UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 702

Genetic and Genomic Analysis of DNA Sequence Variation

PER ERIK LUNDMARK

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2011

ISSN 1651-6206 0346-5462
ISBN 978-91-554-8156-8
urn:nbn:se:uu:diva-158486

Dissertation presented at Uppsala University to be publicly examined in Enghoffsalen, Entrance 50, bottom floor, Uppsala University Hospital, Uppsala. Tuesday, October 25, 2011 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract
Lundmark, P. E. 2011. Genetic and Genomic Analysis of DNA Sequence Variation. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 702. 50 pp. Uppsala. ISBN 978-91-554-8156-8.

The studies in this thesis describe the application of genotyping and allele specific expression analysis to genetic studies. The role of the gene NPC1 in Triglyceride metabolism was explored in mouse models and in humans on the population level in study I. NPC1 was found to affect hepatic triglyceride metabolism, and to be relevant for controlling serum triglyceride levels in mice and potentially in humans. In study II the utility of the HapMap CEU samples was investigated for tagSNP selection in six European populations. The HapMap CEU was found to be representative for tagSNP selection in all populations while allele frequencies differed significantly in the sample from Kuusamo, Finland. In study III the power of Allele specific expression as a tool for the mapping of *cis*-regulatory variation was compared to standard eQTL analysis, ASE was found to be the more powerful type of analysis for a similar sample size. Finally ASE mapping was applied to regions reported to harbour long non-coding RNAs and associated SNPs were compared to published trait-associations. This revealed strong *cis*-regulatory SNPs of long non-coding RNAs with reported trait or disease associations.

Keywords: SNP, NPC1, association study, allele specific expression, tagSNP, non-coding RNA

Per Erik Lundmark, Uppsala University, Department of Medical Sciences, Molecular Medicine, Akademiska sjukhuset, SE-751 85 Uppsala, Sweden.

© Per Erik Lundmark 2011

ISSN 1651-6206 0346-5462
ISBN 978-91-554-8156-8
urn:nbn:se:uu:diva-158486 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-158486)

Till Ulrika och Axel

Supervisors

Prof. Ann-Christine Syvänen, Department of Medical Sciences,
Uppsala University, Sweden

Prof. Karl Michaëlsson, Department of Surgical Sciences,
Uppsala University, Sweden

Chair

Prof. Lars Lind, Department of Medical Sciences,
Uppsala University, Sweden

Faculty Opponent

Prof. Dr. Gert-Jan B. van Ommen, Department of Human Genetics,
Leiden University Medical Center, The Netherlands

Review board

Associate Prof. Jacob Odeberg, Division of Proteomics,
School of Biotechnology, Royal Institute of Technology, Sweden

Assistant Prof. Mattias Jakobsson, Department of Evolutionary
Biology, EBC, Uppsala University, Sweden

Prof. Karin Dahlman-Wright, Department of Biosciences and Nutrition,
Karolinska Institutet, Sweden

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Uronen R-L., **Lundmark, P**., Orho-Melander, M., Jauhiainen M., Larsson, K., Siegbahn, A., Wallentin, L., Zethelius, B., Melander, O., Syvänen, A-C., Ikonen, E. (2010) Niemann-Pick C1 Modulates Hepatic Triglyceride Metabolism and Its Genetic Variation Contributes to Serum Triglyceride Levels. *Arterioscler Thromb Vasc Biol.,* 30(8):1614-20.

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

III Carlsson Almlöf, J., **Lundmark, PE**., Lundmark, A., Maouche, S., Liljedahl, U., Enström, C., Brocheton, J., Sambrook, J., Lloyd-Jones, H., Moore, J., Nelson, CP., Codd, V., Ge, B., Gwilliams, R., Rice, CM., Pastinen, T., Deloukas, P., Goodall, AH., Ouwehand, WH., Cambien, F., Syvänen, AC., Cardiogenics Consortium. The power of allele-specific gene expression analysis for identification of cis-regulatory SNPs. *Manuscript.*

IV **Lundmark, PE.**, Lundmark, A., Ge, B., Liljedahl, U., Enström, C., Adeou, V., Rice, CM., Pastinen, T., Goodall AH., Cambien, F., Deloukas, P., Ouwehand, WH., Syvänen, AC. on behalf of the Cardiogenics Consortium. Identification of trait-associated single nucleotide polymorphisms with cis-regulatory effects on long non-coding RNAs. *Manuscript.*

Reprints were made with permission from the respective publishers.

Additional Publications by the Author

Nordmark, G., Kristjansdottir, G., Theander, E., Appel, S., Eriksson, P., Vasaitis, L., Kvarnström, M., Delaleu, N., **Lundmark, P.**, Lundmark, A., Sjöwall, C., Brun, JG., Jonsson, MV., Harboe, E., Gøransson, LG., Johnsen, SJ., Söderkvist, P., Eloranta, ML., Alm, G., Baecklund, E., Wahren-Herlenius, M., Omdal, R., Rönnblom, L., Jonsson, R., Syvänen, AC. (2011) Association of EBF1, FAM167A(C8orf13)-BLK and TNFSF4 gene variants with primary Sjögren's syndrome. *Genes Immun.,* 12(2):100-9.

Saetre, P., **Lundmark, P.**, Wang, A., Hansen, T., Rasmussen, HB., Djurovic, S., Melle, I., Andreassen, OA., Werge, T., Agartz, I., Hall, H., Terenius, L., Jönsson, EG. (2010) The tryptophan hydroxylase 1 (TPH1) gene, schizophrenia susceptibility, and suicidal behavior: a multi-centre case-control study and meta-analysis. *Am J Med Genet B Neuropsychiatr Genet.,* 153B(2):387-96.

Andreou, D., Saetre. P., **Lundmark, P.**, Hansen, T., Timm, S., Melle, I., Djurovic, S., Andreassen, OA., Werge, T., Hall, H., Agartz, I., Terenius, L., Jönsson, EG. (2009) Tyrosine hydroxylase Val81Met polymorphism: lack of association with schizophrenia. *Psychiatr Genet.,* 19(5):273-4.

Saetre, P., Agartz, I., De Franciscis, A., **Lundmark, P.**, Djurovic, S., Kähler, A., Andreassen, OA., Jakobsen, KD., Rasmussen, HB., Werge, T., Hall, H., Terenius, L., Jönsson, EG. (2008) Association between a disrupted-in-schizophrenia 1 (DISC1) single nucleotide polymorphism and schizophrenia in a combined Scandinavian case-control sample. *Schizophr Res.,* 106(2-3):237-41.

Dahlgren, A., **Lundmark, P.**, Axelsson, T., Lind, L., Syvänen, AC. (2008) Association of the estrogen receptor 1 (ESR1) gene with body height in adult males from two Swedish population cohorts. *PLoS One,* 3(3):e1807.

Warensjö, E., Ingelsson, E., **Lundmark, P.**, Lannfelt, L., Syvänen, AC., Vessby, B., Risérus, U. (2007) Polymorphisms in the SCD1 gene: associations with body fat distribution and insulin sensitivity. *Obesity,* 15(7):1732-40.

Ingelsson, E., Risérus, U., Berne, C., Frystyk, J., Flyvbjerg, A., Axelsson, T., **Lundmark, P.**, Zethelius, B. (2006) Adiponectin and risk of congestive heart failure. *JAMA.,* 295(15):1772-4.

Contents

Abbreviations

ANOVA	Analysis of variance
ASE	Allele specific expression
cDNA	Complementary dideoxy ribonucleic acid
CNV	Copy number variation
eQTL	Expression quantitative trait loci
EST	Expressed sequence tag
FRISC II	Fast revascularization during instability in coronary Artery Disease II study
GWAS	Genome wide association study
HapMap	The international haplotype mapping project
kb	Kilobase
LD	Linkage disequilibrium
Mb	Megabase
MDC	Malmö Diet and Cancer study
ncRNA	Non-coding ribonucleic acid
SNP	Single nucleotide polymorphism
TG	Triglyceride
ULSAM	Uppsala longitudinal study of adult men
VNTR	Variable number tandem repeat

Introduction

The concept of offspring inheriting traits from their parents must have been known since the dawn of the modern human. The domestication of animals and plants with the artificial selection that culture and breeding implies also suggest a notion of inheritance. However it was only in 1866 some fundamental understanding of the underlying mechanisms was achieved. Gregor Mendel published his "Experiments on Plant Hybridization" in Procedings of the Natural History Society of Brünn (Mendel 1866), a paper by many considered the birth of genetics, with his observations on how traits in plants were carried over to the following generations. While Mendel's work went largely unrecognized during his lifetime, it was brought into the research community's eye again at the beginning of the 20th century by researchers who had been working on similar questions in parallel, and Mendel's work was rediscovered. During the same period the term "genetics" was popularised in the scientific community to describe the study of inheritance.

The modern field of molecular genetics, the study on molecular level of the structure and function of genes, is closely linked to the technological advances made in methods for analysing the variation that is a part of what makes us all different. Analysis instruments have progressed from being able to look at a small number of positions a little over a decade ago, to the current systems with capacities to query millions of positions for thousands of samples, and the scale of the questions researchers can ask has expanded along with it. We now work in an exciting time when the sequencing of whole human genomes has become possible, both practically and economically, on a single instrument in a small laboratory. It is truly the beginning of the genomic era.

Variation in the human genome

There are several classes of DNA variation in the human genome. Over the years different types have been the focus of the research community, either as markers in studies to pinpoint traits or disease causation, or as in them self potential causes of disease. Below those of main interest for mapping traits and disease studies are listed.

Variable number tandem repeats

Variable number tandem repeats or VNTRs (Jeffreys et al. 1985) are made up from repeated segments of DNA where the identity of individual alleles is based on the number of repetitions. The class is divided into two main groups based on the length of the repeated segment: minisatellites for a repeat unit of around 10-20 bp, and microsatellites for smaller repeats, usually up to four bp long with an entire repeat segment that tend to be less than 100 bp. Microsatellites are also known as short tandem repeats (STRs). Microsatellites were widely used for linkage studies where the goal is to link a phenotype of interest to a specific chromosome region in a family material. A key advantage of the use of microsatellites in such studies is that there can be a large number of different alleles for each analysed microsatellite, thereby making the variation more informative when the transmission is tracked through the generations in a pedigree, since two separate individuals are less likely to share the same allele by chance rather than relation. Microsatellites are still widely in use for forensic purposes, but for genetic linkage studies the use of single nucleotide polymorphisms has become more common than STRs due to new technology allowing analysis of massive panels of SNPs through different kinds of microarray technologies.

Single nucleotide polymorphisms

Single nucleotide polymorphisms, or SNPs for short, are the main marker variation in use today for genetic studies. It is the variation in a single nucleotide between individuals, present in the population at a frequency larger than 1%. Although they mainly are biallelic, and thereby less informative than for example the microsatellites mentioned earlier, their abundance in the genome and ease of automation in analysis, have made them the marker of choice today. Several million SNPs can be queried in a single sample simultaneously in modern genotyping instruments (Steemers et al. 2006), providing means to take analysis to a genome-wide level, and at the same time generate fine mapping data from the same analysis.They are by number the most common type of variation in the genome, and approximately one nucleotide in 1000 is different between two individuals (Wheeler et al. 2008). The sequencing of James D. Watson's genome to 7.4-fold redun-

dancy (Wheeler et al. 2008) resulted in 3.3 million single base differences compared to the human reference sequence, out of which 10,654 caused amino-acid substitutions in the coding sequences of proteins.

Copy number variation

While the SNP is the variation numbering the highest in the genome, the type affecting the largest proportion of bases are copy number variations (CNVs). These are duplications, insertions and deletions of larger regions up to several mega bases (Mb). Some 5% of the human genome is variable in the normal healthy population due to CNVs. For example Conrad et al. observed a cumulative length of 24Mb (0.78% of the genome) that differed in CNVs between two individuals, and 113Mb (3.7%) across all 41 individuals they assayed with comparative genomic hybridization (CGH) (Conrad et al. 2010). Their tiling CGH approach was mainly able to detect regions of duplications and deletions larger than 1kb.

Massively parallel sequencing can be used to generate CNV data that include also smaller variations. Methods include read depth analysis and paired-end mapping where the ends of DNA fragments of known size are sequenced, and the spacing of the reads in the reference genome is compared with the known fragment length to detect unexpected mappings. Mills et al. reported the results of the Structural Variation Analysis Group of the "1000 genomes project" (Mills et al. 2011) who are sequencing a large number of human genomes and at the time had generated 4.1 terabases of raw sequence data. They found that their sequence based approach detected a far larger proportion of smaller CNVs than the CGH approach of Conrad et al., and that this method allowed them in many cases to map the breakpoints precisely to be able to analyze the driving forces behind CNV formation. They also found that CNVs in many cases disrupted exons and other functional elements, indicating potential functional impact.

Mapping disease genes

When discussing disease that are caused or predisposed for by genetic variation, a division is usually made between mendelian (monogenic) disease and complex (polygenic, multifactorial) disease. A mendelian disease is one that is mainly caused by alterations in one gene, and that displays one of the classical inheritance patterns such as autosomal dominant or recessive inheritance etc. Penetrance is usually high meaning that carriers of an alteration often tend to display the phenotype. The alterations that cause these conditions tend to be rare, often with unique defects in different families. Examples of such diseases are Cystic Fibrosis (autosomal recessive) (Kerem et al.

1989), Huntington's disease (autosomal dominant) (MacDonald 1993) and Duchenne muscular dystrophy (X-linked) (Den Dunnen et al. 1989). Complex disease on the other hand is the opposite in most aspects; many different genes together with environmental effects make small contributions to the outcome. It is hard to predict if a person will get the disease, even if we know that the person is a carrier of known risk factors. Examples of complex disease are many common diseases affecting large parts of the population like cardiovascular disease (Schunkert et al. 2011) and diabetes (Voight et al. 2010).

The fundamental differences in the nature of mendelian and complex disease also mean that different tools are preferable when looking for the causes of disease. The use of genetic linkage studies has been very successful in discovering genes causing mendelian disorders (Botstein et al. 2003). This method works by looking for co-segregation of the phenotype with the genetic markers in pedigrees. The markers in use are microsatellites or SNPs. Parametric (model based) linkage requires certain parameters to be specified such as penetrance and mode of inheritance, these cannot be specified for complex disease, but can usually be estimated for mendelian traits. For complex disorders non-parametric linkage has to be used since this avoids the need for specifying model parameters by not testing a specific inheritance model. Instead the software looks for an increased sharing among affected individuals of specific parts of the chromosomes. Overall linkage has had limited success in finding the causes for complex disease.

The method mainly used in the search for the causes of complex disease is the genome-wide association study (GWAS). This method attempts to associate genotypes across the genome with the phenotype of interest. Association studies are treating the whole population as an extended family for analysis, with the goal of detecting segments of chromosome that is shared in cases as result of distant common ancestry with a person where the defect appeared. Linkage and association analysis are basically complements since they perform best on different types on problems. Association has the most power to detect small effects (Risch et al. 1996), hence its use for complex disease, but since it works on distantly related subjects, more recombination limits the extent of linkage, so a dense marker map is required, but higher resolution is also accomplished. Linkage disequilibrium (LD), the correlation of genotypes in different loci, makes detecting a region easier since each of the correlated markers can work as a proxy to detect a non-genotyped variation. The commonly used measure for LD in this context is r^2. For the purpose of detecting phenotype association through a proxy marker, a level of $r^2 > 0.8$ has been widely used, although the limit would depend on for example effect size. An interpretation of the effect of the r^2 value for an association study is: If the causative SNP could be detected in the study with a

sample size of N, the detection of the same signal in an LD proxy with $r^2 = x$ to the causative SNP, would require N/x samples. In other words, the effective sample size in the LD proxy is xN. Even if association studies should be more powerful for the detection of small effects, finding the cause of complex disease has been difficult. Only in recent years have the well validated findings begun to increase (Speliotes et al. 2010; Voight et al. 2010; Schunkert et al. 2011). Current projects are beginning to amass massive sample sets of in some cases tens and hundreds of thousands of samples. What the future will bring only time will tell.

The HapMap project

The International HapMap project was a key project in enabling straightforward design of genome-wide association studies. It started in October 2002 and its stated goal was to determine the common patterns of DNA sequence variation in the human genome to further medical research (The HapMap Consortium 2003). It would also make this information freely available in the public domain. In 2002, SNP genotyping was a relatively low success rate analysis, due to lacking knowledge of variable positions in the genome and their allele frequencies in different populations. The SNP information available in the central repository for SNP data, dbSNP, was often based on *in-silico* studies of different kinds of sequencing data without experimental validation of the alleles and had various levels of quality control. When a panel of SNPs was genotyped in a sample set, large proportions of markers would often be non-polymorphic due to lack of validation data during the design of the SNP panel and some regions would be poorly covered since few SNPs were known (Gabriel et al. 2002; Carlson et al. 2003; Reich et al. 2003).

The initial phase of the HapMap project included large scale SNP discovery by sequencing, and in just over a year more than doubled the 2.8 million SNPs that were available in the dbSNP at the start of the project. In the first genotyping phase of the project, approximately 1.3 million SNPs across the genome were genotyped in samples of of European, African and Asian origin, 270 samples in total (The HapMap Consortium 2003; The HapMap Consortium 2005). The goal of the first phase was to genotype a SNP at least every five kb in the genome. Phase two then extended the SNP panel by additional genotyping of 2.1 million SNPs in the same samples resulting in an average SNP density of about one per kb (Frazer et al. 2007). Phase two was estimated to cover about 25-35% of all common SNPs, i.e. a minor allele frequency larger than 5%. Finally phase three extended the sample sets to include a more divergent set of populations for example several African populations, African Americans and Gujarati Indians. In the final phase the

sample number increased from the 270 individuals and four populations in phase one and two, to 1301 samples from 11 populations. 1.6 million SNPs were genotyped in the full sample set and ten 100kb regions were sequenced in 692 individuals (Altshuler et al. 2010).

The main reasons for initiating the HapMap project was the lack of a dense SNP map across the genome mentioned above, but also the recognition from work studying linkage disequilibrium (LD) that alleles in SNP markers tended to be correlated to each other (Daly et al. 2001; Patil et al. 2001; Gabriel et al. 2002). Two SNPs could in a sense carry redundant information for a disease study. This indicated the possibility that one could scan a genomic region or the entire genome, without having to query every single SNP for association to a disease or other phenotype. Given prior knowledge of the correlations between markers, a study using a reduced number of SNPs selected to be as representative of the full SNP set as possible, should be able to capture essentially the same information as the full set. This would save money in genotyping or allow larger regions to be investigated. To generate this knowledge of SNP correlation patterns, or LD patterns, was one of the main goals of the HapMap project. SNPs selected to be representatives of a larger group are often called haplotype tagging SNPs (Johnson et al. 2001), or tagSNPs, and this methodology for selection of SNP panels has dominated the association studies since the release of the HapMap data.

The ENCODE project

The ENCODE project or "Encyclopedia of DNA Elements" (http://genome.ucsc.edu/ENCODE/), launched in September 2003, funded by The National Human Genome Research Institute (NHGRI). The ambitious goal was declared to be the discovery of all functional elements in the human genome sequence. It started with a pilot phase where 35 research groups contributed over 200 different datasets focused on a selection of 30Mb, or about 1% of the genome, that was divided between 44 regions. These regions had been selected either because they were well studied with reference data sets available for comparisons, or as randomly selected regions for analysis. This setup would ensure that the analyzed regions were diverse enough to make evaluation of different experimental and computational methods representative for the whole genome. In parallel with this, a technology development effort took place to establish different methods for analysis that could cope with the scale of the project. The pilot phase was completed in 2007 and results from the many types of analysis applied to the data were published (Birney et al. 2007). Since the dataset for the regions was very rich in information on functional elements, and on a very large scale, a number of important findings were made in the analysis.

The project used hybridization of RNA to unbiased tiling arrays and tag-sequencing of cap-selected RNA combined with the previously available DNA and EST sequences to indentify transcripts. The results confirmed previous findings (Bertone et al. 2004; Cheng et al. 2005) that indicated that nearly all of the human genome is transcribed, not only regions containing protein coding genes and closely associated transcripts. ENCODE data indicated that somewhere between 74% and 93% of the genome was covered by primary transcripts depending on the type of validation applied. The project could also confirm the presence of many intercalated transcripts tying together genes thought to be separate units. This had been reported previously (Bertone et al. 2004; Carninci et al. 2006) but been met with some doubt about their functional relevance. A large number of non-coding transcripts were identified, also in regions thought to be transcriptionally void. Regulatory sequences were found to be evenly distributed around transcription start sites instead of having an upstream bias, and analysis revealed that chromatin accessibility and patterns of histone modification could strongly predict location and activity of transcription start sites. It was determined that 5% of the genome was evolutionary conserved within mammals based on the pilot data, and also that 60% of these had been assigned some function by the other types of experiments. Another surprising finding was that many functional regions actually were not conserved among mammals.

As the pilot phases had been successfully completed, NHGRI awarded new funding to move the project into the true production phase, scaling the project to the full human genome. Data sets for the whole human genome have been released continuously during 2011 as well as ENCODE data for mouse to complement the human information.

Non-coding RNA

The “Central dogma of molecular biology” (Crick 1970) where DNA is transcribed into mRNA that is translated into protein, is more and more being challenged. In the classical view, non-coding RNAs were mainly limited to structural RNA like ribosomal RNA and transfer RNAs closely connected to protein production, but as our understanding of transcription in the genome have evolved it is clear that there are more things at play. Non-coding RNA as the name implies are RNAs that do not encode a sequence of amino acids to produce a protein. They are mainly divided in two groups based on the size of the transcript, where RNAs shorter than 200 bp include short or small ncRNAs and micro RNAs (miRNAs). Larger RNAs are usually called long ncRNAs (lncRNA) or sometimes long intergenic RNAs (lincRNAs) if not overlapping with protein coding transcripts. Knowledge is limited of the functions and importance of large ncRNAs in the human genome. A few

lncRNAs are comparatively well studied like XIST, involved in X inactivation (Augui et al. 2011). RNA is expressed from the future inactive X chromosome in females, to coat the chromosome of origin through a *cis* activity that leads to a histone modification that down-regulates expression from the coated chromosome, this has been shown to involve the recruitment of Polycomb repressive complex 2 (Maenner et al. 2010). XIST itself may also be under control of other ncRNAs, indicating an intricate regulatory system involving non-coding transcripts. Also Air (Nagano et al. 2008) and Kcnq1ot1 (Pandey et al. 2008) mediate repression through *cis*-effects, again through chromatin interaction and histone modifications. An example of a ncRNA acting in *trans* is HOTAIR (Rinn et al. 2007), this ncRNA is located in the HOXC gene cluster on chromosome 12. When knockdown experiments were conducted by Rinn et al., they did not detect any regulation of the nearby genes within the same cluster as expected from previously known *cis*-acting ncRNAs, but repression was removed from a 40 kb region in the HoxD cluster on chromosome 2, demonstrating ncRNA *trans*-effects for the first time (Rinn et al. 2007). More recently a study looked for ncRNAs regulated by the transcription factor p53, an important tumor suppressor gene involved in response to DNA damage (Huarte et al. 2010). They discovered among other ncRNAs, lincRNA-p21, a lincRNA that represses multiple genes in different genomic locations in response to p53 activation. They also showed that heterogeneous nuclear ribonucleoprotein K (hnRNP-K), a part of a repressor complex, interacted with lincRNA-p21.

Still little is known of the functions of lncRNAs in the human genome in general, but the pattern that is emerging from the few partially known mechanisms, point to histone modification and epigenetic regulation. Whether this will remain a theme in lncRNA regulation as knowledge in the field expands, time will tell.

Among the shorter forms of non-coding transcripts, miRNA are short ncRNAs on average 22 bp long. They tend to act as post-transcriptional regulators by binding a complementary region on the mRNA target and repress translation or induce target degradation (Bartel 2009). The miRNA database miRBase (Kozomara et al. 2011) release 17, lists over 1400 human entries, and the numbers are growing. Other small RNAs acting in a similar context are small interfering RNAs (siRNA) (Kawasaki et al. 2005) and Piwi-interacting RNAs (piRNAs) (Siomi et al. 2011).

Detecting regulatory variation in the genome

The search for variation that affects gene expression provides a tool to discover regions important for gene regulation, and possibly disease causation.

Two methods used for these types of studies are the association of SNP genotypes to either the total expression of a gene or region here referred to as standard eQTL analysis, or the differential expression of the two chromosomes, known as allele specific expression analysis. While the two methods in many ways accomplish similar things, there are also profound differences due to the phenotype that is measured in the two techniques.

Association to total expression

In standard eQTL analysis (Cheung et al. 2005; Stranger et al. 2005), the genotypes of SNPs in the genome are tested for association to expression levels in much the same way as any continuous phenotype. The main difference compared to tests versus disease or biochemical measurements, is the many different phenotypes tested. The signal from each expression probe or transcript in effect is a phenotype of its own. The technique can be used to detect both *cis*- and *trans*-effects, that is effects located on the same chromosome in proximity to the expressed region, or distant or located on other chromosomes. For attempting to detect *trans*-effects one should be aware of the significant multiple testing problem inherent in testing all markers genotyped across the genome versus all expression phenotypes measured, and the resulting requirements for adjustments to p-values and large sample sizes to attempt to counter these effects. Batch effects and other unknown variation can have a detrimental effect on the measured expression signals (Fare et al. 2003; Akey et al. 2007; Branham et al. 2007) and weaken association signals.

Association to allele specific expression

Allele specific expression analysis is similar to standard eQTL analysis in the sense that genotypes and expression signals are used to look for SNP effects on expression, but here the expression phenotype is the relative expression of the two alleles of the transcript (Yan et al. 2002; Pastinen et al. 2004; Ge et al. 2009; Milani et al. 2009). To measure such a phenotype a system is needed that can discriminate between them, and a relatively cost effective method of analysis is quantitative SNP genotyping of RNA via cDNA. Any heterozygous SNP in the transcript will generate separate signals from each of the alleles since they carry different nucleotides, and these are detected by quantitative genotyping. The genotype signals from genomic DNA are used as a control in the analysis to adjust for any assay efficiency differences between the two alleles such as incorporation rate for the different labelled nucleotides. The fact that comparisons are made within each sample give ASE analysis some unique properties. The two alleles that are compared come from the same cells, were prepared in the same reaction, and were genotyped in the same spot on the chip. Essentially they will have

shared all environmental conditions with each other, this eliminates noise and environmental effects from the analysis. From a biological perspective, the two alleles will also have shared the same exposure to *trans*-acting factors like transcription factors, only leaving the *cis*-effects local to each chromosome to differ from each other. This makes the ASE assay a purely *cis*-detecting assay. It also means that the allele specific signal will be cleaner than the total expression counterpart due to fewer effects from external factors influencing the results, but as a result *trans*-acting factors will not be detected in the assay. Another interesting aspect of SNP ASE analysis is that the assay is sensitive enough to detect the differential expression signal in pre-mRNA in introns, indeed this is the main signal used for ASE calling when using normal whole-genome genotyping arrays, as SNPs in exons are relatively rare. Detecting pre-mRNA also mean that effects of post-translational regulation are smaller, again generating a cleaner *cis*-regulatory association signal. The general workflow of the haplotype based ASE approach (Ge et al. 2009) used in papers III and IV is listed below.

1. Quantitative genotyping of cDNA and DNA. For the analysis SNPs play a double role, both as discreet genotypes when used for grouping the samples in the association test, but also as a continuous expression phenotype for ASE calculations.
2. Normalization of fluorescence signals. The system used for genotyping has an intensity bias in the channel balance, a quadratic normalization is used to remove it.
3. Calculation of ASE values for individual SNPs. ASE is the difference in fraction between RNA (cDNA) and genomic DNA.

$$\frac{A1_{RNA}}{A1_{RNA}+A2_{RNA}} - \frac{A1_{DNA}}{A1_{DNA}+A2_{DNA}}$$

4. Haplotype phasing of the chromosomes. To improve robustness of the ASE signals in analysis windows (generally corresponding to a transcript region or sub region), values are averaged across the window. As the polarity of the ASE value depends on the order of the alleles, the haplotype information is necessary to be able to order alleles according to chromosome and combine data from multiple SNPs in a region. See Figure 1 A and B for an example of the effects of haplotyping on ASE.
5. Combination of ASE in SNPs within the analysis window.
6. Association testing of SNPs in the surrounding regions vs. ASE levels. Haplotypes are ordered according to the alleles in the tested SNP.

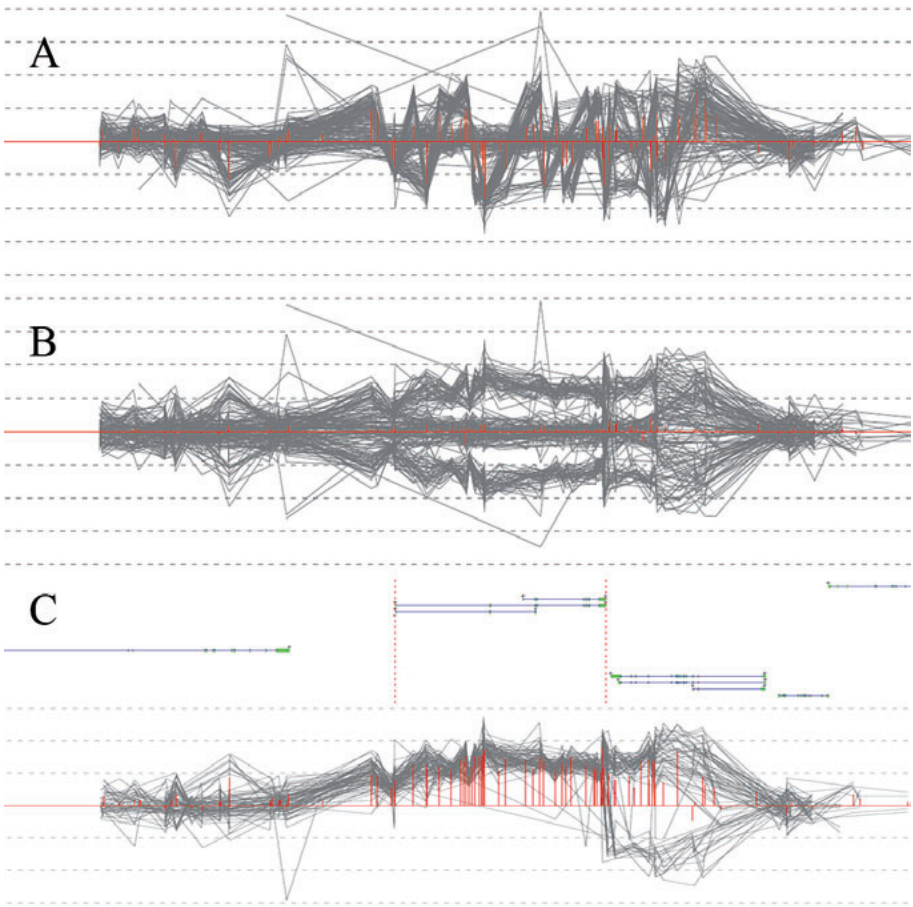

Figure 1. **A.** An example region with strong ASE without haplotype phasing, grey lines are the ASE values of single individuals in heterozygous SNPs along the chromosome, red bars are average ASE in SNPs. **B.** The example region with SNPs haplotype phased, polarity for SNPs within individuals now align. Three groups are visible, positive, negative and no ASE. **C.** Chromosomes ordered according to the alleles of a strongly associated SNP upstream of the displayed region, only heterozygotes for the associated SNP shown. All samples heterozygous for the associated SNP have strong ASE in the analysis window (red dashed lines), the same *cis*-regulatory SNP allele cause over-expression in all samples.

Modern sequencing techniques could also be used to generate data for ASE by counting the relative number of reads for the two haplotypes in each individual by distinguishing the separate alleles by SNPs in the sequence. Most likely this is the direction this type of analysis will take when costs for transcriptome sequencing drop to manageable levels for large sample sets and the deep sequencing necessary to detect rare transcripts.

Genotyping and sequencing technology

The developments in genotyping and sequencing technology during the last decade have been truly staggering. The increase in capacity for instruments are beyond what most people could dream of at the turn of the millennia and have enabled us to move the scale of experiments from genes and regions to complete genomes. Especially in the case of massively parallel sequencing technology, the new instruments have brought sequencing into completely new fields and applications. Below the allele discrimination principles for technologies used in the papers of this thesis are explained.

Genotyping

Minisequencing / single base extension

Minisequencing (Syvanen et al. 1990) also known as single base extension (Fan et al. 2000) among other names, is a DNA polymerase assisted primer extension assay. It works through the extension of a single primer annealing next to the SNP position. Included in the reaction mixture are labelled ddNTPs that will terminate extension after the first nucleotide has been incorporated in the actual SNP position. The SNP genotype can then be read by querying the labels. Labelling methods have included radioisotopes, mass and fluorophores in different instruments and applications, but the main detection method used today on genome-wide scale is fluorescence detection. The assay makes use of the high sequence specificity of the DNA polymerase when discriminating between alleles. It is a testament to the efficiency of the assay that it has been scaled to several million parallel assays through modern array technology (Steemers et al. 2006).
Minisequencing assays were used for genotyping in papers I, II, III and IV. In paper I and II PCR was used for amplification of DNA before the assay, while papers III and IV used whole genome amplification as a part of the Illumina Infinium assay.

Allele specific primer extension

Allele specific primer extension (Wallace et al. 1979) share many features with minisequencing; it is again utilizing the features of the DNA polymerase to discriminate between alleles. In this assay, two primers are used where the last nucleotide of the primers match the two alleles in the SNP position, i.e. two allele specific primers. Each primer will only be extended on the allele where the 3'-end of the primer is a perfect match.
Allele specific primer extension was used in papers III and IV as a part of the Illumina Infinium assay.

Sequencing

The massively parallel sequencing technology has been one of the great catalysts in genomic science. By generating huge datasets in short time they have brought what was only possible to generate in large sequencing factories on classical capillary sequencers into a single benchtop instrument. The power of modern sequencing is not only that it has enabled the rapid and cheap generation of high coverage genomic sequence data, but the many other techniques that it has made possible on a large scale. Two of these techniques that have been extensively used in for example the ENCODE project, are transcriptome sequencing (Nagalakshmi et al. 2008) and chromatin immunoprecipitation (ChIP) sequencing (Johnson et al. 2007). In transcriptome sequencing, cDNA is sequenced, allowing not only the assembly of transcripts present in the sample but also quantification with a very wide dynamic range. In ChIP-sequencing proteins are crosslinked to DNA, the complex is then pulled down with an antibody directed towards a protein of interest. By then releasing the protein from the DNA and sequencing the remaining DNA fragments, a genome-wide protein-DNA interaction map is generated by mapping the reads to the genome. Both these techniques were used for validation purposes in paper IV. In effect sequencers have become the cutting edge also in expression studies and protein-DNA interaction. An important point is that the system captures what is present, regardless of prior assumptions, and can thereby discover new phenomena that were entirely unexpected.

Read length for the types of instruments used in paper IV, the Illumina Genome analyzer II and HiSeq2000 (read lengths in paper IV: 75 and 100 bp), is still short compared to classic capillary sequencing. This brings its own set of problems, especially when used for genome sequencing where e.g. repeat regions may not be possible to span with the shorter reads, it also makes haplotyping difficult since multiple heterozygote positions may not be covered in one read. New types of software had to be created to deal with the data in an optimal way (Zerbino et al. 2008; Langmead et al. 2009; Trapnell et al. 2009), but as the systems have matured, read-length has increased step by step. The current maximum read lengths are now for example 150 bp on the Illumina Genome Analyzer.

The general procedure when sequencing a sample with the Illumina sequencing systems begin with random fragmentation of the sample DNA, followed by an adapter ligation to both ends of the fragments. Single stranded fragments are then randomly bound to the surface of a flow cell channel where additional primers are attached to the surface as well. The fragments anneal to the primers on the surface to form a bridge, attached to the surface in one end, and to the primer in the other. On addition of unlabeled nucleotides and

enzyme the primers are extended to form a double stranded molecule, and when denatured, the result is a copy of the fragment attached to the flow cell by the opposite end as the mother fragment. Multiple cycles of this procedure referred to as to as solid phase bridge amplification, results in an up to 1000-fold amplified colony of clonal sequences on the flow cell surface, and many millions of clusters are formed in the flow cell channel. Sequencing then commence by adding sequencing primers and four labelled reversible terminators. The terminators are incorporated, the fluorescence read i.e. incorporated base, and termination reversed. Then a new cycle will start with another incorporation, read-out, and finally unblocking until the read length has been reached. Analysis of the images from the flow cell is not trivial, as more and more sequences will become out of phase within each cluster, the longer the cycles progress. This is due to incomplete incorporations and deblocking in the cluster that make some sequences lag compared to the optimum, also some terminators may fail and extension continue an extra nucleotide to give sequences that lead the intended pace. As these problems accumulate over the cycles, they are the main limiting factor for read length together with the survival of the DNA polymerase due to long run-times. They are also the reason for the error profile of reads from the technique, where error rate is much higher in the ends of reads, especially those approaching the read length limit of the current chemistry.

The 1000 genomes project

The aim of the 1000 genomes project (http://www.1000genomes.org) is to discover most human variation with at least an allele frequency of 1%. The chosen approach is to lightly sequence the genome of a large set of individuals from diverse populations around the world. The project intends to make use of the presence of LD between many markers in the human genome by using imputation to predict variation that is missing due to the low sequence coverage. This should allow them to make up for low sequencing coverage by identifying the represented haplotypes in the large sample set, and then fill the gaps in each individuals haplotypes based on the data from the large sample set. The project started in January of 2008 with a pilot phase, during which strategies and technology for whole genome shotgun sequencing was evaluated. The pilot consisted of three subprojects: low-coverage sequencing of 179 individuals from four populations, high-coverage sequencing of two trios (parents and one child) and finally exon sequencing of 8140 exons in 697 individuals from seven populations (Table 1). The last of the pilot sequencing effort was completed in June 2009, and the results of the pilot study were published a year later (The 1000 Genomes Project Consortium 2010).

Table 1. *The sub projects of the 1000 genomes pilot.*

Sub pilot	Purpose	Coverage	Samples	Completion
Low coverage samples	Evaluate correlation / imputation strategy	2-4X	179 samples, four populations	Oct. 2008
High coverage trio samples	Evaluate platforms and necessary coverage	20-60X	2 parent-parent-child trios	Oct 2008
High coverage exon capture	Evaluate techniques for region capture and sequencing	50X	697 samples, seven populations	June 2009

In all 4.9 terabases of DNA sequence were generated by nine different sequencing centers, using three different sequencing technologies. Highlights from the results include 15 million SNPs, 1 million small insertion-deletions and 20,000 structural variants of which around 90% were novel depending on class. The project claimed to cover more than 95% of all variation present in any individual in their dataset, and found that an average individual carries 250 – 300 loss of function variants in known genes, with some 50 – 100 previously reported as linked to disease (The 1000 Genomes Project Consortium 2010).

After completing the pilot phase, work moved on to the full production phase. Currently the aim is to sequence around 2,500 individuals from around the world. Data sets are continuously being released from the project, e.g. the latest data release in June 2011 included genotypes from 1,094 individuals.

Aim of this thesis

The aim of this thesis was to apply state of the art genotyping and sequencing technology to the study of human genetic variation and disease.

The present study

The papers included in this thesis cover in many ways the development in the field of human association studies, from a candidate gene study on a single gene, through investigations on the HapMap data set, to whole genome analysis of variation and expression through array technology and massively parallel sequencing. What were once great obstacles of practicality and cost are now only memories in light of the massive throughput of modern genotyping instruments and sequencers. What the future holds no one knows, but with even more powerful technologies around the corner like single molecule sequencing and a budding understanding of transcription and functions in parts of the genome thought to be evolutionary wastelands, it will surely prove interesting.

Paper I: NPC1 and triglyceride metabolism: Mouse studies and human association

Niemann-Pick disease type C (NPC) is in more than 95% of cases caused by mutations in the gene NPC1 on chromosome 18. The disease is a rare autosomal recessive neurodegenerative disorder where a majority of cases suffer from liver failure or milder liver dysfunction depending on the severity of the disease. The liver is the main organ responsible for maintaining lipid balance in the body. NPC1 encodes for an intergral membrane protein and is believed to be involved in export of cholesterol and possibly other lipids from late endocytic organelles. Absence of a functional NPC1 disables the mechanisms that stabilizes cholesterol concentrations in the endoplasmatic reticulum and leads to increased cholesterol biosynthesis and uptake.

The aim of paper I was to study the effects of NPC1 on triglyceride (TG) metabolism, a less studied aspect than the genes influence on cholesterol homeostasis. This was done through the use of mouse models as well as human association studies in multiple cohorts to investigate the effects of NPC1 on blood lipids at the population level. The human part of the study followed the theory that genes involved in severe monogenic disease also can be implicated in milder phenotypic effects on the population level.

The mouse study

The *Npc1*$^{-/-}$ mouse develops a liver disorder that is similar to the severe infant form of human NPC, and it provides a model for the study of hepatic dysfunction related to NPC. In *Npc1*$^{-/-}$ mice a reduced TG content was previously found in liver (Beltroy et al. 2005) and isolated hepatocytes (Kulinski et al. 2007), but the reasons for this are unknown. To investigate the mechanisms of TG imbalance in *Npc1*$^{-/-}$ mouse liver the fatty acid use, TG synthesis and TG secretion were studied in *Npc1*$^{-/-}$ hepatocytes.

The results of the study showed that TG content was significantly reduced in *Npc1*$^{-/-}$ mouse serum, liver and isolated hepatocytes, relative to *Npc1*$^{+/+}$ littermates, in accordance with previous reports. The reduced TG levels for hepatocytes could have three explanations: reduced TG synthesis, increased hydrolysis and/or increased export as very low density lipoprotein (VLDL). Experiments using radio labeled fatty acids and acetate indicated that reduced TG synthesis played an important role, but that hydrolysis could not be ruled out as an additional possibility even if reduced *Npc1*$^{-/-}$ TG excretion spoke against it. Further work with [^{3}H]carbons gave rise to the idea that the enhanced hepatic cholesterol production in *Npc1*$^{-/-}$ mice could be the cause of the TG levels. To test this hypothesis cholesterol synthesis was inhibited with statin. This normalized both the intracellular and secreted TG levels, and suggested that this indeed largely was the cause for the decreased TG levels in *Npc1*$^{-/-}$ hepatocytes. The mechanism is not known, but a plausible explanation is that the upregulation of cholesterol biosynthesis cause less carbon to be available for TG synthesis, hence the reduced TG levels.

NPC1 and blood lipids in humans

To explore if common variation in NPC1 had an influence over serum lipid levels in humans, a panel of 19 SNPs were selected across the gene region of NPC1 on chromosome 18. The average spacing between markers was 3 kb. The sample used for the initial study was the Uppsala Longitudinal Study of Adult Men (ULSAM), a population based cohort of 1053 men (Zethelius et al. 2005). One of the SNPs was P237S, a SNP that had been genotyped in the cohort at an earlier stage, and was included in the study. It had previously been believed to be a recessive NPC-causing mutation, but it had proved to have a frequency of 2% in the general population (Blom et al. 2003). ANOVA was used to test for the influence of SNP genotypes on LDL-cholesterol, HDL-cholesterol, total cholesterol and TG levels with body mass index, lipid medication and diabetes mellitus as covariates. ApoB was tested in a subset of the ULSAM cohort (n=435) where the phenotype had been measured. An association was detected for serum TG levels in a group of five markers in high linkage disequilibrium to each other: rs1788825, rs1788821, rs1788785, rs1788826 and rs1429934. The strongest

p-value ($p < 0.001$) was for for rs1429934. A similar pattern of association was detected for for ApoB in the same SNPs, but p-values were weaker. Based on the high LD between the markers ($r^2 = 0.91$ to 0.98) the associations were most likely based on a same signal. A weak association was detected for P237S to LDL-cholesterol, but the p-value would not survive adjustment for the number of tests that were used.

We genotyped the SNP rs1429934 in two additional Swedish cohorts, in total 8041 individuals, to replicate the TG finding. 2882 individuals were from the FRISC II and SWISCH studies (FRISC II Investigators 1999). FRISC II (the Fragmin and Fast Revascularization during Instability in Coronary Artery Disease II), is a Scandinavian multicenter randomized trial on the treatment of unstable coronary artery disease, and SWISCH (Swedish Women and Men and Ischemic Heart Disese investigation), includes healthy individuals of similar age as FRISC II. 5159 individuals were available for analysis after genotyping in the Malmö Diet and Cancer study (Berglund et al. 1993; Pero et al. 1993) (MDC), a population based prospective cohort from Malmö, Sweden. In this case the samples used were the cardiovascular cohort (MDC-CC), a randomly selected sub sample characterized for cardiovascular and metabolic risk factors.

The association between the A allele in rs1429934 and higher TG levels were detected in both cohorts at p=0.01 in FRISC/SWISCH, and p=0.009 in the MDC-CC. No significant associations were detected for LDL or HDL cholesterol.

Conclusion

The study demonstrated that NPC1, known to be a key protein involved in aspects of cholesterol transport and homestasis, affects hepatic TG metabolism, and that it is relevant for controlling serum TG levels in mice and potentially in humans.

Paper II: An investigation on the utility of the HapMap dataset for six populations of European descent

The HapMap project (The HapMap Consortium 2003; The HapMap Consortium 2005; Frazer et al. 2007; Altshuler et al. 2010) has in retrospect proved to be a great resource for the genetics research community. It simplified much in the field of genotyping and association studies. However during the early years it was not without its critics, both regarding the usefulness of the dataset and general design of the project. Paper II investigates how repre-

sentative the European HapMap sample (CEU) is for six samples of European origin and its performance for the selection of haplotype tagging SNPs.

At the time of the analysis in paper II, the latest release of HapMap data was #22 of March 2007, and contained around 3.8 million validated SNPs. An important question for the application of HapMap data to different studies was how representative the samples used in the HapMap study really were for other sample sets collected around the world. There were some data on tagSNP transferability available (Mueller et al. 2005; Sawyer et al. 2005; Conrad et al. 2006; de Bakker et al. 2006; Ribas et al. 2006), and Gu et al. (Gu et al. 2007) had reported an investigation on ten genomic regions that indicated that definitions of haplotype blocks varied greatly between samples, while tagSNPs were more transferable. De Bakker et al. analysed data from white people from Hawaii and individuals from Finland, their conclusion was that loss of coverage for haplotype tagging was no larger in other samples than the sampling variation within HapMap samples themselves (de Bakker et al. 2006). Other studies like Mueller et al. (Mueller et al. 2005) indicated more mixed results for HapMap data use. Over all there was some discrepancy in results from different studies.

Samples and genotyping

We compared how representative HapMap CEU samples were for six samples from five European populations. They were mostly trios (parents and a child) from Australia (Zhu et al. 2007), Finland (general population and Kuusamo region) (Service et al. 2006), the Netherlands (Boomsma et al. 2006), Sweden (Leon et al. 2005) and the United Kingdom (Spector et al. 2006).

The genomic region used for our comparison was the 1.47 Mb region on chromosome 4 that contains the GRID2 gene. From available HapMap SNPs in the region, 197 SNPs were genotyped with an average spacing of 8 kb. SNPs were genotyped using the GenomeLab SNPStream system (Bell et al. 2002), with manufacturer supplied reagents and software (Beckman Coulter). 11 SNPs failed in genotyping and 13 were non-polymorphic, leaving 173 SNPs for analysis. Checks for Hardy-Weinberg equilibrium and Mendelian inheritance were used as quality controls.

Analysis

The Haploview software (Barrett et al. 2005) was used to calculate LD measures and select tagSNPs via the included Tagger (de Bakker et al. 2005) implementation (pair-wise tagging, $r^2 > 0.8$). TagSNPs were defined in the European HapMap samples, and their ability to capture the variation in the other populations was evaluated. Fisher's exact test was used to test for al-

lele frequency differences between samples and 7000 permutations were used to determine a multiple testing adjusted 5% p-value cut-off for the frequency comparisons. Principal component analysis was used to project all allele frequencies onto a plane to look for any obvious clustering of populations. The Gabriel et al. (Gabriel et al. 2002) blocking algorithm was used to annotate high LD segments in the region. Four blocks with differing r^2 patterns were selected for haplotype reconstruction using Phase version 2.2.1 (Stephens et al. 2001; Stephens et al. 2005), Phase was also used to test for differences in haplotype frequency distributions between samples.

Results

Allele frequencies

To determine if allele frequencies for the SNPs differed between population samples, allele frequencies of the parents were tested pair-wise between populations and the HapMap CEU data. Permutation testing indicated $P < 1.3 \times 10^{-4}$ as the global 5% error level. Only two SNPs in the Kuusamo – HapMap comparison reached this level. To look for overall differences in allele frequencies, without concern for the identity of specific SNPs, an all against all comparison was also performed. The number of differences at an unadjusted $p<0.05$ is shown in Table 2. From this data it is evident that the Kuusamo sample had an elevated number of frequency differences compared to the other samples, given that the number of expected random results below 5% would be between 8 and 9 for each combination of samples. The Kuusamo sample is the only sample that clearly deviates from these numbers. While the Swedish and regular Finnish samples, follow with small potential elevations.

Table 2. *The number of pair-wise allele frequency differences between populations at $p < 0.05$.*

	Aus	FiKu	Fi	Du	UK	Swe	HapMap
Aus		50	8	1	8	10	6
FiKu	50		39	76	46	32	55
Fi	8	39		12	11	11	10
Du	1	76	12		5	18	2
UK	8	46	11	5		16	5
Swe	10	32	11	18	16		5
HapMap	6	55	10	2	5	5	
Average	13.8	49.6	15.2	19.0	15.2	15.3	13.8

PCA was used to project the allele frequencies of the 173 SNPs into two dimensions to summarize differences and similarities (Figure 1). Again the Kuusamo sample was the sample that separated most from the rest along the first component capturing the largest amount of variation. The Swedish sam-

ple with the least amount of differences vs. Kuusamo, was located the closest along the first component. In the second component the Finnish population sample separated from the rest, indicating that in part a separate set of SNPs were responsible for this separation. Examination of the loadings plot, which indicate which variables (SNPs in this case), that are responsible for positions in the score plot, indicated that the SNPs with the strongest effects on separation usually were located close to recombination hotspots.

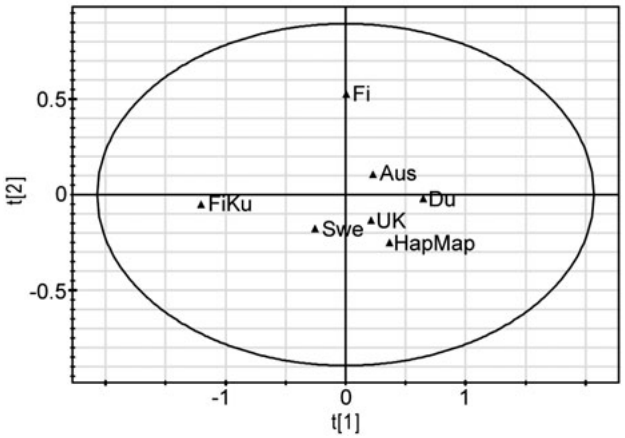

Figure 1. The allele frequency PCA score plot.

Linkage disequilibrium and tagSNP efficiency

Average levels of LD were similar in the different sample-sets and patterns of r^2 and D' were very similar between populations, with some minor differences in the general level of LD. To measure the efficiency of tagSNPs selected in the HapMap CEU sample and then applied to the other populations, a panel of tagSNPs were selected at $r^2 > 0.8$. This resulted in 63 tagSNPs for the region. Around 90% of SNPs in the other populations were captured by the HapMap panel, the highest was Kuusamo with 95% and the lowest the Australian sample with 87% of SNPs captured at $r^2 > 0.8$ (Table 3).

Table 3. *Performance of HapMap tagSNPs.*

	Hap-Map	Aus	FiKu	Fi	Du	UK	Swe
Average r^2 using Hap-Map tagSNPs	0.95	0.93	0.95	0.93	0.94	0.95	0.93
Minimum r^2 using Hap-Map tagSNPs	0.81	0.28	0.27	0.25	0.24	0.27	0.22
% of SNPs with $r^2 > 0.8$ to HapMap tagSNPs	100	87	95	91	90	92	90
Number of native SNPs needed to tag at $r^2 > 0.8$	63	73	58	67	70	61	68

The number of native tagSNPs needed to capture all SNPs in respective population had a similar pattern to the HapMap efficiency, with Kuusamo being the easiest to capture with 58 native SNPs necessary, and Australia the hardest where 73 native tagSNPs were required (Table 3). No increase in SNPs with very low LD to HapMap SNPs were observed in the populations with lower number of SNPs tagged, differences were mainly due to shifts slightly below the hard cut-off at $r^2 > 0.8$.

Haplotypes

Four high D' regions were selected for haplotype reconstruction with Phase to exemplify any differences in haplotype frequencies between the different samples. The case-control functionality of Phase was used to test the HapMap CEU haplotype distribution versus the distribution in the other samples. This test also considers the similarity of tested haplotypes. Generally frequency differences were smaller than 10%, but in some cases they differed as much as 20% (Figure 2).

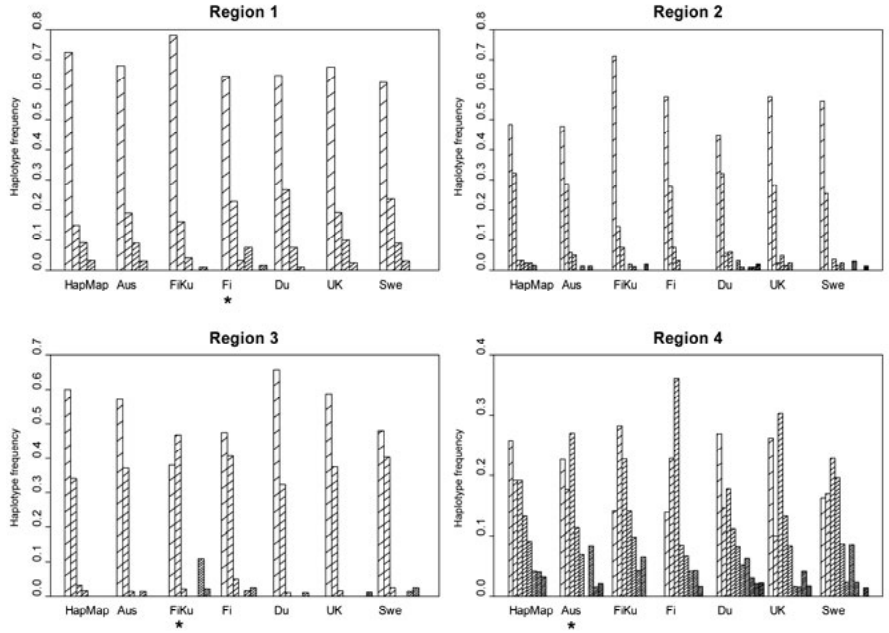

Figure 2. Haplotype frequencies > 1% in the example regions, significant different distributions vs. HapMap marked with an asterix.

Three comparisons were significant in the haplotype distribution tests, these were region 1 Finland ($p = 0.032$), region 3 Kuusamo ($p = 0.002$) and region 4 Australia ($p = 0.039$) (Figure 2). Only the Kuusamo haplotypes would be significant after Bonferroni correction for 6x4 tests ($p<0.0021$). This result

seem to have been driven mainly by a haplotype with frequency 0.11 in Kuusamo while absent in HapMap CEU while having multiple differences compared to haplotypes observed in HapMap.

Conclusion

The aim of our study was to investigate how well the HapMap CEU data represented five specific European population samples, from which SNP genotype data would be subjected to combined / pooled analysis. Results from the study were reassuring with high similarities in LD structure and tagSNP performance, with allele frequency differences only to the Kuusamo sample. Allele frequency findings contrasted with those by Willer et al. (Willer et al. 2006) who noted differences in Finnish allele frequencies vs. HapMap, but their sample size was larger and smaller differences would be detectable. Our conclusion was that allele frequency estimates based on HapMap data were as good as could be expected by the sample size for all included population samples except Kuusamo where additional caution was advised. TagSNP performance was good in all populations, even in Kuusamo where allele frequencies were divergent. Kuusamo performance could have been aided by a reduced number of haplotypes present due to the population history of being founded by 78 people 310 years ago. A limitation in the study was that we only sampled SNPs present in HapMap as rarer variation may have been less well represented. Still we can only conclude in agreement with several other studies (Willer et al. 2006; Service et al. 2007), that the HapMap CEU represent our samples of interest well for the purpose of tagSNP selection and SNP panel design.

Paper III: A comparison of allele specific expression and total expression eQTL mapping for the discovery of *cis*-regulatory SNPs.

The SNPs identified in the surge of genome-wide association studies of later years often reside outside of known protein-coding loci. A common step in trying to provide functional information to such variation is expression quantitative trait loci (eQTL) mapping in relevant tissues. This analysis looks for the association of SNP alleles to the total expression levels of genes. The discovery of a strong association between a SNP and gene would indicate that the tested SNP, or another in high linkage disequilibrium with it, is likely to influence the transcriptional regulation of the gene. Potential function would also mean potential disease implication.

The measurement of the allele-specific expression (ASE) of genes through quantitative genotyping of RNA and DNA, can be a powerful tool for the discovery of *cis*-regulatory variation (Pastinen et al. 2004). By analysing association between SNP genotypes and differential expression, the result is a kind of analysis that inherently filters away environmental and *trans*-effects. This property stems from the fact that the comparisons are made within-sample, between the chromosomes that shared the same environment in the cell and in the lab, as compared to between different samples in standard eQTL mapping. While the theoretical advantages of allele specific analysis for detection of *cis*-regulatory variation is clear, to what degree this translates into real world study power is not. Paper III investigates the relative power of the two techniques.

Samples

The samples used in the study were circulating monocytes from 395 healthy blood donors from Cambridge, UK.

Expression quantitative trait locus mapping

The number of individuals with genotypes and eQTL data was 395. Illumina expression arrays were used for expression profiling. They contained 24526 probes corresponding to 10059 genes. All SNPs within 100kb of the genes were tested for association to expression. As the subset of samples used in ASE analysis was genotyped on a different genotyping chip from the rest of the eQTL samples, only the overlapping 517K SNPs were used in the analysis. Association analysis used linear regression with age and sex as covariates.

Allele specific expression analysis

The samples used for ASE analysis was a subset of 188 individuals from those used for eQTL mapping. Genomic DNA and cDNA from each sample were genotyped on the same chip to reduce variation. ASE was calculated as the difference in allele fractions between cDNA and gDNA.

$(Allele1_{cDNA}/Allele1_{cDNA}+Allele2_{cDNA}) - (Allele1_{gDNA}/Allele1_{gDNA}+Allele2_{gDNA})$

Analysis windows for the ASE were defined as the start and end of transcripts in RefSeq release 47 with sufficient informative SNPs for analysis (five in window). The same set of 517K SNPs was tested for association in the ASE analysis as for the eQTL analysis, and all SNPs within 100kb of the analysis windows were tested for association. Linear regression was used to test for association between SNPs and ASE levels essentially as previously described (Ge et al. 2009).

P-value corrections

Permutation correction of p-values for multiple testing was done using 500 permutations, where the lowest p-value was saved each round. In addition Bonferroni correction was also used for comparisons.

Results and discussion

The total number of monocyte samples with genotypes and eQTL data was 395, of which 188 were analysed for ASE. The eQTL samples were also downsampled to 188, 95 and 50 samples. The ASE samples were downsampled to 95 and 50 samples. Table 4 shows the best P-value for the different sample sizes and the number of significant results for different p-value cutoffs. In all cases the ASE method had a higher number of significant SNPs for the same sample size.

Table 4. *Comparison of the results between the ASE method and the eQTL method. The best p-values and number of significant SNPs are shown for different sample sizes and cutoffs.*

	Method	**Sample size**			
		395	**188**	**95**	**50**
		p-value			
Best p-value	ASE	NA	6.70e-138	8.15e-70[1]	6.60e-37[1]
	eQTL	3.63e-142	1.34e-67[1]	4.26e-49[1]	9.41e-40[1]
Permutated p-value to get into top 5% rank	ASE	NA	1.4e-35	6.1e-19	4.3e-13
	eQTL	1.6e-14	2.3e-9	2.2e-9	2.1e-8
		Number of SNPs			
Corrected permutated p-value < 0.01	ASE	NA	876	157	70
	eQTL	275	69	16	10
Corrected permutated p-value < 0.05	ASE	NA	983	183	144
	eQTL	732	103	40	17
Bonferroni: p-value <= α_{corr}, α=0.01	ASE	NA	19147	1396	607
	eQTL	2704	198	75	27
Bonferroni: p-value <= α_{corr}, α=0.01	ASE	NA	21424	1648	770
	eQTL	3242	259	105	41

[1]*Median of 10 runs.*

To make a fair comparison we used a scheme that is non-parametric and chip independent. We created top lists of the highest ranking p-values from the eQTL and ASE analysis with sizes 100, 500, 1000, 5,000 and 10,000. The change in overlap between the eQTL and ASE top lists when sample size is increased corresponds to the relative power of the two methods. Implicit is the assumption that a hit is more likely to be true if found in both types of analysis. If for a fixed sample size in analysis 1 the sample size of analysis 2 is increased, an increase in top list overlap will reflect more true hits found that was already present in the top list of the first analysis. If then the types of analysis are switched and the procedure repeated but with no or little in-

crease in overlap, this would indicate that analysis 1 was more powerful since the common true hits were already present in this top list at smaller sample sizes. In the non-parametric comparison using top lists, shown in Table 5 and Figure 3, the overlap between ASE and eQTL increase as eQTL sample size becomes larger (Figure 3A), while the overlap remains almost constant when ASE sample size increases (Figure 3B). The average slope in Figure 3A is 1.4 and 0.12 in Figure 3B. The increased power of the larger eQTL sample sizes generate more true hits in the toplists, but many of these exist in the ASE top list already in small sample sizes. However, when we increase the power in the ASE analysis through an increased sample size, overlap does not increase indicating that hits are absent from the eQTL top list, eQTL mapping has not been able to pick up weaker hits. The numbers of significant hits based on the permutation p-values indicate that the ASE analysis need below half of the samples for the same statistical power as the eQTL mapping. Expressed as number of hits, ASE analysis finds more than five times more regulatory SNPs using the same sample size.

Table 5. *Overlap between p-value toplists from the ASE and eQTL analysis. Where there are down sampled data the overlap is calculated as an average over 10 runs. Both counts and percentages are shown.*

Size of top list		100	500	1000	5000	10000
Number of samples						
ASE	eQTL					
50	50	2 (1.9%)	20 (4.0%)	42 (4.2%)	267 (5.3%)	713 (7.1%)
50	95	2 (2.1%)	24 (4.9%)	58 (5.8%)	370 (7.4%)	914 (9.1%)
50	188	2 (2.0%)	27 (5.3%)	66 (6.6%)	472 (9.4%)	1134 (11.3%)
50	395	2 (2.0%)	28 (5.5%)	71 (7.1%)	540 (10.8%)	1323 (13.2%)
95	50	5 (4.5%)	19 (3.8%)	42 (4.2%)	275 (5.5%)	745 (7.4%)
95	95	3 (3.2%)	23 (4.6%)	59 (5.9%)	376 (7.5%)	941 (9.4%)
95	188	3 (3.2%)	25 (5.1%)	69 (6.9%)	482 (9.6%)	1177 (11.8%)
95	395	3 (3.2%)	26 (5.2%)	73 (7.3%)	545 (10.9%)	1380 (13.8%)
190	50	3 (3.2%)	19 (3.9%)	43 (4.3%)	264 (5.3%)	732 (7.3%)
190	95	4 (4.2%)	22 (4.3%)	58 (5.8%)	390 (7.8%)	971 (9.7%)
190	188	4 (4.1%)	25 (4.9%)	70 (7.0%)	508 (10.2%)	1215 (12.1%)
190	395	4 (4.0%)	26 (5.2%)	74 (7.4%)	577 (11.5%)	1429 (14.3%)

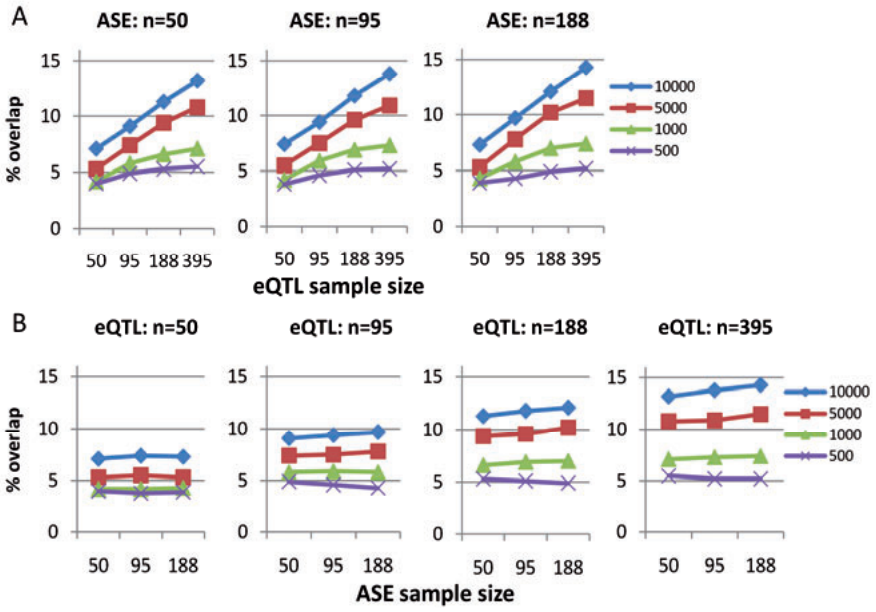

Figure 3. The overlap between top lists from the ASE and eQTL analysis. In A the ASE sample size is fixed at different levels, and the eQTL sample size increased. In B the eQTL sample size is fixed while the ASE sample size is increased. The legend contains top list sizes. The slope is clearly higher in A than in B showing that the ASE method is more powerful.

For Bonferroni correction the assumption is that all tests are independent. For SNP data this tends to be far from the truth with correlation introduced by linkage disequilibrium. Still, the correction of the permutation p-value is even more conservative both in the eQTL and ASE analysis. This could be explained by the fact that we only keep the lowest value from each permutation run, suitable if we expect few real signals in the data. This fits a study into disease associated SNPs for example where few strong hits are expected. In our case we consider all kinds of differential expression which we expect to be plentiful. So, a more suitable procedure could be obtained by keeping an equal number of p-values in each permutation as the expected number of differentially expressed genes. This number is hard to estimate. Therefore, a conservative choice is recommended. This does not matter for the method comparison as this would affect both methods in the same manner.

In addition to this non-parametric comparison we also use p-value cutoffs obtained from permutations and Bonferroni corrections to measure the relative power of the two methods. In summary, all strategies used to compare the two methods point in the same direction; that the ASE methodology is more powerful than the eQTL analysis using the same sample size.

Paper IV: A novel application of ASE analysis.

Long non-coding RNAs are a class of transcripts in the genome that we know relatively little about compared to their protein-coding relatives. A few have been deeply studied, with increasing knowledge of activity and mechanisms, like XIST mentioned under the section on ncRNA in the thesis introduction. Relatively little is known about the larger group of transcripts that belong to this class. In paper IV we apply the technique of ASE mapping to regions reported to harbour long non-coding RNAs (lncRNAs). We then compare the list of lncRNA *cis*-regulatory SNPs to published trait and disease associated SNPs, to look for potentially disease related alterations in ncRNA expression.

Allele specific expression

Allelic specific expression (ASE) was measured in monocytes from 188 healthy Caucasian individuals from the United Kingdom National Blood Service in Cambridge, UK. RNA and DNA were extracted and complementary DNA (cDNA) synthetised from the RNA. The Illumina Infinium II assay was used to genotype cDNA and genomic DNA (gDNA) on 1.2M Duo custom chips and gDNA and cDNA from each sample were genotyped on the same chip to reduce variation. ASE was calculated as the difference in allele fractions between cDNA and gDNA, where the genomic fraction for a heterozygote would be 0.5 in the ideal case:

$(Allele1_{cDNA}/Allele1_{cDNA}+Allele2_{cDNA})$ - $(Allele1_{gDNA}/Allele1_{gDNA}+Allele2_{gDNA})$

Impute 2 (Howie et al. 2009) was used to impute the missing genotypes and phase the chromosomes. ASE in each individual was averaged across the heterozygous SNPs in each analysis window to form the phenotype for the association test. The regions tested for ASE was retrieved from public sources. In all over 5000 potential ncRNA regions were used. SNPs located within 250kb of the regions were tested for association to ASE essentially as preciously described (Ge et al. 2009).

SNP expression and validation

SNP-expression, the summed signals of both alleles in the cDNA genotyping, were used to evaluate expression, as this measure was available for the full sample set. A summed signal of 1000 was used as a SNP-expression cutoff, based on the level of background signal detected in the channels (alleles) not present according to genotyping. The transcriptome of a monocyte sample was subjected to massively parallel sequencing for validation purposes. It was sequenced with one lane of reads on the Illumina Genome Analyzer II, and one lane on an Illumina HiSeq2000 machine. Reads were aligned with BWA (Li et al. 2009) for full reads and Bowtie/Tophat

(Langmead et al. 2009; Trapnell et al. 2009) to map spliced reads. A region was called as expressed if more than 5% had an average coverage larger than 2.1 to allow for low coverage of rare transcripts and suboptimal windows. To characterize transcribed regions for high-scoring ASE hits, chromatin immunoprecipitation followed by high-throughput sequencing was used (ChIP-Seq). Antibodies to detect regions affected by histone H3 lysine 4 trimethylation (H3K4me3) were used. This histone modification is a mark of an actively transcribed promoter (Barski et al. 2007; Mikkelsen et al. 2007). The same sample used for transcriptome sequencing was subjected to ChIP-Seq, with one lane of Genome Analyzer II sequencing.

Results and discussion

The goal of our study was to explore the effect of *cis*-variation on lincRNAs, and especially their overlap with published associations to trait and disease. The first step was to look for expression in the regions indicated as ncRNAs by our source materials. Transcriptome sequencing and SNP-expression were used to define the expressed set of regions. The general expression levels in ncRNA regions were determined to be slightly lower than introns, but higher than intergenic regions (Figure 4) and roughly a quarter of the regions were called as expressed.

Our attention then turned to the ASE analysis and the search for variation with an influence on expression of ncRNAs. SNPs with significant effect on transcription in the regions of interest were observed in close to three hundred regions with varying patterns of association to surrounding genes. We observed both isolated ASE in intergenic regions, and ASE that extended across neighbouring protein-coding genes. Finally comparisons were made to the SNPs with published trait associations. In all 29 SNPs with disease or trait overlap had an effect on the differential expression in ncRNA regions, in many cases transcription could be confirmed with RNA-Seq reads mapping to the region and H3K4Me3 peaks flanking the ASE region.

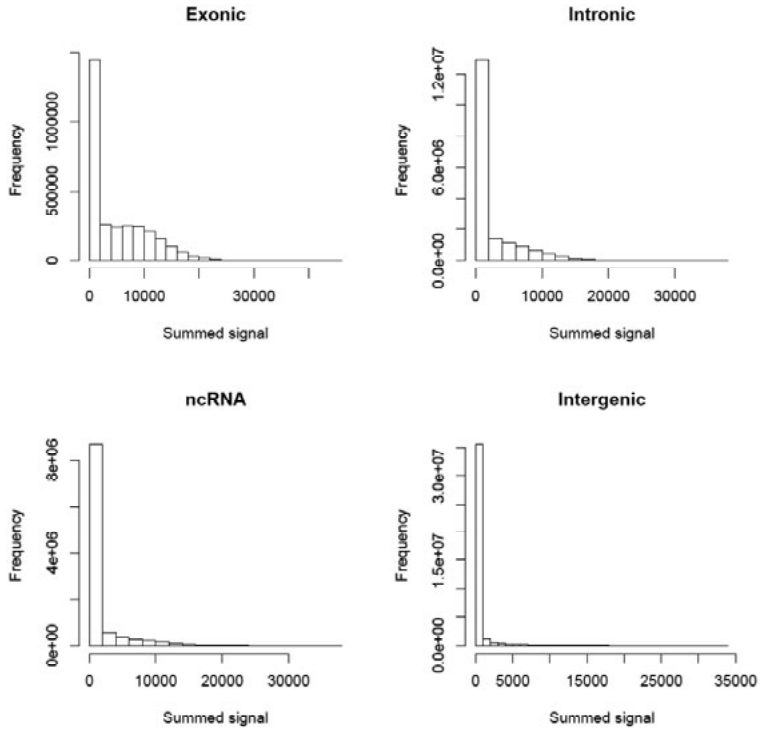

Figure 4. SNP expression signals (summed alleles) for different annotation classes.

Conclusion

Mapping of allele specific expression proved to be a robust technique for the discovery of *cis*-regulatory variation. We detected highly significant associations between SNPs and expressed regions outside of known protein-coding transcripts. Our study has generated interesting leads where further studies are needed to characterize the functional aspects of the findings.

Concluding remarks

I started my work in human genetics at a time when the human genome draft sequence had recently become available, and the first thing needed when looking at a region of interest was to try to evaluate how big the holes were in the sequence. We now move into times when the human genome sequence is taken for granted. In this thesis the progress of the field of human genetic studies in the last decade is mirrored in many ways. The first study uses association analysis in a single gene, based on prior knowledge of involvement in monogenic disease. We then move on to a larger chromosomal region using haplotype tagging SNPs, now entering the post HapMap phase of disease studies, when marker selection is much simpler, but still using a genotyping system with relatively low multiplexing of assays. Finally we arrive at techniques comparing the expression from the two chromosomes across the whole genome while trying to link differences to genetic variation, with genotyping technology that can query millions of SNPs in a sample, and parallel sequencing technology that allows assembly of transcriptomes from single runs. More than just the scale I think it is the integration of many types of information that mark a new epoch in the field of genetics. Many large data sets from projects like ENCODE and 1000 genomes are becoming freely available and are integrated with genome browsers and other tools that make them easily accessible to researchers across the world. Reference sets of epigenetic modifications, transcripts, structural variants etc. will be available for many tissues and thousands of genomes as new sequencing technologies generate genome-wide datasets at rates and costs never seen before. With massive amounts of data available, the problem may shift from producing the data, to extracting the information. As the knowledge of the mechanisms that support life increase, and ever more complexity and clever regulation is discovered, the only thing that is sure is that we will not run out of mysteries anytime soon.

Acknowledgements

The studies in this thesis were performed in the group for Molecular Medicine, Department of Medical Sciences at Uppsala University.

A big thank you to all the people who have contributed to this work over the (many) years I have been involved in projects in the research group. Since I worked in parallel to my PhD studies, it took me more years than most to get to the point of writing this thesis. A lot of great people passed through the group during the years, and all of you brought something good to the mix.

Chrisse, many thanks for the interesting projects and opportunity to combine the PhD with my job!

A sincere thank you to all collaborators on different projects, both local in Uppsala like the ULSAM group and many others, and in the international projects like GenomEUtwin and Cardiogenics. I met many interesting people and learned a lot at project meetings at home and abroad.

Trying to buy nose spray for a cold on my way home from a meeting in Paris comes to mind. School-French from 15 years ago got the job done, my head almost exploded on the flight home (no I'm not allergic, I had a cold), good phenotyping is important!

Many thanks to Tomi and the group in Montreal, for your hospitality and help during my visit, and for interesting discussions, collaborations and ideas for the ASE projects.

Thanks to the entire current MolMed group for all the good discussions and great atmosphere, you are all aces. Anders for hacking all that R, Jonas for working hard on the comparison, Johanna for all sorts of things including getting the book done, Jessica on expression, Chuan on ChIP-Seq and Camilla for the genotyping. Lillebil thanks for keeping it all working, and all the people at the platform, IT group and in the research group for making it such a nice and friendly place to work and study.

Last but not least, thanks to Axel for dragging me out of the study into the sandbox from time to time this summer, and to Ulrika for dragging Axel out of the study at other times, and for all your support and help giving me some time to get the thesis done.

References

Akey, J. M., S. Biswas, et al. (2007). "On the design and analysis of gene expression studies in human populations." Nat Genet **39**(7): 807-808; author reply 808-809.

Altshuler, D. M., R. A. Gibbs, et al. (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

Augui, S., E. P. Nora, et al. (2011). "Regulation of X-chromosome inactivation by the X-inactivation centre." Nat Rev Genet **12**(6): 429-442.

Barrett, J. C., B. Fry, et al. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." Cell **136**(2): 215-233.

Bell, P. A., S. Chaturvedi, et al. (2002). "SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery." Biotechniques **Suppl**: 70-72, 74, 76-77.

Beltroy, E. P., J. A. Richardson, et al. (2005). "Cholesterol accumulation and liver cell death in mice with Niemann-Pick type C disease." Hepatology **42**(4): 886-893.

Berglund, G., S. Elmstahl, et al. (1993). "The Malmo Diet and Cancer Study. Design and feasibility." J Intern Med **233**(1): 45-51.

Bertone, P., V. Stolc, et al. (2004). "Global identification of human transcribed sequences with genome tiling arrays." Science **306**(5705): 2242-2246.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blom, T. S., M. D. Linder, et al. (2003). "Defective endocytic trafficking of NPC1 and NPC2 underlying infantile Niemann-Pick type C disease." Hum Mol Genet **12**(3): 257-272.

Boomsma, D. I., E. J. de Geus, et al. (2006). "Netherlands Twin Register: from twins to twin families." Twin Res Hum Genet **9**(6): 849-857.

Botstein, D. and N. Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet **33 Suppl**: 228-237.

Branham, W. S., C. D. Melvin, et al. (2007). "Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements." BMC Biotechnol **7**: 8.

Carlson, C. S., M. A. Eberle, et al. (2003). "Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans." Nat Genet **33**(4): 518-521.

Carninci, P., A. Sandelin, et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Cheng, J., P. Kapranov, et al. (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution." Science **308**(5725): 1149-1154.

Cheung, V. G., R. S. Spielman, et al. (2005). "Mapping determinants of human gene expression by regional and genome-wide association." Nature **437**(7063): 1365-1369.

Conrad, D. F., M. Jakobsson, et al. (2006). "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome." Nat Genet **38**(11): 1251-1260.

Conrad, D. F., D. Pinto, et al. (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-712.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-232.

de Bakker, P. I., N. P. Burtt, et al. (2006). "Transferability of tag SNPs in genetic association studies in multiple populations." Nat Genet **38**(11): 1298-1303.

de Bakker, P. I., R. Yelensky, et al. (2005). "Efficiency and power in genetic association studies." Nat Genet **37**(11): 1217-1223.

Den Dunnen, J. T., P. M. Grootscholten, et al. (1989). "Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications." Am J Hum Genet **45**(6): 835-847.

Fan, J. B., X. Chen, et al. (2000). "Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays." Genome Res **10**(6): 853-860.

Fare, T. L., E. M. Coffey, et al. (2003). "Effects of atmospheric ozone on microarray data quality." Anal Chem **75**(17): 4672-4675.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

FRISC II Investigators (1999). "Invasive compared with non-invasive treatment in unstable coronary-artery disease: FRISC II prospective randomised multicentre study. FRagmin and Fast Revascularisation during InStability in Coronary artery disease Investigators." Lancet **354**(9180): 708-715.

Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-2229.
Ge, B., D. K. Pokholok, et al. (2009). "Global patterns of cis variation in human cells revealed by high-density allelic expression analysis." Nat Genet **41**(11): 1216-1222.
Gu, S., A. J. Pakstis, et al. (2007). "Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations." Eur J Hum Genet **15**(3): 302-312.
Howie, B. N., P. Donnelly, et al. (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.
Huarte, M., M. Guttman, et al. (2010). "A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response." Cell **142**(3): 409-419.
Jeffreys, A. J., V. Wilson, et al. (1985). "Hypervariable 'minisatellite' regions in human DNA." Nature **314**(6006): 67-73.
Johnson, D. S., A. Mortazavi, et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.
Johnson, G. C., L. Esposito, et al. (2001). "Haplotype tagging for the identification of common disease genes." Nat Genet **29**(2): 233-237.
Kawasaki, H., K. Taira, et al. (2005). "siRNA induced transcriptional gene silencing in mammalian cells." Cell Cycle **4**(3): 442-448.
Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.
Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids Res **39**(Database issue): D152-157.
Kulinski, A. and J. E. Vance (2007). "Lipid homeostasis and lipoprotein secretion in Niemann-Pick C1-deficient hepatocytes." J Biol Chem **282**(3): 1627-1637.
Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.
Leon, D. A., I. Koupil, et al. (2005). "Fetal, developmental, and parental influences on childhood systolic blood pressure in 600 sib pairs: the Uppsala Family study." Circulation **112**(22): 3478-3485.
Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.
MacDonald, M. (1993). "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group." Cell **72**(6): 971-983.
Maenner, S., M. Blaud, et al. (2010). "2-D structure of the A region of Xist RNA and its implication for PRC2 association." PLoS Biol **8**(1): e1000276.

Mendel, J. G. (1866). "Versuche über Pflanzenhybriden." Verhandlungen des naturforschenden Vereines in Brünn(Bd. IV für das Jahr, 1865): 3-47.
Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.
Milani, L., A. Lundmark, et al. (2009). "Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation." Genome Res **19**(1): 1-11.
Mills, R. E., K. Walter, et al. (2011). "Mapping copy number variation by population-scale genome sequencing." Nature **470**(7332): 59-65.
Mueller, J. C., E. Lohmussaar, et al. (2005). "Linkage disequilibrium patterns and tagSNP transferability among European populations." Am J Hum Genet **76**(3): 387-398.
Nagalakshmi, U., Z. Wang, et al. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-1349.
Nagano, T., J. A. Mitchell, et al. (2008). "The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin." Science **322**(5908): 1717-1720.
Pandey, R. R., T. Mondal, et al. (2008). "Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation." Mol Cell **32**(2): 232-246.
Pastinen, T. and T. J. Hudson (2004). "Cis-acting regulatory variation in the human genome." Science **306**(5696): 647-650.
Patil, N., A. J. Berno, et al. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." Science **294**(5547): 1719-1723.
Pero, R. W., A. Olsson, et al. (1993). "The Malmo biological bank." J Intern Med **233**(1): 63-67.
Reich, D. E., S. B. Gabriel, et al. (2003). "Quality and completeness of SNP databases." Nat Genet **33**(4): 457-458.
Ribas, G., A. Gonzalez-Neira, et al. (2006). "Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes." Hum Genet **118**(6): 669-679.
Rinn, J. L., M. Kertesz, et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell **129**(7): 1311-1323.
Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-1517.
Sawyer, S. L., N. Mukherjee, et al. (2005). "Linkage disequilibrium patterns vary substantially among populations." Eur J Hum Genet **13**(5): 677-686.
Schunkert, H., I. R. Konig, et al. (2011). "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nat Genet **43**(4): 333-338.

Service, S., J. DeYoung, et al. (2006). "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies." Nat Genet **38**(5): 556-560.
Service, S., C. Sabatti, et al. (2007). "Tag SNPs chosen from HapMap perform well in several population isolates." Genet Epidemiol **31**(3): 189-194.
Siomi, M. C., K. Sato, et al. (2011). "PIWI-interacting small RNAs: the vanguard of genome defence." Nat Rev Mol Cell Biol **12**(4): 246-258.
Spector, T. D. and F. M. Williams (2006). "The UK Adult Twin Registry (TwinsUK)." Twin Res Hum Genet **9**(6): 899-906.
Speliotes, E. K., C. J. Willer, et al. (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index." Nat Genet **42**(11): 937-948.
Steemers, F. J., W. Chang, et al. (2006). "Whole-genome genotyping with the single-base extension assay." Nat Methods **3**(1): 31-33.
Stephens, M. and P. Scheet (2005). "Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation." Am J Hum Genet **76**(3): 449-462.
Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-989.
Stranger, B. E., M. S. Forrest, et al. (2005). "Genome-wide associations of gene expression variation in humans." PLoS Genet **1**(6): e78.
Syvanen, A. C., K. Aalto-Setala, et al. (1990). "A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E." Genomics **8**(4): 684-692.
The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.
The HapMap Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-796.
The HapMap Consortium (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.
Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.
Wallace, R. B., J. Shaffer, et al. (1979). "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch." Nucleic Acids Res **6**(11): 3543-3557.
Wheeler, D. A., M. Srinivasan, et al. (2008). "The complete genome of an individual by massively parallel DNA sequencing." Nature **452**(7189): 872-876.
Willer, C. J., L. J. Scott, et al. (2006). "Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database." Genet Epidemiol **30**(2): 180-190.

Voight, B. F., L. J. Scott, et al. (2010). "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis." Nat Genet **42**(7): 579-589.

Yan, H., W. Yuan, et al. (2002). "Allelic variation in human gene expression." Science **297**(5584): 1143.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zethelius, B., H. Lithell, et al. (2005). "Insulin sensitivity, proinsulin and insulin as predictors of coronary heart disease. A population-based 10-year, follow-up study in 70-year old men using the euglycaemic insulin clamp." Diabetologia **48**(5): 862-867.

Zhu, G., G. W. Montgomery, et al. (2007). "A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions." Eur J Hum Genet **15**(1): 94-102.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 702

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-158486

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2011