DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
225

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
225

TÕNU ESKO

Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies

UNIVERSITY OF TARTU
PRESS
1632

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

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in Gene Technology) on 14.09.2012 by the Council of the Institute of Molecular and Cell Biology, University of Tartu

Supervisor: Professor Andres Metspalu, MD, PhD
Department of Biotechnology, Institute of Molecular and Cell Biology, and the Estonian Genome Center of University of Tartu,
University of Tartu, Estonia

Opponent: Professor Tiina Paunio, MD, PhD
National Institute of Health and Welfare,
University of Helsinki, Finland

Commencement: Room No 105, Institute of Molecular and Cell Biology, University of Tartu; 23b Riia Str., Tartu, on October 19th, 2012, at 16.00

The University of Tartu grants the publication of this dissertation.

This research is financially supported by FP7 programs (ENGAGE, OPENGENE, BBMRI, ECOGENE, LifeSpan), Estonian Government SF0180142s08, Estonian Science Foundation (7859), Estonian Research Roadmap through Estonian Ministry of Education and Research, Center of Excellence in Genomics and University of Tartu (SP1GVARENG).

ISSN 1024–6479
ISBN 978–9949–32–116–2 (print)
ISBN 978–9949–32–117–9 (pdf)

University of Tartu Press
www.tyk.ee
Order No. 451

In memory of my beloved mother

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

Ref.I 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.

Ref.II **Esko T***, Mezzavilla M*, Nelis M, Borel C, Debniak T, Jakkula E, Julia A, Karachanak S, Khrunin A, Kisfali P, Krulisova V, Kučinskienė Z, Rehnström K, Traglia M, Nikitina-Zake L, Zimprich F, Antonarakis S, Estivill X, Glavač D, Gut I, Klovins J, Krawczak M, Kučinskas V, Lathrop M, Macek M, Marsal S, Meitinger T, Melegh B, Limborska S, Lubinski J, Paolotie A, Schreiber S, Toncheva D, Toniolo D, Wichmann E, Zimprich A, Metspalu M, Gasparini P*, Metspalu A*, D'Adamo P*. (2012). "Genetic diversity of northeastern Italian population isolates". *Eur J Hum Genet* in press.

Ref.III McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, **Esko T**, Corre T, Davies G, Kaakinen M, Lyytikäinen LP, Kristiansson K, Havulinna AS, Gögele M, Vitart V, Tenesa A, Aulchenko Y, Hayward C, Johansson A, Boban M, Ulivi S, Robino A, Boraska V, Igl W, Wild SH, Zgaga L, Amin N, Theodoratou E, Polašek O, Girotto G, Lopez LM, Sala C, Lahti J, Laatikainen T, Prokopenko I, Kals M, Viikari J, Yang J, Pouta A, Estrada K, Hofman A, Freimer N, Martin NG, Kähönen M, Milani L, Heliövaara M, Vartiainen E, Räikkönen K, Masciullo C, Starr JM, Hicks AA, Esposito L, Kolčić I, Farrington SM, Oostra B, Zemunik T, Campbell H, Kirin M, Pehlic M, Faletra F, Porteous D, Pistis G, Widén E, Salomaa V, Koskinen S, Fischer K, Lehtimäki T, Heath A, McCarthy MI, Rivadeneira F, Montgomery GW, Tiemeier H, Hartikainen AL, Madden PA, d'Adamo P, Hastie ND, Gyllensten U, Wright AF, van Duijn CM, Dunlop M, Rudan I, Gasparini P, Pramstaller PP, Deary IJ, Toniolo D, Eriksson JG, Jula A, Raitakari OT, Metspalu A, Perola M, Järvelin MR, Uitterlinden A, Visscher PM, Wilson JF; on behalf of the ROHgen Consortium. (2012). "Evidence of inbreeding depression on human height". *PLoS Genet* 8(7): e1002655.

Ref.IV Allebrandt K*, Amin N*, Müller-Myhsok B, **Esko T**, Teder-Laving M, Azevedo R, Hayward C, van Mill J, Vogelzangs N, Green E, Melville S, Lichtner P, Wichmann E, Oostra B, Janssens A, Campbell H, Wilson J, Hicks A, Pramstaller P, Dogas Z, Rudan I, Merrow M, Penninx B, Kyriacou C, Metspalu A, van Duijn C, Meitinger T, Roenneberg T. (2011). "A K(ATP) channel gene effect on sleep duration: from genome-wide association studies to function in Drosophila". *Mol Psychiatry* doi: 10.1038/mp.2011.142. [Epub ahead of print]

Ref.V Day-Williams AG, Southam L, Panoutsopoulou K, Rayner NW, **Esko T**, Estrada K, Helgadottir HT, Hofman A, Ingvarsson T, Jonsson H, Keis A, Kerkhof HJ, Thorleifsson G, Arden NK, Carr A, Chapman K, Deloukas P, Loughlin J, McCaskie A, Ollier WE, Ralston SH, Spector TD, Wallis GA, Wilkinson JM, Aslam N, Birell F, Carluke I, Joseph J, Rai A, Reed M, Walker K; arcOGEN Consortium, Doherty SA, Jonsdottir I, Maciewicz RA, Muir KR, Metspalu A, Rivadeneira F, Stefansson K, Styrkarsdottir U, Uitterlinden AG, van Meurs JB, Zhang W, Valdes AM, Doherty M, Zeggini E. (2011). "A variant in MCF2L is associated with osteoarthritis". *Am J Hum Genet* 89(3): 446–50

*These authors contributed equally to this work.

Author's contributions:

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

Ref. IV participated in the Estonian Biobank specific study design, analyzed the the Estonian Biobank data, performed in part the meta-analyses, participated in preparation of the paper

Ref. III, V participated in the Estonian Biobank specific study design, analyzed the Estonian Biobank data, participated in the critical review of the paper

LIST OF ABBREVIATIONS

GWAS	genome-wide association study
LD	linkage disequilibrium
MAF	minor allele frequency
OR	odds ratio
PC	principal component
PCA	principal component analysis
SNP	single nucleotide polymorphism
tagSNP	tagging SNP

INTRODUCTION

A decade ago, the first draft sequence of the human genome was published. Rather than being an endpoint to the human genome sequencing project, however, this event became a stepping stone for further refinement of biological information and for seeking the medically relevant implications of such data. In the earliest attempt to understand the role of the genomic sequence in biological characteristics, genetic variation was catalogued by single nucleotide polymorphisms (SNPs) by the international SNP Consortium. This initiative was followed by the International HapMap Project, which sought to determine the haplotype structure of the human genome. Comprehensive cataloging of DNA sequence variants, in turn, helped to further evolve the genome sequencing and bioinformatic technologies. Development of high-throughput genotyping arrays enabled cost-effective genotyping of millions of SNPs in a large number of samples. Advanced statistical methods and software tools opened the door for genome-wide association study (GWAS) to effectively seek the genetic variants that underlie the dynamic complexity of human phenotypes.

All these advances made it possible to analyze the genome without any biological priors and enabled the discovery of new pathways and biological mechanisms, which not only provided insights into human traits but also into disease etiology. Most often, the former is achieved by GWAS of a continuous phenotype in a population-based sample, while the latter is achieved by GWAS comparison of genetic variant allele frequency between disease cases and matched healthy controls. Since genetic effect sizes are relatively small and diseases are often heterogeneous, extremely large sample sizes (up to tens and hundreds of thousands) are needed to attain the statistical power necessary to detect sequence variants affecting trait variance or disease susceptibility.

Over the last five years, the number of validated complex human trait-associated loci has exceeded 3,000 independent genetic variants related to more than 600 distinct traits and diseases. However, even after doubling the number of disease associated genes and discovering new underlying biological pathways of disease pathogenesis, the potential of genome-wide association studies has not yet been fully realized.

The thesis work presented herein begins with a literature overview, which will address the important milestones that have been reached in understanding the genetic architecture of complex human phenotypes. First, an overview of the Estonian Biobank developed by the Estonian Genome Center of the University of Tartu will be presented. Then, a review will follow that outlines the value of appropriate statistical power and study design to GWAS and presents important findings from large-scale genome-wide association studies. Finally, the causes of hidden heritability in GWAS and the approaches used to demonstrate the full impact of human genetic variation on a phenotype will be discussed. The research portion of this thesis will focus on the following issues: 1) to fill in the gaps of the genetic structuring of northeastern European populations; 2) to evaluate the genetic structure of different European populations; 3) to identify

novel DNA sequence variants that confer phenotype variability and disease predisposition and 4) to investigate the problem of hidden heritability in the genetics of complex traits.

1. REVIEW OF LITERATURE

1.1. The biobank of the Estonian Genome Center

1.1.1. Population based biobanks

After publishing of the draft sequence of human genome (Lander et al., 2001; Venter et al., 2001), the emphasis moved towards understanding the structure, organization and function of genomes and the biological basis of complex human traits. Genetics has been traditionally associated with rare hereditary Mendelian disorders and studies involving linkage analysis, positional cloning and search for mutations in single genes. The new knowledge obtained from the human sequencing project and new emerging technologies created an opportunity to study DNA sequence variation on a genome-wide scale and opened the field to studies on common complex diseases (Lander, 2011).

The need to extend the genetic studies to the population-wide scale in order to conduct genetic research on common complex diseases was first recognized by Risch and Merikangas (Risch and Merikangas, 1996). A population-based cohort design would give several advantages over familial studies or case-control design in discovering the DNA sequence variants that have an effect on normal phenotype variability or increased disease susceptibility. For example, a prospective population-based cohort would enable to design several nested case-control analyses and to study many different conditions and endpoints, as comprehensive phenotype data is available for all samples. Furthermore, these studies incorporate information about environmental exposure prior to development of the disease. Lastly, a population-based large-scale cohort would provide large enough sample sizes for achieving sufficient statistical power to find genetic variants even with subtle effects (Collins et al., 2003).

In order to effectively conduct genetic research a population-wide data collection is needed with both collections of human biological samples and associated comprehensive clinical and lifestyle information. This and the opportunity to use new technologies, electronic health records and IT solutions lead to the establishment of new population based biobanks (Kohane, 2011). So far many of the traditional epidemiological cohorts did not collect DNA or did not have a proper informed consent. The governing ethics of such collections soon became a highly debated topic. The subsequent requirement of a new, rather broad informed consent (Knoppers, 2001; Deschênes et al., 2001) emerged as a challenge for biobanks. A large international consortia, named The Public Population Project in Genomics, was formed to lead, catalyze and coordinate the international efforts and expertise in developing and setting up the legal, ethical, and infrastructural frameworks for population-wide biobanks (www.p3gobservatory.org). The United Kingdom Biobank (UK Biobank, 2011), Icelandic Biobank (deCODE Genetics, 2010) and the Estonian Biobank were established to become some of the first "new-generation" biobanks. Most of the European biobanks are part of Biobanking and Biomolecular Resources

Research Infrastructure consortium (www.bbmri.eu), which facilitates the pan-European collaboration (Wichman et al., 2011).

The biobanks have been recognized as a powerful platform for health innovation and knowledge generation, thus having a pivotal role in elucidating disease etiology, translation, and advancing public health (Harris et al., 2012). In a longer perspective the biobanks are seen as the cornerstones in leading the paradigm change in healthcare, mostly known as 4P medicine – Predictive, Preventive, Personalized and Participatory (Hood et al 2004; Bousquet et al., 2011).

1.1.2. Study design and sample collection

The Estonian Biobank is part of the Estonian Genome Center, which is an institution of the University of Tartu and whose mission is to create a large biobank composed of a wide range of health information, biological samples, and high-resolution genomics data from the Estonian population (www.biobank.ee). The principal objectives of the Biobank are to advance genetic knowledge through scientific research and to promote general public health through genome-based medicine (Metspalu et al., 2011). The epidemiological and clinical sample collections gathered in Estonia prior to the Estonian Biobank were relatively small or did not meet the legal requirements that were borne from the genomic era. Unfortunately, creating an amalgamated collection of the different cohorts was not possible due to differences in the originating study design, the wide range of assessment methodologies used and for subjective reasons. These aspects would have limited the possibility to analyze different cohorts together and could have introduced a systematic bias into any results obtained (Metspalu, 2004). Yet, a large Estonian population-based sample collection was needed to effectively answer the genetics-driven questions about the common complex diseases that arose when the entire sequence of the human genome was determined.

The Estonian Biobank was designed as a prospective, longitudinal, population-based database of large numbers of health records and accompanying biological samples. It was established in 2001 and the legal, ethical, and infrastructural frameworks were carefully designed to meet the public requirements (Metspalu, 2004). The Estonian Biobank questionnaire was developed according to the framework of the European Prospective Investigation into Cancer and Nutrition (Riboli and Kaaks, 1997; Kaaks et al., 1997). The 320 questions of the questionnaire are designed to obtain personal, genealogical and lifestyle data, as well as educational and occupational history. Medical history and current health status are recorded in accordance with the World Health Organization's 10^{th} release of the International Classification of Diseases (www.who.int/classifications/icd) together with diagnosis reliability scores (EGCUT, 2012). Figure 1 outlines the Estonian Biobank questionnaire modules.

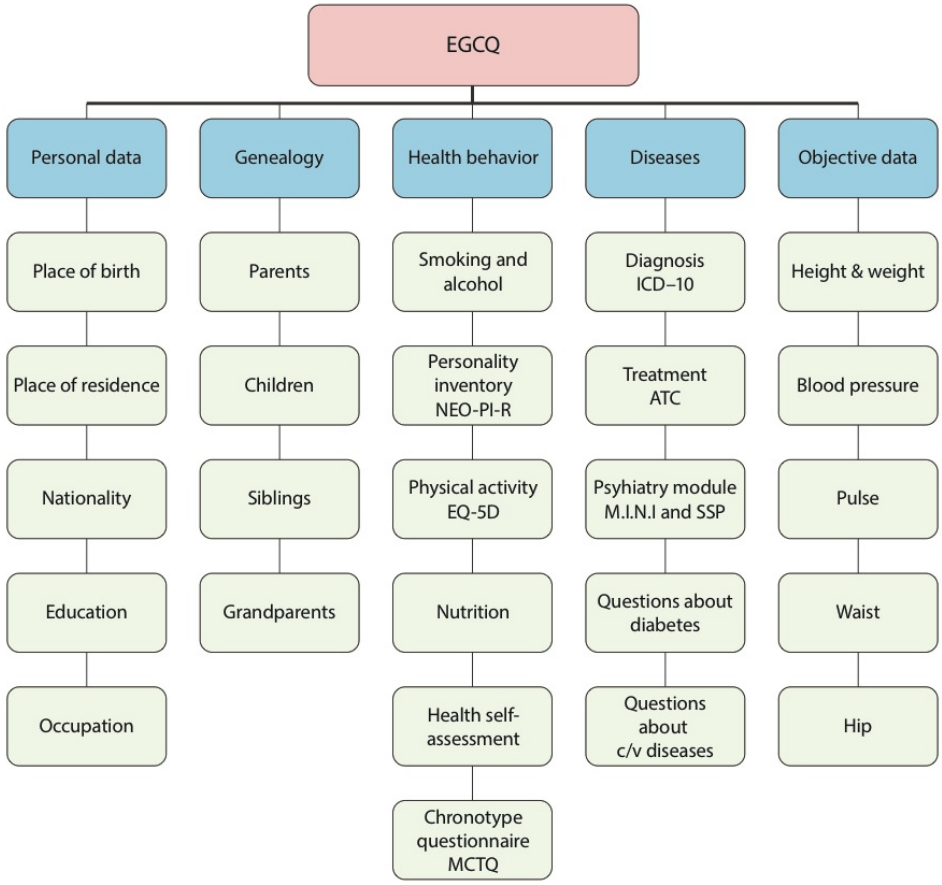

Figure 1. Structure and content of the Estonian Biobank questionnaire. The questionnaire gathers information in five main categories (blue): personal data, genealogy, health behavior, diseases, and objective data (EGCUT, 2012).

For the recruitment of the Biobank participants, a unique network of data collectors was established. This network consisted of family physicians (involving around half of the general practitioners in Estonia) and other medical personnel in private practices, hospitals or recruitment offices. Engaging experienced medical professionals was expected to ensure the highest possible quality data and to allow for incorporation of pre-existing medical records, thereby further increasing the accuracy of the collected data.

To date, the Biobank phenotype information is periodically updated by accessing various databases of healthcare institutions and registries, or by re-contacting the participants directly. The broad informed consent for participation provides ethical and legal rights to verify and supplement the database in such a manner (EGCUT, 2012).

1.1.3. Brief description of the Estonian Biobank cohort

The phase of active recruitment for the Biobank participants was completed by the end of 2010, and yielded more than 50,000 donors aged 18 years or older. While the Biobank represents the Estonian population quite well, the male-to-female ratio is not reflecting that of the population and some age groups are under- or overrepresented (Figure 2). Altogether, there are currently a total of 372,892 diagnoses, which translates to an average of 7.6 diagnoses per participant. Almost all of the diagnosed diseases in biobank have approximately the same prevalences as reported for the general Estonian population.

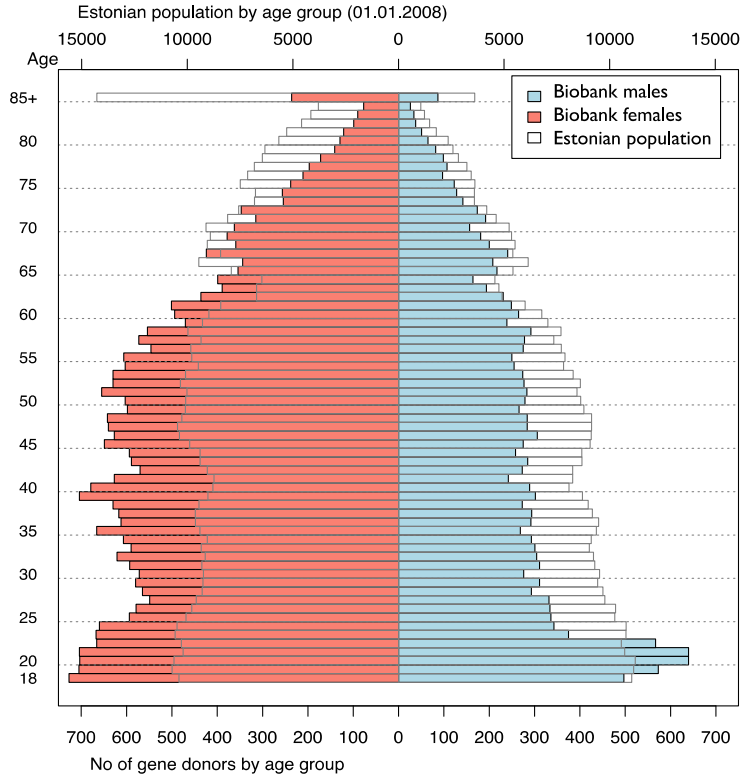

Figure 2. Age and sex distribution of the Estonian Biobank participants at recruitment, compared to the general Estonian adult population. Counts at the top of the graph indicate the number of individuals in the general Estonian adult population. Counts at the bottom indicate the number of biobank participants (EGCUT, 2012).

The 51,534 biobank participants of the Estonian Biobank encompass approximately 5% of the adult population of Estonia (EGCUT, 2012). The overall size of the study cohort is not exceptional compared to the other biobanks. The Biobank Japan Project is composed of ~200,000 participants, while the Californian Kaiser Permanente Study in the USA includes ~400,000, and both

the United Kingdom Biobank (Kohane, 2011) and the China Kadoorie Biobank each have ~500,000 (Chen et al., 2011). However, the Estonian Biobank is one of the few large-scale population-based collections, which is collected according to the same protocol. Several of the earlier population-based cohorts were collections of smaller studies with different protocols applied (Kohane, 2011). The deCODE Genetics biobank (www.decode.com), which incorporates information for about 40% of the Iceland population (deCODE Genetics, 2010), is an excellent example how a comprehensively designed biobank enables important discoveries (>200 published papers), technological advances (product beta-tester for the sequencing-by-synthesis platform), and development of sophisticated analysis methods (Kong et al., 2010; Kong et al., 2012).

1.2. Genome-wide association studies

The ultimate goals of human genetics are to understand the genetic architecture of complex traits and to translate the genetic findings into the medical field in order to improve diagnosis and treatment. These data are also expected to aid in the development of more efficient drugs and optimal dosages, as well as to facilitate proactive measures based on risk prediction and prevention strategies. Using the sequence information of the human genome, large-scale and high-throughput studies in human genetics are necessary to achieve these goals (Guttmacher and Collins, 2003).

Common diseases and complex traits, such as height, blood pressure, or plasma lipids levels, are difficult to study since they result from numerous genetic and environmental factors. Although these traits cluster in families and show considerable levels of heritability (Boomsma et al., 2002), they do not follow the typical Mendelian heritance patterns and are referred to as complex traits. The analyses of these traits are complicated further by the fact that many of them follow a polygenic model, in which tens, if not hundreds (or even thousands) of genes regulate the end phenotype (Gambaro et al., 2000).

Single nucleotide polymorphisms (SNPs) have proven useful for linking the genetic background with certain phenotypic conditions. Unlike many other genetic markers (e.g. restriction fragment length polymorphisms, microsatellites and minisatellites, or structural variants), SNPs are the easiest to genotype and well suited to high-throughput detection methods (Wang et al., 1998). This is because SNPs are bi-allelic, occur approximately once per 300 base pair (Sachidanandam et al., 2001), and have a substantially low mutation rate (Jorde et al., 2000). These estimates have been verified by research studies over the past decade and specified by the latest high-throughput sequencing study (1000 Genomes Project Consortium, 2010).

Linkage mapping in families, as well as in genome-wide association analysis of unrelated samples, accounts for the fact that DNA is inherited in blocks of sequence, and that within a single block there exists a strong allelic association and linkage disequilibrium (LD) between the genetic variants (Chapman and

Wijsman, 1998). Analyses of chromosome-wide SNP genotype data confirmed that the genome has a block-like structure (Daly et al., 2001; Patil et al., 2001; Dawson et al., 2002). Detected haplotype blocks were characterized as sizable regions in the genome with low recombination rates, and in most cases a limited number of haplotypes was found to be present in a particular population (Gabriel et al., 2002). The haplotype block structure enabled selection and genotyping of only a fraction of the SNPs (known as tagSNPs) to identify haplotypes as representatives of all the underlying SNP genotypes, while several algorithms were developed to select the most appropriate tagSNPs to genotype (Gabriel et al., 2002; Carlson et al., 2004; de Bakker et al., 2005).

Over the last decade the scientific community has invested heavily into describing the genetic landscape of the human genome. The HapMap Project (www.hapmap.com) genotyped 2.4 million SNPs in three large ethnic groups (International HapMap Consortium, 2005 and 2007). The ENCODE Project (www.genome.ucsc.edu/ENCODE) completed deep sequencing of approximately 1% of the human genome in an attempt to discover all functional elements present in those regions (ENCODE Project Consortium, 2007). Recent discoveries from the ENCODE Project highlight that up-to 80% of the non-coding portion of human genome is full of functional elements and regulatory motifs (ENCODE Project Consortium, 2012 [and see references within]; Gerstein et al., 2012; Neph et al., 2012). Most recently, the 1000 Genomes Project (www.1000genomes.org) set forth to sequence the whole-genome of 2500 samples from diverse populations, and the pilot phases have already been completed (1000 Genomes Project Consortium, 2010; Marth et al., 2011). The data from these projects represent rich sources from which to select the optimal panel of tagSNPs and manufacture high-throughput and cost-effective genotyping arrays, which effectively cover at least 80% of the genome (Barret and Cardon, 2006; Pe'er et al., 2006; Mägi et al., 2007).

A genome-wide association study can be considered an extension of the classical candidate gene study, where the difference in allele frequency is being tested between cases and controls. While GWAS approaches the genome without any prior information, the classical approach relies on previous knowledge about underling biological pathways. Therefore, candidate gene studies of diseases and traits with poorly described or unknown biological mechanisms can be markedly biased (Reich and Lander, 2001). The GWAS approach of genotyping hundreds of thousands or even millions of SNPs in well-characterized, large cohorts overcomes this limitation and allows for the hypothesis-free discovery of genetic variants that modulate complex traits in humans.

1.2.1. Sample size and power

Commercial genotyping arrays remain cost-effective alternatives to traditional methods, but are still considerably expensive; therefore, it is crucial to generate an optimal study design (Spencer et al., 2009). Each experimental study should gain sufficient statistical power (usually 80%) to identify an association

between a SNP and the trait of interest. The power of a GWAS is influenced by a host of factors, including study sample size, the susceptibility locus, minor allele frequency of the effect variant, LD strength between the tagSNP and the causative variant, and the burden of multiple testing (Cardon and Bell, 2001). In allele frequency-based tests, a clear reverse correlation exists between the study sample size and LD (measured by r^2) for a tagSNP and a causative variant that is required to achieve a certain level of power (Pritchard and Przeworski, 2001). When there is a perfect correlation (r^2 =1.0) between the tested and causative SNP, a sample size of N is needed; however, perfect correlation is rarely found and the sample size required scales up exponentially (N/r^2) (Wang et al., 2005). In case-control studies, the effect of the susceptibility variant is measured by the odds ratio (OR), which is defined as the odds of a case being exposed to the susceptible genetic variant compared with that in controls. Figure 3 illustrates the effects of allele frequency on the required sample size. Several software tools have been developed to estimate the required sample size for different analytical scenarios (Skol et al., 2006; Menashe et al., 2008).

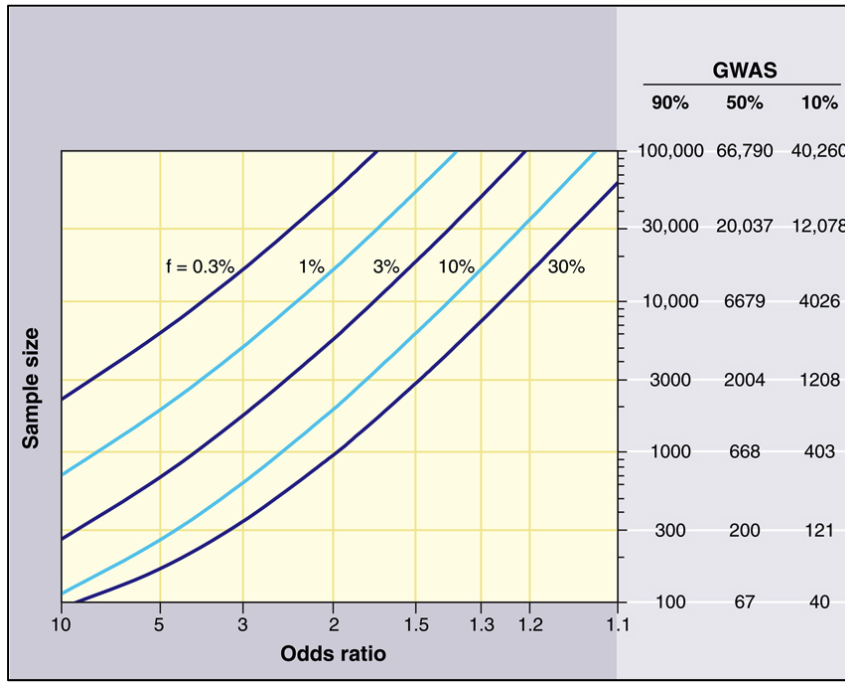

Figure 3. The number of cases required in an association study for ranges of allelic ORs with statistical power of 90%, 50% and 10% at a significance level of $P = 1x10^{-8}$ (adapted from Altshuler et al., 2008). The extremely low significance level is due to the multiple testing-burden of analyzing hundreds of thousands of markers. The significance level of $P = 1x10^{-8}$ represents a finding expected by chance once per 20 GWASs (Altshuler et al., 2008). *f* indicates the minor allele frequency for a tested DNA sequence variant.

It is important to note that a study's power is mostly affected by the OR of the underling disease-susceptibility variant (Wang et al., 2005). There is much speculation as to the underlying allele frequency spectrum of causative alleles and the according effect size distribution (Reich and Lander, 2001; Terwilliger and Weiss, 2003). Functionally replicated candidate gene studies of common complex diseases have shown that the ORs are in the order of 1.1 to 1.5 and the distribution is biased towards smaller effects (Ionnadis et al., 2003; Lohmueller et al., 2003). Theoretical estimations and empirical data of GWASs have verified that tens and hundreds of thousands of samples are needed to robustly detect and replicate common disease susceptibility variants (Hindorff et al., 2009; NHGRI GWAS Catalog, 2012).

Finally, the required sample size can be further increased when several suboptimal study conditions are present, such as weak effects, rare alleles in incomplete LD with a tagSNP, ascertainment bias, improper selection of controls, and population stratification (Wang et al., 2005).

1.2.2. Population stratification

Very large sample sizes are required to detect SNPs with modest effects, and population-based cohorts have been used to scale up the sample size (Risch and Merikangas, 1996). In large cohorts, the presence of substructure, while undetected, can mimic the signal of association and lead to false positive associations or masking of the real signals (Cardon and Bell, 2001; Freedman et al., 2004). When a studied sample includes subpopulations that differ both genetically and on the disease prevalence, then the proportions of cases and controls sampled from each of the subpopulations can be different and the allele frequencies will be systematically different in any loci where the two subpopulations differ (Marchini et al., 2004).

The effect of stratification was demonstrated when analysis on height was carried out in samples of European ancestry and a lactose intolerance associated variant (Enattah et al., 2002) was showing a strong association (Campbell et al., 2005). Both taller individuals and lactose tolerance are more frequent in Northern Europe (Bersaglieri et al. 2004). However, the association was lost when the potential confounding factor of grandparental ancestry was corrected for (Campbell et al., 2005).

In many GWAS studies, the cases are systematically characterized but the controls are not, and may even be obtained from a public databases (Nelson et al., 2008). Even a small fraction of stratification (10% of controls) can cause bias (Marchini et al., 2004). This problem increases with lower minor allele frequencies (<5%) and is pronounced in rare-variant analyses (Morris and Zeggini, 2010; Mathieson and McVean, 2012).

Stratification and presence of cryptic relatedness leads to inflated type I error (Voight and Pritchard, 2005). Several mathematical models and software tools have been developed to correct for hidden population structure; the most conservative of which is known as the genomic control method (Devlin and

Roeder, 1999). This method assumes that stratification changes the null distribution of the test statistic by a multiplicative factor λ, and therefore all statistics are uniformly corrected. However, this approach can overcorrect any loci not affected by stratification (Price et al., 2006). The most used method to correct stratification in GWAS is principal component analysis (PCA), which enables systematic correction of only the loci with different allele frequencies between the subpopulations (Price et al., 2006; Patterson et al., 2006). In this method, the covariance due to past demographical events is captured by a few eigenvectors, so that all of the other covariates reflect sampling noise (Price et al., 2006; Roeder and Luca, 2009). Discriminant analyses of principal components (Jombart et al., 2010) and spatial ancestry analyses (Yang et al., 2012) have recently been developed for fine scale population structure analyses. Finally, the non-hierarchical cluster analysis (Pritchard et al., 2000) and unsupervised maximum likelihood-based clustering algorithms (Alexander et al., 2009) are used mostly to study population demographic history (Behar et al., 2010; Metspalu et al., 2011) but less so for correcting stratification in association analyses.

One of the requirements for GWAs studies has been a replication in an independent, equally powered sample (Cardon and Bell, 2000). A population that is closest genetically to the test sample holds the highest probability to achieve a successful replication (Marchini et al., 2004). Therefore, it is crucial to know the genetic structure of and genetic distance between the discovery and replication populations. The availability of high-density genotypes for many individuals sampled from geographically diverse populations has made it possible to precisely estimate such distances. PCA and unsupervised clustering-based methods have unambiguously demonstrated a high correlation between the genetic clustering of studied populations and their respective geographical distances. The structure of the genetic variation has been analyzed on global (Jakobsson et al., 2008; Li et al., 2008) and continental scales (Novembre et al., 2008; Lao et al., 2008; Heath et al., 2008; Tian et al., 2008; Tishkoff et al., 2009), as well as among ethnic groups (such as Jewish (Behar et al., 2010)), in population isolates (Jakkula et al., 2008; Price et al., 2009) and general populations (O'Dushlaine et al., 2010). The genetic structure maps illustrate that under the spatial models in which migration and gene flow occur in a homogeneous manner over short distances, the similarity between estimated genetic distances and geography is high. This regularity is known already from the seminal studies using a limited number of genetic markers (Menozzi et al., 1978; Cavalli-Sforza et al., 1994) but at the same time the genome-wide allele frequency data provides the necessary resolution for detecting the subtle structuring within a community or geographical region (Wang et al., 2012).

Figure 4 illustrates the study with the most European populations (37) included to date (Novembre et al., 2008). This particular study has two limitations: 1) many populations (18) were represented by fewer than 10 samples each; and 2) some northeastern European populations, such as Estonians and Lithuanians, were not presented at all, while others, such as Finns and Latvians were represented by only one sample. This has biased the spatial structuring

estimates in northeastern and central Europe (Jakkula et al., 2008; Lao et al., 2008; Heath et al., 2008). Recent effort to systematically quantify the geographic structure of human genetic variation worldwide have shown that a larger dataset and more genetic markers are required to charactherize the relatively homogeneous population structure in Europe (Wang et al., 2012).

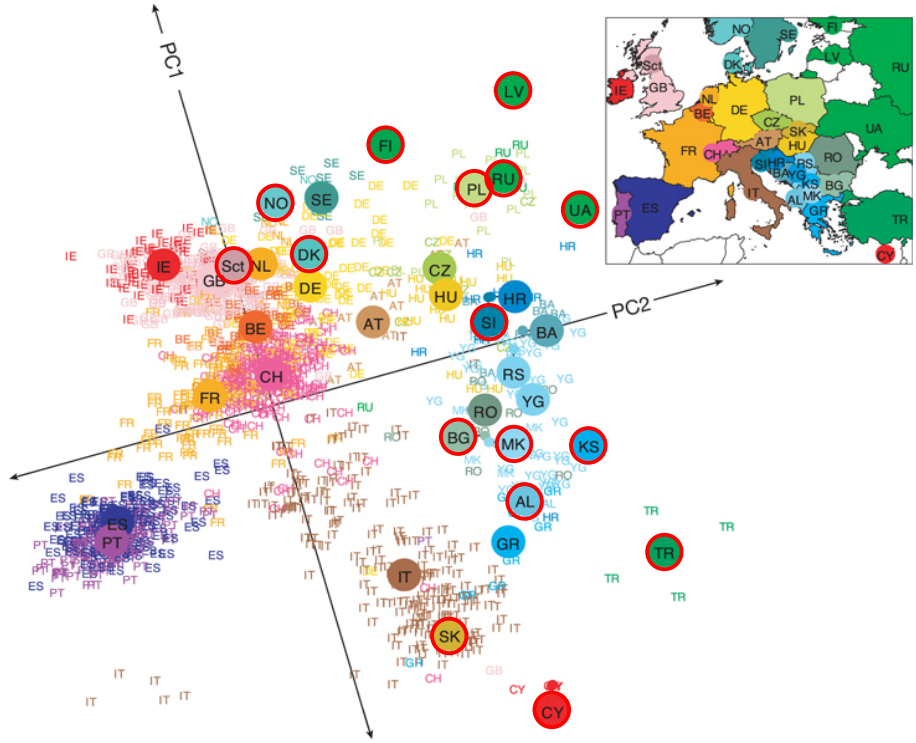

Figure 4. PCA plot of European ancestry populations. The first two principle components (PC1 and PC2) are plotted and demostrate a strong correlation between genetic and geographic distances. Small colored labels represent individuals, and large colored circles represent the median PC1 and PC2 values for each country. Colored circles with red line indicate populations that are represented with less than 6 samples. Label coloration corresponds to the geographic location on the map (inset). 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. Adapted from Novembre et al., 2008.

1.2.3. General findings from GWAS

Over past 30 years, and before the "GWAS era", the studies on complex human diseases had identified and irrefutably replicated only 50 of associated genes and respective allelic variants (Ioannidis et al., 2003; Lohmueller et al., 2003). During the first years of GWAS, the field was lead by the common disease/ common variant hypothesis, which supposed that common diseases are the result of a limited number of common alleles with moderate effect sizes that are shared among the cases (Reich and Lander, 2001). This hypothesis was partially proven by association analyses of age-related macular degradation (Klein et al., 2005; Dewan et al., 2006). However, other disease studies, with very limited numbers of analyzed cases and controls, did not find such evidence (McCarthy et al., 2008; Altshuler et al., 2008).

By the year 2007, all of the necessary theoretical models, analytical tools, and high-throughput genotyping technologies for analyzing thousands of DNA samples in a cost-effective manner were available. One of the seminal works, on which future gene discovery studies were modeled, was conducted by the Welcome Trust Case Control Consortium. This study comparatively analyzed 14,000 cases drawn from seven common diseases with 3,000 healthy controls (WTCCC, 2007). This landmark study showed that with sufficient sample size the GWAS approach is a powerful tool to robustly replicate already known risk loci (Ioannidis et al., 2003; Lohmueller et al., 2003) and to discover new ones. For only two of the diseases, bipolar disorder and hypertension, no risk variants were found; these negative results may be explained by the presence of controls that were not well-characterized, possibly including unidentified cases (Burton et al., 2009), or different effect sizes and allele frequency spectrums of risk variants between diseases (Manolio et al., 2009; Gershon et al., 2011).

The Welcome Trust Case Control Consortium study demonstrated that unrealistically large sample sizes (retrodiction-based estimation taken from Wang et al., 2005) are needed to uncover disease genes. Combining the available datasets through meta-analysis was proposed as a solution to this problem (de Bakker et al., 2008; Mägi et al., 2010). Since several commercial genotyping arrays with partially non-overlapping SNPs were used in the different studies, genotype prediction algorithms were developed that would be able to infer the missing genotypes, thereby making the different datasets comparable (Marchini et al., 2007; Willer et al., 2008; Browning and Browning, 2009). These bioinformatic methods rely on reference populations obtained from public databases (such as the HapMap Project and the 1000 Genomes Project) for imputation to infer the missing genotypes and relying upon the underling haplotype structure. This approach increased the power of meta-analyses because in many instances the tagSNPs are not the causative variants (Marchini et al., 2007). Although, the imputation accuracy depends upon SNP density as well as the similarity of LD patterns between the data used and the reference population (Marchini et al., 2007). Use of the HapMap European reference panel for imputation in Estonians (Montpetit et al., 2006) and other European populations (Marchini and Howie, 2010) is accepted as an appropriate strategy.

Familial linkage analysis studies have identified more than 2,700 genes and their respective genetic variants associated with human diseases and phenotypes (OMIM, 2012), and that number is steadily continuing to grow due to the ever-advancing sequencing technologies (Bamshad et al., 2011). By April 2012, more than 1,200 successful GWAS have been published, accounting for the identification of more than 3,000 distinct SNPs for over 600 diseases and individual traits (such as height, blood pressure, and eye color) (NHGRI GWAS Catalog, 2012).

The loci targeted by GWASs to date appear to be evenly distributed among the autosomes (Figure 5), with fewer involving the sex chromosomes (Voight et al., 2009). The sex chromosomes present unique methodological difficulties (Marchini et al., 2007), and the published studies of them lack power (Elks et al., 2010). In particular, the individual effect sizes of the associated variants are modest (OR = 1.1–1.5) and skewed towards the lower end (Hindorff et al., 2009). Regardless, most of the associated variants discovered cluster outside of exons (Hindorff et al., 2009), are significantly enriched in functional elements (Ernst et al., 2011) and are concentrated in euchromatic non-coding regulatory regions of the human genome (Maurano et al., 2012), where they often act as expression quantitative trait loci (QTL) (Fehrmann et al., 2011) and show signs of recent positive selection (Casto and Feldman, 2011; Nicholson et al., 2011).

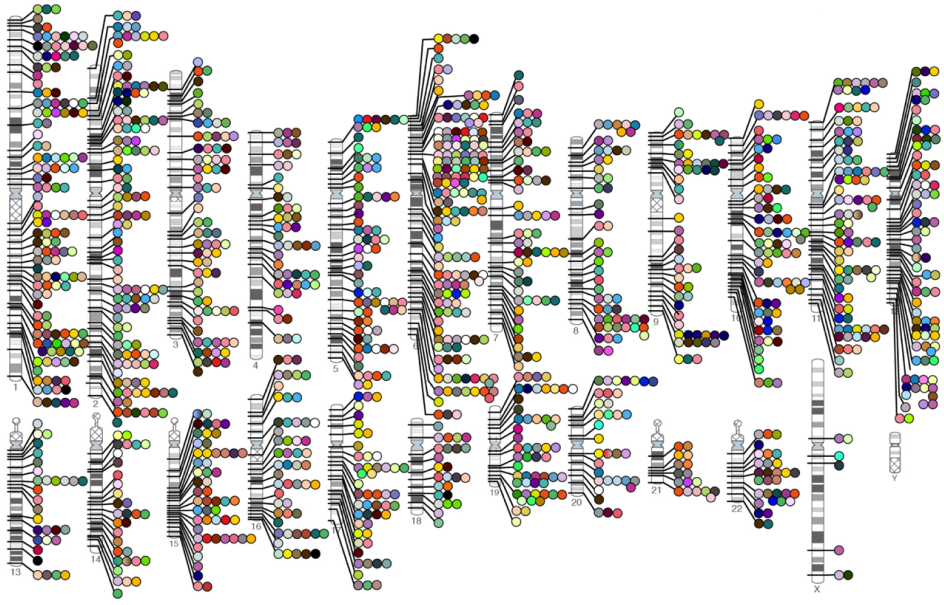

Figure 5. Karyotype plot presenting the loci identified through GWAS. The 22 autosomal and two sex chromosomes are shown. Tick marks on the chromosomes indicate the location of trait-associated loci, and the linked colored circles refer to the respective trait. The extensive legend for the trait color-coding can be found on the National Human Genome Research Institute web page (www.genome.gov/gwastudies).

Over the last five years, there has been a constant endeavor to increase the sample sizes of meta-analyses. These efforts are based on the clear linear correlation that exists between sample size and the number of newly detected associated loci; for example, doubling the sample size can lead to at least twice as many hits (Visscher et al., 2012). Three initial GWASs on human height (Weedon et al., 2008; Lettre et al., 2008; Gudbjartsson et al., 2008) identified a total of 54 robustly associated loci, some of which were found by all three studies and others were unique to each study. However, when the study samples were combined (each having ~25,000) and newly genotyped cohorts added, then the discovery sample size of more than 130,000 samples yielded 180 new loci. All of these hits were robustly replicated in an independent sample of 50,000 (Lango-Allen et al., 2010). The same tendency was found in meta-analyses of plasma lipid levels (Teslovich et al., 2010), Crohn's disease (Franke et al., 2010), and diabetes mellitus type 2 (Voight et al., 2010). Moreover, the GWASs carried out in populations of non-European ancestry have verified known loci and lead to discovery of new loci; the studies of diabetes mellitus type 2 are good examples of this (Cho et al., 2011; Saxena et al., 2012). Thus, adhering to a careful and strict study design and a stringent level for statistical significance is important to achieve robust and replicable findings (Cardon and Bell, 2000).

1.2.4. Medical applicability of established GWAS loci

The identification of disease-associated alleles may have two major implications for clinical medicine: 1) prediction of future outcomes or disease risks; or 2) revealing underlying biological pathways that may be used to develop therapeutic interventions (Hirschhorn and Gajdos, 2011). The large-scale GWASs have identified tens or even hundreds of loci for some diseases, such as Crohn's disease (Franke et al., 2010) and diabetes mellitus type 2 (Voight et al., 2010), or only a couple for others, such as schizophrenia (Ripke et al., 2011) and bipolar disorder (Sklar et al., 2011). The failures to pinpoint causal genes in neuropsychiatric disorders have been explained by differences in genetic architecture (Owen et al., 2009; Gershon et al., 2011).

The predictive values of identified genetic variants for disease outcome are improving as more loci are found (Jostins and Barrett, 2011; Wray et al., 2010) (shown in Figure 5) and are already comparable to the traditional lifestyle-driven models, such as the Framingham risk score for coronary artery disease (Kraft and Hunter, 2009). For example, in age-related macular degeneration only a limited number of variants with strong effects in complement factor H explain the majority of genetic risk (Maller et al., 2006). Although the predictive power is strong this has not yet impacted clinical management of this disease, since an effective treatment remains to be developed (Hirschhorn and Gajdos, 2011). Genetic variants usually have only small individual effects (Hindorff et al., 2009), and thus explain less than 1% of the disease risk in most

cases (Altshuler et al., 2008). In the case of inflammatory bowel disease, where different treatment is applied according to different disease subtypes, the genetic risk scores generated from more than 100 common risk variants that each have relatively modest effect sizes, but which allow for effective distinction between ulcerative colitis and Crohn's disease patients (Franke et al., 2010) and even between subclasses of these two disorders (Inflammatory Bowel Disease Genetics Consortium, unpublished data). The latest GWASs of plasma lipid levels, a major risk factor for myocardial infarction, have identified 95 phenotype-modulating loci, which in combination may explain ~25% of the genetic variance of lipid levels. When individuals were grouped according their genetic risk scores, the top quartile group showed a 44-fold increased risk of hypertriglyceridemia compared to the bottom quartile group (Teslovich et al., 2010).

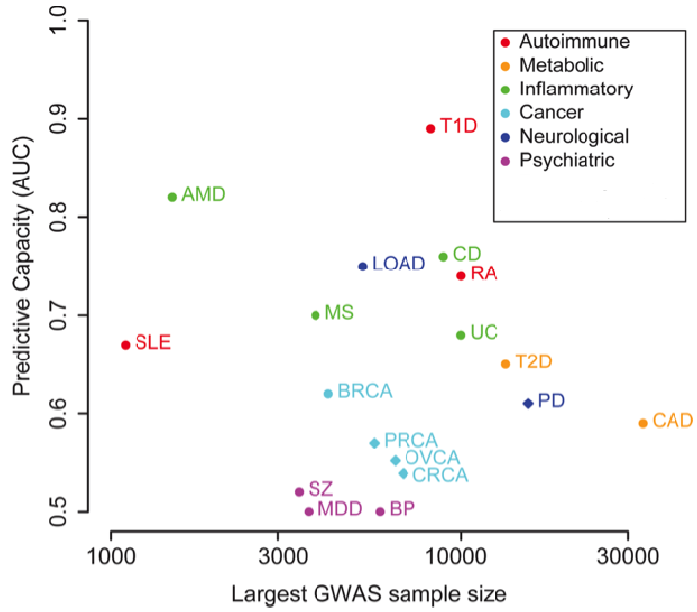

Figure 6. Disease outcome prediction using all genetic variants identified by pre- and post-GWAS era studies. PD: Parkinson's disease; AMD, age-related macular degeneration; T1D, type 1 diabetes; T2D, type 2 diabetes; UC, ulcerative colitis; CD, Crohn's disease; RA, rheumatoid arthritis; CAD, coronary artery disease; BRCA, breast cancer; LOAD, late-onset Alzheimer's disease; MS, multiple sclerosis; MDD, major depressive disorder; BP, bipolar disorder; SLE, systemic lupus erythematosus; SZ, schizophrenia; CRCA, colorectal cancer; PRCA, prostate cancer; OVCA, ovarian cancer. Adapted from Jostins and Barrett, 2011.

Stratified medicine could be carried out in the field of pharmacogenomics to predict and avoid adverse reactions (Harrison, 2012) or for example to prevent the development of diabetes mellitus type 2 in a cost-effective manner by

treating only the group with elevated genetic risk (Hirschorn and Gajdos, 2011). Such prediction-based measures are expected to improve when the real causal variants are identified because current genotyping arrays were designed to capture the haplotype variability with tagSNPs and incomplete LD decreases the effect estimation (Cardon and Bell, 2000; Wang et al., 2005; Visscher et al., 2012).

The GWAS findings have widened the conception of a disease and shed light on the causal biological mechanisms (Altshuler et al., 2008). For example, diseases with similar clinical features, such as Crohn's disease and ulcerative colitis, or autoimmune diseases, tend to share some associated risk variants, which make the effects pleiotropic. However, in many other cases, the associated variants originate from different haplotypes, suggesting different regulatory mechanisms that may mediate divergence in disease pathogenesis (Franke et al., 2010; Zhernakova et al., 2009). The regulatory balance of a gene can be interrupted in several ways, as has been indicated by some regions having allelic heterogeneity and some gene loci being affected by multiple independent signals (Voight et al., 2010; Elks et al., 2010, Lango-Allen et al., 2010) – up to seven in the case of human stature (GIANT Consortium Height Working Group, unpublished data). GWAS results have revealed that many of the genes for which rare variants cause familial forms of disease also harbor common alleles that modulate the normal variability of a trait (Lango-Allen et al., 2010; Teslovich et al., 2010). There are also opposite examples, where established GWAS loci (Teslovich et al., 2010) have guided the identification of mutations in the monogenic form of a common disease, such as in the case of hypolipidemia (Musunuru et al., 2010).

The GWAS prioritizes the DNA sequence variants without any prior biological information, and this approach enables the identification of novel pathways not yet linked to a specific disorder or trait (Hirschhorn, 2009). The functions of some of the genes that have been associated with diabetes mellitus type 2 risk suggest involvement of many new mechanisms, including melatonin secretion and circadian rhythms, beta cell dysfunction and zinc transport, and regulation of cell proliferation by modifying the mass of the pancreatic Langerhans islets (Visscher et al., 2012). Genetic variants that have been associated with perturbed fasting glucose and fasting insulin levels in healthy non-diabetic individuals suggest several mechanisms that may be good therapeutic targets to regulate abnormal glucose homeostasis (Dupuis et al., 2010). This idea is justified by the fact that several sites of action of known therapies have been highlighted through GWAS. A good example is the 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGGR*) gene, which represents the primary target for a class of cholesterol synthesis inhibitors, known as statins. The common variants in the *HMGCR* gene explain only a fraction of variance in low-density lipoprotein levels (~5%) for which statin-based treatment is highly efficient (30% of redaction) (Altshuler et al., 2008).

Thus, it can be concluded that the variation explained on the population level by a common genetic variant is not an appropriate measure to evaluate the

relevance of a GWAS finding. It is important to remember, however, that the regions identified through GWASs are enriched for regulatory elements, which helps to make the design of new drugs easier since targeting a biologically buffered regulatory mechanism is more efficient, less laborious, and less dangerous than repairing a loss-of-function or gain-of-function mutation (Aartsma-Rus et al., 2010).

1.3. Problems of hidden heritability

Despite the fact that GWASs have doubled the number of known disease susceptibility associated DNA sequence variants and, therefore, have guided the initiation of numerous new functional and molecular biology studies to uncover the underlying biological pathways, broaden our understanding of disease etiology, and identify new potential drug targets, several concerns still exist about the relevance and feasibility these types of studies (Maher, 2008; McClellan and King, 2010; Crow, 2011). This general discontent with GWAS arises from the fact that even when tens and hundreds of thousands of samples have been pooled in GWAS meta-analyses and thousands of potential causal genetic variants have been described, only a small fraction (estimation ranges from less than 1% to more than 50%) of phenotypic variance or genetic predisposition of genes have been explained (Lander, 2011; Visscher et al., 2012).

Follow-up experimental studies are necessary to understand why the current GWAS findings have only been able to explain so little, and to determine where the remaining hidden heritability lies. Several strategies, next to GWAS, have been proposed for finding the hidden heritability of complex traits but no consensus has been reached (Gilbert, 2012).

It is important to note that the phenotypic variance due to genes can never be completely understood because of practical limitations in detecting common and rare variants with extremely low effects, in predicting *de novo* mutations, and in modeling all complex interactions between genes and environmental factors (Altshuler et al., 2011).

1.3.1. Phenotypic variability and concept of heritability

In quantitative genetics, the phenotype (P) is a function of both genetic regulation (G) and environmental exposure (E). Likewise, the variance seen at the population level in a phenotype (var[P]) is the sum of variance due to genotype (var[G]) and variance due to environment (var[E]). Heritability, the part of phenotypic variance due to genetic effects, is divided into broad-sense heritability (H^2) and narrow-sense heritability (h^2) (Strachan and Read, 2011). In the case of broad-sense heritability, all of the genetic contributions are considered, including the additive, dominant, epistatic and imprinting effects; such a measure is relevant for clinical risk assessment, as it gives the maximum estimation of how well a phenotype can be predicted from a genotype (Zuk et

al., 2012). The additive effects explain the majority of the phenotypic variance in a population. In contrast, the narrow-sense heritability indicates only the additive effects of genes, and represents the maximum variance that can be explained by a linear combination of the allelic counts. In GWAS, the explained heritability refers to the fraction of narrow-sense heritability accounted for by the associated genetic variants (Zuk et al., 2012).

Twin studies have been used to quantify the contribution of genes, shared environment, individual-specific environment, and their interactions to complex human traits. The estimation improves when genetically identical (monozygotic) twin pairs are raised in different environments and genetically discordant (dizygotic) pairs share an identical environment (Boomsma et al., 2002). When a trait is assumed to be strictly additive the h^2 can be calculated as twice the difference of the phenotype correlation between mono- (r_{MZ}) and dizygotic (r_{DZ}) twins, as follows: $h^2 = [2 \times (r_{MZ} - r_{DZ})]$ (Strachan and Read, 2011). Table 1 shows the heritability estimates for some of the human common diseases and complex traits. However, the heritability estimates derived from twin studies may be inaccurate due to limited sample sizes (Yang et al., 2010).

Table 1. Proportion of explained additive variance in complex traits. For each phenotype three estimates are shown: 1) the proportion of phenotype variability in a population due to additive genetic variants estimated from pedigree studies; 2) the proportion of phenotypic variance or variance in liability to a disease explained by significant and validated SNPs of GWAS; and 3) the proportion of phenotypic variance or variance in liability to a disease explained when all GWAS SNPs are considered simultaneously (those for diabetes mellitus type 2 are not yet available). Adapted from Visscher et al., 2012.

Trait or Disease	Pedigree studies h^2	GWAs Hits h^2	All GWAs SNPs h^2
Height	0.80	0.10	0.50
Obesity (BMI)	0.40-0.60	0.01-0.02	0.20
QT interval	0.37-0.60	0.07	0.20
Diabetes mellitus type 2	0.30-0.60	0.05-0.10	NA
Diabetes mellitus type 1	0.90	0.60	0.30
Crohn's Disease	0.60-0.80	0.10	0.40
Schizophrenia	0.70-0.80	0.01	0.30
Bipolar disorder	0.60-0.70	0.02	0.40

The hidden heritability is defined as the proportion between explained additive variance and the total additive variance, and is calculated as follows: $[1 - (h^2_{explained} / h^2_{total})]$ (Zuk et al., 2012). The amount of additive variance explained for a complex trait or disease was reported to range between 1% and 25% when

classical genetic variants in the human leukocyte antigen region are not considered (Lander, 2011). It has been proposed that the hidden heritability could lie in gene-gene and gene-environment interactions (Frazer et al., 2009), but according to the narrow-sense heritability definition, non-additive effects are not a relevant explanation (Yang et al., 2010). A substantial amount of additive genetic variance is explained when all of the GWAS SNPs are considered simultaneously (shown in Table 1) (Lee et al., 2011; Yang et al., 2011). The explained heritability for GWAS loci and cumulative estimation can also be underestimated if either or both of the following conditions exist: 1) the GWASs have not identified the causal variant and instead only identified the LD block where the causal variant is expected to be located; and 2) inherent uncertainty in the imputation algorithms. If the real causative variant is not known, the effect of a certain variant is decreased by the factor of r^2. The same holds true for imperfect genotype predictions (Visscher et al., 2012). For example, when both mentioned variables are taken into account, essentially the entire additive genetic heritability of height was explained by common variants in height but only half of the variability of body mass index was explained (Yang et al., 2010). This type of finding suggests the involvement of rare sequence variants (Gibson, 2012).

1.3.2. Next steps in GWA studies

Large-scale meta-analyses of continuous traits, such as height and obesity, have estimated that more than half a million samples are needed to double the currently explained heritability (Lango-Allen et al., 2010; Speliotes et al., 2010; Heid et al., 2010). Moreover, calculations indicate that approximately half of the additive heritability would be explained when all GWAS SNPs are considered simultaneously (Visscher et al., 2012) (Table 1). Two key parameters must be changed to improve the discovery yield of GWAS. First, even larger sample sizes are needed for common variants with weak effects to reach the genome-wide significance, and this is especially pronounced for neuropsychiatric disorders. Second, the imputation reference panels need to be improved to be able to pinpoint the real causal variants and to test the variants of lower allele frequencies (McCarthy et al., 2008; Manolio et al., 2009). DNA sequence variants of different scales on allele frequency and effect sizes are explained in Figure 7.

Active genotyping with genome-wide arrays over the past years have increased the discovery sample size of human stature from 130,000 to 250,000 (GIANT Consortium Height Working Group, unpublished data), and from more than 20,000 to 40,000 cases with twice as many controls for coronary artery disease. (CARDIoGRAMPlus Consortium, unpublished data). As predicted by Visscher et al. (2012), in both undertakings the number of trait-associated independent genetic variants was doubled. As the yet to be discovered signals lie in the GWAS "grey zone" (the *P*-value range from 10^{-5} to 10^{-8} (Naukkarinen

et al., 2010)), two custom-made arrays, Immunochip and Cardio-Metabochip, were designed to analyze these regions in large samples in a cost-effective manner. Both arrays contain roughly 250 loci (total of 200,000 SNPs) of nominal significance from immune-related (Immunochip) and metabolic or anthropometric (Cardio-Metabochip) traits (Voight et al., 2012). This has enabled to cost-effectively genotype more that 500,000 samples (CardioMetabochip Consortium and ImmunoChip Consortium, unpublished data). The combined results from GWAS and the Immno- or Cardio-Metabochip studies explained more than 50% of the heritability in celiac disease (Trynka et al., 2011) and increased sample size in GWASs of human stature to more than 320,000, yielding 700 independent variants (GIANT Consortium Height Working Group, unpublished data). The custom-made arrays had been supplemented with new variants derived from the 1000 Genomes Project, which enabled fine mapping of the association signal in several previously validated loci (Trynka et al., 2011; Morris et al., 2012; Scott et al., 2012). Conditioning out the main-effect has shown that multiple independent variants are present for one-third of the loci (Altshuler et al., 2008; Trynka et al., 2011; Wood et al., 2011). Only recently, step-wise conditioning of meta-analyses summary statistics was developed (Yang et al., 2012), which has enabled the discovery of up to seven independent variants in an associated loci (GIANT Consortium Height Working Group, unpublished data). The high level of allelic heterogeneity is ignored when calculating the narrow-sense heritability, but may improve the estimations when modeled in (Yang et al., 2012).

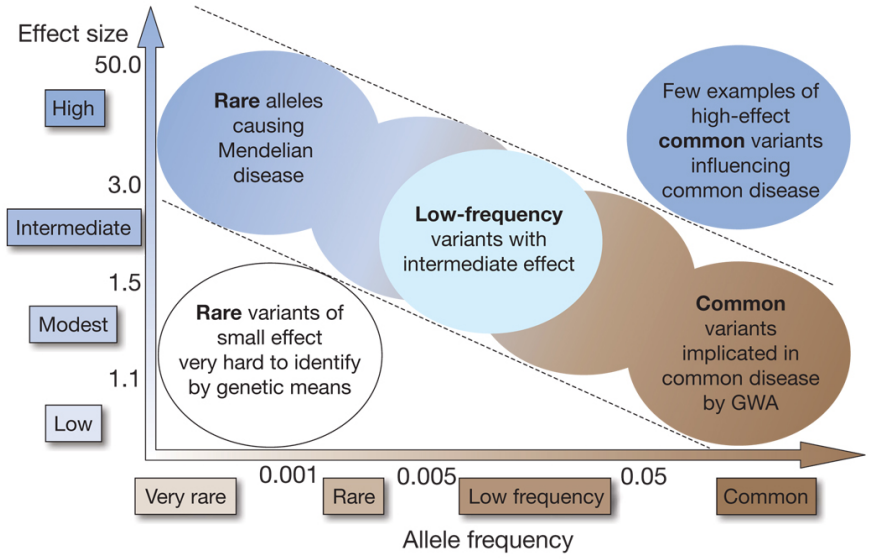

Figure 7. Feasibility of identifying a trait-associated genetic variant by allele frequency and strength of genetic effect (odds ratio). Most of the genetic variants discovered to date lie within the area between the dotted diagonal lines (Manolio et al., 2009).

The March 2012 release of the 1000 Genomes Project is composed of 40 million genetic variants, which includes 2.4 million short insertions and deletions (1000 Genomes Project, 2012). Thus, the reference panel is now 16 times denser than the previous HapMap panel. Moreover, the enriched reference panel is capable of analyzing markers with minor allele frequency, down to half a percent. The European subpanel contains 500 samples in total representing five geographically distant regions, which helps to account for the allele frequency changes in Europe. The entire reference panel currently contains more than 1,600 samples from 19 populations (www.1000genomes.org). Use of this combined sample increases power and enables more accurate prediction of haplotypes that are extremely rare in one population but relatively common in others (Howie et al., 2011).

So far, the new panel has been used to verify the presence of non-synonymous substitutions in GWAS loci (Heid et al., 2010; Speliotes et al., 2010), and very recently for imputation, which yielded new signals and fine-tuning of known loci (Huang et al., 2012). The true power of the 1000 Genomes Project reference panel will not be realized, however, until tens and hundreds of thousands of samples are imputed and pooled as was done in the previous HapMap imputation-based meta-analyses. Although it is computationally laborious, preliminary results from large consortia indicate that tens of new loci can be found with modest (40,000) sample sizes (ENGAGE Consortium, unpublished data). It is expected to take another year or two before such an approach is applied to all the existing GWAS data sets.

1.3.3. Proposed approaches to find the hidden heritability

The ongoing GWAS efforts of common variants and improved reference panels are expected to explain a substantial amount of narrow-sense heritability. Even then, it is likely that a fraction of the heritability will remain hidden (Gibson, 2012). Several approaches have been proposed to help guide the process of finding hidden heritability. In the first, SNPs with frequencies lower than 1% are targeted, since the current GWASs are not designed to detect these types of variants (McCarthy et al., 2008). In the second, structural variants, such as deletions, duplications and inversions are targeted, that are not robustly detectable by the current SNP genotyping arrays (Altshuler et al., 2008). In the third, imprecise phenotypes and heterogeneous patient groups are targeted (Manolio et al., 2009). In the fourth and final proposed approach, the non-sequence based heritability and complex interactions are targeted for study (Eichler et al., 2010).

1.3.3.1. Low-frequency variants

The common disease/common variant hypothesis, which states that a limited number of genetic variants with intermediate effects underlie common disease,

turned out to be not entirely true, as there are hundreds and most probably thousands of common and many less frequent genetic variants that contribute to the trait variability (Altshuler et al., 2008). If the common allele associations were solely caused by underlying low frequency and rare variants, then a greater percentage of heritability would have been explained than has been estimated from the pedigree studies to date (Visscher et al., 2012). The infinitesimal model of many variants, both common and rare, with small effects fits theoretically and empirically (Gibson, 2012). Since rare alleles with large effects have been implicated in many rare familial disorders, it is reasonable that many other rare alleles with modest or low effects exist (Gibson, 2012). This presumption is further supported by the fact that several GWAS loci harbor rare variants (Musunuru et al., 2010; Johansen et al., 2010; Rivas et al., 2011).

Advances in sequencing technology have made it possible to sequence whole genomes and exomes, but it still remains an expensive undertaking for large-scale studies, as extremely large sample sizes are needed to achieve the necessary statistical power (Figure 3; Manolio et al., 2009). The following options have been proposed to overcome these two limitations: 1) imputing rare alleles using existing GWAS datasets and the 1000 Genomes reference panel; 2) sequencing only a small sample from the extreme cases selected from a large population, since these individuals would be expected to be enriched for rare variants (Chan et al., 2011; Guey et al., 2011); 3) sequencing of isolated populations, since rare alleles may have drifted to higher frequencies; and 4) development of a cost-effective custom-made genotyping array to detect rare sequence variants in very large samples (Zeggini et al., 2011). By combining these strategies, a risk variant with allele frequency of 0.38% and OR of 12.5 was found for sick sinus syndrome, a collection of hearth rhythm disorders, in the Icelandic population by analyzing 40,000 samples (Holm et al., 2011). The carriers of the non-synonymous mutation had a 50% chance of developing the disease, but since the variant was exclusive to Icelanders the finding could not be validated or used for prediction in non-Icelandic populations (Holm et al., 2011). This study indicated that identification of a rare risk variant requires a large and homogeneous population due to the fact that rare variants have arisen recently and tend to cluster geographically. Recent population structure associated with rare variants can bias the results since current methods are not capable of correcting for this type of stratification (Mathieson and McVean, 2012; Graves et al., 2011). Analyzing rare coding variants is complicated and even puzzling. Indeed, by estimation, every individual genome carries more than 100 protein truncating or stop loss-of-function variants, of which ~30 exist in the homozygous state (MacArthur et al., 2012), as well as numerous loss-of-function compound heterozygotes (Gibson, 2012).

To achieve the vast sample size that is needed for rare variant analysis, a custom-made array called the “Exomechip” has been developed. The Exomechip contains ~240,000 rare non-synonymous coding sequence variants that have been reported at least three times among the 12,000 exomes and whole-

genomes sequenced to date (Exome Chip Design, 2012). Array-based genotyping is very accurate and cost-effective compared to next-generation sequencing. The product came to market in May 2012 and it is expected that at least 1.5 million samples will be genotyped (Illumina Inc., personal communication). As the effect sizes for non-synonymous variants are large (OR >2), a study composed of 5,000 cases and an equal number of genetically matched controls should have enough power to detect an association when the effect variant frequency is higher than 0.5% (Figure 3; Wang et al., 2005). When effect sizes are smaller or risk variant frequencies are lower, larger sample sizes are needed. Substantial power can be gained by analyzing only individuals selected from the tails of the phenotype distribution in a large (50,000) homogeneous population (Guey et al., 2011). It is expected that by the year 2014, we will know how much of the heritability in complex traits is attributable to less common (minor allele frequency > 0.5%) DNA sequence variants located both in protein-coding genes and flanking regulatory regions.

1.3.3.2. Structural variants

Structural variations, such as copy number variants (CNVs; duplications and deletions) and copy neutral variation (such as inversions and translocations), of ~1000 base pairs in size are detectable by SNP genotyping arrays, although they are analytically challenging (Pinto et al., 2011). Common copy number polymorphisms have been associated with common diseases, but due to strong LD with flanking SNPs these associations were found through GWAS (Manolio et al., 2009). Even using a high-density custom tiling array to genotype 19,000 samples for eight common diseases did not reveal any new trait-associated CNVs (WTCCC, 2010). For neuropsychiatric disorders, ~5% of schizophrenia and autism cases are explained by a couple of associated structural variants (Gibson, 2012), while the unexplained case-population is highly enriched for rare copy number events (International Schizophrenia Consortium, 2008). Although trait-associated CNVs tend to have large effects, the effect is not sufficient to explain much of the hidden heritability on a population level since such events are extremely rare and in most cases occur *de novo* (Walters et al., 2010; Gibson, 2012).

1.3.3.3. Incomplete phenotype

The ability to measure genotypes currently exceeds the quality of phenotyping. For example, a disorder diagnosis is usually made when the majority of the symptoms are present (Manolio et al., 2009). Recent GWAS findings have demonstrated that genetic risk scores enable dissection of a general diagnosis into smaller subclasses, which is complicated by clinical diagnosis (Franke et al., 2010). The same holds true for tumors, which can share a single dysfunctional mutation but vary significantly in their clinical presentation

according to the affected cell type (Stratton, 2011); the shared mutation, however, may facilitate a common response to anticancer therapies (Garnett et al., 2012).

Most of the molecular mechanisms defined for complex traits analyzed to date, including height and blood pressure, are very distant from the causal effect of a primary DNA sequence variant, which complicates the efficiency of a method to detect an association (Figure 8). A GWAS using high-throughput profiling of serum metabolite levels can be used as a proof of principle, since only a thousand samples are needed to statistically robustly identify tens of new loci with sequence variants of strong effect (Gieger et al., 2010; Suhre et al., 2011). The same concept has been shown for fractionated lipid compounds (Kettunen et al., 2012) and expression QTL mapping (Fehrmann et al., 2011; Fu et al., 2012). Likewise, GWASs with very accurate phenotypes for a limited number of samples can explain a large fraction of trait variability (up to 50%) (Fairfax et al., 2012). Finally, brain imaging was shown to aid in the discovery of sequence variants that regulate the normal anatomical variability, thereby providing insights into the biological cause of neurodevelopmental disorders (ENIGMA Network; www.enigma.loni.ucla.edu).

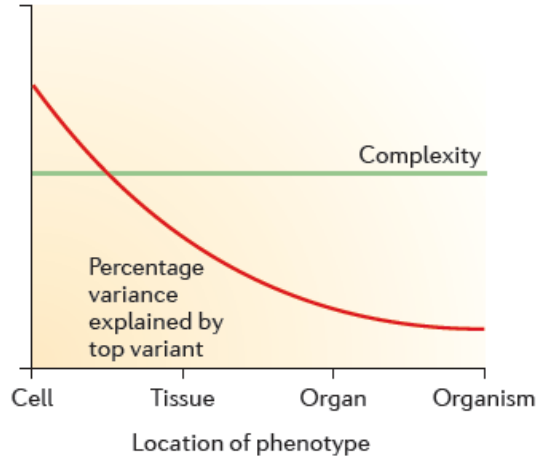

Figure 8. Expectation of phenotypic variation for different organismal levels. When the complexity in each system is taken as constant, the effect of a sequence variant declines when moving away from primary molecular effect and so the statistical power is smaller to find the association (adapted from Dermitzakis, 2012).

It is important to understand the biological processes as a continuum. The systems biology approach enables such an endeavor by combining several "–Omics" datasets (i.e. genomics, transcriptomics, proteomics, and metabolomics) (Ala-Korpela et al., 2011; Inouye et al., 2010). Furthermore, this type of comprehensive approach is expected to open the gateway to personalized

medicine by two strategies: 1) an individual is screened on many platforms over a long period of time and 2) "-omics" profiling on single cell level (Chen et al., 2012).

1.3.3.4. Epigenetic effects and complex interactions

Narrow-sense heritability can be overestimated since dominance and interactions are assumed to not exist (or at least to only account for a very minor fraction of genetic variance [var(G)] at the population level) and h^2 is basically equal to H^2 (Zuk et al., 2012). A large GWAS of 130,000 samples did not detect any deviation from additivity (Lango-Allen et al., 2010). It was later estimated that a sample of 500,000 is necessary to detect any underlying interactions between genetic variants (Zuk et al., 2012). In twin studies, narrow-sense heritability can also be underestimated because the concordance of methylation patterns between monozygotic twins decreases over time in case of their different environmental exposures, which ultimately may make the individuals more phenotypically discordant (Bell and Saffery, 2012).

deCODE Genetics has shown that parent-of-origin and imprinting has an important effect in common disease, as genetic risk was found to be increased only when inherited from one parent but to have a neutral or protective effect when inherited from the other parent (Kong et al., 2009). It is well known that environmental factors can alter methylation patterns (Bell and Saffery, 2012). A good example of this effect is the obesity-associated *FTO* gene, for which a sequence variant generates a methylation site that can be differentially methylated according to a change in environment (Bell et al., 2010).

Gene-gene interaction represents an important form of *trans*-regulation of gene expression. Through this mechanism, a regulatory sequence may mediate the transcription of a gene located several million base pairs away or even on a different chromosome. Approximately 10% of the genetic variants detected through GWASs appear to exert a *trans* effect on gene expression in different tissues (Ferhmann et al., 2011; Fu et al., 2012; Fairfax et al., 2012). Most recent studies indicate that up to 25% of the associated sequence variants map within euchromatic functional motifs and reveal the involvement of complex regulatory networks in disease etiology (Maurano et al., 2012). The underling mechanisms of *trans*-regulation, however, are so far largely unknown and require further research (Fairfax et al., 2012).

During my PhD studies I have co-authored approximately 40 research articles (full list is presented in the LIST OF PUBLICATIONS section), mostly in the field of GWAS but also on population genetics and analyses on lifestyle factors that shape the human health. These papers reflect the trends in modern human genetics. Namely, the need to combine a single cohort association results through GWAS meta-analyses to reach the statistical power to robustly identify DNA sequence variants that increase the susceptibility to disease or modulate complex traits in humans.

The current Ph.D thesis incorporates the results of five research articles that have been prioritized for the following reasons. I am the shared-first author in articles Ref I and Ref II and participated in the study design, performed in part the experiments, analyzed the data, participated in the preparation and writing of the papers. The results of the Ref I study were important in enabling to incorporate the Estonian Biobank samples to international large-scale association analyses. The work of Ref II confirmed the status of a genetc isolate for a set of Italian village communities, thus providing a powerful tool for genetic epidemiology studies.

Articles Ref III, Ref IV and Ref V are presenting novel analytical strategies that lead to unsolving some of the hidden heritability in complex traits (height, sleep duration, and osteoarthritis respectively). The chosen papers are in many ways proof-of-principle studies and demonstrate that several of the proposed approaches to find the hidden heritability (discussed in detail in section 1.3.3) are justified and are going to explain at least in part the phenotypic variability of complex traits.

2. AIMS OF THE PRESENT STUDY

1. To fill in the gaps of the genetic structuring of northeastern European populations and more specifically to estimate the spatial positioning of Estonians, Latvians, Lithuanians and northwestern Russians using the whole-genome SNP allele frequency data.

2. To evaluate the effect of the sample size and the geographical range of sampling in assessing the genetic structuring within different European populations and isolated communities from northeastern region of Italy by utilizing the SNP array data.

3. To discover novel DNA sequence variants that modulate complex traits or affect the susceptibility to common diseases in the framework of large-scale collaborative studies.

4. To investigate the problem of hidden heritability in complex trait variability by applying novel analytical approaches.

3. RESULTS AND DISCUSSION

3.1. Studied populations

In the study I and II, the whole-genome genotype data for 1,090 the Estonian Biobank samples were used. Eighty samples (40 males and 40 females) were selected randomly by place of birth, and represented 13 Estonian counties (Harju, Ida-Viru, Jõgeva, Järva, Lääne-Viru, Põlva, Pärnu, Rapla, Saaremaa, Tartu, Valga, Viljandi, and Võru). Fifty samples (25 males and 25 females) were selected from the combined Hiiumaa and Läänemaa counties (Ref. I, Figure 1). While the Estonian territory is relatively small (~45,300 km^2), it is located in a geopolitically important region of Northern Europe. Situated on the eastern coast of the Baltic Sea, Estonia has experienced several immigration waves from neighboring areas over the last 800 years. The current population size is estimated at 1.3 million, of which approximately 1 million are ethnic Estonians.

In the study I, the whole-genome genotype data was supplemented with 3,112 individuals analyzed by Illumina HumanHap 300K/370CNV chips. These samples represented a total of 19 cohorts from the following 16 countries: Austria (Vienna), Bulgaria (entire country), Czech Republic (Prague, Moravia, and Silesia), Estonia (entire country, detailed description in previous section), Finland (Helsinki, and a subisolate of Kuusamo), France (Paris), Germany (two cohorts: Schleswig-Holstein region (north) and the Augsburg region (south)), Hungary (entire country), Italy (two cohorts: Borbera Valley (north) and a 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. I, Table 1). In addition, HapMap data was retrieved from public databases for the following four populations: –United States' Utah residents with ancestry from Northern and Western Europe (CEU), Yoruban people of Ibadan, Nigeria (YRI), unrelated individuals from Beijing, China (CHB), and unrelated individuals from Tokyo, Japan (JPT). After quality control procedures, 273,464 SNPs were available for analysis.

For the study II, data for all of the samples, except the CEU, YRI, CHB and JPT HapMap populations, from the study I were pooled along with the following three datasets: 1,310 Italians (collected from six small villages in northeastern Italy) analyzed with Illumina 370CNV chips, 96 Slovenians (entire country) genotyped with Illumina OmniExpress chip, and 2,421 international samples (publicly available data from across the globe) genotyped with Illumina arrays (Ref. II, Supplementary Table 1). After applying quality control procedures, 145,000 SNPs and 3,091 samples were available for analysis.

Studies III, IV and V were designed as large, collaborative meta-analysis efforts, in which several single on-site analyzed cohorts were pooled to generate summary association statistics. The study III pooled results from 21 studies, which yielded a total sample size of 35,945 (the full list of cohorts and sample

sizes are given in Ref. III, Supplementary Table 1). In addition, a total of 2,395 Estonian Biobank samples, genotyped with the Illumina 370CNV chip, were included in the final analysis. The study IV was divided into three stages: the discovery stage, (including 4,251 samples from seven cohorts, and 924 samples from the Estonian Biobank), the *in silico* validation stage (made up exclusively of 536 Estonian Biobank samples), the *de novo* validation stage (made up exclusively of three SNPs that had been genotyped in 5,949 Estonian Biobank samples) (the full list of cohorts and sample sizes are given in Ref. IV, Supplementary Table 1). The study V was similarly divided into three stages: the discovery phase (including 3,177 cases and 4,894 controls collected from British isles), the *de novo* validation stage (using 9,620 cases and 9,177 controls collected from the British isles), and the *in silico* or *de novo* validation stage (using 6,604 cases and 10,393 controls collected from four non-British cohorts, including 2,617 cases and 2,619 matched controls from the Estonian Biobank) (Ref. V, Table 1).

For all the studies the Estonian Biobank genotype data was generated and analyzed by the Estonian Genome Center personnel.

3.2. Genetic structure in Europe (Refs. I and II)

Analyses of complex human traits have shown that the effects of common sequence variants are usually modest (see Review of literature, paragraph 1.2.3 "General findings from GWAS"). Thus, a single population or sample collection cannot provide sufficient statistical power to detect these associations. The GWAS meta-analysis approach has emerged as an efficient way to combine large datasets, although differences in ancestry-derived heterogeneity in the association signal can decrease its statistical power. An association found for a sequence variant may not be replicated because the haplotype structure and allele frequencies may occur with different properties.

3.2.1. Genetic distances between European populations

In the study I the European genetic structure were analyzed with a focus on Estonian and other Northeast European populations. Autosomal genotype data of more than 260,000 genetic markers was available for 3,112 samples from 19 cohorts from 16 European populations (Ref. I, Table 1; Ref. II, Supplementary Table 1) with a wide geographical range, from South Italy to North Finland and from Spain to Northwest Russia. Three different parameters were used to describe the genetic structure, namely principle component loadings, pair-wise fixation index (Fst) and pair-wise inflation factor (λ).

Principal component analysis is the most commonly applied method in GWAS to correct for hidden population structure, but is also applied to estimate spatial structure of the genetic variation of world populations. In the study I, the two first principal components of the genetic variability corresponded to the

northeast to southwest gradient. All of the populations positioned according to their geographical location. (Ref. I, Figure 2B). There was almost no overlap in the clustering between the populations of the northeastern region, whereas the populations of central and western Europe formed an unvarying continuum (Ref. I, Figure 2B). The results revealed that Finns position distantly from Swedes and other northeastern Europeans, while Estonians cluster next to their geographical neighbours (Latvians, Lithuanians and northwestern Russians).

Fst is used to determine how much of the genetic variability between individuals from different populations is due to inter-, and not intra-, population variation. By correlating the genetic Fst with geographic distances (Ref. I, Supplementary Table 2), a barrier was revealed between Finns, Italians, and other populations. A barrier was also detected between Swedes and northeastern European populations (Estonians, Latvians, Lithuanians, northwestern Russians and Poles) (Ref. I, Supplementary Figure 4). In both cases, geographical obstacles, such as the Baltic Sea and the Alps mountain range, has interrupted the genetic continuum.

Genomic control is used to estimate the deviation of the observed test-statistic distribution from the expected under null hypothesis, for which λ represents the scaling factor. It is possible to estimate the similarity in minor allele frequency by modeling an allelic association test between two populations and estimating λ. The resultant value reflects the genetic similarity and can be taken as a proxy for selecting the best population for validating a genetic association with a phenotype. The λ values were smallest between populations in geographical proximity (Ref. I, Table 2). The within-group λ values were the smallest in the northeastern European region (λ_{mean} = 1.23) and central and western European region (λ_{mean} = 1.22). Moreover, the number of loci with statistically significant allele frequencies, which may have confounded the results of the association analyses, was also lower in the two population clusters (Ref. I, Table 2).

The results of study I are in line with those of previous analyses of European population structure (Novembre et al., 2008; Lao et al., 2008; Heath et al., 2008) and, for the first time, position the northeastern European populations on the high-density autosomal genetic structure map of Europe. The modeling results indicate that by carrying out genetic association study in populations of geographical proximity, the loss in statistical power is minimized and the probability to validate an association is maximized by the overall genetic similarity. This regularity is already known from seminal studies using limited number of genetic markers (Menozzi et al., 1978; Cavalli-Sforza et al., 1994), but the effect of population structure to GWAS studies was not studied in depth. Such knowledge about the genetic distances between different populations is crucial for designing an optimal GWAS that will be capable of evaluating the contribution of specific cohorts without the risk of generating false positive or false negative findings.

3.2.2. Genetic structure within single populations

A structured population composed of subpopulations that differ both genetically and in disease prevalence have proportions of cases and controls that can differ in each subpopulation, thereby causing a systematic difference in allele frequencies in any loci where the genetic ancestry does not match and leading to spurious associations. Furthermore, a complex trait arises from new mutations as well as from the interplay between existing genetic variants and exposure to environmental conditions, thus it is desirable to study genetically homogeneous populations, such as isolated populations, as more power is gained for genetic association mapping studies.

In the study II, six linguistically and culturally diverse village populations (Ref. II, Figure 1A; Supplementary Note) sampled from the northeastern part of Italy (region of Friuli-Venezia Giulia (FVG)) were analyzed. The FVG village samples were genetically compared with publicly available genomic datasets with the emphasis on well-known geographical and cultural population isolates in order to evaluate if any of the village populations represented a genuine population isolate.

At first model-based structure-like analyses were applied to estimate the hypothetical ancestry proportion distributions among the FVG village samples that in general, were very similar to the other populations in the same geographical region. In contrast, for higher K values almost all of the FVG populations became dominated by a single component largely specific to that particular village (Ref. II, Figure 1C). The village-specific components were present in the background profile of all European populations, representing a fraction of the overall genetic variability and being an indication of a pronounced random genetic drift (Ref. II, Figure 1C). Furthermore, substantial levels of intra-population structure were revealed in the FVG populations in elevated variability in membership to the village-specific ancestry component (Figure 1C).

For further analysis the FVG populations were split into subpopulations according to the ancestry estimations at K = 10: a) general set, when village-specific ancestry loading was smaller than 30% and b) more isolated set, when loading exceeded 30%. The clustering of the FVG general set samples by both the principal component analysis (Ref. II, Supplementary Figure 2) and the spatial ancestry analysis (Ref, II, Figure 2) were roughly representative to their geographical location. At the same time, the FVG isolated set samples showed more extreme values for all considered measures of isolation, such as genomic homozygosity, inbreeding coefficient and the extent of LD, compared to the known population isolates (Ref II, Figure 4; Figure 5; Supplementary Figure 7; Supplementary Table 2). This indicates that the village-specific ancestral components arise from the increased genetic similarity within the specific subsets of samples from the respective villages and not from the differences in genetic origin.

Significant differences in haploblock structure and haplotype diversity were detected between populations of European ancestry in both the studies I and II (Ref. I, Figure 1; Ref II, Figure 4; Figure 5). Northern populations had longer haploblocks than the Southern populations. The genomic homozygosity was used as the proxy for haplotype diversity. A strong correlation between population haploblock length and level of genomic homozygosity was detected. Both of these features have been proposed as the cause of relatively small effects in ancestral population (Service et al., 2006; McQullian et al., 2008; Kirin et al., 2010).

In the study I, several sample collections where recruited over the entire country and for some multiple cohorts were available from a single population. The inter-population variability was extremely elevated in the more isolated sub-population of Finns (Ref. I, Figure 2B), which was similar to the results for FVG village populations in the study II. The intra-population structure was also detectable in several general populations, such as Estonian, German and Czech (Ref. I, Supplementary Figure 3) when sample collections from multiple geographical regions were compared. Plotting the Estonians by county of birth produced sub-cohort clusters that were largely overlapping (Ref. I, Supplementary Figure 1) whereas the median PC value per county showed an almost perfect resemblance to the regional map of Estonia (Ref. 1, Figure 2C and Supplementary Figure 1).

The collective results from phases I and II of this thesis study illustrate that geographic proximity does not always translate to genetic similarity and that population structuring can be detected in small countries, such as Estonia, and even in a small but topographically variable region, as was demonstrated for Northeast Italy. Thus, analyses should always be corrected for population structure and both study I and II highlight the need to analyze a large and representative sample to precisely estimate the intragroup variability within a population as the random genetic drift may lead to elevated differences in allele frequencies.

3.3. Search for hidden heritability in GWAS (Refs. III, IV, and V)

Over the past five years, GWAS meta-analyses of complex human phenotypes have identified more than 3,000 sequence variants that associate with genetic predisposition to a disease or contribute to normal variability of a continuous phenotype. One limitation of these large-scale analyses is that only a small fraction of trait variability accounted for by genetic factors is explained. Several approaches have been proposed to find the hidden heritability and three of them were assessed in this thesis work.

3.3.1. Genomic homozygosity and recessive effects

Rare sequence variants represent one source for finding the missing heritability. Rare variants are expected to have occurred recently or to have been selected against, when the mutation is deleterious or reduces fitness, which can explain their low allele frequency in the population. If these are not *de novo* events, a specific haplotypic background is linked to each variant in a population. Family-based linkage analyses have exploited this property of the genome to identify thousands of disease causing mutations. Since many mutations are recessive, the effect is only revealed when the variant is in a homozygous state. Recent GWAS meta-analyses that combined the results for more than 130,000 samples increased the number of height variation-modulating loci to 250, but the sequence variants explained only 10% of the trait variance in population and focused only on additive effects (Lango-Allen et al., 2010). However, when all the SNPs were considered simultaneously, up to 50% of the heritability was explained.

The study III aimed to estimate the effect of recessive genetic component on complex phenotypes and used human stature as a model to reveal the genetic architecture of the causative allele frequency spectrum. The recessive genetic component was estimated by genomic homozygosity, which was calculated as the fraction of the genome that is covered with long runs of only homozygous genotypes (designated as F_{ROH}). A cohort-specific linear regression between F_{ROH} and height was carried out while correcting for age, sex, and socio-economical status, and the summary statistics were combined through meta-analyses. A small but statistically strong ($P = 1.23 \times 10^{-11}$) inverse correlation was observed. The signal achieved even stronger statistical significance ($P = 1.23 \times 10^{-88}$) when ancestral haplotype sharing was removed and only recent parental effect was considered. The 1% increase in genomic homozygosity was estimated to equal 0.6 cm decrease in body height, which translated to a reduction of 3 cm in offspring of first cousins. Interestingly, when adjusting the full results for recent parental effects, the observed inverse association remained significant, indicating that ancestral recessive variants also play an important role in this phenotype effect.

Collectively, the results from the study III demonstrate that the effect is not associated with any specific or single genome region, but instead reflects the overall polygenic recessive component's contribution to the genetic architecture of human height. Strong effects for rare variants were detected when explained variance of validated common SNPs were analyzed for the extreme cases drawn from a large population. The predicted mean height for extremely short stature was previously found to be smaller than expected and statistically different from the other height groups (Chen et al., 2011). This finding is in line with the results from the study III and with previously reported observations that indicated rare variants are enriched for coding variants (Li et al., 2010). The non-synonymous substitutions tend to have adverse or deleterious effects and

may cause deviation from the expected outcome only in cases of extremely short stature.

3.3.2. Confounding by environment

In the study IV, large-scale GWAS meta-analyses were performed to find genetic variants that affected the variability of human sleeping behavior. Disturbed normal sleep-patterns can lead to metabolic syndrome, cardiovascular disease, and psychiatric disorders. Individualized sleep requirements underlie the different experiences of "social" jetlag among a population, as each person's inner circadian rhythm does not precisely match the generalized one that is socially dictated.

To find sequence variants that affect sleep behavior, 4,251 samples from seven cohorts (Ref. IV, Supplementary Table 1) were tested for linear associations of SNPs and sleep duration (adjusted for age, sex, and body mass index). The results were combined through fixed-effect inverse variance-weighted meta-analysis. One SNP (rs11046205), located in intron 27 of the ATP-binding cassette, sub-family C member 9 (*ABCC9*) gene achieved genome-wide significance ($P = 3.68 \times 10^{-8}$). The association was strongest in a directly genotyped cohort; therefore, *de novo* genotyping of the variant was carried out in the three largest cohorts (including the Estonian Biobank samples). After repeating the meta-analysis, the association remained significant ($P = 3.99 \times 10^{-8}$). The validation by independent *in silico* (536 samples) and *de novo* genotyped (5949 samples) cohorts drawn from the Estonian Biobank did not show any statistically significant replication ($P > 0.05$). When detailed phenotype modeling was performed in the *de novo* cohort, a systematic difference in sleep duration was observed between the samples recruited during winter and summer months, as well as between individuals with early and late chronotypes. The same was observed for two of the discovery cohorts (Ref. IV, Supplementary Table 7). As clear confounding by season and chronotype was detected (Ref. IV, Supplementary Table 8), the early half of winter collection was considered as a valid replication sample, considering that they would be less sleep deprived. After combining the discovery, *in silico* and *de novo* early winter cohorts, a borderline genome-wide significant association ($P = 7.9 \times 10^{-8}$) was detected for rs11046205 (Ref. IV, Figure 1a). This single variant explained 3% of trait variability, which translated to a sleep duration difference of 16 minutes (Ref. IV, Figure 1b). Since the heritability of sleep duration was estimated to be 40%, 12% of the variability due to the additive component was explained by rs11046205.

The study IV demonstrated how important is to have uniform, precise and sophisticated phenotype information for a large set of samples in order to be able detect a stratifying effect of environmental exposure. The confounding effect was found to be higher in the Estonian Biobank samples, which is likely due to the fact that the difference in winter and summer photoperiods in Estonia is greater than eight hours. While SNP rs10046205 has an additive (allele

dosage) effect on sleep duration (Ref. IV, Figure 1b), the modulating effect can be reversed by differences in environmental exposure (length of the photo-period) and, if not corrected for, the association signal is averaged out.

3.3.3. Improved reference panel for imputation

In the study V, large-scale GWAS meta-analyses were carried out to find unknown sequence variants that increase the susceptibility to osteoarthritis, a degenerative joint disease that affects articular cartilage and subchondral bone. Only a limited number of genetic associations have been previously reported for osteoarthritis.

Imputation of the 1000 Genomes Project reference panel has been proposed as an approach that would increase the power to find new genetic associations of a given trait because the dataset includes several times more SNPs than the previous HapMap reference. Thus, the enriched reference panel is considered to have better resolution. In addition, it includes haplotypes that have not yet been tested for an association. A sample of 3,177 osteoarthritis cases and 4,894 matched controls were imputed with the 1000 Genomes pilot 1 reference built from sequences of 60 CEU individuals. A total of 7.2 million variants were tested by logistic regression and the six most promising loci were selected for validation in UK and non-UK cohorts (study design shown in Ref. V, Figure 1). After several validation steps, a SNP on 13q34 achieved genome-wide significance under an inverse variance-weighted fixed effect meta-analysis with a *P*-value of 2.07×10^{-8} and an OR of 1.17. The effect direction was generally uniform, with the exception of one cohort, and the risk allele frequency was around 0.92 (Ref. V, Figure 2). The identified sequence variant is located in intron 4 of the MCF.2 cell-line-derived transforming sequence-like (*MCF2L*) gene, which encodes a guanine nucleotide exchange factor. The functional impact of the variant may lay in altering the splicing or changing a regulatory sequence motif, thus reducing the binding affinity of a protein, which in turn can influence gene expression. In human cells, *MCF2L* regulates neurotrophin-3 (a member of nerve growth factor family) induced cell migration in Schwann cells, and treatment of osteoarthritis patients with humanized MCF2L antibody that inhibits nerve growth factors reduces pain and improves joint function (Lane et al., 2010

The detailed comparison of the regional association plots at chromosome 13-associated loci showed that directly genotyped and HapMap imputed analyses identified only a singleton variant with borderline significance. Only after 1000 Genomes imputation did the region show a broad pattern of association (Ref. V, Figure 3). Therefore, the results from this study demonstrate that an improved reference panel strengthens the power to detect novel susceptibility variants. As the reference panel grows larger over time (by March 2012 the panel consisted of 1,600 samples) it will become more feasible to impute and test for association. In addition, the increased amount of low-frequency variants will increase the fraction of explained variability for any given trait.

4. CONCLUSIONS

The following conclusions can be drawn from the current Ph.D. thesis study:

1. By applying the principal component analysis on genotype data of more than 270,000 SNPs of samples from 19 European populations the gaps in the genetic structuring of northeastern European populations were filled in. The analyses produced a genetic structure map, in which the populations were positioned according to their approximate geographic locations. The results revealed that Finns position distantly from Swedes and other northeastern Europeans, while Estonians cluster next to their geographical neighbours (Latvians, Lithuanians and northwestern Russians). The estimated Fst distances and λ values demonstrate, that the Estonian Biobank samples can be analyzed together with the other cohorts of European ancestry in large-scale gene discovery studies.

2. The results from model-based structure-like analyses on a set of linguistically and culturally diverse village communities sampled from the north-eastern part of Italy demonstrated that substantial genetic structure can be found even in communities considered earlier as largely genetically homogeneous populations. The genetic comparisions between well-known population isolates and an isolated fraction of the Italian village communities revealed (for the latter) more extreme values for all considered measures of isolation. The current findings emphasize that a representative sample must be analyzed in order to achive a substantial enough power to reveal the structuring within a population and to avoid false positive findings from genetic association studies.

3. Through the studies of this thesis two novel associations between a complex trait and a DNA sequence variant were established. Two genes, *MCF2L* for osteoarthritis and *ABCC9* for normal variation in sleep duration, were prioritized for follow-up studies.

4. In this thesis, three different approaches were used to uncover the hidden heritably in complex phenotypes. First, using the human stature as an example it was demonstrated that recessive genetic effects on top of the additive play an important role in the phenotype variability. Second, by controlling for confounding factors by environment a sequence variant was revealed that explained 12% of the narrow-sense heritability in sleep duration. Last, by applying improved reference panels to genotype prediction the power to find novel complex trait associated sequence variants is going to be improved. These studies illustrate that the proposed sources of hidden heritability are justified and that the current genome-wide datasets will keep providing insights into the biological mechanisms behind complex human traits.

REFERENCES

1000 Genomes Project Consortium. (2010). "A map of human genome variation from population-scale sequencing". *Nature* **467**(7319): 1061–73.

Aartsma-Rus A, den Dunnen JT, et al. (2010). "New insights in gene-derived therapy: the example of Duchenne muscular dystrophy". *Ann N Y Acad Sci* **1214**:199–212.

Alexander DH, Novembre J, et al. (2009). "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res* **19**(9): 1655–64.

Altshuler D, Daly MJ, et al. (2008). "Genetic mapping in human disease". *Science* **322**(5903): 881–8.

Bamshad MJ, Ng SB, et al. (2011). "Exome sequencing as a tool for Mendelian disease gene discovery". *Nat Rev Genet* **12**(11): 745–55.

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

Behar DM, Yunusbayev B, et al. (2010). "The genome-wide structure of the Jewish people". *Nature* **466**(7303): 238–42.

Bell CG, Finer S, et al. (2010). "Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus". *PLoS One* **5**(11): e14040.

Bell JT, Saffery R. (2012). "The value of twins in epigenetic epidemiology". *Int J Epidemiol* **41**(1): 140–50.

Bersaglieri T, Sabeti PC, et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene". *Am J Hum Genet* **74**(6): 1111–20.

Biobank, UK. (2011). "UK Biobank – improving the health of future generations". Retrieved April 2012, from http://www.ukbiobank.ac.uk/

Boomsma D, Busjahn A, et al. (2002). "Classical twin studies and beyond". *Nat Rev Genet* **3**(11): 872–82.

Bousquet J, Anto JM, et al. (2011). "Systems medicine and integrated care to combat chronic noncommunicable diseases". *Genome Med* **3**(7): 43.

Browning BL, Browning SR. (2009). "A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals". *Am J Hum Genet* **84**: 210–223.

Burton PR, Hansell AL, et al. (2009). "Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology". *Int J Epidemiol* **38**: 263–73

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

Carlson CS, Eberle MA, et al. (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium". *Am J Hum Genet* **74**(1): 106–20.

Casto AM, Feldman MW. (2011). "Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations?" *PLoS Genet* **7**(1): e1001266.

Cavalli-Sforza LL, Menozzi P, et al. (1994). "The History and Geography of Human Genes". *Princeton: Princeton University Press*.

Chan Y, Holmen OL, et al. (2011). "Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals". *PLoS Genet* **7**(12): e1002439.

Chapman NH, Wijsman EM. (1998). "Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility". *Am J Hum Genet* **63**(6): 1872–85.

Chen R, Mias GI, et al. (2012). "Personal omics profiling reveals dynamic molecular and medical phenotypes". *Cell* **148**(6): 1293–307.

Chen Z, Chen J, et al. (2011). "China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up". *Int J Epidemiol* **40**(6): 1652–66.

Cho YS, Chen CH, et al. (2011). "Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians". *Nat Genet* **44**(1): 67–72.

Collins FS, Green ED, et al. (2003). "A vision for the future of genomics research". *Nature* 422(6934): 835–847.

Crow TJ. (2011). "The missing genes: what happened to the heritability of psychiatric disorders?". *Mol Psychiatry* **16**(4): 362–4

Daly MJ, Rioux JD, et al. (2001). "High-resolution haplotype structure in the human genome". *Nat Genet* **29**(2): 229–32.

Dawson E, Abecasis GR, et al. (2002). "A first-generation linkage disequilibrium map of human chromosome 22". *Nature* **418**(6897): 544–8.

de Bakker PI, Ferreira MA, et al. (2008). "Practical aspects of imputation-driven meta-analysis of genome-wide association studies". *Hum Mol Genet* **17**(R2): R122–8.

de Bakker PI, Yelensky R, et al. (2005). "Efficiency and power in genetic association studies". *Nat Genet* **37**(11): 1217–23.

deCODE. (2010). "deCODE Genetics – The population approach". Retrieved April 2011, from http://www.decode.com/research/

Deschênes M, Cardinal G, et al (2001). "Human genetic research, DNA banking and consent: a question of 'form'?" *Clin Genet* **59**(4): 221–39.

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

Dewan A, Liu M, et al. (2006). "HTRA1 promoter polymorphism in wet age-related macular degeneration". *Science* **314**(5801): 989–92.

Dupuis J, Langenberg C, et al. (2010). "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk". *Nat Genet* **42**(2): 105–16.

Eichler EE, Flint J, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease". *Nat Rev Genet* **11**(6): 446–50.

Elks CE, Perry JR, et al. (2010). "Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies". *Nat Genet* **42**(12): 1077–85.

ENCODE Project Consortium. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". *Nature* **447**(7146): 799–816.

ENCODE Project Consortium. (2012). "An integrated encyclopedia of DNA elements in the human genome." *Nature* **489**(7414): 57–74.

Exome Chip Design web-resource: Retrieved April 2012, from http://genome.sph.umich.edu/wiki/Exome_Chip_Design

Fairfax BP, Makino S, et al. (2012). "Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles". *Nat Genet* **44**(5): 502–10.

Fehrmann RS, Jansen RC, et al. (2011). "Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA". *PLoS Genet* **7**(8): e1002197.

Franke A, McGovern DP, et al. (2010). “Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci”. *Nat Genet* **42**(12): 1118–25.
Frazer KA, Murray SS, et al. (2009). “Human genetic variation and its contribution to complex traits”. *Nat Rev Genet* **10**(4): 241–51.
Freedman ML, Reich D, et al. (2004). “Assessing the impact of population stratification on genetic association studies”. *Nat Genet* **36**(4): 388–93.
Fu J, Wolfs MG, et al. (2012). “Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression”. *PLoS Genet* **8**(1): e1002431.
Enattah NS, Sahi T, et al. (2002). “Identification of a variant associated with adult-type hypolactasia”. *Nat Genet* **30**(2): 233–7.
Ernst J, Kheradpour P, et al. (2011). “Mapping and analysis of chromatin state dynamics in nine human cell types”. *Nature* **473**(7345): 43–9.
Estonian Genome Center, University of Tartu. (2012). “Annual report 2001 to 2011”. Retrieved April 2012, from http://www.biobank.ee
Gabriel SB, Schaffner SF, et al. (2002). “The structure of haplotype blocks in the human genome”. *Science* **296**(5576): 2225–9.
Gambaro G, Anglani F, et al. (2000). “Association studies of genetic polymorphisms and complex disease”. *Lancet* **355**(9200): 308–11.
Garnett MJ, Edelman EJ, et al. (2012). “Systematic identification of genomic markers of drug sensitivity in cancer cells”. *Nature* **483**(7391): 570–5.
Gershon ES, Alliey-Rodriguez N, et al. (2011). “After GWAS: searching for genetic risk for schizophrenia and bipolar disorder”. *Am J Psychiatry* **168**(3): 253–6.
Gerstein MB, Kundaje A, et al. (2012). “Architecture of the human regulatory network derived from ENCODE data.” *Nature* **489**(7414): 91–100.
Gibson G. (2012). “Rare and common variants: twenty arguments”. *Nat Rev Genet* **13**(2): 135–45.
Gieger C, Radhakrishnan A, et al. (2011). “New gene functions in megakaryopoiesis and platelet formation”. *Nature* **480**(7376): 201–8.
Gravel S, Henn BM, et al. (2011.) “Demographic history and rare allele sharing among human populations”. *Proc Natl Acad Sci U S A* **108**(29): 11983–8.
Gudbjartsson DF, Walters GB, et al. (2008). “Many sequence variants affecting diversity of adult human height”. *Nat Genet* **40**(5): 609–15.
Guey LT, Kravic J, et al. (2011). “Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants”. *Genet Epidemiol* doi: 10.1002/gepi.20572.
Guttmacher AE, Collins FS. (2003). “Welcome to the genomic era”. *N Engl J Med* **349**(10): 996–998.
Harris JR, Burton P, et al. (2012). “Toward a roadmap in global biobanking for health”. *Eur J Hum Genet* doi: 10.1038/ejhg.2012.96. [Epub ahead of print]
Harrison C. (2012). “Adverse drug reactions: Computational model predicts side effects.” *Nat Rev Drug Discov* 11(**8**): 602.
Heath SC, Gut IG, et al. (2008). “Investigation of the fine structure of European populations with applications to disease association studies”. *Eur J Hum Genet* **16**(12): 1413–29.
Heid IM, Jackson AU, et al. (2010). “Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution”. *Nat Genet* **42**(11): 949–60.

Hindorff LA, Sethupathy P, et al. (2009). “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. *Proc Natl Acad Sci USA* **106**(23): 9362–7.
Hirschhorn JN. (2009). “Genomewide association studies – illuminating biologic pathways”. *N Engl J Med* **360**(17): 1699–701.
Hirschhorn JN, Gajdos ZK. (2011). “Genome-wide association studies: results from the first few years and potential implications for clinical medicine”. *Annu Rev Med* **62**:11–24.
Holm H, Gudbjartsson DF, et al. (2011). “A rare variant in MYH6 is associated with high risk of sick sinus syndrome”. *Nat Genet* **43**(4): 316–20.
Hood L, Heath JR, et al. (2004). “Systems biology and new technologies enable predictive and preventative medicine”. *Science* **306**(5696): 640–3.
Huang J, Ellinghaus D, et al. (2012). “1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data”. *Eur J Hum Genet* **20**(7): 801–5.
Inouye M, Kettunen J, et al. (2010). “Metabonomic, transcriptomic, and genomic variation of a population cohort”. *Mol Syst Biol* **6**:441.
International HapMap Consortium. (2005). “A haplotype map of the human genome”. *Nature* **437**(7063): 1299–320.
International HapMap Consortium. (2007). “A second generation human haplotype map of over 3.1 million SNPs”. *Nature* **49**(7164): 851–61.
International Schizophrenia Consortium. (2008). “Rare chromosomal deletions and duplications increase risk of schizophrenia”. *Nature* **455**(7210): 237–41.
Ioannidis JP, Trikalinos TA, et al. (2003). “Genetic associations in large versus small studies: an empirical assessment”. *Lancet* **361**(9357): 567–71.
Jakkula E, Rehnstrom K, et al. (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, et al. (2008). “Genotype, haplotype and copy-number variation in worldwide human populations”. *Nature* **451**(7181): 998–1003
Johansen CT, Wang J, et al. (2010). “Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia”. *Nat Genet* **42**(8): 684–7.
Jombart T, Devillard S, et al. (2010). “Discriminant analysis of principal components: a new method for the analysis of genetically structured populations”. *BMC Genet* 2010; 11:94
Jorde LB. (2000). “Linkage disequilibrium and the search for complex disease genes”. *Genome Res* **10**(10): 1435–44.
Jostins L, Barrett JC. (2011). “Genetic risk prediction in complex disease”. *Hum Mol Genet* **20**(R2): R182–8.
Kaaks R, Slimani N, et al. (1997). “Pilot phase studies on the accuracy of dietary intake measurements in the EPIC project: overall evaluation of results”. *Int J Epidemiol* **26 (**Suppl 1): S26–36.
Kettunen J, Tukiainen T, et al. (2012). “Genome-wide association study identifies multiple loci influencing human serum metabolite levels”. *Nat Genet* **44**(3): 269–76.
Kirin M, McQuillan R, et al. (2010). “Genomic runs of homozygosity record population history and consanguinity”. *PLoS One* **5**(11): e13996.
Klein RJ, Zeiss C, et al. (2005). “Complement factor H polymorphism in age-related macular degeneration”. *Science* **308**(5720): 385–9.
Knoppers BM. (2001). “Cancer genetics: a model for multifactorial conditions?” *Med Law* **20**(2): 177–82.

Kohane IS. (2011). “Using electronic health records to drive discovery in disease genomics.” *Nat Rev Genet* 12(**6**): 417–28.
Kong A, Frigge ML, et al. (2012). “Rate of de novo mutations and the importance of father's age to disease risk.” *Nature* **488**(7412):471–5.
Kong A, Steinthorsdottir V, et al. (2009). “Parental origin of sequence variants associated with complex diseases”. *Nature* **462**(7275): 868–74.
Kong A, Thorleifsson G, et al. (2010). “Fine-scale recombination rate differences between sexes, populations and individuals”. *Nature* **467**(7319): 1099–103.
Kraft P, Hunter DJ. (2009). “Genetic risk prediction – are we there yet?” *N Engl J Med* **360**(17): 1701–3.
Lander ES. (2011). “Initial impact of the sequencing of the human genome”. *Nature* **470**(7333): 187–97.
Lander ES, Linton LM, et al. (2001). “Initial sequencing and analysis of the human genome”. *Nature* **409**(6822): 860–921.
Lane NE, Schnitzer TJ, et al. (2010). “Tanezumab for the treatment of pain from osteoarthritis of the knee”. *N Engl J Med* **363**(16): 1521–31.
Lango Allen H, Estrada K, et al. (2010). “Hundreds of variants clustered in genomic loci and biological pathways affect human height”. *Nature* **467**(7317): 832–8.
Lao O, Lu TT, et al. (2008). “Correlation between genetic and geographic structure in Europe”. *Curr Biol* **18**(16):1241–8.
Lee SH, Wray NR, et al. (2011). “Estimating missing heritability for disease from genome-wide association studies”. *Am J Hum Genet* **88**(3): 294–305.
Lettre G, Jackson AU, et al. (2008). “Identification of ten loci associated with height highlights new biological pathways in human growth”. *Nat Genet* **40**(5): 584–91.
Li JZ, Absher DM, et al. (2008). “Worldwide human relationships inferred from genome-wide patterns of variation”. *Science* **319**(5866): 1100–4.
Li Y, Vinckenbosch N, et al. (2010). “Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants”. *Nat Genet* **42**(11): 969–72.
Lohmueller KE, Pearce CL, et al. (2003). “Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease”. *Nat Genet* **33**(2): 177–82.
MacArthur DG, Balasubramanian S, et al. (2012). “A systematic survey of loss-of-function variants in human protein-coding genes”. *Science* **335**(6070): 823–8.
Maher, B. (2008). “Personal genomes: The case of the missing heritability”. *Nature* **456**(7218): 18–21.
Manolio TA, Collins FS, et al. (2009). “Finding the missing heritability of complex diseases”. *Nature* **461**(7265): 747–53.
Marchini J, Cardon LR, et al. (2004). “The effects of human population structure on large genetic association studies”. *Nat Genet* **36**(5): 512–7.
Marchini J, Howie B. (2010). “Genotype imputation for genome-wide association studies”. *Nat Rev Genet* **1**(7): 499–511.
Marchini J, Howie B, et al. (2007). “A new multipoint method for genome-wide association studies by imputation of genotypes”. *Nat Genet* **39**(7): 906–13.
Marth GT, Yu F, et al. (2011). “The functional spectrum of low-frequency coding variation”. *Genome Biol* **12**(9): R84.
Maurano MT, Humbert R, et al. (2012). “Systematic localization of common disease-associated variation in regulatory DNA.” *Science* **337**(6099): 1190–5.
McCarthy MI, Abecasis GR, et al. (2008). “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. *Nat Rev Genet* **9**(5): 356–69.

McClellan J, King MC. (2010). “Genetic heterogeneity in human disease”. *Cell* **141**; 210–217.
McQuillan R, Leutenegger AL, et al. (2008). “Runs of homozygosity in European populations”. *Am J Hum Genet* **83**(3): 359–72.
Menashe I, Rosenberg PS, et al. (2008). “PGA: power calculator for case-control genetic association analyses”. *BMC Genet* **9**:36.
Menozzi P, Piazza A, et al. (1978). “Synthetic maps of human gene frequencies in Europeans”. *Science* **201**(4358): 786–92.
Metspalu, A. (2004). “The Estonian Genome Project”. *Drug Development Research*, **62**: 97–101.
Metspalu A, Leitsalu L, et al. (2011). “The Estonian biobank – the gateway for the stratified medicine. Research in Estonia. Present and future”. Estonian Academy of Sciences 283–301.
Metspalu M, Romero IG, et al. (2011). “Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia”. *Am J Hum Genet* **89**(6): 731–44.
Montpetit A, Nelis M, et al. (2006). “An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population”. *PLoS Genet* **2**(3): e27.
Morris AP, Voight BF, et al. (2012). “Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes”. *Nat Genet* **44**(9): 981–990.
Morris AP, Zeggini E. (2010). “An evaluation of statistical approaches to rare variant analysis in genetic association studies”. *Genet Epidemiol* **34**(2): 188–93.
Mägi R, Lindgren CM, et al. (2010). “Meta-analysis of sex-specific genome-wide association studies”. *Genet Epidemiol* **34**(8): 846–53.
Mägi R, Pfeufer A, et al. (2007). “Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation”. *BMC Genomics* **8**:159.
Musunuru K, Pirruccello JP, et al. (2010). “Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia”. *N Engl J Med* **363**(23): 2220–7.
Naukkarinen J, Surakka I, et al. (2010). “Use of genome-wide expression data to mine the “Gray Zone” of GWA studies leads to novel candidate obesity genes”. *PLoS Genet* **6**(6): e1000976.
Nelson MR, Bryc K, et al. (2008). “The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research”. *Am J Hum Genet* **83**(3): 347–58.
Neph S, Vierstra J, et al. (2012). “An expansive human regulatory lexicon encoded in transcription factor footprints.” *Nature* **489**(7414): 83–90.
Nicholson G, Rantalainen M, et al. (2011). “A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection”. *PLoS Genet* **7**(9): e1002270.
NHGRI GWAS Catalog. (2012). "National Human Genome Research Institute GWAS Catalog". Data retrieved April 2011 from http://www.genome.gov/gwastudies/
Novembre J, Johnson T, et al. (2008). “Genes mirror geography within Europe”. *Nature* **456**(7218): 98–101.
O'Dushlaine CT, Morris D, et al. (2010). “Population structure and genome-wide patterns of variation in Ireland and Britain”. *Eur J Hum Genet* **18**(11): 1248–54.
OMIM. (2012). “Online Mendelian Inheritance in Man”. Data retrieved April 2011 from http://www.omim.org/statistics/entry

Patil N, Berno AJ, et al. (2001). “Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21”. *Science* **294**(5547): 1719–23.
Patterson N, Price AL, et al. (2006). “Population Structure and Eigenanalysis”. *PLoS Genet* **2**: e190.
Pe'er I, de Bakker PI, et al. (2006).” Evaluating and improving power in whole-genome association studies using fixed marker sets”. *Nat Genet* **38**(6): 663–7.
Pinto D, Darvishi K, et al. (2011). “Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants”. *Nat Biotechnol* **29**(6): 512–20.
Price AL, Helgason A, et al. (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, et al. (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, Stephens M, et al. (2000). “Inference of population structure using multi-locus genotype data”. *Genetics* **155**(2): 945–59.
Reich DE, Lander ES. (2001). “On the allelic spectrum of human disease”. *Trends Genet* **17**(9): 502–10.
Riboli E, Kaaks R. (1997). “The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition”. *Int J Epidemiol* **26**(Suppl 1): S6–14.
Ripke S, Sanders AR, et al. (2011). “Genome-wide association study identifies five new schizophrenia loci”. *Nat Genet* **43**(10): 969–76.
Risch N, Merikangas K. (1996). “The future of genetic studies of complex human diseases”. *Science* **273**(5281): 1516–17.
Rivas MA, Beaudoin M, et al. (2011). “Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease”. *Nat Genet* **43**(11): 1066–73.
Roeder K, Luca D. (2009). “Searching for disease susceptibility variants in structured populations”. *Genomics* **93**(1): 1–4.
Sachidanandam R, Weissman D, et al. (2001). “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms”. *Nature* **409**(6822): 928–33.
Saxena R, Elbers CC, et al. (2012). “Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci”. *Am J Hum Genet* **90**(3): 410–425.
Scott RA, Lagou V, et al. (2012). “Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways”. *Nat Genet* **44**(9): 991–1005.
Service S, DeYoung J, et al. (2006). “Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies”. *Nat Genet* **38**(5): 556–60.
Sklar P, Ripke S, et al. (2011). “Large-scale genome-wide association analysis of bi-polar disorder identifies a new susceptibility locus near ODZ4”. *Nat Genet* **43**(10): 977–83.
Skol AD, Scott LJ, et al. (2006). “Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies”. *Nat Genet* **38**(2): 209–13.
So HC, Gui AH, et al. (2011). “Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases”. *Genet Epidemiol* **35**(5): 310–7

Speliotes EK, Willer CJ, et al. (2010). “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index”. *Nat Genet* **42**(11): 937–48.
Spencer CC, Su Z, et al. (2009). “Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip”. *PLoS Genet* **5**(5): e1000477.
Strachan T, Read A. (2010). “Human Molecular Genetics: 4th edition”. *Taylor & Francis, Inc.* Page: 81.
Stratton MR. (2011). “Exploring the genomes of cancer cells: progress and promise”. *Science* **331**(6024): 1553–8.
Suhre K, Shin SY, et al. (2011). “Human metabolic individuality in biomedical and pharmaceutical research”. *Nature* **477**(7362): 54–60.
Sullivan P. (2012). “Don't give up on GWAS”. *Mol Psychiatry* **17**(1): 2–3.
Tian C, Kosoy R, et al. (2008). “Analysis of East Asia genetic substructure using genome-wide SNP arrays”. *PLoS One* **3**(12): e3862.
Terwilliger JD, Weiss KM. (2003). “Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'”. *Ann Med* **35**(7): 532–44.
Teslovich TM, Musunuru K, et al. (2010). “Biological, clinical and population relevance of 95 loci for blood lipids”. *Nature* **466**(7307): 707–13.
Thorlacius S, Olafsdottir G, et al. (1996). “A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes”. *Nat Genet* **13**(1): 117–9.
Tishkoff SA, Reed FA, et al. (2009). “The genetic structure and history of Africans and African Americans”. *Science* **324**(5930): 1035–44.
Trynka G, Hunt KA, et al. (2011). “Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease”. *Nat Genet* **43**(12): 1193–201.
Venter JC, Adams MD, et al. (2001). “The sequence of the human genome”. *Science* **291**(5507): 1304–51
Visscher PM, Brown MA, et al. (2012). “Five years of GWAS discovery”. *Am J Hum Genet* **90**(1): 7–24.
Voight BF, Kand HM, et al. (2012). “The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits”. *PLoS Gene*t **8**(8): e1002793.
Voight BF, Pritchard JK. (2005). “Confounding from cryptic relatedness in case-control association studies”. *PLoS Genet* **1**(3): e32.
Voight BF, Scott LJ, et al. (2010). “Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis”. *Nat Genet* **42**(7): 579–89.
Wang C, Zöllner S, et al. (2012). “A quantitative comparison of the similarity between genes and geography in worldwide human populations”. *PLoS Genet* **8**(8): e1002886.
Wang DG, Fan JB, et al. (1998). “Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome”. *Science* **280**(5366): 1077–82.
Wang WY, Barratt BJ, et al. (2005). “Genome-wide association studies: theoretical and practical concerns”. *Nat Rev Genet* **6**(2): 109–18.
Weedon MN, Lango H, et al. (2008). “Genome-wide association analysis identifies 20 loci that influence adult height”. *Nat Genet* **40**(5): 575–83.
Wellcome Trust Case Control Consortium. (2007). “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. *Nature* **447**(7145): 661–78.

Wellcome Trust Case Control Consortium. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls". *Nature* **464**(7289): 713–20.
Wichmann HE, Kuhn KA, *et al*. (2011). "Comprehensive catalog of European biobanks". *Nat Biotechnol* **29**(9): 795–7.
Willer CJ, Sanna S, et al. (2008). "Newly identified loci that influence lipid concentrations and risk of coronary artery disease". *Nat Genet* **40**(2): 161–9.
Wood AR, Hernandez DG, et al. (2011). "Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association". *Hum Mol Genet* **20**(20): 4082–92.
Wray NR, Yang J, et al. (2010). "The genetic interpretation of area under the ROC curve in genomic profiling". *PLoS Genet* **6**(2): e1000864.
Yang J, Benyamin B, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". *Nat Genet* **42**(7): 565–9.
Yang J, Ferreira T, et al. (2012). "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits". *Nat Genet* **44**(4): 369–75.
Yang J, Manolio TA, et al. (2011). "Genome partitioning of genetic variation for complex traits using common SNPs". *Nat Genet* **43**(6): 519–25.
Yang WY, Novembre J, et al. (2012). "A model-based approach for analysis of spatial structure in genetic data". *Nat Genet*. 2012; **44(6):** 725–31.
Walters RG, Jacquemont S, et al. (2010). "A new highly penetrant form of obesity due to deletions on chromosome 16p11.2". *Nature* **463**(7281): 671–5.
Zeggini E. (2011). "Next-generation association studies for complex traits". *Nat Genet* **43**(4): 287–8.
Zhernakova A, van Diemen CC et al. (2009). "Detecting shared pathogenesis from the shared genetics of immune-related diseases". *Nat Rev Genet* **10**(1): 43–55.
Zuk O, Hechter E, et al. (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability". *Proc Natl Acad Sci USA* **109**(4): 1193–8.

WEB RESOURCES

1000 Genomes Project – http://www.1000genomes.org/
BBMRI – Biobanking and Biomolecular Resources Research Infrastructure – http://www.bbmri.eu
ENCODE Project – http://genome.ucsc.edu/ENCODE
ENIGMA Network – http://www. enigma.loni.ucla.edu
Estonian Genome Center, University of Tartu – http://www.biobank.ee
Exome Chip Design web-resource – http://genome.sph.umich.edu/wiki/Exome_Chip_Design
deCODE Genetics biobank – http://www.decode.com
HapMap Project – http://www.hapmap.com
National Human Genome Research Institute GWAS Catalog – http://www.genome.gov/gwastudies/
OMIM – Online Mendelian Inheritance in Man – http://www.omim.org/statistics/entry
P3G – The Public Population Project in Genomics – http://www.p3gobservatory.org/
WHO – World Health Organization's 10th release of the International Classification of Diseases – http://www.who.int/classifications/icd

SUMMARY IN ESTONIAN

Genotüpiseerimiskiibi andmete uudsed rakendused Euroopa geneetilise struktuuri analüüsil ning geneetilistes assotsiatsioonuuringutes

Inimese genoomi täisjärjestuse avaldamine on viinud genotüpiseerimis- ja sekveneerimistehnoloogiate kiirele arengule ning teinud võimalikuks tuvastada sadades DNA proovides samaaegselt miljoneid järjestusvariatsioone ja määrata inimese genoomi täisjärjestus vähem kui kahe nädalaga. Laiapõhjaliste genoomiuuringute tulemusena on rohkem kui 3000 DNA järjestusvariatsiooni seostatud enam kui 600 erineva komplekstunnusega. Kuna üksikud järjestusvariatsioonid kirjeldavad enamasti ära vähem kui 1% tunnuse pärilikust komponendist on ülegenoomsetes assotsiatsioonuuringutes vaja kombineerida paljude kohortide andmestikke, et oleks võimalik formuleerida statistiliselt usutavaid järeldusi.

Käesolevas doktoritöös on käsitletud mitmeid ülegenoomsete assotsiatsioonuuringutega seonduvaid aspekte ning eksperimentaalne osa tugineb peamiselt Tartu Ülikooli Eesti Geenivaramu biopanga andmestikule.

1) Hinnati Kirde-Euroopa populatsioonide paiknemist Euroopa alleelisageduste geneetilise struktuuri kaardil. Doktoritöö raames keskenduti eestlaste, lätlaste, leedulaste ning loode-venelaste analüüsile. Kasutades peakomponent analüüsi koostastati geneetilise struktuuri kaart, kus populatsioonide paiknemine oli selges korrelatsioonis geograafilise asendiga. Selgus, et soomlased distantseeruvad nii rootslastest kui ka teistest Loode-Euroopa populatsioonidest, samas kui eestlased paiknevad lähestikku lätlaste, leedulaste ning loode-venelastega. Hinnatud geneetilise distantsi parameetrite väärtused näitavad, et Tartu Ülikooli Eesti Geenivaramu andmete kaasamine suuremahulistesse assotsiatsioonuuringutesse koos teiste Euroopa päritolu kohortidega on õigustatud.

2) Uuriti geneetilise struktureerituse olemasolu kuues nii keeleliselt kui kultuuriliselt eristuvas Loode-Itaalia külakogukonnas. Mudelipõhine struktuuranalüüs näitas, et tugev geneetiline struktureeritus võib esineda isegi kogukondades, mida eelnevalt on peetud geneetiliselt väga ühtseks populatsioonideks eelkõige nende geograafilise eraldatuse tõttu. Võrreldes tuntud populatsiooni isolaatidega, nagu näiteks sardiinlased, tuvastati Loode-Itaalia külakogukondade suurem geneetilise islolatsiooni tase. Lisaks sellele rõhutavad antud uuringu tulemused selgelt suure ning esindusliku valimi olulisust geneetilise struktureerituse analüüsides.

3) Antud doktoritöö raames viidi läbi kaks assotsiatsioonuuringut ning tuvastati uudsed DNA järjestusvariatsioonid vastavalt *MCF2L* geenis, mis suurendavad riski haigestuda osteoartriiti, ning vastavalt *ABCC9* geenis, mis reguleerivad une kestvust.

4) Käesoleva doktoritöö raames rakendati kolme uutset meetodit, et suurendata kirjeledatud päriliku kompondi osakaalu komplekstunnuse varieeruvuses.

Esiteseks, demonstreeriti kehapikkuse näitel, et lisaks aditiivsele komponendile kirjeldavad olulise osa tunnuse varieeruvusest ära ka retsessiivsed alleelid. Teiseks, võttes arvesse uuritavate uneprofiili koostamise aastaega tuvastati DNA järjestusvariant, mis kirjeldas ära 12 % une kestvuse geneetilisest komponendist. Kolmandaks, täpsemate ning suurema katvusega järjestusvariatsioonide andmestikke rakendamine genotüübiandmete parendamiseks võimaldab effektiivsemalt tuvastada uusi komplekstunnnusega seotud lookuseid. Tuginedes antud uuringutele võib väita, et mitmed kirjanduses väljapakutud meetodid komplekstunnuste veel väljaselgitamata pärilikkuse komponentide leidmiseks on õigustatud.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest gratitude to my supervisor Professor Andres Metspalu for his guidance, constructive criticism, thoughtful discussions and support through the entire time of collaboration. His everlasting enthusiasm has been a real inspiration to me.

My deep gratitude goes to Mari Nelis for her valuable collaboration and for technical guidance in array based genotyping methods. Also, thanks to Reedik Mägi and Krista Fischer for guidance in genetic bioinformatics tools and epidemiology methods. I am very grateful to Mait Metspalu for thoughtful discussions on population genetics and for his valuable collaboration.

I thank all my good friends and colleagues from Estonian Genome Center and from Department of Biotechnology for creating a motivating and friendly atmosphere. I am especially thankful to Helene Alavere, Krista Liiv, Merike Leego, Annely Allik, Mari-Liis Tammesoo, Kairit Mikkel, Katrin Männik, Tiit Nikopensius, Toomas Haller, Evelin Mihailov, Merli Hass, Steven Smit, Maris Teder-Laving, Eva Reinmaa, Viljo Soo and Heidi Saulep for advice, technical help and continuous support.

Besides I would like to thank all the collaborators throughout the world for their effort in collecting the samples, performing the analyses, providing the results and sharing even the valuable raw data when needed. My deepest gratitude goes to Massimo Mezzavilla, Pio D'Adamo, Prof. Paolo Gasparini for sharing their valuable FVG village population dataset and also to Karla Allebrandt, Eleftheria Zeggini and James Wilson for the opportunity to participate in their research projects and for their valuable collaboration.

I appreciate the guidance and technical support given by Lauri Anton and Martin Loginov from High Performance Computing Center of University of Tartu.

I would like to thank graduate school in Biomedicine and Biotechnology. Estonian Genome Foundation, ARCHIMEDES and OPENGENE for fellowship nominations that allowed me to participate at conferences, practical courses and workshops.

I would also thank all the data collectors and gene donors in the Estonian Genome Center for their dedication.

Last but not least, I would like to thank my family and friends for supporting me at all times. My warmest hugs go to my beloved wife Kaija for her continuous understanding and everlasting patience.

PUBLICATIONS

CURRICULUM VITAE

Name: Tõnu Esko
Date of birth: 27.01.1985
Citizenship: Estonian
Phone: +372 737 4028
E-mail: tonu.esko@ut.ee

Education:

2001–2004	Nõo Reaalgümnaasium, *Honorable mention (silver medal)*
2004–2007	B.Sc. in Gene Technology, Faculty of Biology and Geography, University of Tartu, *Cum Laude*
2007–2009	M.Sc. in Gene Technology, Faculty of Science and Technology, University of Tartu, *Cum Laude*
2009–2012	Ph.D. student, Faculty of Science and Technology, University of Tartu

Professional employment:

2007–…	Estonian Biocenter, Genotyping specialist
2008–…	Estonian Genome Center of University of Tartu, Specialist
2010–…	Estonian Genome Center of University of Tartu, Member of Council

Scientific work:

During my M.Sc and Ph.D studies I have performed whole-genome genotyping experiments and participated in data analysis of genome-wide genotype data-sets. My research focus has been on the population genetics, precisely I have examined the genetic variation in human genome and its influence on complex human traits. Since 2008 I have been the leading analyst for the Estonian Biobank genome-wide datasets and responsible for coordinating the Biobank collaboration with international genetic consortia.

ELULOOKIRJELDUS

Nimi: Tõnu Esko
Sünniaeg: 27.01.1985
Kodakondsus: Eesti
Telefon: 737 4028
E-post: tonu.esko@ut.ee

Hariduskäik:

2001–2004	Nõo Reaalgümnaasium, *Lõpetatud kiitusega (hõbemedal)*
2004–2007	Bakalaureuse kraad geenitehnoloogia erialal, Bioloogia-Geograafiateaduskond, Tartu ülikool, *Cum Laude*
2007–2009	Magistri kraad geenitehnoloogia erialal, Loodus- ja Tehnoloogiateaduskond, Tartu ülikool, *Cum Laude*
2009–2012	Doktorant, Loodus- ja Tehnoloogiateaduskond, Tartu ülikool

Erialane teenistuskäik:

2007–…	Eesti Biokeskus, Genotüpiseerimise spetsialist
2008–…	Tartu Ülikooli Eesti Geenivaramu, Spetsialist
2010–…	Tartu Ülikooli Eesti Geenivaramu, Nõukogu liige

Teadustegevus:
Magistrantuuri ning doktorantuuri õpingute ajal olen teostanud ülegenoomseid genotüpiseerimise eksperimente ja osalenud hilisemas andmete korrastamises ning analüüsis. Oma senines teadustöös olen keskendunud peamiselt populatsioonigeneetikale – uurinud inimese genoomi järjestusvariatsioone ning hinnanud nende mõju komplekstunnustele. Alates 2008 aastatst olen olnud vastutav uurija TÜ Eesti Geenivaramu suuremahuliste assotsiatsioonanalüüsides ning organiseerinud koostööd ja osalemist rahvusvahelistes konsortsiumites.

LIST OF PUBLICATIONS

Esko T*, Mezzavilla M*, Nelis M, *et al*. (2012). "Genetic diversity of north-eastern Italian population isolates". *Eur J Hum Genet* in press

Yang J, Loos RJF, Powell JE, /…/, **Esko T** *et al*. (2012). "Genetic effects on variability: *FTO* genotype is associated with phenotypic variance of body mass index". *Nature* [Epub ahead of print]

Mõttus R, Realo A, Allik J, **Esko T**, Metspalu A. (2012). "History of the diagnosis of a sexually transmitted disease is linked to normal variation in personality traits." *J Sex Med* [Epub ahead of print]

Scott RA, Lagou V, Welch RP, /.../, **Esko T** *et al*. (2012). "Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways". *Nat Genet* **44**(9): 991–1005.

Morris AP, Voight BF, Teslovich TM, /.../, **Esko T** *et al*. (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". *Nat Genet* **44**(9): 981–990.

Boraska V, Jeroncic A, Colonna V, /…/, **Esko T** *et al*. (2012). "Genome-wide meta-analysis of common variant differences between men and women". *Hum Mol Genet* [Epub ahead of print]

McQuillan R, Eklund N, Pirastu N, /…/, **Esko T** *et al*. (2012). "Evidence of inbreeding depression on human height". *PLoS Genet* **8**(7): e1002655.

arcOGEN Consortium and arcOGEN Collaborators, /…/, **Esko T** *et al*. (2012). "Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study". *Lancet* **380**(9844): 815–23.

Verhoeven VJ, Hysi PG, Saw SM, /…/, **Esko T** *et al*. (2012). "Large scale international replication and meta-analysis study confirms association of the 15q14 locus with myopia". *Hum Genet* **131**(9):1467–80.

Ellinghaus D, Ellinghaus E, Nair RP, /…/, **Esko T** *et al*. (2012). "Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci". *Am J Hum Genet* **90**(4): 636–47.

Pattaro C, Köttgen A, Teumer A, /…/, **Esko T** *et al*. (2012). "Genome-wide association and functional follow-up reveals new loci for kidney function". *PLoS Genet* **8**(3): e1002584

Alavere H, Fischer K, **Esko T**, Leitsalu-Moynihan L, Metspalu A. (2012). "The Estonian Genome Center of the University of Tartu at the disposal of scientists". *Est Med J* **91**(4): 190–198.

Mõttus R, Realo A, Allik J, Deary IJ, **Esko T**, Metspalu A. (2012). "Personality traits and eating habits in a large sample of Estonians". *Health Psychol* [Epub ahead of print]

Stolk L, Perry JR, Chasman DI, /…/, **Esko T** *et al*. (2011). "Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways". *Nat Genet* **44**(3): 260–8.

Luciano M, Lopez LM, de Moor MH, /…/, **Esko T** *et al*. (2011). "Longevity candidate genes and their association with personality traits in the elderly". *Am J Med Genet B Neuropsychiatr Genet* **159B**(2): 192–200.

Ellinghaus E, Stuart PE, Ellinghaus D, /…/, **Esko T** *et al*. (2011). "Genome-wide meta-analysis of psoriatic arthritis identifies susceptibility locus at REL". *J Invest Dermatol* **132**(4): 1133–40.

Gieger C, Radhakrishnan A, Cvejic A, /…/, **Esko T** *et al*. (2011). "New gene functions in megakaryopoiesis and platelet formation". *Nature* **480**(7376): 201–8.

Allebrandt KV, Amin N, Müller-Myhsok B, **Esko T** *et al*. (2011). "A K(ATP) channel gene effect on sleep duration: from genome-wide association studies to function in Drosophila". *Mol Psychiatry* [Epub ahead of print]

Surakka I, Isaacs A, Karssen LC, /…/, **Esko T** *et al*. (2011). "A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol". *PLoS Genet* **7**(10): e1002333.

Chambers JC, Zhang W, Sehmi J, /…/, **Esko T** *et al*. (2011). "Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma". *Nat Genet* **43**(11): 1131–8.

Wain LV, Verwoert GC, O'Reilly PF, /…/, **Esko T** *et al*. (2011). "Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure". *Nat Genet* **43**(10): 1005–11.

Jacquemont S, Reymond A, Zufferey F, /…/, **Esko T** *et al*. (2011). "Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus". *Nature* **478**(7367): 97–102.

Strawbridge RJ, Dupuis J, Prokopenko I, /…/, **Esko T** *et al*. (2011). "Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes". *Diabetes* **60**(10): 2624–34.

Day-Williams AG, Southam L, Panoutsopoulou K, /…/, **Esko T** *et al*. (211). "A variant in MCF2L is associated with osteoarthritis". *Am J Hum Genet* **89**(3): 446–50.

Schumann G, Coin LJ, Lourdusamy A, /…/, **Esko T** *et al*. (2011). "Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption". *Proc Natl Acad Sci U S A* **108**(17): 7119–24.

Speliotes EK, Yerges-Armstrong LM, Wu J, /…/, **Esko T** *et al*. (2011). "Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits". *PLoS Genet* **7**(3): e1001324.

Middeldorp CM, de Moor MH, McGrath LM, /…/, **Esko T** *et al*. (2011). "The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data". *Transl Psychiatry* **1**:e50.

Terracciano A, **Esko T**, Sutin AR *et al*. (2011). "Meta-analysis of genome-wide association studies identifies common variants in CTNNA2 associated with excitement-seeking". *Transl Psychiatry* **1**:e49.

Panoutsopoulou K, Southam L, Elliott KS, /…/, **Esko T** *et al*. (2011). "Insights into the genetic architecture of osteoarthritis from stage 1 of the arcOGEN study". *Ann Rheum Dis* **70**(5): 864–7.

de Moor MH, Costa PT, Terracciano A, /…/, **Esko T** *et al*. (2011). "Meta-analysis of genome-wide association studies for personality". *Mol Psychiatry* **17**(3): 337–49.

Männik K, Parkel S, Palta P, /…/, **Esko T** *et al*. (2011). "A parallel SNP array study of genomic aberrations associated with mental retardation in patients and general population in Estonia". *Eur J Med Genet* **54**(2): 136–43.

Allik J, Realo A, Mõttus R, **Esko T**, Pullat J, Metspalu M. (2010). "Variance determines self-observer agreement on the big five personality traits". *J Res Person* **44**(4): 421–26.

Elks CE, Perry JR, Sulem P, /…/, **Esko T** *et al*. (2010). "Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies". *Nat Genet* **42**(12): 1077–85.

Speliotes EK, Willer CJ, Berndt SI, /…/, **Esko T** *et al*. (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index". *Nat Genet* **42**(11): 937–48.

Heid IM, Jackson AU, Randall JC, /…/, **Esko T** *et al*. (2010). "Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution". *Nat Genet* **42**(11): 949–60.

Lango Allen H, Estrada K, Lettre G, /…/, **Esko T** *et al*. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height". *Nature* **467**(7317): 832–8.

Thorgeirsson TE, Gudbjartsson DF, Surakka I, /…/, **Esko T** *et al*. (2010). "Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior". *Nat Genet* 42(5): 448–53.

Ellinor PT, Lunetta KL, Glazer NL, /.../, **Esko T** *et al*. (2010). "Common variants in KCNN3 are associated with lone atrial fibrillation". *Nat Genet* **42**(3):240–4.

Walters RG, Jacquemont S, Valsesia A, /.../ **Esko T** *et al*. (2010). "A new highly penetrant form of obesity due to deletions on chromosome 16p11.2". *Nature* **463**(7281): 671–5.

Nelis M*****, **Esko T***, Mägi R *et al*. (2009). "Genetic structure of Europeans: a view from the North-East". *PLoS ONE* **4**(5): e5472.

*These authors contributed equally to this work // Antud autorid panustasid võrdselt.

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 phylo-geographic 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. 112 p.
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. **Velda 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.
176. **Mari Nelis.** Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov.** Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster.** Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap.** Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar.** Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül.** Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.

182. **Arto Pulk.** Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü.** Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla.** Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey.** The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp.** Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber.** The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro.** The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold.** Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert.** Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu.** Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik.** ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber.** Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper.** Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak.** Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo.** Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel.** Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus.** Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius.** Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk.** Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.

202. **Arno Põllumäe.** Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht.** Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus.** Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov.** PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova.** Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar.** The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild.** Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p**.**
211. **Pauli Saag.** Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla.** Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve.** Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar.** Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp.** Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero.** Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram.** Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents.** Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.

221. **Katrin Männik.** Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous.** Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu.** Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg.** Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.