs that act as proxies for nearby variants. TagSNPs can be used to capture the vast majority of SNP variation in a region, thereby substantially reducing the cost of genotyping (Johnson et al., 2001).

Several strategies exist for tagSNP selection. One possibility is to use the greedy search algorithm, which selects the minimum number of tagSNPs necessary to monitor the remaining non-tagSNPs above a threshold level of correlation, usually set as $r^2>0.8$ (Carlson et al., 2004). TagSNP are selected so the SNP is directly assayed or exceed a threshold level of LD ($r^2>0.8$) with the assayed SNP. Another method is to select the tagSNPs from the pre-defined haplotype blocks (Gabriel et al., 2002). The limitation of this method is that it requires full knowledge of recombination events in the region. In addition, some SNPs may be associated with only one haplotype, while others may represent clades of related haplotypes. Also, block boundaries are not always consistent within or between populations (Crawford et al., 2004; TheInternationalHap MapConsortium, 2005). Thirdly, unlike haplotype blocks, which are defined as contiguous groups of SNPs, the SNPs may make up a bin that can be interdigitated with SNPs that are part of other bins (Hinds et al., 2005).

1.2 Concept of the genome-wide association study

Genetic disorders can be clustered into two classes, monogenetic and complex disorders. Monogenetic disorders are a direct consequence of a single defect in a locus, and linkage mapping is commonly used to identify the probable locus for these disorders (Kruglyak, 1999). Monogenetic disorders show Mendelian inheritance, so large pedigrees are the basis for studying these diseases by comparing the frequency of studied markers between disease carrying, and non-carrying family members. A difference in the frequency of a studied marker allele between the two groups indicates the likelihood of association with the disease (Risch and Merikangas, 1996).

Many genes may contribute to a disease, however, and a mutation at a single locus may have different phenotypic expressions. For example, the solution to cystic fibrosis was expected to follow the identification of the cystic fibrosis transmembrane conductance regulator (CFTR) gene in 1989 (Kerem et al., 1989). However, today we know that over 1000 mutations are found in this single gene, and many variations of the disease within the same CFTR genotype are known (Cystic Fibrosis Mutation Database, www.genet.sickkids.on.ca/cftr/app). In addition, other conditions such as infertility, diarrheal diseases and asthma have been associated with mutations in the CFTR gene.

Complex disorders such as cardiovascular disease, hypertension, diabetes or cancer are difficult to study, since many genes contribute to the disease, and the effect of any single locus is weak. Complex diseases are also studied in families, but most studies are performed between groups of unrelated disease-carrying individuals and non-carriers. Using association studies, markers can be directly associated with a disease or a disease-causing marker (Cardon and Bell, 2001). Traditionally, candidate genes were used in association studies, but technological developments have made whole-genome association studies possible. As of 2009, over 300 GWAS have been performed (www.genome.gov/gwastudies).

Study design is most challenging aspect of association studies, with many factors to consider, including the loci to genotype and the number of individuals to study.

1.2.1 Sample size and power

The individuals enrolled in an association study must be thoroughly phenotyped. Often, variations in a single population are of insufficient frequency to provide an adequate number of individuals for a study. Therefore, access to several different biobanks is crucial.

A critical question is the number of samples needed to achieve the study power of 80%? In association studies, a clear correlation is seen between the sample size and the between-marker correlation (r^2) that is required to achieve a

certain level of power (Pritchard and Przeworski, 2001). The relationship specifies that N subjects are needed to achieve a certain level of power when the disease variant is directly observed and analyzed, but the number of required subjects increases (N/r^2) when a tagSNP for the disease variant with a correlation of r^2 is analyzed. Thus, the power of allelic association studies depends primarily upon sample size, the effect of the susceptibility locus, the strength of LD with the marker, and the frequencies of susceptibility and marker alleles (Zondervan and Cardon, 2004). Figure 1 illustrates the sample size needed to detect disease variants with allelic odds ratios (OR) ranging from 1.2 to 2, at 80% power and significance level of $p<10^{-6}$, assuming a multiplicative model for the effects of alleles, and perfect correlative LD between alleles of the test markers and disease variants. The number of individuals increases as OR decreases, and additional study subjects are required to detect low frequency alleles, as complex diseases are assumed to be influenced by many genetic loci that each has a modest effect on the trait (Reich and Lander, 2001; Wang et al., 2005).

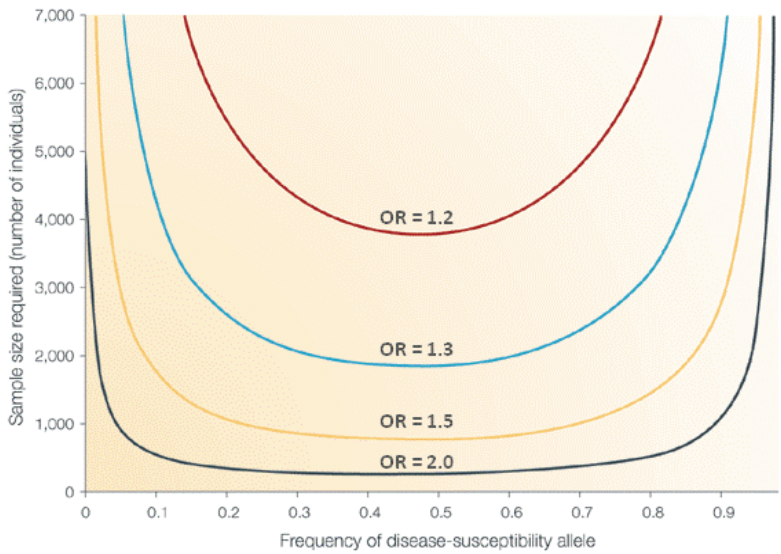

Figure 1. The numbers of cases and controls that are required in an association study to detect disease variants with different allelic odds ratios with the study power of 80% at a significance level of $p<10^{-6}$ (Wang et al., 2005).

Limitations in the design of early GWAS, such as imprecise phenotyping and the use of control groups of questionable comparability, may have affected the identification of variants associated with common diseases (Manolio et al., 2009). Furthermore, variants with small effects are not penetrant enough to show Mendelian segregation, and are therefore not detected using traditional

linkage approaches or association studies, especially when commercial SNP panels are used (Figure 2). Thus, the genetic components of many common complex diseases still remain unknown, as the "dark matter" of genetics (McCarthy and Hirschhorn, 2008). Identification of such variants is challenging, but next-generation sequencing technology should allow the study of rare variants (frequency <5%), which will extend our understanding of the genetic components of common complex diseases.

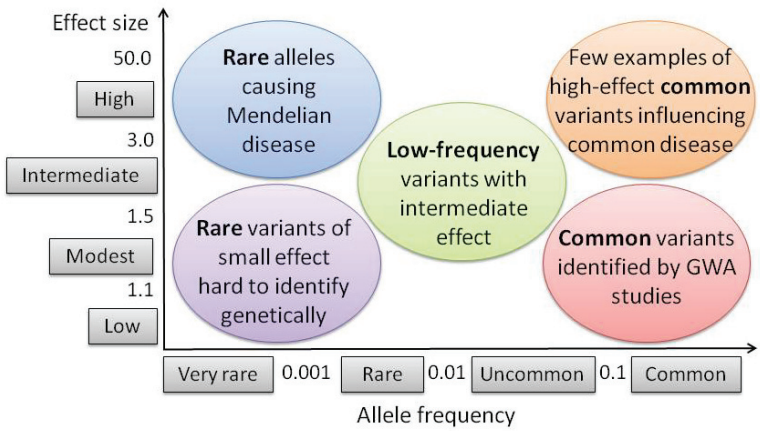

Figure 2. The allele frequency and effect of susceptibility locus that increases the risk to develop a disease (McCarthy et al., 2008).

1.2.2 Population stratification

Population stratification may affect the results of large-scale association studies, and therefore should be tested carefully before data analysis. Specifically, population stratification refers to differences in allele frequencies between cases and controls that is caused by systematic differences in ancestry, rather than by the association of genes with the disease (Figure 3) (Marchini et al., 2004). Disease prevalence often changes with geography and ethnic origin, and allele frequencies can vary widely throughout the world. Confounding factors of association studies such as allelic, population, and phenotype heterogeneity may lead to false positives or negatives and should be considered carefully before data analysis.

One biobank or sample collection is usually not large enough to contain the required sample size for a study power of 80%, so collaboration between biobanks has become common, as has meta-analysis involving large-scale analysis of data-sets genotyped at several institutions. Since the number of cases may be limited, the number of controls can be increased. Large sets of control

genotypes are publicly available, but care must be taken to assessing the appropriateness of a set of controls.

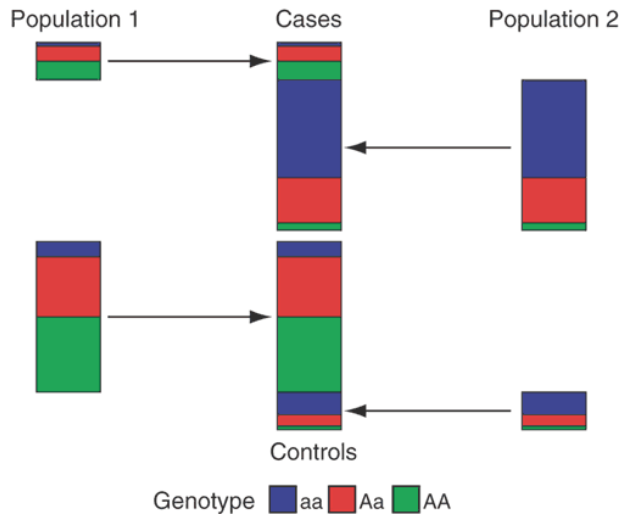

Figure 3. Population stratification in case-control association studies. In population 2, the percentage of cases is larger so the frequency of A allele is higher and could cause false positive associations in studies (Marchini et al., 2004).

1.2.3 Parameters to detect the population stratification

Population stratification can be detected using several methods. One is genotyping >300 unlinked markers that are scattered throughout the genome, are not linked to any phenotypes, and whose allele frequencies differ between populations (Freedman et al., 2004). Study subjects are considered to be correctly selected if the unlinked maker alleles are similar in cases and controls (Pritchard and Rosenberg, 1999).

In the genomic control method, the inflation factor lambda (λ) is calculated based on marker allele frequencies, and is used to reevaluate a χ^2 test (Devlin and Roeder, 1999). The value of λ is estimated as the median of the observed chi-square statistics divided by the median of the central chi-square distribution with one degree of freedom (i.e. 0.455), and is expected to be constant across the genome. Population stratification often inflates the null distribution of the test statistic, so the genomic control method applies a correction by dividing the association test chi-square values by λ. The disadvantage of this method is that it corrects for stratification using a uniform overall inflation factor to adjust association statistics at each marker. However, some markers differ more than others in their allele frequencies across ancestral populations. Thus, this

uniform adjustment may be insufficient for some markers and superfluous for others, leading to a loss in power.

A third method takes advantage of clustering. The structured association method also uses information from unlinked markers, but instead of generating an estimated scaling factor, it uses the marker information to divide the population into homogenous subpopulations (Pritchard et al., 2000a). The disease-marker association test is then performed for each subgroup, and the final results generated by combining the previously analyzed data of each subgroup (Pritchard et al., 2000b). The Bayesian clustering method (the STRUCTURE program) described by Pritchard et al. (2000a) is often used to detect genetic substructure of a population. However, analyzing genome-wide data with the STRUCTURE program is computationally intensive, as it is a Markov chain Monte Carlo (MCMC) sampling-based approach.

Lately, many studies have used principal component analysis (PCA) to detect population stratification. PCA was first applied to genetic data by Cavalli-Sforza and colleagues to infer the worldwide axes of human genetic variation from the allele frequencies of various populations (Menozzi et al., 1978). This approach is applied in the EIGENSOFT software, which focuses on individual genotype data, and assigns a statistical significance to each axis (Patterson et al., 2006; Price et al., 2006; Roeder and Luca, 2009). Most of the eigenvalues of the theoretical covariance will be "small," or nearly equal, arising from sampling noise, while only a few eigenvalues will be "large," reflecting past demographic events. The first two eigenvalues usually describe the greatest variance between individuals. In contrast to the genomic control method, only markers that vary because of population admixture or ancestry are corrected in the association study (Price et al., 2006).

Multidimensional scaling (MDS) of pair-wise identity by state (IBS) sharing data can also be used to visualize population genetic structure, and is available in PLINK software (Purcell et al., 2007). In the first phase, an IBS matrix of distance is conducted containing each pair-wise combination of all individuals, and secondly the classical MDS analysis is applied to explore the similarities in the matrix.

Finally, in addition to the described methods to detect population stratification, a measure often used in population genetics is Wright's F_{st}, which describes the proportion of total variation in allele frequency that is due to differences between populations (Wright, 1969). On the basis of HapMap data, the F_{st} is lowest between populations with European and Asian descent (F_{st}=0.07), and highest between populations with African and Asian descent (F_{st}=0.12) (TheInternationalHapMapConsortium, 2005).

1.2.4 Evaluation of the performance of commercial genotyping panels

Current whole-genome DNA chips allow the genotyping of 100,000 to 1 million SNPs from one individual. Two main companies, Illumina Inc. (www.illumina.com) and Affymetrix Inc. (www.affymetrix.com) offer DNA chips at different scales. Although both companies originally offered smaller scale chips, they have currently reached 1 million SNPs per chip, which is approximately 10% of the common SNPs with MAF≥0.05 in the human genome.

Chips are often evaluated by estimating the proportion of SNPs captured at a certain level, usually $r^2>0.8$, which is referred to as the coverage of the chip (Barrett and Cardon, 2006). The strategy of SNP selection has an impact on the coverage, and differs between the Illumina and Affymetrix chips. SNPs on Illumina chips are selected using a pair-wise correlation-based algorithm applied to genotype data from HapMap samples. In contrast, SNPs on the Affymetrix chips are preselected primarily on the basis of technical quality and distribute evenly across the human genome (Pe'er et al., 2006).

Evaluating the Illumina and Affymetrix chips that have been most commonly used for GWAS shows that increasing the sample size is likely to have a larger effect on power than increasing the chip SNP density (Spencer et al., 2009). The sample size needed to study common human diseases is usually large, with more than 2000 cases and 2000 controls required for a relative risk of 1.3–1.5 to detect the causal variant. For common alleles, the most effective chips have been the Affymetrix 100 K and 500 K chips, and the Illumina 300 K chip (Table 1). One way to improve the SNP coverage of the genome is to predict or impute the missing genotypes using, for example, HapMap data (Marchini et al., 2007). The imputation has an even greater effect on low frequency SNPs that are usually missing from the commercial SNP chips (Spencer et al., 2009).

Using the budget of a study as a fixed variable, we can estimate the sample size and power of a given study. The table 2 illustrates a situation with a fixed budget and relative risk of 1.5, MAF of at least 0.05 and a p-value threshold of $5x10^{-7}$. The same tendency is seen if more samples are included to obtain an optimum power of 0.821 with the lowest price per chip. Compared to the Illumina 1M chip, which reaches a power of 0.635, the Illumina 300 K has 17% greater power with three times fewer SNPs (Spencer et al., 2009).

Thus, the design of a GWAS is important and should be performed carefully. A calculation of study power must take into account the set of SNPs on the selected chips, the sample size, the effect size and the budget.

Table 1. The power for each commercial chip type for an equal number of cases and controls (2000 vs. 2000) and a relative risk of 1.3 at the causal SNP and p-value threshold of $5x10^{-7}$. "Complete" chip – when all SNP available in HapMap database are genotyped (Spencer et al., 2009).

Chip	Chip SNP tests	MultiMarker tests	Tests including imputed SNPs
Affymetrix 100K	0.178	0.212	0.242
Affymetrix 500K	0.363	0.378	0.450
Illumina 300K	0.392	0.424	0.467
Illumina 610K	0.439	0.455	0.488
Illumina 650K	0.443	0.458	0.492
Affymetrix 6.0	0.420	0.433	0.478
Illumina 1M	0.457	0.461	0.493
"Complete"	0.499	0.499	0.499

Table 2. Power achieved by different chips with a budget of $2,000,000, and assuming a disease causing allele with a relative risk of 1.5, a minor allele frequency of at least 0.05, and a p-value threshold of $5x10^{-7}$. The last line of the table shows the power that would be obtained using the "Complete" chip that types all the SNPs in HapMap database (Spencer et al., 2009).

Chip	Average price ($) of a chip	Number of cases/controls	Power
Affymetrix 500K	420	2381	0.767
Illumina 300K	377	2653	0.821
Illumina 610K	452	2212	0.818
Affymetrix 6.0	505	1980	0.772
Illumina 1M	750	1257	0.635
"Complete"	-	2653	0.881

1.3 Genetic structure maps

Genome-wide genotyping of hundreds of thousands of autosomal markers has enabled the construction of accurate genetic structure maps of populations that correlate to some extent with geographical maps. In association studies, the

ancestry of each study participant prior to data analysis must be determined to reduce false positive associations with a trait, or reducing the power to detect such an association.

The ancestry of each study subject can be determined using ancestry-informative markers. These are typically SNPs that show large allele frequency differences between populations. The markers with the most allele frequency variability between populations are usually at loci that determine the skin pigmentation, hair morphology and coloration, or are associated with dietary adaptation and the immune system (Coop et al., 2009). Examples include the skin pigmentation locus *SLC24A5*, with an allele with a nonsynonymous SNP that is strongly associated with lighter skin color, and has a high frequency within European, Middle East and South Asian populations (Lamason et al., 2005); the hair and skin coloration locus *MC1R*, with an allele with a non-synonymous SNP that shows a high frequency in East Asian and American populations (Rana et al., 1999); a nonsynonymous SNP in the *EDAR* gene that affects hair morphology and shows a similar geographic pattern as the *MC1R* gene (Sabeti et al., 2007); the *KITLG* locus that leads to lighter skin pigmentation and one haplotype of that gene is present only in non-African populations, with a high frequency across Eurasia, the Americas and Oceania (Miller et al., 2007); and SNPs in *lactase* (Bersaglieri et al., 2004), and in the adaptive immune system *Toll-like receptor 6* (Todd et al., 2007) gene region, which have a marker frequency gradient across Europe.

Genetic structure maps can be classified by dimensions using global, European, and single population genetic maps. This thesis concentrates on the genetic structure of European populations, so the relevant European genetic structure maps are considered in detail, and only studies using more than 100,000 markers are reviewed.

1.3.1 Global genetic structure maps

Genetic and geographic distances correlate well in the global genetic structure map (Li et al., 2008), which shows a boomerang-like curve with Africa on one side and Oceania in the other, with Europe as the turning point (Figure 4). Between Africa and Europe are the Middle East and Mexico (Nelson et al., 2008), and between Europe and Oceania are Central-, South- and East Asia, and Native Americans (Jakobsson et al., 2008). The genetic variability between subpopulations is characterized by F_{st}, which varies from 0.1 to 0.15 for world populations and is much smaller within continental populations. The F_{st}-value also depends on the population history. Within Europe, the F_{st}-value is 0.006, while Amerindians show a much higher level of diversity, with F_{st}=0.04 or greater (Cavalli-Sforza et al., 1994).

Figure 4. Global genetic structure map. The genetic variance (based on 512,762 SNP genotypes) between different world populations (443 unrelated HGDP-CEPH individuals) was analyzed using the multidimensional scaling of allele-sharing distance between individuals (Jakobsson et al., 2008). C/S Asia – Central/South Asia; HGDP-CEPH – Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain.

1.3.2 European genetic structure maps

The first studies to use whole-genome genotype data of European-Americans identified a clear gradient from northwest to southeast in Europe (Bauchet et al., 2007; Price et al., 2008; Tian et al., 2008b). Analysis of whole-genome data on a larger number of individuals sampled from multiple European populations extended these findings (Heath et al., 2008; Lao et al., 2008; Novembre et al., 2008). The European genetic structure map published by Novembre et al. (2008), is the most precise map so far, covering 37 European populations, although 25 are represented by fewer than 20 individuals (Figure 5). The Baltic countries are not included in these studies and only one individual from Latvia is represented on the genetic map published by Novembre et al. As expected, the first two principal components (PC1 and 2) correlate well with the geographic axes (r^2 = 0.71 for PC1 versus latitude; r^2 = 0.72 for PC2 versus longitude) (Lao et al., 2008; Novembre et al., 2008). In addition, PC1 aligns northwest/southeast with an eigenvalue of 4.09, and PC2 aligns northeast/southwest with an eigenvalue of 2.04. European genetic structure maps show that Spain, Italy and the Balkan Peninsula are in the south, and that Finland and East-Europe are in the north, as the Western- and Central European populations are in the middle of the genetic structure maps.

The genetic structure of Asia (Tian et al., 2008a) and Africa (Tishkoff et al., 2009) have been also studied.

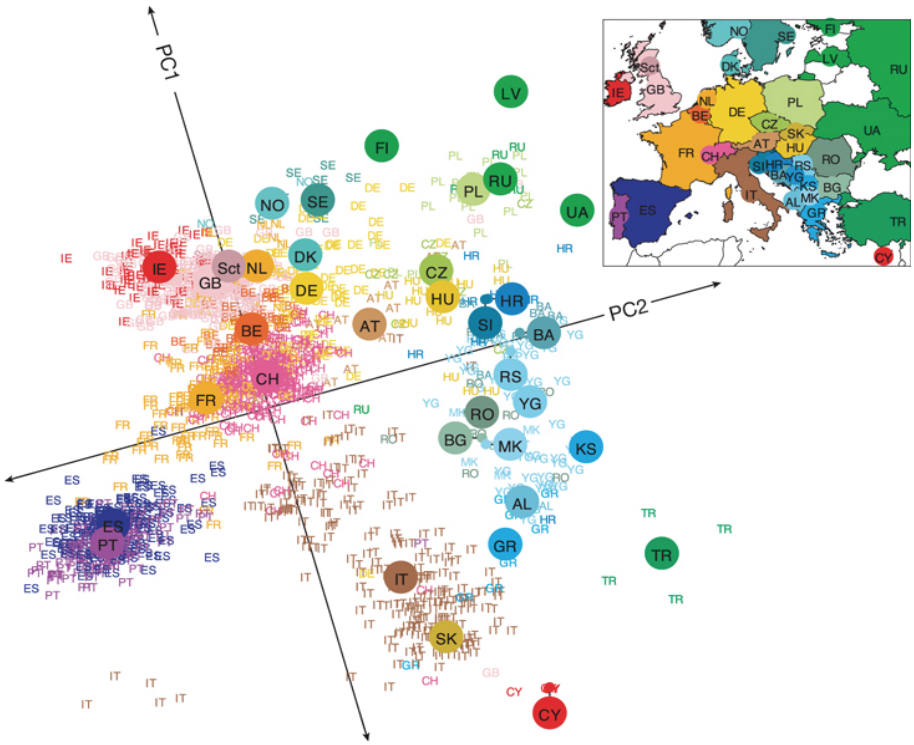

Figure 5. European genetic structure map. The genetic structure in Europe is illustrated with the scatter plot of the two first PCs from the PC analysis that was performed using genotype data (Affymetrix 500K chip) of 1387 Europeans. Small colored labels represent individuals and large colored points represent the median PC1 and PC2 values for each country: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia (Novembre et al., 2008).

1.3.3 Single population genetic structure maps

On the global map, single populations generally appear to be homogenous, but analysis at the population level usually shows slight genetic structure. The genetic structure of several larger populations, for example Japan (Yamaguchi-Kabata et al., 2008) and Mexico (Silva-Zolezzi et al., 2009), have been studied. Furthermore, the genetic variation in genetically isolated Ashkenazi Jews was compared to the HapMap CEU population (Olshen et al., 2008), showing small but significant differences in measures of genetic diversity (mean F_{st}=0.009).

Much interest has focused on European populations, for example Iceland (Price et al., 2009), Finland (Jakkula et al., 2008), Northern Europe (Salmela et al., 2008; McEvoy et al., 2009), Germany (Steffens et al., 2006), and the United Kingdom (WTCCC, 2007). The Icelandic population arose from an admixture of Norse and Gaelic ancestors around 1100 years ago. The plot of the first two PCs from the PCA shows remarkable concordance with Icelandic geography, following a ring-shape topology (Figure 6). The population structure is minimal (average F_{st}=0.0026) and the difference between regions is due to recent genetic drift that occurred in the 1100 years since the first settlement of Iceland (Price et al., 2009).

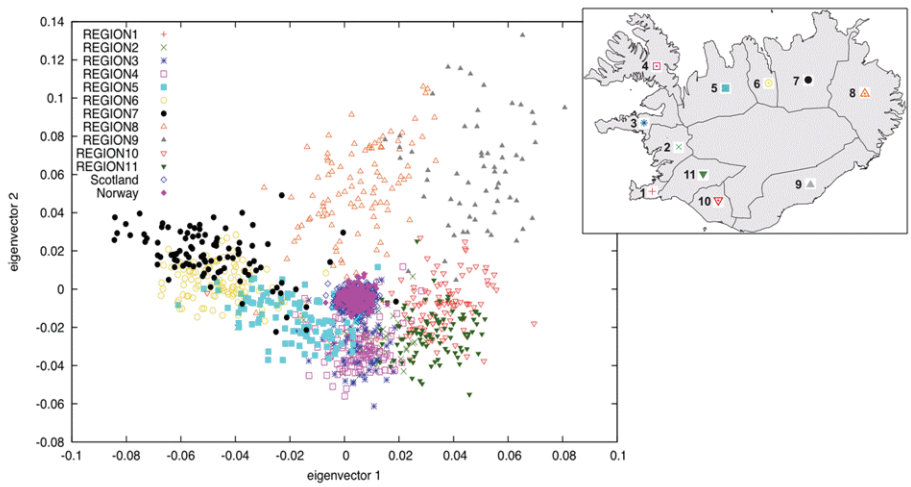

Figure 6. The genetic structure of Iceland. PCA plot of 877 samples with most of their ancestry from 11 regions of Iceland. The PCA analysis was ran using genotype data of 292,289 SNPs. The samples from Norway and Scotland added on the map show that the varying contributions from ancestral populations are not a major determinant of genetic differences between Icelandic regions, rather is the recent genetic drift in the Icelandic gene pool (Price et al., 2009).

Another well-studied founder population is in Finland, which has been inhabited for 10,000 years. Two major migration waves have had the greatest influence on the genetic structure of the population. The presence of subisolates make it a good country for identifying Mendelian disease genes (Varilo and Peltonen, 2004). The genetic structure map of Finland corresponds well with east-west and north-south geographies, and demonstrates internal migration from country-sides to the capital city (Figure 7). Differences in F_{st}-values between Helsinki and recent subisolates (F_{st}=0.004) are comparable with the F_{st}-values between northwest and southeast European populations (Heath et al., 2008). As somewhat smaller genetic structure has been described within Germany (Northern Germany / Southern Germany, F_{st}=0.00017 (Steffens et al., 2006)), and United Kingdom (WTCCC, 2007).

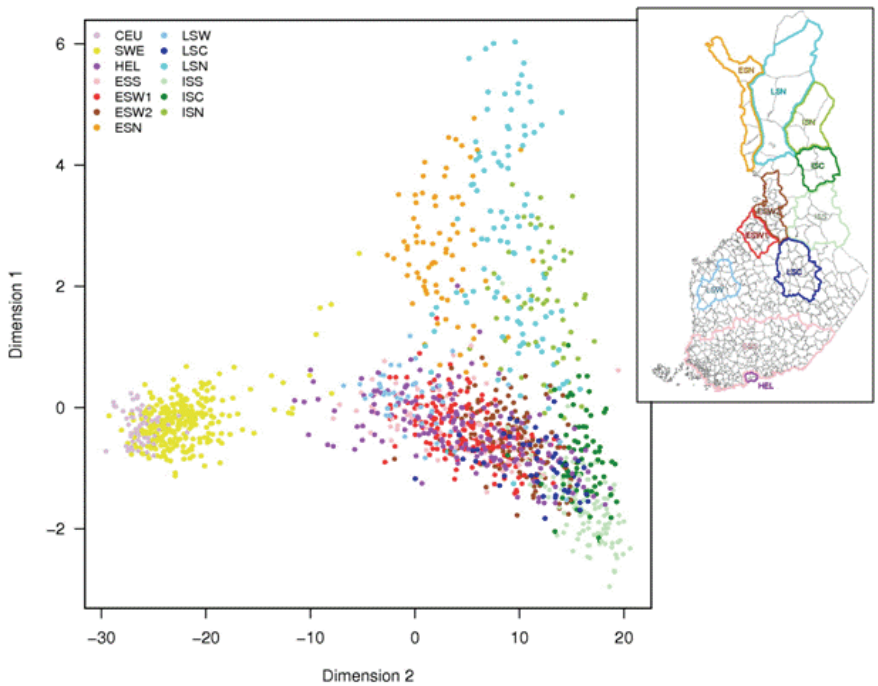

Figure 7. The genetic structure of Finland. The genetic variance between Finnish subisolates (901 samples) as well as the Helsinki (162), Swedish (302) and CEU (60 samples) populations is represented on the multidimensional scaling plot. Genotype data of 231,116 SNPs was used to perform the PC analysis. CEU – Utah residents with ancestry from Northern and Western Europe, SWE – Sweden, HEL – Helsinki, ESS – early-settlement south, the south coastal region, ESW1 – early-settlement west, South Oulu, ESW2 – North Oulu, ESN – early-settlement north, the Tornio-River valley in West Lapland, LSW – late-settlement west, South Ostrobothnia, LSC – late-settlement central, Central Finland, LSN – late-settlement north Central Lapland, ISS – isolate south, South Kainuu, ISC – isolate central, North Kainuu, and ISN – isolate north, East Lapland (Jakkula et al., 2008).

Nonetheless, the picture of the genetic structure of European populations is still incomplete. Efforts have been made to describe the genetic structure of Europe, with an emphasis on the western- and central countries. In summary, analyses of genetic structure using different methods show minor genetic differences between nearby populations, and larger variation with more distant and isolated populations.

2. AIMS OF THE PRESENT STUDY

1. To describe tagSNP transferability from HapMap populations to the Estonian population, and to determine how SNP density, minor allele frequency, and sample size influence tagSNP selection.

2. To evaluate the performance of commercial SNP panels – Illumina 300 K and 550 K, and Affymetrix 100 K and 500 K.

3. To characterize the genetic structure of Estonia and compare it to other European populations.

3. RESULTS AND DISCUSSION

3.1 Studied populations and regions

Populations

Throughout this analysis, the genotype information of 1090 DNA samples, selected from 10,317 samples from the Estonian biobank in 2005 were used. Eighty samples (40 males and 40 females) were selected randomly by place of birth, from each of 13 Estonian counties (Harju, Ida-Viru, Jõgeva, Järva, Lääne-Viru, Põlva, Pärnu, Rapla, Saaremaa, Tartu, Valga, Viljandi, Võru), and 50 samples (25 males and 25 females) were selected from the combined Hiiumaa and Läänemaa counties (Ref. I, Figure 1). Estonia is a small country (43,400 km^2) in the Baltic region of Northern Europe. As a coastal area, it has been the recipient of several migration waves from neighboring countries, and the population is approximately 1 million Estonians, 300,000 Russians and other nationalities. In all studies, individuals with Estonian descent were analyzed. As demonstrated in the III study, the genetic structure of Estonians, although minimal, is still detectable and might be used in future studies, if correlation with a disease pattern can be demonstrated.

Genotype data from two ENCODE regions (described below) were used in the I and II studies. From each ENCODE region, 768 selected SNPs were genotyped at McGill University and Genome Quebec Innovation Centre as part of the HapMap project, using the Illumina GoldenGate assay. An overview of the studied populations and regions is in table 3.

In the III study, whole-genome genotype data from 3112 individuals analyzed with Illumina 318K/370CNV chips was used. The samples represent a total of 19 cohorts from 16 countries: Austria (Vienna), Bulgaria (entire country), Czech Republic (Prague, Moravia and Silesia), Estonia (entire country), Finland (Helsinki, and a young internal subisolate of Kuusamo), France (Paris), two cohorts from Germany (Schleswig-Holstein (north), and the Augsburg region (south)), Hungary (entire country), two cohorts from Italy (Borbera Valley (north), and Region of Apulia (south)), Latvia (Riga), Lithuania (entire country), Poland (West-Pomerania), Russia (Andreapol district of the Tver region), Spain (entire country), Sweden (Stockholm) and Switzerland (Geneva) (Ref. III, Table 1). Approximately half of the samples were genotyped at the Estonian Biocentre and raw data was obtained from the centers for other populations (Ref. III, Table 1). After quality control procedures, 273,464 SNPs remained and were used for further analyses.

The HapMap data used throughout these studies comprised four populations, namely CEU – U.S. Utah residents with ancestry from Northern and Western Europe, YRI – Yoruba people of Ibadan, Nigeria, CHB – unrelated individuals from Beijing, China, and JPT – unrelated individuals from Tokyo, Japan.

Table 3. Studied populations and regions.

Study	Studied populations	# of individuals	Studied regions	Illumina genotyping assay
Ref. I, II	Estonia	1090	ENCODE regions	GoldenGate assay
	HapMap samples		ENr112 (500 kb)	
	CEU	60	ENr131 (500 kb)	
	CHB	45		
	JPT	44		
	YRI	60		
Ref. III	Austria (Vienna)	88	whole genome	CNV370
	Bulgaria	48		CNV370
	Czech Republic (Prague and Moravia)	94		CNV370
	Estonia	1090		CNV370
	Finland (Helsinki)	100		CNV370
	Finland (Kuusamo)	84		CNV370
	France (Paris)	100		HumHap300
	Northern Germany (Schleswig-Holstein)	210		HumHap300
	Southern Germany (Augsburg region)	473		CNV370
	Hungary	50		CNV370
	Northern Italy (Borbera Valley)	96		CNV370
	Southern Italy (Region of Apulia)	95		CNV370
	Latvia (Riga)	95		CNV370
	Lithuania	95		CNV370
	Poland (West-Pomerania)	48		CNV370
	Russia (Andeapol district of Tver region)	96		CNV370
	Spain	200		HumHap300
	Sweden (Stockholm)	100		HumHap300
	Switzerland (Geneva)	216		HumHap550
	HapMap samples			
	CEU	60		HumHap300
	CHB	44		HumHap300
	JPT	44		HumHap300
	YRI	55		HumHap300

Regions

Two 500 kb ENCODE regions on chromosome 2 (ENr112 on 2p16.3 (ENCODE 1) and ENr131 on 2p37.1 (ENCODE 2)) were analyzed in studies I and II. These regions were previously resequenced in entirety in 48 individuals from various origins, and were genotyped in four populations as part of the HapMap project. The regions differ in their average recombination rates (0.8 cM/Mb for ENCODE 1 and 2.1 cM/Mb for ENCODE 2). Overall, 2431 SNPs in ENCODE 1, and 2067 SNPs in ENCODE 2 have been successfully genotyped as part of HapMap. From each of the ENCODE regions, 768 random SNPs were selected and genotyped in 1090 samples from Estonia. Of these, 721 SNPs in ENCODE 1, and 699 SNPs in ENCODE 2, passed all genotyping quality criteria and were used in studies I and II.

3.2 Evaluation of tagSNPs derived from HapMap in Estonian population sample (I)

The HapMap project aimed to describe the LD structure of four populations – Europeans (U.S. residents with ancestry from Northern and Western Europe), Chinese (from Beijing), Japanese (from Tokyo) and Africans (Yoruba people of Ibadan, Nigeria), with a goal of minimizing the number of SNPs required for association studies. The four population samples were proposed as references for selecting tagSNPs for other world populations. We performed several analyses to determine the transferability and performance of tagSNPs from HapMap to the Estonian population. We used data for 1536 SNPs from two 500 kb ENCODE regions on chromosome 2, genotyped for 1090 individuals from Estonia.

The Estonian sample appeared to be similar to the CEU sample (Ref. I, Figure 2) by comparison of MAF distributions between studied populations. This was also found for LD block distribution, but the LD structure appeared to match the CHB/JPT structure (Ref. I, Figure 3). Using median-joining network analysis, common haplotypes shared in all studied populations were found. As expected, the Estonian samples most often shared haplotypes with the CEU. However, depending on the region studied, the two European samples (CEU and Estonia) occasionally had different haplotype frequencies, some of which were present only in the Estonian samples, or were shared with CHB/JPT.

STRUCTURE program analysis could not detect the population substructure within Estonia, possibly because the number of available markers was too low or the two ENCODE regions too narrow to detect a population substructure.

To describe the transferability of tags across the populations, we selected tagSNPs from the HapMap sample and tested them on other population samples. Pair-wise algorithm of the Tagger (de Bakker et al., 2005) was used to select tags with different allele frequencies, and with a high correlation co-efficient measure, r^2=0.8. The tagSNPs selected from the CEU HapMap samples captured most of the variation in the Estonian sample (90–95% of SNPs with MAF>5%) (Ref. I, Figure 5). The CHB/JPT tags performed less well

on CEU or Estonian sample, capturing only 80% of the SNPs. The YRI tags worked surprisingly well on other samples, but required the use of two to three times more tagSNPs. Many more tagSNPs were required for lower LD regions, and tagging performance dropped sharply for SNPs with MAFs of 10% or higher. The latter could be explained by the presence of many SNPs with high allele frequencies in the target population, but with frequencies lower than the selected MAF threshold in the population from which the tagSNPs were selected. Nonetheless, our analysis showed that for the low-recombination ENCODE 1 region, one tag every 6 kb was sufficient to capture all common alleles (MAF>5%), while one tag every 4 kb was sufficient for the high-recombination ENCODE 2 region.

Allele frequency, sample size and SNP density in the dataset used to select the tags, all had important effects on tagging performance and therefore must be taken into account when designing association studies.

To illustrate the MAF distribution, tagSNPs were selected from the CEU sample and tested on Estonian sample (Ref. I, Figure 6). As expected, markers with higher allele frequencies in the Estonian sample tended to correlate better with tagSNPs selected from the CEU sample. However, markers with MAF<5% were poorly captured by the CEU tags. Thus, association study planning must consider that low-frequency, population-specific SNPs might not be covered by tagSNPs selected from the HapMap.

To test the sample size needed for optimal tagging, tagSNPs were selected from different sets of 10 to 1000 Estonian samples and tested on CEU sample (Ref. I, Figure 7). The optimal tagging of SNPs with MAF<5% required at least 90–100 independent samples, and sample size was a more crucial factor for the less frequent SNPs (MAF<5%).

To test the effect of SNP density on selection of tagSNPs, the 500 kb ENCODE region was divided into equal-sized (1.3–10 kb) windows, and one polymorphic SNP from the CEU population was selected for each. Performance of the selected tagSNPs was measured using the Estonian sample (Ref. I, Figure 8). A clear decline in tagging performance was observed for with each decrease in density studied. The best performance was obtained when selected SNPs occurred every 1.3 kb, which mimics HapMap Phase II.

LD patterns and the transferability of tagSNPs from HapMap data have been studied previously on samples from different global geographical regions (de Bakker et al., 2006; Gonzalez-Neira et al., 2006; Xing et al., 2008). Several studies on European populations have shown that tagSNPs selected from the HapMap CEU samples perform well in samples from a wide variety of European populations, with ~75–95% of non-tagSNPs in LD with CEU selected tagSNPs at a level of $r^2 \geq 0.8$ (Mueller et al., 2005; Ribas et al., 2006; Willer et al., 2006; Lundmark et al., 2008). The same tendency was seen with the Estonian sample, with 90–95% of common SNPs captured by the tagSNPs selected from the CEU HapMap sample. TagSNPs selected from the HapMap data also performed well in population isolates (Service et al., 2007; Lundmark et al., 2008).

3.3 Evaluation of commercial genotyping panels (II)

Two main companies, Illumina Inc. and Affymetrix Inc., produce whole-genome genotyping chips with different levels of genome coverage. To better understand how well these chips capture common variations in the human genome, we selected two SNP panels from each company for performance analysis. The HumanHap 300 and HumanHap 550 Array Sets from the Illumina Infinium series, and the Mapping 100 K and Mapping 500 K Array Sets from the Affymetrix GeneChip series were tested in four HapMap populations, and in Estonian population samples.

Tagging performance was tested using the same genotyping data as the I study, from the two ENCODE regions ENr112 on 2p16.3, and ENr131 on 2p37.1. The number of SNPs present in each HapMap population is in Ref. II, Table 1.

For each marker in the HapMap data, the best tagging SNP from each commercial SNP panel was determined. The percentage of SNPs covered at $r^2 \geq 0.8$ was calculated, along with the mean r^2 between each marker and its optimal tagging SNP. This was done at MAF cut-offs of 0.01 and 0.05. The best performance for all tested SNP panels was seen in the non-African HapMap populations. The HumanHap 550 panel showed the highest coverage, with 80–90% in European and Asian populations at a MAF cut-off of 1% (Ref. II, Figure 1 A). At a MAF cut-off of 5%, the HumanHap 550 showed the most SNP coverage, but the HumanHap 300 also showed good performance for European (89%) and Asian (70%) populations (Ref. II, Figure 1 B). The higher number of SNPs on the HumanHap 550 panel did not result in a large advantage over the HumanHap 300 panel, since the selected tagSNPs showed nearly the same values.

Whole-genome coverage of the commercial panels was also tested. Here again, the HumanHap 550 had the best results: CEU – 86%, JPT/CHB – 83% and YRI – 48%. These results also confirmed that the ENCODE regions accurately reflected the whole genome. The results were consistent with previous analysis of coverage by these commercial panels, with the exception of the HumanHap 550. Our obtained values were identical to an earlier study, in spite of some differences in data (Barrett and Cardon, 2006).

The performance of commercial SNP sets was tested using the Estonian samples. Since the number of genotyped SNPs was lower in the Estonian population (1420) than in the HapMap samples (CEU – 4670; CHB/JPT – 4495; YRI – 4540), markers were selected according to those genotyped in the Estonian samples (Ref. II, Table 2). Calculations were carried out for the HapMap and Estonian populations, and the results expressed as a fraction of the coverage of the CEU sample (Ref. II, Figure 2 A-D). The results showed that non-African populations and Yoruba sample were covered equally well by the commercial products.

The universal and population-specific SNPs on the commercial panels were also analyzed. For each SNP in the HapMap population, the highest-performing

tagSNP on the commercial panel was determined. We then determined whether the commercial SNP was the best describer for one, two or all three populations (Ref. II, Figure 3). A strong bias towards CEU-specific SNPs was seen for the Illumina HumanHap 550 and particularly for the HumanHap 300. This could be because the SNPs were chosen from the HapMap database, which ensured that the HapMap populations had the best coverage. In contrast, the GeneChip 100 K and GeneChip 500 K described population-specific markers from all populations fairly equally. The results show that universal markers constitute 63–82% of all markers, for all studied commercial platforms studied, and approximately 10% of the SNPs describe SNPs from only a single population sample.

In this study, we have shown that commercial SNP panels, particularly Illumina panels with SNPs chosen from the HapMap database, can capture most of the common SNPs from non-reference European population samples. The Illumina HumanHap 550 whole-genome coverage reached 86%, and HumanHap 300 coverage was 76%. Thus, for performance and chip price considerations in study design, the HumanHap 300 could be selected without compromising much information. However, new, improved SNP panels with one million or more markers are becoming available, offering the possibility of denser genome-wide coverage.

3.4 European genetic structure map (III)

In the I and II studies, we examined the transferability of tagSNPs from HapMap to other populations, and evaluated commercial panels from two major companies. Before performing an association study, however, cases and controls must be carefully selected. Since one population may not be able to provide sufficient samples for studying a complex disease, samples from different population might be combined. In this III study, we examined the population structure within Estonia, and within other European countries, and compared the genetic structure between these populations.

To describe the genetic structure across Europe, we used whole-genome genotype data of more than 270,000 SNPs (the number of SNPs remaining after quality control (QC)), genotyped with Illumina 318K/370CNV chips. SNPs found to be out of Hardy-Weinberg equilibrium at $p<10^{-5}$, or missing more than 1% of genotypes, or with a MAF<0.01 were removed from the dataset during the QC procedures (WTCCC, 2007). Samples from 3112 individuals, comprising 19 cohorts across 16 European countries, were used. The large number of markers enabled us to study the MAF spectrum between Estonia and neighboring countries. We found that the correlation coefficient r^2 for MAFs of the studied SNPs varied markedly between Estonia and other countries, including 0.9247 for Latvia, 0.8913 for Finland, and 0.7312 for Southern Italy. The genome-wide LD, as expected, was more extensive in isolated than in outbred populations (Ref. III, Figure 1). In addition, the cohorts diverged more

clearly at larger LD distances (above 75 kb), and the LD extent decreased from northern to southern countries.

Genetic structure was determined using PCA in three dimensions, on intercontinental and intracontinental scales, and within a single country (Ref. III, Figure 2). First, genetic structure within Europe was analyzed. A major gradient from northwest to southeast was identified as the first PC, with an eigenvalue of 8.7. The second PC identified a gradient from Finland to the Southern European countries, with an eigenvalue of 4.9. When the Asian and African HapMap populations were added to the European samples, the eigenvalue of the first PC increased to 36.6, and second to 23.8. The Asian and African populations were distant from the European populations, while the CEU samples overlapped with the other European populations in the analysis. When the data of Asian and African populations was included, the PC analysis indicated that Europe is quite uniform.

We also studied the internal genetic structure of several of the studied populations, for which two or more cohorts were available (Ref. III, Figure S3) and found pronounced intrapopulation genetic differences in Finland and Italy. Kuusamo was previously shown to be a young population isolate, with a genetic structure that differs from that of Helsinki (Varilo et al., 2000). The two Italian cohorts represented a small mountain village in Northern Italy and the Apulia region in Southern Italy. Somewhat smaller genetic diversity was seen in an analysis of three cohorts from Czechia, and an analysis of Southern and Northern Germany. PC analysis of 966 Estonians, representing 14 counties, revealed the fine-structure of the population, with eigenvalues for the first two PCs of 1.9 and 1.5. The spread of individuals was relatively wide, with sub-regions overlapping on an individual level, but the median PC values, calculated for each county, correlated remarkably with the regional map of Estonia.

The genetic variance between populations was also studied using the fixation index (F_{st}) and the inflation factor lambda (λ). The values of F_{st} correlated considerably with geographic distance (r^2=0.382, $p<<0.01$). Values ranged from $\leq$0.001 for neighboring countries, to 0.023 between Southern Italy and a young subisolate of Finland (Kuusamo) (Ref. III, Table S2). The intrapopulation variability was also measured by F_{st}, and mean F_{st} was 0.001 for the 14 Estonian counties, 0.005 for Finland, 0.000 for Germany and 0.005 for Italy.

The pair-wise inflation factor lambda (λ) was calculated for studied samples, using the genomic control method (Devlin and Roeder, 1999). Values of λ ranged from unity (between samples from the same country) to 4.21 (between Spain and the Kuusamo region). The overall average λ value was 1.82; in separate clusters, it was 1.23 for the Baltic Region, Western Russia and Poland; 1.54 for Italy and Spain; 1.22 for Central and Western Europe; and 1.86 for Finland. The correlation coefficient between geographic distance and λ was r^2=0.386 ($p<<0.01$). This value is probably an underestimate of the European-wide relationship due to the inclusion of samples from the young population subisolate Kuusamo, and the highly heterogeneous international metropolis Geneva.

Marker-wise significance tests for allelic differences in pair-wise comparisons of the studied samples (i.e., simulated association studies between populations) resulted in 2263 loci that were significantly different between the studied populations. As our sample included some genetically and geographically distant cohorts, such as Finns and Italians, where strong founder effects and isolation-driven genetic drift has changed allele frequencies, only loci that were present in non-Italian and non-Finnish comparisons were considered. Furthermore, comparing only loci that had at least two significant hits in at least two pair-wise comparisons caused the total number of significantly different loci to decrease to 18 (Ref. III, Table S4). Four genes were within the LCT (lactase) loci, a haplotype block covering more than 1 Mb (Bersaglieri et al., 2004)) that differentiates not only between European populations (Heath et al., 2008), but also within a given population (WTCCC, 2007).

Since we examined the function of the twenty-two most variable SNPs between populations identified by PCA (11 SNPs for the first PC and 11 SNPs for the second), the SNPs were expectedly from the genes which allele frequencies varied the most between populations (Ref. III Table S3). The three genetically most variable SNPs revealed by PC analysis represented loci that were also present in the list of 18 loci from the marker-wise significance test.

Based on the analysis performed in this III study, we conclude that using neighboring populations in association studies is meaningful, because their genetic similarities minimize the loss of power. For example, the Baltic countries may be analyzed together with Western Russia and Poland, and data from Central and Western Europe can be analyzed together. Interestingly, language similarity does not always match the genetic background of the two populations. For example, Estonia and Finland show more differences in genetic structure than either does with Latvia, even though the Finnish language is from the same Finno-Uralic group as Estonian, and Latvian belongs to Balto-Slavic language group. Knowledge of genetic distances between different populations is helpful in defining which biobanks can reasonably contribute samples and data to a GWAS.

CONCLUSIONS

Following conclusions can be drawn from the current Ph.D. thesis:

1) TagSNPs selected from the CEU HapMap sample, representing two 500 kb ENCODE regions, capture most of the variation in the Estonian sample (90–95% of the SNPs with a MAF >5%). In addition, the allele frequency, sample size and SNP density in the dataset used to select the tags, all have important effects on tag performance, and must be considered in designing association studies.
2) From the four evaluated commercial SNP panels: Illumina 300 K and 550 K, and Affymetrix 100 K and 500 K, all SNP sets have the coverage of approximately 50% on HapMap Yoruban population, whereas the coverage of HapMap CEU and Asian populations can reach to 80–90% on Illumina 500 K. The results show that the Estonian population is tagged with the same efficiency as the HapMap CEU population sample, as the coverage of Illumina 550 K reaches up to 86%, and Illumina 300 K coverage is 76% in these populations of European descent.
3) PC analysis of genotype data of more than 270,000 SNPs of 3112 individuals from Europe yielded a genetic structure map of Europe in which two first PCs highlighted genetic diversity corresponding to a northwest to southeast gradient, and positioned the populations according to their approximate geographic origin. The results of this thesis demonstrate that Estonian samples can be analyzed with most other European samples, with the exception of the isolates (Kuusamo) identified here and the southernmost Europeans, without great loss of power. Using the estimated values of F_{st} and λ, we can now calculate how much power would be lost by combining populations in a study, and the precise benefits of increasing the number of subjects using samples from other European biobanks.

REFERENCES

Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. Nat Genet 38(6):659–62.

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. 2007. Measuring European population stratification with microarray genotype data. Am J Hum Genet 80(5):948–56.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111–20.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis

GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816.

Buchanan JA, Scherer SW. 2008. Contemplating effects of genomic structural variation. Genet Med 10(9):639–47.

Cardon LR, Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet 2(2):91–9.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74(1):106–20.

Cavalli-Sforza L, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton University Press, Princeton, NJ.

Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH. 1984. Nonuniform recombination within the human beta-globin gene cluster. Am J Hum Genet 36(6):1239–58.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. 2009. Origins and functional impact of copy number variation in the human genome. Nature.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. PLoS Genet 5(6):e1000500.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74(4):610–22.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat Genet 29(2):229–32.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature 418(6897):544–8.

de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, Onofrio RC, Lyon HN, Stram DO, Haiman CA, Freedman ML, Zhu X, Cooper R, Groop L, Kolonel LN, Henderson BE, Daly MJ, Hirschhorn JN, Altshuler D. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 38(11):1298–303.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37(11):1217–23.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55(4):997–1004.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F,

Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–61.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet 36(4):388–93.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296(5576):2225–9.

Gonzalez-Neira A, Ke X, Lao O, Calafell F, Navarro A, Comas D, Cann H, Bumpstead S, Ghori J, Hunt S, Deloukas P, Dunham I, Cardon LR, Bertranpetit J. 2006. The portability of tagSNPs across populations: a worldwide survey. Genome Res 16(3):323–30.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. Pathogenetics 1(1):4.

Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D,

Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Genin E, Cardon LR, Lathrop M. 2008. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16(12):1413–29.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307(5712):1072–9.

Jakkula E, Rehnstrom K, Varilo T, Pietilainen OP, Paunio T, Pedersen NL, deFaire U, Jarvelin MR, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L. 2008. The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83(6):787–94.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181): 998–1003.

Jeffreys AJ. 1979. DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. Cell 18(1):1–10.

Jeffreys AJ, Holloway JK, Kauppi L, May CA, Neumann R, Slingsby MT, Webb AJ. 2004. Meiotic recombination hot spots and human DNA diversity. Philos Trans R Soc Lond B Biol Sci 359(1441):141–52.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29(2):233–7.

Jorde LB. 2000. Linkage disequilibrium and the search for complex disease genes. Genome Res 10(10):1435–44.

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245(4922):1073–80.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22(2):139–44.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'Donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782–6.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Ruther A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Krawczak M, Kayser M. 2008. Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241–8.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human

relationships inferred from genome-wide patterns of variation. Science 319(5866): 1100–4.

Lundmark PE, Liljedahl U, Boomsma DI, Mannila H, Martin NG, Palotie A, Peltonen L, Perola M, Spector TD, Syvanen AC. 2008. Evaluation of HapMap data in six populations of European descent. Eur J Hum Genet 16(9):1142–50.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747–53.

Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. Nat Genet 36(5):512–7.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–13.

Marshall E. 1999. A high-stakes gamble on genome sequencing. Science 284(5422): 1906–9.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356–69.

McCarthy MI, Hirschhorn JN. 2008. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17(R2):R156–65.

McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, Spector TD, Cherkas L, Ahmadi KR, Boomsma D, Willemsen G, Hottenga JJ, Pedersen NL, Magnusson PK, Kyvik KO, Christensen K, Kaprio J, Heikkila K, Palotie A, Widen E, Muilu J, Syvanen AC, Liljedahl U, Hardiman O, Cronin S, Peltonen L, Martin NG, Visscher PM. 2009. Geographical structure and differential natural selection among North European populations. Genome Res 19(5):804–14.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. Science 201(4358):786–92.

Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM. 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell 131(6):1179–89.

Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T. 2005. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76(3):387–98.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746): 321–4.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156(1):297–304.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83(3):347–58.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456(7218):98–101.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294(5547):1719–23.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2(12):e190.

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38(6):663–7.

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, Seligsohn U, Waliszewska A, Schirmer C, Ardlie K, Ramos A, Nemesh J, Arbeitman L, Goldstein DB, Reich D, Hirschhorn JN. 2008. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4(1):e236.

Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, Andreassen OA, Reich D, Kong A, Stefansson K. 2009. The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet 5(6):e1000505.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–9.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1–14.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65(1):220–8.

Pritchard JK, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. Genetics 155(2):945–59.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured populations. Am J Hum Genet 67(1):170–81.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–75.

Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M, Jenkins T, Li WH. 1999. High polymorphism at the human melanocortin 1 receptor locus. Genetics 151(4):1547–57.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411(6834):199–204.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17(9):502–10.

Ribas G, Gonzalez-Neira A, Salas A, Milne RL, Vega A, Carracedo B, Gonzalez E, Barroso E, Fernandez LP, Yankilevich P, Robledo M, Carracedo A, Benitez J. 2006. Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. Hum Genet 118(6):669–79.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516–7.

Roeder K, Luca D. 2009. Searching for disease susceptibility variants in structured populations. Genomics 93(1):1–4.

Roses AD. 2000. Pharmacogenetics and the practice of medicine. Nature 405(6788): 857–65.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–8.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928–33.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, Sistonen P, Savontaus ML, Schreiber S, Kere J, Lahermo P. 2008. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS One 3(10):e3519.

Service S, Sabatti C, Freimer N. 2007. Tag SNPs chosen from HapMap perform well in several population isolates. Genet Epidemiol 31(3):189–94.

Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, Balam-Ortiz E, del Bosque-Plata L, Velazquez-Fernandez D, Lara C, Goya R, Hernandez-Lemus E, Davila C, Barrientos E, March S, Jimenez-Sanchez G. 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proc Natl Acad Sci U S A 106(21):8611–6.

Spencer CC, Su Z, Donnelly P, Marchini J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5(5):e1000477.

Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A, Konig IR, Kohler K, Ludemann J, Diaz Lacava A, Fimmers R, Lichtner P, Ziegler A, Wolf A, Krawczak M, Nurnberg P, Hampe J, Schreiber S, Meitinger T, Wichmann HE, Roeder K, Wienker TF, Baur MP. 2006. SNP-based analysis of genetic substructure in the German population. Hum Hered 62(1):20–9.

Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. Nat Rev Genet 5(2):89–100.

TheInternationalHapMapConsortium. 2005. A haplotype map of the human genome. Nature 437(7063):1299–320.

Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, Seldin MF. 2008a. Analysis of East Asia genetic substructure using genome-wide SNP arrays. PLoS One 3(12):e3862.

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK, Seldin MF. 2008b. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4(1):e4.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM. 2009. The genetic structure and history of Africans and African Americans. Science 324(5930): 1035–44.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgoviste C, Simmonds MJ, Heward JM, Gough SC, Dunger DB, Wicker LS, Clayton DG. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39(7):857–64.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320(5875):539–43.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280(5366):1077–82.

Wang WY, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6(2):109–18.

Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L. 2000. Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. Eur J Hum Genet 8(8):604–12.

Varilo T, Peltonen L. 2004. Isolates and their potential use in complex gene mapping efforts. Curr Opin Genet Dev 14(3):316–23.

Weiss KM, Clark AG. 2002. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18(1):19–24.

Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF, Kinnunen L, Mohlke KL, Valle TT, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2006. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet Epidemiol 30(2): 180–90.

Wright S. 1969. Evolution and the Genetics of Populations Volume 2: the Theory of Gene Frequencies. 294–295 (Univ. of Chicago Press, Chicago).

WTCCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–78.

Xing J, Witherspoon DJ, Watkins WS, Zhang Y, Tolpinrud W, Jorde LB. 2008. HapMap tagSNP transferability in multiple populations: general guidelines. Genomics 92(1):41–51.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. 2008. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83(4):445–56.

SUMMARY IN ESTONIAN

Eesti populatsiooni geneetiline struktuur ja geneetiline kaugus teistest Euroopa päritolu populatsioonidest

Kogu genoomi hõlmavaid juht-kontrolluuringuid on läbi viidud juba mitmeid aastaid. Seda on võimaldanud kõrgtihedusega DNA kiipide tehnoloogia, mis võimaldab üle kogu genoomi otsida uusi komplekshaiguseid põhjustavaid regioone. Oluline juht-kontrolluuringute läbiviimisel on algne uuringu disain, mille käigus valitakse genotüpiseeritavad SNP markerid (ühenukleotiidsed DNA polümorfismid) ja uuritavad indiviidid. Käesolevas doktoritöös on käsitletud mõlemaid aspekte.

1) Hindasin HapMap andmebaasist valitud tagSNPde (SNP markerit "esindav" SNP) sobivust kirjeldamaks sagedasi markereid (minoorse alleeli sagedus >5%) Eesti populatsioonis. Uuritavate piirkondadena kasutasin kahe 500 kb pikkuse ENCODE regiooni genotüpiseerimise andmeid. Saadud tulemused näitasid, et HapMap CEU (Lääne- ja Põhja-Euroopa päritolu Utah's, Ameetika Ühenriikides elavad indiviidid) populatsiooni põhjal valitud tagSNPd kirjeldavad 90–95% Eesti populatsioonis esinevatest sagedastest markeritest. Samuti mõjutavad tagSNPde valimise efektiivsust markeri alleelisagedus, SNPde tihedus ja indiviidide arv.
2) Kuna sobiva kiibi valik uuringuks on väga oluline, hindasin nelja kommertsionaalse kiibi sobivust Eesti populatsiooni kirjeldamiseks. Valitud kiibid olid järgmised: Illumina HumanHap 300 K ja 550 K ning Affymetrix'i Mapping 100 K ja 500 K. Parima kattuvusega olid Illumina poolt pakutavad kiibid – HumanHap 550 K kattuvus oli 86% ja HumanHap 300 K korral 76%.
3) Uurisin Eesti geneetilist struktuuri ja hindasin geneetilist kaugust teistest Euroopa populatsioonidest. Kasutades peakomponentanalüüsi, mis hindab indiviidide geneetilist erinevust üksteisest, koostasin Euroopa geneetilise struktuuri iseloomustava kaardi. Peakomponentanalüüsi kaks esimest komponenti lahutavad Euroopa populatsioonid loode-kagu suunaliselt vastates populatsioonide geograafilisele asendile. Eesti populatsiooni suhteliselt ühtlane geneetiline struktuur annab hea eelise komplekshaiguste kaardistamiseks juht-kontrolluuringus. Samuti on hea koostööd teha teiste Euroopa biopankadega, sest suurem osa Euroopa populatsioonidest, välja arvatud geneetilised isolaadid ning Euroopa lõunapoolsemad populatsioonid, on geneetiliselt sarnased Eestiga ning tulenevalt populatsioonide erinevuste korrigeerimisest andmete analüüsimisel on uuringu võimsuse kadu minimaalne.

ACKNOWLEDGEMENTS

I would like to thank my supervisor prof. Andres Metspalu who has given the opportunity to be part of the world science. His enthusiasm and encouragement to use new technologies, methods and collaborate with other scientists have been valuable. Special thanks to my first supervisor Maris Teder-Laving who has taught the basic things about the molecular diagnostics. Also I would like to thank all the people from the department of biotechnology and bioinformatics, especially Krista Liiv, Merike Leego, prof. Ants Kurg, Reedik Mägi and prof. Maido Remm. Thanks to Viljo Soo and Tõnu Esko who have widened my knowledge about science and fiction.

Besides I would like to thank all the collaboration partners from Canada and from different European countries. Their efforts in collecting the samples, assistance with genotyping, sharing the genotyping data and help in data analysis have been valuable. I would also like to thank all the gene donors of Estonian Genome Project.

Hugs to Signe, Monika, Pille, Kristiina, Ingrid, Kadri, Siim, Tõnis, Kersti, Ervin, Veronika, Peeter, Ott, Riin, Kristi, Eva.

Finally, I want to thank my family for their love and support through my everlasting studies.

PUBLICATIONS

CURRICULUM VITAE

Mari Nelis

Date of birth:	15 October 1979, Tartu, Estonia
Address:	Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu 23 Riia street, 51010, Tartu, Estonia
Phone:	+3727375880
E-mail:	mari.nelis@ut.ee

Education

1986–1998	Leisi Secondary School
1998–2002	B.Sc. in Molecular Diagnostics, Faculty of Biology and Geography University of Tartu
2002–2004	M.Sc. in Molecular Diagnostics, Faculty of Biology and Geography University of Tartu
2004–2009	Ph.D. student, Faculty of Science and Technology, University of Tartu

Professional employment

2001–2002	Asper Biotech, laboratory assistant
2002–2003	Estonian Genome Project Foundation, laboratory assistant
2002–2007	Institute of Molecular and Cell Biology, University of Tartu, genotyping manager of the collaboration project of ARCAGE
2006–…	Manager of the Genotyping Core Facility, Estonian Biocentre

Scientific work

During my M.Sc. studies I participated in a large-scale association study, where potential candidate genes involved in the susceptibility of upper aerodigestive tract and lung cancer were studied. In my further research I have concentrated on the population genetics, precisely I have examined the genetic variation in human genome and its influence on phenotypes. Also, I have studied the genetic structure in European populations and its influence on whole-genome association studies.

Publications

Ellinor PT, Lunetta KL, Glazer NL, Pfeufer A, Alonso A, Chung MK, Sinner MF, de Bakker PI, Mueller M, Lubitz SA, Fox E, Darbar D, Smith NL, Smith JD, Schnabel RB, Soliman EZ, Rice KM, Van Wagoner DR, Beckmann BM, van Noord C, Wang K, Ehret GB, Rotter JI, Hazen SL, Steinbeck G, Smith AV, Launer LJ, Harris TB, Makino S, **Nelis M**, Milan DJ, Perz S, Esko T, Kottgen A, Moebus S, Newton-Cheh C, Li M, Mohlenkamp S, Wang TJ, Linda Kao WH, Vasan RS, Nothen MM, Macrae CA, Ch Stricker BH, Hofman A, Uitterlinden AG, Levy D, Boerwinkle E, Metspalu A, Topol EJ, Chakravarti A, Gudnason V, Psaty BM, Roden DM, Meitinger T, Wichmann HE, Witteman JC, Barnard J, Arking DE, Benjamin EJ, Heckbert SR, Kaab S. Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet 42(3):240–4.

Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chevre JC, Lecoeur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N, Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Beri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, **Nelis M**, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjostrom L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Mannik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463(7281):671–5.

Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, McKay JD, Lathrop M, McLaughlin J, Wang Y, Tsao MS, Spitz MR, Wang Y, Krokan H, Vatten L, Skorpen F, Arnesen E, Benhamou S, Bouchard C, Metspalu A, Vooder T, **Nelis M**, Välk K, Field JK, Chen C, Goodman G, Sulem P, Thorleifsson G, Rafnar T, Eisen T, Sauter W, Rosenberger A, Bickeböller H, Risch A, Chang-Claude J, Wichmann HE, Stefansson K, Houlston R, Amos CI, Fraumeni JF Jr, Savage SA, Bertazzi PA, Tucker MA, Chanock S, Caporaso NE. 2009. A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome

5p15 Associated with Risk for Adenocarcinoma. Am J Hum Genet 85(5):679–91.

Nelis M*, Esko T*, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskackova T, Balascak I, Peltonen L, Jakkula E, Rehnstrom K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgar N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julia A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. 2009. Genetic structure of Europeans: a view from the North-East. PLoS ONE 4(5):e5472.

Canova C, Hashibe M, Simonato L, **Nelis M**, Metspalu A, Lagiou P, Trichopoulos D, Ahrens W, Pigeot I, Merletti F, Richiardi L, Talamini R, Barzan L, Macfarlane GJ, Macfarlane TV, Holcatova I, Bencko V, Benhamou S, Bouchardy C, Kjaerheim K, Lowry R, Agudo A, Castellsague X, Conway DI, McKinney PA, Znaor A, McCartan BE, Healy CM, Marron M, Brennan P. 2009. Genetic associations of 115 polymorphisms with cancers of the upper aerodigestive tract across 10 European countries: the ARCAGE project. Cancer Res 69(7):2956–65.

Lagiou P, Georgila C, Minaki P, Ahrens W, Pohlabeln H, Benhamou S, Bouchardy C, Slamova A, Schejbalova M, Merletti F, Richiardi L, Kjaerheim K, Agudo A, Castellsague X, Macfarlane TV, Macfarlane GJ, Talamini R, Barzan L, Canova C, Simonato L, Lowry R, Conway DI, McKinney PA, Znaor A, McCartan BE, Healy C, **Nelis M**, Metspalu A, Marron M, Hashibe M, Brennan PJ. 2009. Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. Eur J Cancer Prev 18(1):76–84.

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 8:159.

Montpetit A*, **Nelis M***, Laflamme P, Mägi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A. 2006. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet 2(3):e27.

*These authors contributed equally to this work.

ELULOOKIRJELDUS

Mari Nelis

Sünniaeg ja koht:	15. oktoober 1979, Tartu
Aadress:	Biotehnoloogia õppetool, Molekulaar- ja Rakubioloogia Instituut, Tartu Ülikool Riia 23, 51010, Tartu, Eesti
Telefon:	+3727375880
E-mail:	mari.nelis@ut.ee

Hariduskäik

1986–1998	Leisi Keskkool
1998–2002	Bakalaureuse kraad molekulaardiagnostika erialal, bioloogia-geograafiateaduskond, Tartu ülikool
2002–2004	Teadusmagistrikraad molekulaardiagnostika erialal, bioloogia-geograafiateaduskond, Tartu ülikool
2004–2009	Doktorant, loodus- ja tehnoloogiateaduskond, Tartu ülikool

Erialane teenistuskäik

2001–2002	Asper Biotech, laborant
2002–2003	Eesti Geenivaramu, laborant
2002–2007	Molekulaar- ja Rakubioloogia Instituut, Tartu ülikool, koostööprojekti ARCAGE Eesti poolne koordinaator
2006–…	Eesti Biokeskus, genotüpiseerimise tuumiklabori juhataja

Teadustegevus

Magistrantuuri õpingute ajal osalesin suuremahulise juht-kontrolluuringu genotüpiseerimise protsessis ning hilisemas andmete analüüsimises. Antud projekti eesmärgiks oli leida seoseid kandidaatgeenide ja ülemiste hingamisteede-, söögitoru- ning kopsukasvaja tekkimise vahel. Edasises teadustöös olen tegelenud peamiselt populatsioonigeneetikaga – uurinud inimese genoomi variatsioone ja nende mõju fenotüübile. Olen põhjalikumalt uurinud Euroopa populatsioonide geneetilist kaugust ja selle mõju kogu-genoomi hõlmavatele assotsiatsioonuuringutele.

Publikatsioonid

Ellinor PT, Lunetta KL, Glazer NL, Pfeufer A, Alonso A, Chung MK, Sinner MF, de Bakker PI, Mueller M, Lubitz SA, Fox E, Darbar D, Smith NL, Smith JD, Schnabel RB, Soliman EZ, Rice KM, Van Wagoner DR, Beckmann BM, van Noord C, Wang K, Ehret GB, Rotter JI, Hazen SL, Steinbeck G, Smith AV, Launer LJ, Harris TB, Makino S, **Nelis M**, Milan DJ, Perz S, Esko T, Kottgen A, Moebus S, Newton-Cheh C, Li M, Mohlenkamp S, Wang TJ, Linda Kao WH, Vasan RS, Nothen MM, Macrae CA, Ch Stricker BH, Hofman A, Uitterlinden AG, Levy D, Boerwinkle E, Metspalu A, Topol EJ, Chakravarti A, Gudnason V, Psaty BM, Roden DM, Meitinger T, Wichmann HE, Witteman JC, Barnard J, Arking DE, Benjamin EJ, Heckbert SR, Kaab S. Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet 42(3):240–4.

Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chevre JC, Lecoeur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N, Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Beri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, **Nelis M**, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjostrom L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Mannik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463(7281):671–5.

Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, McKay JD, Lathrop M, McLaughlin J, Wang Y, Tsao MS, Spitz MR, Wang Y, Krokan H, Vatten L, Skorpen F, Arnesen E, Benhamou S, Bouchard C, Metspalu A, Vooder T, **Nelis M**, Välk K, Field JK, Chen C, Goodman G, Sulem P, Thorleifsson G, Rafnar T, Eisen T, Sauter W, Rosenberger A, Bickeböller H, Risch A, Chang-Claude J, Wichmann HE, Stefansson K, Houlston R, Amos CI, Fraumeni JF Jr, Savage SA, Bertazzi PA, Tucker MA, Chanock S, Caporaso NE. 2009. A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome

5p15 Associated with Risk for Adenocarcinoma. Am J Hum Genet 85(5):679–91.

Nelis M*, Esko T*, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskackova T, Balascak I, Peltonen L, Jakkula E, Rehnstrom K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgar N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julia A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. 2009. Genetic structure of Europeans: a view from the North-East. PLoS ONE 4(5): e5472.

Canova C, Hashibe M, Simonato L, **Nelis M**, Metspalu A, Lagiou P, Trichopoulos D, Ahrens W, Pigeot I, Merletti F, Richiardi L, Talamini R, Barzan L, Macfarlane GJ, Macfarlane TV, Holcatova I, Bencko V, Benhamou S, Bouchardy C, Kjaerheim K, Lowry R, Agudo A, Castellsague X, Conway DI, McKinney PA, Znaor A, McCartan BE, Healy CM, Marron M, Brennan P. 2009. Genetic associations of 115 polymorphisms with cancers of the upper aerodigestive tract across 10 European countries: the ARCAGE project. Cancer Res 69(7):2956–65.

Lagiou P, Georgila C, Minaki P, Ahrens W, Pohlabeln H, Benhamou S, Bouchardy C, Slamova A, Schejbalova M, Merletti F, Richiardi L, Kjaerheim K, Agudo A, Castellsague X, Macfarlane TV, Macfarlane GJ, Talamini R, Barzan L, Canova C, Simonato L, Lowry R, Conway DI, McKinney PA, Znaor A, McCartan BE, Healy C, **Nelis M**, Metspalu A, Marron M, Hashibe M, Brennan PJ. 2009. Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. Eur J Cancer Prev 18(1):76–84.

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 8:159.

Montpetit A*, **Nelis M***, Laflamme P, Mägi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A. 2006. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet 2(3):e27.

* Jagatud autorlus.

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro*. Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous crassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan.** Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.

41. **Sulev Ingerpuu.** Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.
42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) — induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and erotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu 2000. 88 p.

61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu 2000. 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu 2000. 122 p.
63. **Jonne Kotta.** Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu 2000. 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000. 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000. 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu 2001. 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu 2001. 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu 2001. 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu 2001. 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002. 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002. 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002. 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002. 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002. 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003. 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003. 168 p.
79. **Viljar Jaks.** p53 — a switch in cellular circuit. Tartu, 2003. 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003. 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003. 159 p

82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003. 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003. 109 p.
84. **Ivika Ostonen.** Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003. 158 p.
85. **Gudrun Veldre.** Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003. 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004. 159 p.
87. **Aare Abroi.** The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004. 135 p.
88. **Tiina Kahre.** Cystic fibrosis in Estonia. Tartu, 2004. 116 p.
89. **Helen Orav-Kotta.** Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004. 117 p.
90. **Maarja Öpik.** Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004. 175 p.
91. **Kadri Tali.** Species structure of *Neotinea ustulata*. Tartu, 2004. 109 p.
92. **Kristiina Tambets.** Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004. 163 p.
93. **Arvi Jõers.** Regulation of p53-dependent transcription. Tartu, 2004. 103 p.
94. **Lilian Kadaja.** Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004. 103 p.
95. **Jaak Truu.** Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004. 128 p.
96. **Maire Peters.** Natural horizontal transfer of the *pheBA* operon. Tartu, 2004. 105 p.
97. **Ülo Maiväli.** Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004. 130 p.
98. **Merit Otsus.** Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004. 103 p.
99. **Mikk Heidemaa.** Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004. 167 p.
100. **Ilmar Tõnno.** The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.

102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.
103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.
105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.
106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.
109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 106 p.
114. **Mait Metspalu.** Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005. 138 p.
115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006. 124 p.
116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006. 126 p.
117. **Heili Ilves.** Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006. 120 p.
118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006. 126 p.
119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006. 90 p.
120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006. 94 p.
121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006. 186 p.

122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007. 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007. 143 p.
126. **Aveliina Helm.** Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007. 89 p.
127. **Leho Tedersoo.** Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007. 233 p.
128. **Marko Mägi.** The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007. 135 p.
129. **Valeria Lulla.** Replication strategies and applications of Semliki Forest virus. Tartu, 2007. 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007. 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007. 171 p.
132. **Tatjana Krama.** Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007.
133. **Signe Saumaa.** The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007. 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007. 96 p.
135. **Priit Kilgas.** Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007. 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007. 95 p.
137. **Kärt Padari.** Protein transduction mechanisms of transportans. Tartu, 2008. 128 p.
138. **Siiri-Lii Sandre.** Selective forces on larval colouration in a moth. Tartu, 2008. 125 p.
139. **Ülle Jõgar.** Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008. 99 p.
140. **Lauri Laanisto.** Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008. 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008. 105 p.

142. **Birgot Paavel.** Bio-optical properties of turbid lakes. Tartu, 2008. 175 p.
143. **Kaire Torn.** Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg.** Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd.** Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag.** Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu.** Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm.** Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks.** Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu.** Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš.** Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko.** Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast.** Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats.** Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova.** The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik.** Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2008, 112 p.
158. **Kaja Kiiver.** Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja.** Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast.** Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.

161. **Ain Vellak.** Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel.** Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe.** Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe.** Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa.** Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.
166. **Pille Säälik.** The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil.** Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik.** Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark.** Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap.** Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan.** Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe.** Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi.** Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Venda Lauringson.** Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts.** Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.