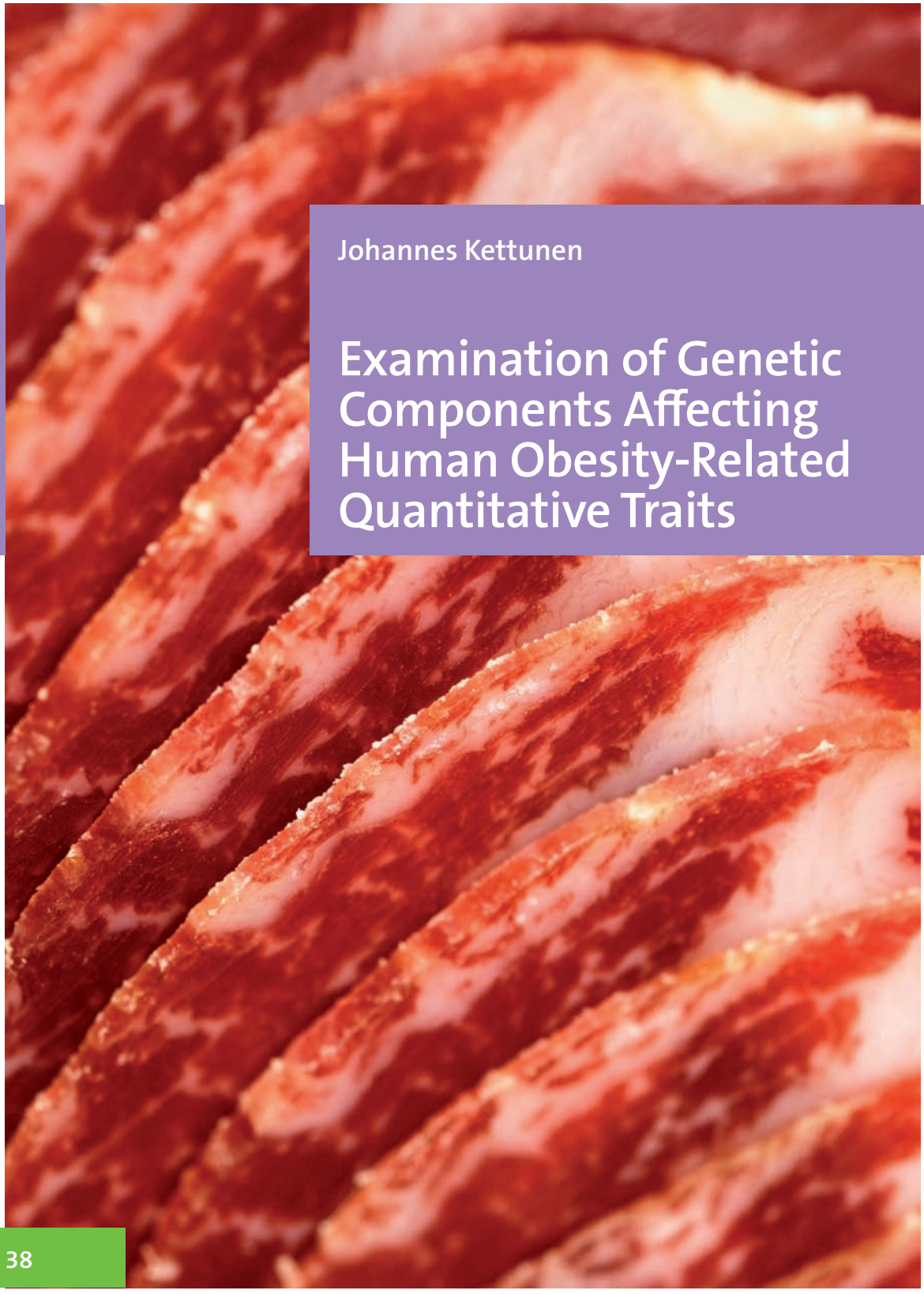
NATIONAL INSTITUTE
FOR HEALTH AND WELFARE



RESEARCH

Johannes Kettunen

Examination of Genetic Components Affecting Human Obesity-Related Quantitative Traits

38

Johannes Kettunen

Examination of Genetic Components Affecting Human Obesity-Related Quantitative Traits

ACADEMIC DISSERTATION

To be presented with the permission of the Medical Faculty,
University of Helsinki, for public examination in lecture hall 3 at Biomedicum 1,
on September 17th, at 12 noon.

National Institute for Health and Welfare, Finland,
Institute for Molecular Medicine, Finland
and
Wellcome Trust Sanger Institute, Cambridge, United Kingdom

RESEARCH 38
Helsinki 2010

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

ISBN 978-951-245-317-4 (printed)
ISSN 1798-0054 (printed)
ISBN 978-951-245-318-1 (pdf)
ISSN 1798-0062 (pdf)

Helsinki University Print
Helsinki, Finland 2010

Supervised by

Academician of Science Leena Peltonen-Palotie
Wellcome Trust Sanger Institute,
Cambridge,
United Kingdom
National Institute for Health and Welfare,
Unit of Public Health Genomics and
University of Helsinki,
Institute for Molecular Medicine Finland
Helsinki, Finland

Adjunct Professor Markus Perola
National Institute for Health and Welfare,
Unit of Public Health Genomics and
University of Helsinki,
Department of Medical Genetics
Helsinki, Finland

Reviewed by

Harald H. H. Göring, Ph.D.
Department of Genetics
Southwest Foundation for Biochemical Research
San Antonio, Texas, USA

Professor Johanna Schleutker
Institute of Medical Technology
University of Tampere
Tampere, Finland

Opponent

Professor Juni Palmgren
Mathematical Statistics
Medical Epidemiology and Biostatistics
Stockholm University
Karolinska Institutet
Stockholm, Sweden

“Missä tässä on nyt se pullonkaula?”

In Memoriam, Leena Peltonen-Palotie

To my family

ABSTRACT

Johannes Kettunen, Examination of Genetic Components Affecting Human Obesity-Related Quantitative Traits. National Institute for Health and Welfare (THL), Research 38, 152 pages. Helsinki 2010.
ISBN 978-951-245-317-4 (printed), ISBN 978-951-245-318-1 (pdf)

Obesity increases the risk for several conditions, including type 2 diabetes mellitus, cardiovascular disease, hypertension, osteoarthirits and certain types of cancer. Twin- and family studies have shown that there is a major genetic component in the determination of body mass. In recent years several technological and scientific advance have been made in obesity research. For instance, novel replicated loci have been revealed by a number of genome wide association studies.

This thesis aimed to investigate the association of genetic factors and obesity-related quantitative traits. The first study investigated the role of the lactase gene in anthropometric traits. We genetically defined lactose persistence by genotyping 31 720 individuals of European descent. We found that lactase persistence was significantly correlated with weight and body mass index but not with height.

In the second study we performed the largest whole genome linkage scan for body mass index to date. The sample consisted of 4401 twin families and 10 535 individuals from six European countries. We found supporting evidence for two loci (3q29 and 7q36). We observed that the heritability estimate increased substantially when additional family members were removed from the analyses, which suggests reduced environmental variance in the twin sample.

In the third study we assessed metabonomic, transcriptomic and genomic variation in a Finnish population cohort of 518 individuals. We formed gene expression networks to portray pathways and showed that a set of highly correlated genes of an inflammatory pathway associated with 80 serum metabolites (of 134 quantified measures). Strong association was found, for example, with several lipoprotein subclasses. We inferred causality by using genetic variation as anchors. The expression of the network genes was found to be dependent on the circulatory metabolite concentrations.

Keywords: Obesity, Body mass index, lipoprotein, metabolome, transcriptome, inflammation

TIIVISTELMÄ

Johannes Kettunen, Examination of Genetic Components Affecting Human Obesity-Related Quantitative Traits [Ihmisen lihavuuteen liittyvien kvantitatiivisten ominaisuuksien geneettinen tutkimus]. Terveyden ja hyvinvoinnin laitos (THL), Tutkimus 38, 152 sivua. Helsinki 2010.
ISBN 978-951-245-317-4 (painettu), ISBN 978-951-245-318-1 (pdf)

Lihavuus on huomattava, lisääntyvä ongelma maailmassa. Lihavuus lisää riskiä sairastua sydän- ja verisuonitautiin, tyypin 2 diabetekseen, nivelrikkoon ja tietyn tyyppisiin syöpiin. Perhe- ja kaksostutkimukset ovat osoittaneet että suuri osa ruumiinpainon vaihtelusta selittyy perinnöllisillä tekijöillä.

Tämän työn tarkoituksena oli tutkia lihavuuteen liittyvien jatkuvien muuttujien ja perinnöllisten komponenttien vuorovaikutusta. Ensimmäisessä osatyössä tarkasteltiin laktaasigeenin vaikutusta ruumiin rakenteeseen. Määritimme geneettisesti laktoosi-intoleranssin 31 720 Eurooppalaisessa henkilössä. Havaitsimme, että laktoosi-intolerantikoilla oli tilastollisesti merkittävästi pienempi ruumiinpaino, sekä painoindeksi kuin laktoosia sietävillä henkilöillä. Laktoosi-intoleranssin ei havaittu vaikuttavan loppupituuteen.

Toisessa osatyössä tutkimme painoindeksiä toistaiseksi suurimmalla kaksosperheistä koostuvalla kytkentätutkimuksella. Tutkimusaineistona oli 10 535 eurooppalaista henkilöä 4 401 perheestä, kuudesta eri maasta. Havaitsimme kromosomeissa 3q29 ja 7q36 aikaisempia tutkimuksia tukevia löydöksiä. Lisäksi havaitsimme että heritabiliteetti kasvoi, kun jätimme analyyseistä pois muut perheenjäsenet, joka viittaisi ympäristöstä johtuvan vaihtelun pienenemiseen kaksosaineistossa.

Kolmannessa osatyössä tutkimme aineenvaihdunta-, geeniekspressio- ja geenimerkkidataa suomalaisessa väestöotoksessa joka koostui 518 suomalaisesta henkilöstä. Muodostimme geeniverkkoja keskenään vahvasti korreloivista geeneistä ja havaitsimme että tulehdukseen liittyvä geeniverkko korreloi vahvasti 80 seerumin aineenvaihduntatekijän kanssa 134:stä mitatusta. Erittäin vahvoja korrelaatioita löytyi esimerkiksi lipoproteiinien alaluokista. Arvioimme myös syy-seuraussuhdetta käyttämällä geenimerkkejä suuntaavina pisteinä verkkoanalyysissä. Geeniverkon ilmentymisen eheyden todettiin olevan riippuvainen aineenvaihduntatekijöiden pitoisuudesta veressä.

Avainsanat: lihavuus, painoindeksi, lipoproteiniinit, tulehdus, aineenvaihdunta, geenin ilmentyminen

CONTENTS

ABBREVIATIONS

^{1}H NMR	Proton nuclear magnetic resonance
Apo	Apolipoprotein
ATBC	The Alpha- Tocopherol, Beta-Carotene Cancer Prevention study
BWHHS	The British Women's Heart and Health Study
BMI	Body mass index
CI	Confidence intervals
CNV	Copy number variation
DILGOM	The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study
DNA	Deoxyribonucleic acid
DZ	Dizygotic twin
eQTL	Expression quantitative trait locus
ERF	The Erasmus Rupchen Family Study
GWAS	Genome wide association study
HDL	High density lipoprotein
Health2000	The Health 2000 Health Examination Survey
HUGO	The Human Genome Organization
HWE	Hardy-Weinberg equilibrium
IBD	Identity-by-descent
IBS	Identity-by-state
KORA	Cooperative health research in the Region of Augsburg, Southern Germany
LD	linkage disequilibrium
LDL	Low density lipoprotein
LL	The Lipid-Leukocyte module

LOD	Logarithm-of-odds
MLOD	Multipoint logarithm of odds score
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
N	Number of individuals
NA	Not available
NEO	Network Edge Orientation
NFBC 1966	The North Finland Birth Cohort 1966
NR	Not reported
OR	Odds ratio
PCA	Principal components analysis
QTL	Quantitative trait locus
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SD	Standard deviation
SNP	Single nucleotide polymorphisms
TFBS	Transcription factor binding site
tRNA	Transfer ribonucleic acid
VC	Variance components
VLDL	Very low density lipoprotein
WHO	World health organization
YF	The Cardiovascular Risk of Young Finns Study

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles referred to in the text by their Roman numerals:

I **Kettunen J**, Silander K, Saarela O, Amin N, Müller M, Timpson N, Surakka I, Ripatti S, Laitinen J, Hartikainen AL, Pouta A, Lahermo P, Anttila V, Männistö S, Jula A, Virtamo J, Salomaa V, Lehtimäki T, Raitakari O, Gieger C, Wichmann EH, Van Duijn CM, Smith GD, McCarthy MI, Järvelin MR, Perola M, Peltonen L: European lactase persistence genotype shows evidence of association with increase in body mass index. Hum Mol Genet. 2010 Mar 15;19(6):1129-36

II **Kettunen J**, Perola M, Martin NG, Cornes BK, Wilson SG, Montgomery GW, Benyamin B, Harris JR, Boomsma D, Willemsen G, Hottenga JJ, Slagboom PE, Christensen K, Kyvik KO, Sørensen TI, Pedersen NL, Magnusson PK, Andrew T, Spector TD, Widen E, Silventoinen K, Kaprio J, Palotie A, Peltonen L; GenomEUtwin-project: Multicenter dizygotic twin cohort study confirms two linkage susceptibility loci for body mass index at 3q29 and 7q36 and identifies three further potential novel loci. Int J Obes (Lond). 2009 Nov;33(11):1235-42

III Inouye M[#], **Kettunen J**[#], Soininen P, Silander K, Ripatti S, Kumpula LS, Hämäläinen E, Jousilahti P, Kangas AJ, Männistö S, Savolainen MJ, Palotie A, Salomaa V, Perola M, Ala-Korpela M, and Peltonen L: Metabonomic, transcriptomic, and genomic variation of a population cohort. Manuscript submitted.

[#] These authors contributed equally

These articles are reproduced with the kind permission of their copyright holders.

1 INTRODUCTION

In the last two decades our knowledge of the human genome has increased dramatically. Looking back at the technological advances that have been made is astonishing. Large numbers of "Mendelian" disease genes have been found, but the hunt for complex trait genes is still ongoing. Obesity, defined as excessive fat accumulation, is one of those complex traits. Obesity increases risk for many diseases, including type 2 diabetes mellitus, cardiovascular disease, certain types of cancer and osteoarthritis. Body mass is defined by genes and environmental influences acting individually and in concert. A large number of genes involvel in obesity have been identified, but they still explain only a fraction of the variance attributed to genetic factors identified by family and twin studies. Advances have been made in finding some rare high-effect variants, such as the *MC4R*-mutations. On the other end of spectrum lie the common, low-effect variants such as those in the *FTO* gene region. Most of the variation between these two ends of the spectrum is still unfound and more work needs to be done in finding the rest of the genetic variance explaining the trait's heritability. Genetic studies have proceeded from single gene association to linkage and now whole genome association. Quite possibly the next step will be sequencing the whole genome in several individuals in order to identify trait-related alleles.

It seems that we have now picked the low hanging fruits from a large number of phenotypes. Now, genome-wide association studies for obesity require samples in the order of 200 000 to find new loci with minute effect sizes. It is time for us to take another point of view and explore different aspects. One possibility is to venture into the world of gene expression, which is studied using challenging data sets revealing several details influencing the expression level of a gene. These factors include the tissue of interest, environmental exposures, genetic variation and developmental time. In this study we approached this complexity in the simplest possible way by measuring gene expression levels from whole blood, which is the same functional matrix where the phenotypes of interest, metabolites, are found.

Another way to reduce the complexity is to refine our phenotypes of interest. How well does a simple measure of fat accumulation describe the predisposition to cardiovascular disease? It certainly plays its role but lipoproteins are closer to the site of action and are actually causing the foam cells to emerge in the first place. Lipoproteins have been studied widely before this study. We introduce an effective way to gain information on the lipoprotein fractions using proton nuclear magnetic resonance (^{1}H NMR). ^{1}H NMR gives more detailed information on both blood metabolites and different sub classes of lipoproteins.

2 REVIEW OF THE LITERATURE

2.1 Genomics

The term genome describes the entire hereditary material of an organism. It consists mostly of deoxyribonucleic acid (DNA), but in some viruses it consists of ribonucleic acid (RNA). Genomics is a term describing the study of genomes. It covers everything from sequencing the whole genome of an organism to fine scale genetic mapping. It also includes study of intragenomic interplay of genetic effects, such as pleiotropy (one gene can modify several phenotypic traits) and epistasis (a function of a gene is modified by one or several others).

2.1.1 Structure and variation of human genome

Most of the heritable genetic material in humans is packed in the nuclei of cells. It is a tightly packed structure where a double helical string of DNA is wound around histone beads. Coiling of histones and further scaffolding forms a superstructure called a chromosome (Figure 1). The human genome is diploid in most nuclei (haploid in germ cells) and consists of 22 autosomal chromosome pairs and a sex chromosome pair (Figure 2), which is either XX (female) or XY (male). In addition to the DNA in nuclei, there is a ~16 kilo-base circular DNA molecule in mitochondria, which is a cell organelle mostly responsible for cellular energy metabolism.

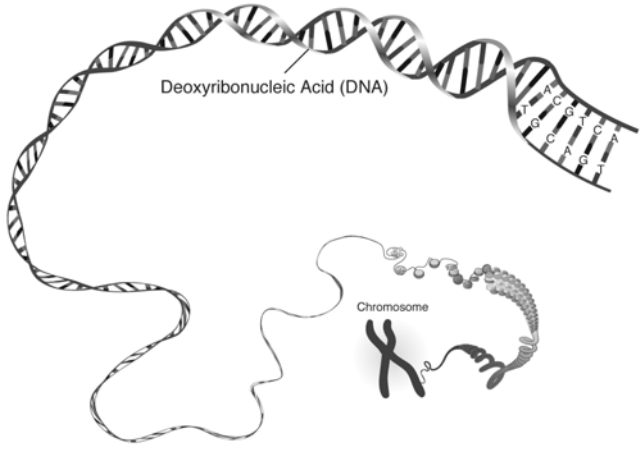


Figure 1. The DNA organization into a chromosome (www.genome.gov/Glossary)

If the number of base pairs of all chromosomes were added together it would total over 3 billion base pairs. Of this grand total only about 1.5% codes for proteins and is called the exome. The human genome contains over 20 000 protein coding genes. The definition of a gene has become rather vague in the recent years. A gene is now broadly defined as a "union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein *et al.*, 2007). The function, if any, of the remaining 97% of the human genome remains unknown. It consists largely (~50%) of repeat sequences, such as short and long interspersed nuclear elements, tandem repeats and non-coding RNA genes.

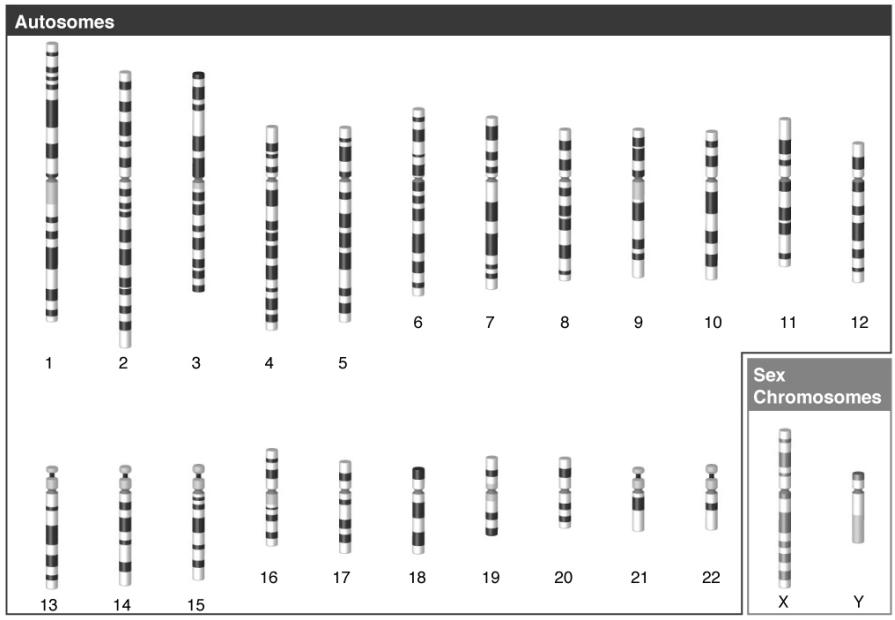


Figure 2. The human chromosomes (www.genome.gov/Glossary)

Genetic mapping utilizes genetic variation between individuals. The smallest type of variation in the nucleotide sequence is the single nucleotide polymorphism (SNP), in which a single nucleotide is substituted by another. The average genomic mutation rate for a base is $\sim 2.5 * 10^{-8}$ per meiosis (Nachman & Crowell, 2000).

Mutation throughout the genomic sequence, including single nucleotides, causes variation. However, some mutations produce changes that render the organism non-viable, thus these variations are not detected in population. Pieces of DNA can be transferred to another location in the genome, in a process called translocation. A

fragment of DNA can be either inserted (insertion) or removed (deletion) from genomic sequence (described in Figure 3 and Figure 4). Alleles are the different forms of a given DNA sequence.

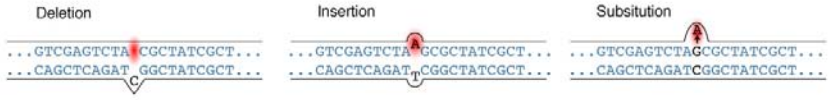


Figure 3 . Mutation events of SNPs (modified from:www.genome.gov/Glossary)

Microsatellites are tandem repeats of 1-6 nucleotide sequences. The number of repeats can vary between individuals and is quantifiable as illustrated in Figure 5. Microsatellites have higher mutation rate ($4.5*10^{-4}$) than SNPs (Whittaker *et al.*, 2003). The most commonly used genetic markers in mapping studies are microsatellites and SNPs.

Other types of mutation in the human genome are copy number variation (CNV) and chromosomal rearrangements (Figure 4). They include large insertions, deletions and translocations in the genome. Current technologies enable as small as 443 base pair CNVs to be detected, as described in the CNV mapping effort by Conrad and colleagues (Conrad *et al.*, 2010). The largest copy number variation can include the whole chromosome (trisomy) leading to severe developmental problems. The introduction of CNV assays on genome-wide SNP arrays has enabled the genotyping of CNVs in large population cohorts, which has dramatically increased the studies of CNV association with several traits of interest, including obesity (Bochukova *et al.*, 2010).

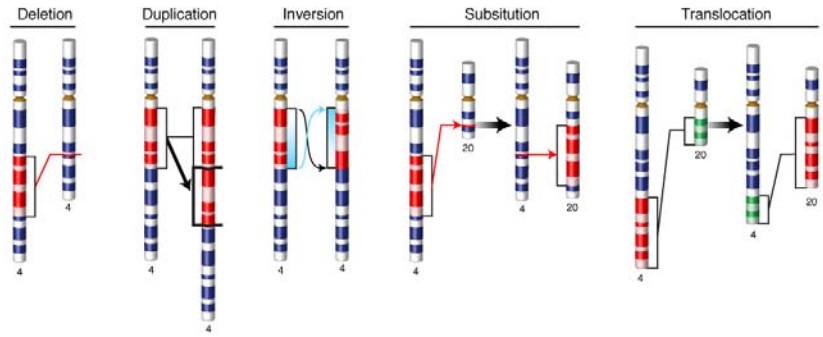


Figure 4. Description of different large scale mutations in the genome (modified from: www.genome.gov/Glossary)

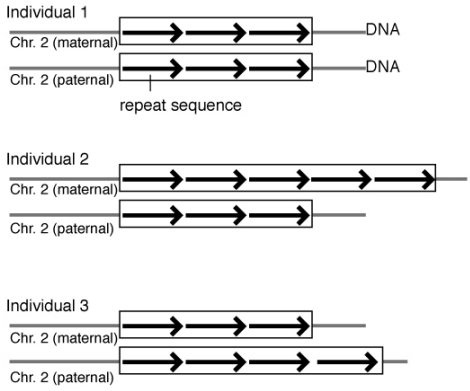


Figure 5. Copy number variation or microsatellites, each arrow represents a copy of a nucleotide sequence (modified from: www.genome.gov/Glossary)

Recombination is a major force, which reorganizes genetic material. It is a common phenomenon in which two chromosomes exchange segments of nucleotide sequence. It may occur between similar DNA sequences (homologous recombination) or different sequences (non-homologous end joining). When recombination within a chromosome happens in meiosis, it is the result of a chromosomal crossover which occurs between paired chromosomes (Figure 6). One recombination event appears roughly once in every one hundred million base pairs. This adds up to over 30 recombination events within chromosomes per meiosis on average. Meiosis is a source of genetic shuffling on genomic level. The haploid sex cells of an individual can contain 2^{23} combinations of the individual's chromosome pairs even without intra-chromosomal recombination. An egg after fertilization can have any of possible 4^{23} ($\sim 7*10^{13}$) combinations of the parental chromosomes. If one adds recombination to the shuffling of genomic information in meiosis and fertilization, the numbers in combinatorics become immense.

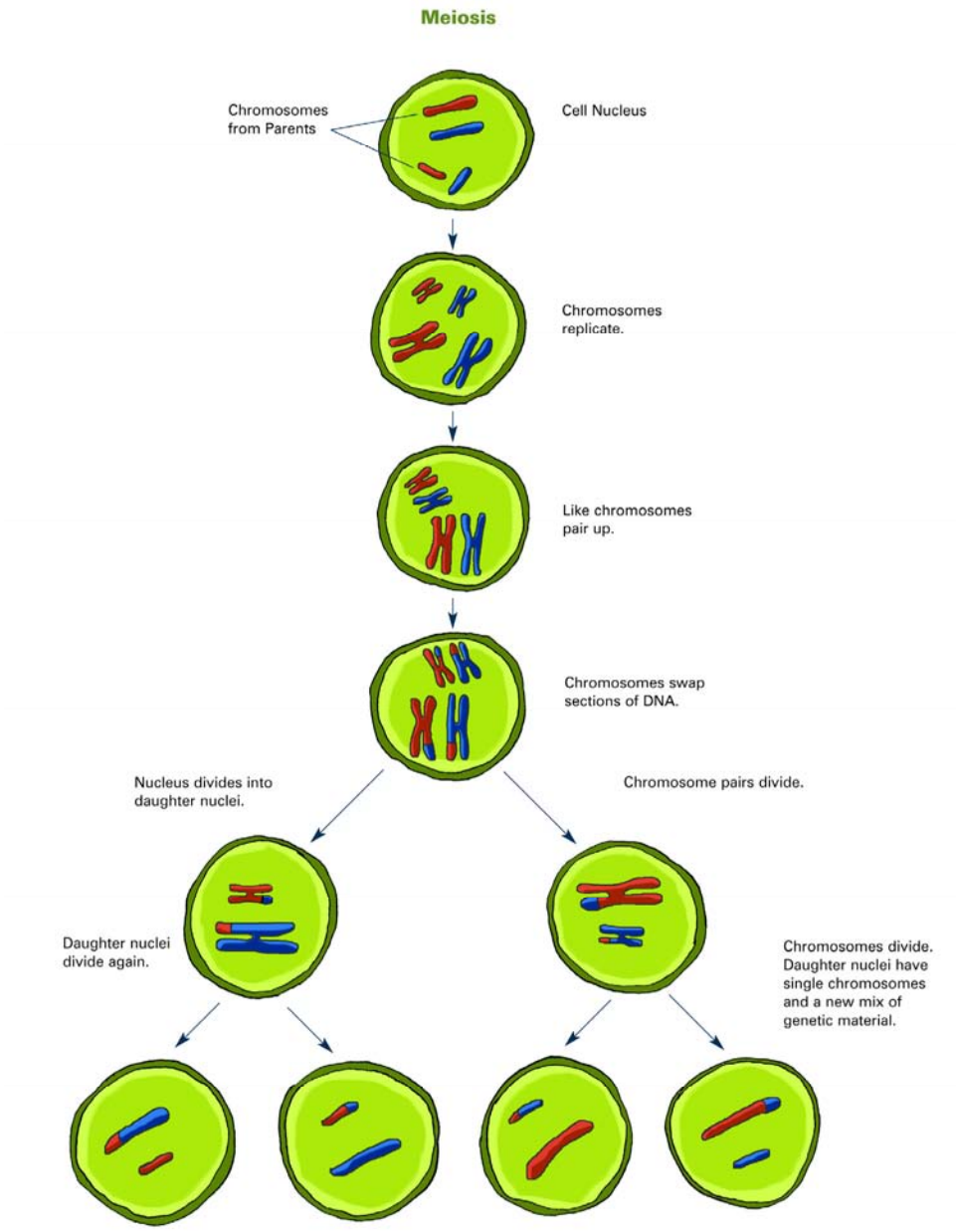


Figure 6. Meiosis and recombination (http://images.nigms.nih.gov/)

2.1.1.1 Sequencing the human genome

The human genome project was an international collaborative project where the genome of one individual was sequenced. It took ten years for the world's best genome centers to produce 90% of the total sequence of that individual (Collins *et al.*, 2003). After the publication of the human genome, the international HapMap project started (The International HapMap Consortium, 2003). Its goal was to catalogue patterns of common variation (common SNPs) in the human genome in different populations. These patterns are described as correlations between SNPs. The measure is called linkage disequilibrium (LD). It describes a phenomenon where the alleles of two loci occur together more often in the population than what would be expected by the random formation of haplotypes. This means that there is limited recombination between the two loci or that one or both of the alleles arose recently.

We are now in the wake of the 1000 genomes project (www.1000genomes.org) which aims to sequence a total of 2500 individuals from diverse population backgrounds, including 100 Finns, within the next few years. The project was made possible by the advances made in sequencing technology. This effort will revolutionize human genomics since even after the first pilot release of 60 individuals at rather low resolution, over 8 million SNPs have been revealed. Parallel to the 1000 genomes project, other groups are performing whole exome sequencing, in which all known RNA-coding regions are sequenced. Several conference presentations have reported novel coding variants on the order of several hundred per individual. As these projects are running, new ones are being planned. The UK10k project aims to sequence 10 000 individuals from United Kingdom in the near future.

2.1.2 Mapping quantitative trait loci

Phenotypes used in trait mapping can be classified broadly into two categories. They can be either dichotomous (yes/no) or continuous (quantitative). Dichotomous phenotypes are usually, for example, disease status such as type 2 diabetes mellitus or Crohn's disease. Quantitative phenotypes can be of any value, but the distribution of the phenotype sets limitations on the analysis options. Sometimes quantitative phenotypes are dichotomized by using a certain threshold to classify individuals as affected/not affected (Figure 7). The loss of information using this approach has been shown to reduce power in genetic analysis (Duggirala *et al.*, 1997). The power of a study is the probability that it rejects a false null hypothesis. There are numerous factors that affect the power of a study including: sample size, multiple

testing, effect size of the functional variant, linkage disequilibrium between the studied marker and the functional variant, penetrance, allele frequency of the functional variant and age of onset.

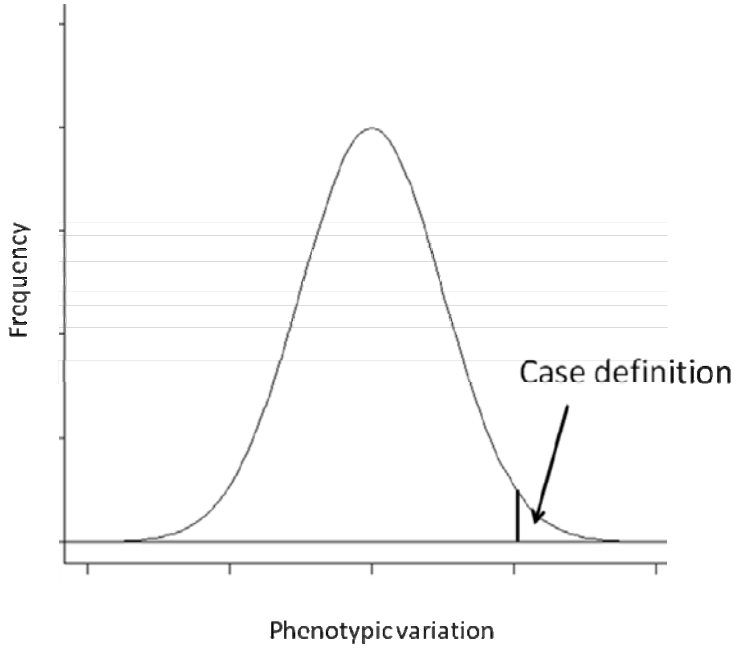


Figure 7. Dichotomizing a quantitative trait, such as BMI, to obesity classes loses a significant amount of information

The study of quantitative genetic variation in humans goes back to the work of Fisher and Galton at the turn of 20th century. Galton introduced the term correlation and made observations about the relationship of parent and offspring height (Galton, 1889, Galton, 1890, Galton, 1897). Fisher introduced the notion of quantitative genetic variation where multiple genes, in addition to environmental variance, could produce continuous variation (Fisher, 1918). This study by Fisher combined two rival fields, Mendelian and Galtonian, of genetics and introduced a new theory that could be applied to both.

2.1.2.1 Basic principles of quantitative trait mapping

An individual's phenotypic value is a combination of a genetic value (G) and an environmental value (E):

$P = G + E$

The genetic value represents the influence of all genetic variation and is therefore often divided into additive (A) and dominance (D) components. The additive effect describes the cumulative effect of individual genes. The dominance component refers to dominant gene action in a locus where the effect of one allele dominates the effect of the other. The environmental factor can be subdivided into pure environmental components and the interaction component (I) that describes the epistatic interactions between genes:

$P = A + D + I + E$

These can be expanded from an individual to a population where one can estimate the variance components of each:

$V_p = V_a + V_d + V_i + V_e$

2.1.2.2 Heritability

Broad sense heritability (H^2) is defined as the proportion of the total trait variance (V_p) that can be explained by genetic variation (V_g [V_a+V_d]):

$$H^2 = \frac{V_g}{V_p}$$

The narrow sense (h^2) of heritability is defined by the proportion of cumulative genetic effects (V_a) from the total phenotypic variance (V_p):

$$h^2 = \frac{V_a}{V_p}$$

Heritability is estimated from a sample where familial relationships are known or can be readily estimated. In family-based studies the familial relationships are known (though perhaps with errors). The genome-wide SNP data can be utilized to estimate familial relationship in population cohorts (Visscher *et al.*, 2006).

Heritability is an estimate derived from the study sample and gives a crude measure of the possible genetic contribution to the phenotype.

2.1.2.3 Linkage mapping

Linkage is a measure of whether two loci are inherited together. It can be utilized to position markers into chromosomal positions or to map positions in the genome that influence a certain trait. The linkage method in trait mapping aims to identify genetic loci where an allele segregates in a family along with a locus affecting a trait of interest. The method can be only performed in samples with known familial relationships. Linkage commonly utilizes microsatellite markers since they are very informative. They can be selected so that they have a large number of common alleles, and thus produce a maximum number of informative meioses (full heterozygous parents) in pedigrees. The relatively large mutation rate of microsatellites is usually not a problem because the segregation is only measured in a few generations per pedigree and the probability of a mutation is low. Linkage can be performed either by testing one marker at a time against the trait of interest (singlepoint), or by testing several markers at a time in small regions (multipoint).The multipoint linkage method increases power to detect linkage because it increases the information content by incorporating haplotype information across several markers. On the downside, the multipoint method is sensitive to miss-specified marker distances, which in turn reduces power (Halpern & Whittemore, 1999). In addition to locus mapping, linkage has been successfully utilized in Mendelian disorders leading to causal gene discovery, exemplified by Meckel syndrome (Kyttala *et al.*, 2006) and Aspartylglucosaminuria (Ikonen *et al.*, 1991).

Mendelian disorders commonly have few high penetrance mutations, clear clinical phenotypes and they show distinct hereditary patterns in pedigrees. The mode of inheritance is also usually easily identifiable if the disease requires two causal alleles for onset (recessive form) or if one causal allele is sufficient for the trait manifestation (dominant form). The recessive form is exemplified by RAPADILINO-syndrome (Kaariainen *et al.*, 1989), which is a rare recessive disease and requires two causal alleles for the disease to manifest. The dominant form is exemplified by lactase persistence, where only one mutated transcription factor binding site is sufficient to cause the trait (Enattah *et al.*, 2002). In contrast, multiple genes and environmental factors underlie complex traits, each having a relatively small effect on the trait. This is thought to be the reason why linkage studies have not been able to consistently detect genomic regions affecting complex traits. Linkage, as a method, is powerful in detecting large effect size variants, even though they may be relatively rare. This is exemplified by monogenic forms of severe

obesity, such as mutations in the *POMC* (Krude *et al.*, 1998, Comuzzie *et al.*, 1997) and *MC4R* (Duggirala *et al.*, 1996, Yeo *et al.*, 1998, Vaisse *et al.*, 1998) genes. Genes reported to cause monogenic forms of obesity are listed in Table 1. Although obesity is clearly a complex trait and linkage has not been able to identify any common variants affecting human obesity, these rare Mendelian forms of obesity are exceptions.

Table 1. Genes reported to cause monogenic forms of obesity in mouse or human studies

Gene	HUGO gene name	First Author(s)
Leptin	*LEP*	(Montague *et al.*, 1997, Strobel *et al.*, 1998, Ozata *et al.*, 1999)
Leptin receptor	*LEPR*	(Clement *et al.*, 1998)
Pro-opiomelanocortin	*POMC*	(Krude *et al.*, 1998, Challis *et al.*, 2002)
Melanacortin 4 receptor	*MC4R*	(Vaisse *et al.*, 1998, Yeo *et al.*, 1998, Farooqi *et al.*, 2003, Lubrano-Berthelier *et al.*, 2003)
Melanacortin 3 receptor	*MC3R*	(Lee *et al.*, 2002)
Agouti signaling protein	*ASIP*	*(Bultman *et al.*, 1992)
Agouti-related peptide	*AGRP*	*(Huszar *et al.*, 1997), *(Ollmann *et al.*, 1997)
prohormone convertase 1	*PC1*	(Jackson *et al.*, 1997, Jackson *et al.*, 2003), *(Naggert *et al.*, 1995)
G protein-coupled receptor 24	*GPR24*	(Gibson *et al.*, 2004)
Corticotropin-releasing hormone receptor-1	*CRHR1*	(Challis *et al.*, 2004)
Corticotropin-releasing hormone receptor-2	*CRHR2*	(Challis *et al.*, 2004)
bHLH-PAS transcription factor	*SIM1*	(Holder *et al.*, 2000, Faivre *et al.*, 2002)

HUGO = Human Genome Organization, * = evidence from studies in mice

When a linkage study is performed on a dichotomous trait, one can use parametric linkage, in which disease locus parameters, penetrance and disease allele frequency are predefined. Penetrance is defined as the proportion of individuals who actually manifest the trait divided by the proportion of individuals who should get the disease based on their genetic composition. In nonparametric linkage these parameters are not required, because the nonparametric method utilizes sib-pairs (Penrose, 1935) or affected sib-pairs (Bishop & Williamson, 1990) and simply tests for IBD sharing. In

quantitative trait mapping, one tests for the change of phenotypic covariance with respect to the change in the sharing of genetic material per locus. The modification of the nonparametric linkage method to study quantitative traits was originally done by Haseman and Elston (Haseman & Elston, 1972). The mode of inheritance is generally assumed to be additive, such that each causal allele contributes equally to the trait. Genetic sharing is usually defined by identity-by-descent (IBD). IBD is a measure, between two individuals, of sharing an allele which is segregated from a common ancestor. The Haseman-Elston method was originally intended for sib-pairs where the squared trait value difference was regressed against the IBD estimate of each locus. When larger pedigrees are at hand this method is no longer applicable. The variance components (VC) method takes the expected genetic covariance into account, enabling larger pedigrees to be handled. The resulting model is a function of additive genetic components based on the expected genetic covariance (relatedness) and estimated IBD sharing of the locus of interest. The VC method also takes into account the environmental variance, but it is assumed that there is no correlation between environmental deviations and genotypic values and that there is no interaction between environment and genes. This is estimated by maximum likelihood and is usually represented as and the logarithm-of-odds (LOD) score. Likelihood estimation is computationally very intensive.

$$LOD = \log_{10} \frac{\max_{V_q^2, V_g^2, V_e^2} L(V_q^2, V_g^2, V_e^2)}{\max_{V_g^2, V_e^2} L(V_q^2 = 0, V_g^2, V_e^2)}$$

2.1.2.4 Association analysis

To overcome the problem of large linked regions with usually close to a hundred genes for further study under the linkage peak, one can do an association analysis. Linkage tests for joint segregation of a genetic marker and a trait whereas association tests if the allele segregating in the population is the same. Association testing can be performed in families as well as in a population sample. The model in population samples excludes the expected genetic covariance and makes an assumption of unrelatedness. Association testing is generally performed with less computationally intensive methods in population samples. Commonly used methods are logistic regression for dichotomous traits and linear regression for quantitative

traits. There is an information portal, GENSTAT, which is designed to aid in the planning and analysis of a genetic association study (Ripatti *et al.*, 2009).

The term genome-wide association study (GWAS) refers to association testing where a sample is genotyped on a dense genome-wide SNP panel, typically between 100 000 and 1 000 000 markers, and each SNP is correlated against the phenotype of interest. The SNP panels are commercially available by two main manufacturers, Affymetrix and Illumina, and contain fixed marker sets. Custom design panels are also available. Illumina has selected the markers on their platforms using the LD structure of the 90 European ancestry HapMap individuals whereas Affymetrix has used an evenly spaced approach. The SNP panels currently available contain approximately one million SNPs across the whole genome. The first major GWAS for an obesity-related trait was published by Sladek for type 2 diabetes mellitus (Sladek *et al.*, 2007). After the study by Sladek *et al*, numerous GWASs of a large number of traits have been collected by the National Human Genome Research Institute website (http://genome.gov/26525384). The website is not focused on any trait and all reported p-values under $1*10^{-5}$ are represented as circles in Figure 8. Table 2 summarizes GWAs for total cholesterol levels, Table 3 for HDL cholesterol levels, Table 4 for LDL cholesterol levels and Table 5 for triglycerides and obesity-related blood measures other than metabolomic levels. Metabolomic GWAS are summarized in more detail in Table 11 and Table 12. Obesity-related anthropometric measures, other than BMI, are summarized in Table 6. BMI studies are summarized in more detail in Table 9. Additional studies have been reported for blood measures (Zemunik *et al.*, 2009, Kathiresan *et al.*, 2007) and anthropometric traits (Kang *et al.*, 2010, Kiel *et al.*, 2007, Liu *et al.*, 2010, Liu *et al.*, 2009b, Norris *et al.*, 2009, Polasek *et al.*, 2009) but these studies have not shown association at genome-wide significance level; probably due to their small sample sizes.

Table 2. Published GWASs for total cholesterol

Author	Trait	Chromosomal position reported
(Igl *et al.*, 2010)	Cholesterol, total	16q13
(Ma *et al.*, 2010)	Cholesterol, total	2q21.3, 12q21.2
(Aulchenko *et al.*, 2009)	Cholesterol, total	1p13.3, 1p31.3, 1p36.11, 2p21, 2p24.1, 5q13.3, 7p15.3, 8q24.13, 11q12.2, 11q23.3, 15q22.1, 18q21.1, 19p13.11, 19p13.2, 19q13.32

Table 3. Published GWASs for HDL cholesterol

Author	Trait	Chromosomal position reported
(Aulchenko *et al.*, 2009)	HDL cholesterol	2p24.1, 8p21.3, 9q31.1, 11p11.2, 11q12.2, 15q22.1, 16q13, 16q22.1, 18q21.1, 19q13.32
(Chambers *et al.*, 2008)	HDL cholesterol	8p21.3, 16q13
(Chasman *et al.*, 2008)	HDL cholesterol	15q21.3, 16q13
(Heid *et al.*, 2009)	HDL cholesterol	8p21.3, 16q13, 18q21.1
(Kathiresan *et al.*, 2009)	HDL cholesterol	1q42.13, 8p21.3, 9p22.3, 9q31.1, 11q12.2, 11q23.3, 12q24.11, 15q22.1, 16q13, 16q22.1, 18q21.1, 19p13.2, 20q13.12, 20q13.12
(Kathiresan *et al.*, 2008)	HDL cholesterol	1q42.13, 8p21.3, 9q31.1, 15q22.1, 16q13, 18q21.1
(Ridker *et al.*, 2009)	HDL cholesterol	16q13
(Sabatti *et al.*, 2009)	HDL cholesterol	11p11.2, 16q13, 16q22.1, 15q22.1, 17p13.3
(Willer *et al.*, 2008)	HDL cholesterol	1q42.13, 8p21.3, 9q31.1, 12q24.11, 15q22.1, 16q13, 16q22.1, 18q21.1

HDL = high density lipoprotein

Table 4. Published GWASs for LDL cholesterol

Author	Trait	Chromosomal position reported
(Aulchenko *et al.*, 2009)	LDL cholesterol	1p13.3, 1p31.3, 2p21, 2p24.1, 5q13.3, 7p15.3, 8q24.13, 11q12.2, 11q23.3, 19p13.11, 19p13.2, 19q13.32
(Burkhardt *et al.*, 2008)	LDL cholesterol	5q13.3, 19q13.32
(Chasman *et al.*, 2008)	LDL cholesterol	1q13.3, 2p24.1, 19p13.2, 19q13.31
(Hiura *et al.*, 2009)	LDL cholesterol	16q13
(Kathiresan *et al.*, 2009)	LDL cholesterol	1p13.3, 1p32.3, 2p21, 2p24.1, 5q13.3, 5q33.3, 12q24.31, 19p13.11, 19p13.2, 19q13.32, 20q12
(Kathiresan *et al.*, 2008)	LDL cholesterol	1p13.3, 1p32.3, 2p24.1, 5q13.3, 19p13.11, 19p13.2, 19q13.32
(Sabatti *et al.*, 2009)	LDL cholesterol	1p13.3, 1q32.2, 2p24.1, 11q12.2, 19p13.2, 19q13.32, Xq12
(Sandhu *et al.*, 2008)	LDL cholesterol	1p13.3, 2p24.1, 19q13.32
(Wallace *et al.*, 2008)	LDL cholesterol	1p13.3, 2p23.3, 11q23.3
(Willer *et al.*, 2008)	LDL cholesterol	1p13.3, 1p32.3, 2p24.1, 6p21.32, 19p13.11, 19p13.2, 19q13.32

LDL = low density lipoprotein

Table 5. GWAs for triglycerides and obesity-related blood measures other than metabolomic levels

Author	Trait	Chromosomal position reported
(Heid *et al.*, 2009)	Adiponectin levels	3q27.3
(Ling *et al.*, 2009)	Adiponectin levels	3q27.3, 5q35.2
(Richards *et al.*, 2009)	Adiponectin levels	3q27.3, 5q11.2
(SUN *et al.*)	Soluble *LEPR* levels	1p31.3
(Hicks *et al.*, 2009)	Sphingolipid concentrations	4p12, 11q12.3, 14q23.2, 19p13.2, 20p12.1
(Aulchenko *et al.*, 2009)	Triglycerides	1p31.3, 2p23.3, 2p24.1, 7q11.23, 8p21.3, 11q23.3, 19p13.11, 19q13.32
(Chambers *et al.*, 2008)	Triglycerides	2p23.3
(Chasman *et al.*, 2008)	Triglycerides	2p23.3, 8p21.3, 11q23.3
(Kamatani *et al.*, 2010)	Triglycerides	11q23.3
(Kathiresan *et al.*, 2009)	Triglycerides	1p31.3, 2p23.3, 2p24.1, 7q11.23, 8p21.3, 8p23.1, 8q24.13, 11q12.2, 11q23.3, 19p13.11, 20q13.12
(Kathiresan *et al.*, 2008)	Triglycerides	1p31.3, 1q42.13, 2p23.3, 2p24.1, 7q11.23, 8p21.3, 8q24.13, 11q23.3, 19p13.11
(Kooner *et al.*, 2008)	Triglycerides	7q11.23, 8p21.3, 11q23.3
(Lowe *et al.*, 2009)	Triglycerides	11q23.3
(Pollin *et al.*, 2008)	Triglycerides	11q23.3
(Sabatti *et al.*, 2009)	Triglycerides	2p23.3, 2p24.1, 8p21.3, 15q14
(Saxena *et al.*, 2007)	Triglycerides	2p24.1, 8p21.3, 16q13, 19q13.32
(Willer *et al.*, 2008)	Triglycerides	1p31.3, 1q42.13, 2p23.3, 7q11.23, 8p21.3, 8q24.13, 11q23.3, 15q22.1, 19p13.11

Table 6 . Published GWASs for obesity-related anthropometric measures, other than BMI

Author	Trait	Chromosomal position reported
(Freathy *et al.*, 2010)	Birth weight	3q21.1, 3q25.31
(Liu *et al.*, 2009a)	Body mass (lean)	8q23.1
(Scuteri *et al.*, 2007)	Hip	16q12.2
(Herbert *et al.*, 2006)	Obesity	2q14.2
(Liu *et al.*, 2008)	Obesity	20q11.32
(Meyre *et al.*, 2009)	Obesity	10p13, 16q12.2, 16q23.2, 18q11.2, 18q21.32
(Hinney *et al.*, 2007)	Obesity (early onset extreme)	16q12.2
(Cotsapas *et al.*, 2009)	Obesity (extreme)	2p16.1, 2q33.3, 3p24.2, 3p24.3, 4q26, 5q23.3, 6p21.31, 10p11.21, 10q22.1, 11p14.2, 16q12.2, 20p12.1
(Scherag *et al.*, 2010)	Obesity (extreme)	16q12.2, 18q21.32
(Chambers *et al.*, 2008)	Waist circumference	18q21.32
(Heard-Costa *et al.*, 2009)	Waist circumference	5p14.3, 6p12.2, 11p15.4, 12q13.13, 14q31.1, 16q12.2, 18q21.32,
(Lindgren *et al.*, 2009)	Waist circumference	1q42.3, 6p12.3, 8p23.1
(Cho *et al.*, 2009)	Waist-to-hip ratio	12q24.13
(Lindgren *et al.*, 2009)	Waist-to-hip ratio	1q41 (in women)
(Johansson *et al.*, 2009)	Weight	5q35.3
(Thorleifsson *et al.*, 2009)	Weight	1p21.3, 1p31.1, 1q25.2, 2p25.3, 3q27.2, 5q23.2, 6p21.33, 11p14.1, 12q13.13, 13q12.2, 16p11.2, 16q12.2, 18q21.32, 19q13.11

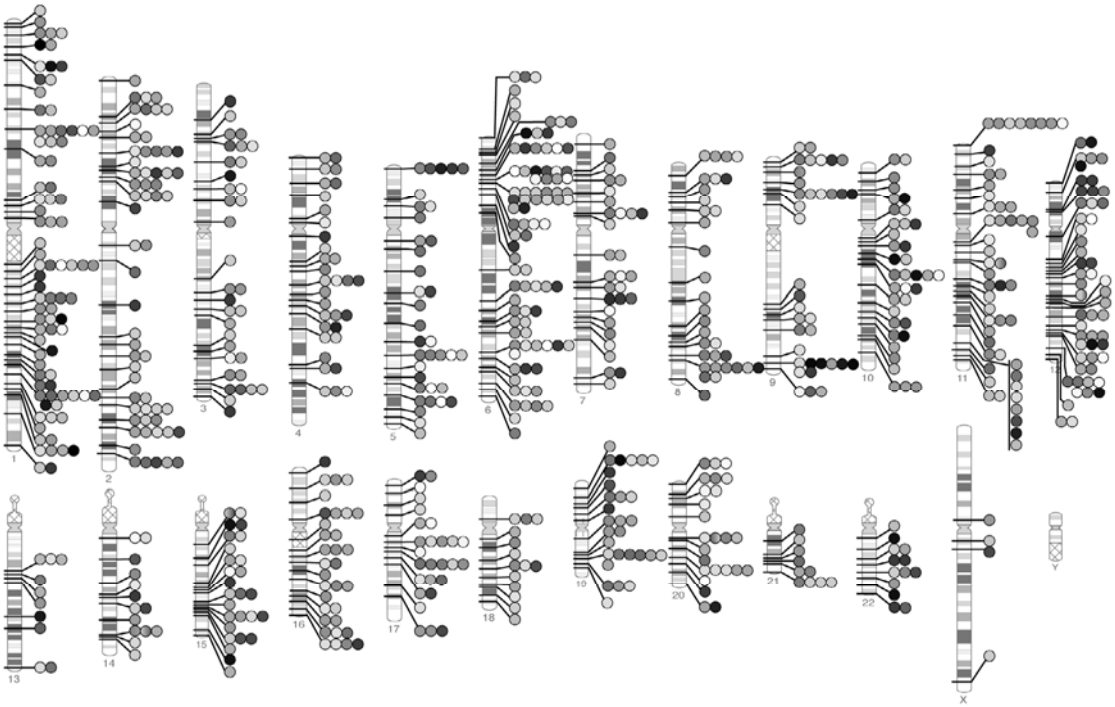


Figure 8. All reported p-values under $1*10^{-5}$ in GWAS up to 3/2010 depicted as circles (modified from http://www.genome.gov/images/illustrations/GWAS2010-3.ppt)

2.1.3 Population genetics

A population is defined as group of individuals in which there is random mating. This usually means that the individuals are of the same species and near each other geographically. There are four main forces that influence the genetic diversity of a population of interest: natural selection, genetic drift, mutation and gene flow.

Table 7. A theoretical example of changes in haplotypic frequencies under positive selection. This lead to increased linkage disequilibrium in the region under selection. Occurrence of the new beneficial mutation is marked in bold.

	Original ancestral haplotypes			Frequency of the beneficial haplotype increases	
a)	Haplotypes	Frequency	**b)**	Haplotypes	Frequency
	TGGCGGGTGG	0.2		TGGCGGGTGG	0.03
	AGTCGCGAGG	0.15		AGTCGCGAGG	0
	AGGCGGGTTG	0.21		AGGCGGGTTG	0.03
	ACTCGGGTGG	0.34		ACTCGGGTGG	0.09
	TGGCCCGTTG	0.1		TGGCCCGTTG	0
	TGG**T**CCGTTG	New mutation		TGG**T**CCGTTG	0.85

Natural selection is a phenomenon where certain phenotypes benefit their carriers enabling them to have more offspring. This leads to the increase of the phenotype in frequency and may even fix the beneficial causative allele. Selection leaves distinct and identifiable pattern in the genome around the variant under selection. The haplotype that carries an allele under positive selection increases rapidly in frequency as shown in Table 7. This leads to large LD around the variant and may lead to decreased haplotypic variation. Selection may be due to the occurrence of a new beneficial mutation or change of environment. Hollox *et al* gave four possible explanations for the positive selection of the *LCT* region: i) extra nutrition from the lactose component of fresh milk, ii) additional water source for desert nomads iii) improved calcium absorption from calcium rich milk and iv) protection from malaria due to riboflavin in milk (Hollox *et al.*, 2001).

Genetic drift is a random process where allele frequencies fluctuate by chance over time. The genetic drift is large in small populations and small in large populations. The Finnish population history shows a good example of genetic drift. Finland was settled slowly and the population underwent several bottle necks. Finland has the most extensive LD structures in Europe but also very large genetic differences between regions (Jakkula *et al.*, 2008, Service *et al.*, 2006). The regional differences

describe the Finnish population history where each village formation has been a bottleneck and source for rapid genetic drift.

Genetic variation is increased in the population by new mutations and gene flow. Every human germ line contains a host of new mutations. Different kinds of mutations are described in Figure 4. Gene flow occurs in migration into and out of the population. Immigration may introduce new genetic variation into the population. On the other hand emigration may reduce the variation.

2.2 Body mass index, lipids and obesity

2.2.1 Obesity

The definitions of overweight and obesity [OMIM 601665] are defined as abnormal or excessive fat accumulation that presents a risk to health. A crude population measure of obesity is the body mass index (BMI), a person's weight (in kilograms) divided by the square of his/her height (in meters). A person with a BMI of 30 or more is generally considered obese. A person with a BMI equal to or more than 25 is considered overweight. More detailed cut-off points for obesity are presented in Table 8. The World Health Organization (WHO) has gathered facts about obesity on their website (www.who.int/mediacentre/factsheets/fs311/en/index.html). According to WHO, approximately 1.6 billion adults (age 15+) were overweight and 400 million were obese in 2005. In the organization's 2015 projection 2.6 billion are overweight and 700 million are obese. The WHO predicts that diabetes-related deaths will increase by more than 50 % in the next 10 years.

Other anthropometric measures of obesity include waist circumference, hip circumference and waist-to-hip ratio. Waist-to-hip ratio is considered the best way to measure abdominal fat. More laborious, but also generally more accurate, measures are exemplified by bioelectric impedance analysis, underwater weighing, computed tomography and magnetic resonance imaging (Bell *et al.*, 2005). These techniques are not routinely available in sample collections whereas weight and height are common measures and therefore BMI is readily available for analysis.

2.2.1.1 Obesity as a risk factor

Overweight and obesity are major risk factors for a number of chronic diseases, including diabetes, cardiovascular diseases and certain types of cancer. Once considered a problem only in high income countries, overweight and obesity can be considered pandemic due to the rapid increase in developing countries, particularly in urban areas. The increase in obesity prevalence and obesity-associated diseases will produce a major increase in medical costs.

BMI values do not depend on age or gender for adults. But on the other hand, BMI acts differently as a predictor for risk of obesity-related disorders in different populations. For instance, the proportion of Asian people with a high risk of type 2 diabetes and cardiovascular disease is substantial at BMI's lower than the existing WHO cut-off points for overweight (WHO, 2004). This may result from the descriptive properties of BMI for percentage of body fat and body fat distribution among different populations. The question has been raised that should the obesity limits, such as those proposed in Table 8, be derived for different populations individually rather than from global cut-off points.

2.2.1.2 Obesity and environment

Environment plays a role in the development of obesity. Changes in the economic climate during last century have consequently changed the balance between physical exercise and energy intake. Food availability has increased globally. These factors have led to an “obesogenic” environment where we eat too much and exercise too little. The environmental influence is best detected in the increase of obesity prevalence in the world.

The change in obesity prevalence from mid 20th century is well described in the Swedish army conscripts. The 2.4 times increase in overweight prevalence was reported by Rasmussen between 1971 and 1995 in 18 year old males (Rasmussen *et al.*, 1999). The obesity prevalence increased from 6.9% to 16.3%.

Table 8. The International Classification of adult underweight, overweight and obesity according to BMI (kg/m^2) Source: Adapted from WHO, 1995, WHO, 2000 and WHO 2004.

Classification	BMI
	Principal cut-off points
Underweight	**<18.50**
Severe thinness	<16.00
Moderate thinness	16.00 - 16.99
Mild thinness	17.00 - 18.49
Normal range	**18.50 - 24.99**
Overweight	**≥25.00**
Pre-obese	25.00 - 29.99
Obese	**≥30.00**
Obese class I	30.00 - 34-99
Obese class II	35.00 - 39.99
Obese class III	≥40.00

2.2.1.3 Genes affecting body mass index

Obesity has been a well-studied trait due to its increasing risk for adverse health conditions. Every chromosome, except Y, has been linked to an obesity-related trait during the last few decades. The human obesity gene map catalogued all association and linkage findings up to year 2005 (Rankinen *et al.*, 2006). Numerous twin, family and adoption studies have been conducted for BMI and they report estimates of heritability that range roughly between 30 and 70 percent (Bouchard *et al.*, 1990, Allison *et al.*, 1996, Comuzzie *et al.*, 1996, Price & Gottesman, 1991, Stunkard *et al.*, 1986, Maes *et al.*, 1997, Stunkard *et al.*, 1990, Sorensen *et al.*, 1992a, Sorensen *et al.*, 1992b, Vogler *et al.*, 1995, Rice *et al.*, 1999). Therefore, a relatively large genetic contribution to the total phenotypic variance exists.

The obesity-related quantitative trait GWAS era started in the last decade with a paper published in Science by Frayling (Frayling *et al.*, 2007). They reported an association of the *FTO* gene with BMI. Interestingly, they did not study BMI initially but type 2 diabetes mellitus. They found a hit on chromosome 16 that did not replicate in other type 2 diabetes studies, which baffled them. After investigation they found that they had not controlled the association study with BMI as the other studies had, and thus their association to type 2 diabetes mellitus was mediated

through BMI (which increases the risk for type 2 diabetes). Several GWASs have reported BMI-associated loci during the last few years (Table 9). The effort is still alive to grow the sample sets even larger to identify more of the loci with smaller effect sizes.

Table 9. Published GWAS for BMI

Region	Reported Gene(s)	Strongest SNP-Risk Allele	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Author
1p21.3	*NR*	rs10783050-C	0.36	4×10^{-6}	2.6 [1.50-3.70] % SD	(Thorleifsson *et al.*, 2009)
1p31.1	*NEGR1*	rs2568958-A	0.58	1×10^{-11}	3.77 [2.67-4.87] % SD	(Thorleifsson *et al.*, 2009)
1p31.1	*NEGR1*	rs2815752-A	0.62	6×10^{-8}	.1 [0.04-0.16] kg/m2 increase	(Willer *et al.*, 2009)
1q25.2	*SEC16B, RASAL2*	rs10913469-C	0.2	6×10^{-8}	3.36 [2.14-4.58] % SD	(Thorleifsson *et al.*, 2009)
2p25.3	*TMEM18*	rs7561317-G	0.84	4×10^{-17}	6.12 [4.69-7.55] % SD	(Thorleifsson *et al.*, 2009)
2p25.3	*TMEM18*	rs6548238-C	0.84	1×10^{-18}	.26 [0.19-0.34] kg/m2 increase	(Willer *et al.*, 2009)
3q27.2	*SFRS10, ETV5, DGKG*	rs7647305-C	0.77	7×10^{-11}	4.42 [3.09-5.75] % SD	(Thorleifsson *et al.*, 2009)
4p13	*GNPDA2*	rs10938397-G	0.45	3×10^{-16}	.19 [0.13-0.25] kg/m2	(Willer *et al.*, 2009)
10p13	*PTER*	rs10508503-C	0.91	2×10^{-7} (children)	1.56 [1.10-2.78]	(Meyre *et al.*, 2009)
11p11.2	*MTCH2*	rs10838738-G	0.34	5×10^{-9}	.07 [0.01-0.13] kg/m2 increase	(Willer *et al.*, 2009)
11p14.1	*BDNF*	rs6265-G	0.85	5×10^{-10}	4.58 [3.07-6.09] % SD	(Thorleifsson *et al.*, 2009)
11p14.1	*BDNF*	rs925946-T	0.34	9×10^{-10}	3.85 [2.62-5.08] % SD	(Thorleifsson *et al.*, 2009)
11p14.1	*BDNF*	rs7481311-T	0.24	8×10^{-6}	3.15 [1.78-4.52] % SD	(Thorleifsson *et al.*, 2009)
11p15.4	*STK33*	rs10769908-C	0.53	1×10^{-6}	NR	(Willer *et al.*, 2009)
12q13.13	*BCDIN3D, FAIM2*	rs7138803-A	0.37	1×10^{-7}	3.28 [2.06-4.50] % SD	(Thorleifsson *et al.*, 2009)
15q25.2	*RKHD3*	rs12324805-C	0.31	7×10^{-6}	NR	(Willer *et al.*, 2009)
16p11.2	*SH2B1, ATP2A1*	rs7498665-G	0.44	3×10^{-10}	3.63 [2.49-4.77] % SD	(Thorleifsson *et al.*, 2009)
16p11.2	*SH2B1*	rs7498665-G	0.41	5×10^{-11}	.15 [0.08-0.21] kg/m2 increase	(Willer *et al.*, 2009)
16q12.2	*FTO*	rs8050136-A	0.41	1×10^{-47}	8.04 [6.96-9.12] % SD	(Thorleifsson *et al.*, 2009)
16q12.2	*FTO*	rs6499640-A	0.41	4×10^{-13}	5.25 [3.82-6.68] % SD	(Thorleifsson *et al.*, 2009)
16q12.2	*FTO*	rs9939609-A	0.41	4×10^{-51}	.33 [0.27-0.39] kg/m2 increase	(Willer *et al.*, 2009)
16q12.2	*FTO*	rs1121980-?	NR	4×10^{-8}	.06 [0.04-0.08] unit increase in log(BMI)	(Loos *et al.*, 2008)

16q12.2	*FTO*	rs9939609-A	0.39	2 x 10^{-20}	.36 [NR] kg/m2 per copy in adults	(Frayling *et al.*, 2007)
16q12.2	*FTO*	rs1421085-C	0.4	1 x 10^{-28} (children)	1.39 [1.27-1.51]	(Meyre *et al.*, 2009)
16q23.2	*MAF*	rs1424233-A	0.43	4 x 10^{-13} (children)	1.12 [1.00-1.24]	(Meyre *et al.*, 2009)
18q11.2	*NPC1*	rs1805081-A	0.56	3 x 10^{-7} (children)	1.33 [1.08-1.75]	(Meyre *et al.*, 2009)
18q21.32	*MC4R*	rs12970134-A	0.3	1 x 10^{-12}	4.38 [3.16-5.60] % SD	(Thorleifsson *et al.*, 2009)
18q21.32	*MC4R*	rs17782313-C	0.21	5 x 10^{-18}	.2 [0.12-0.28] kg/m2 increase	(Willer *et al.*, 2009)
18q21.32	*MC4R*	rs17782313-C	0.24	3 x 10^{-15}	.05 [0.04-0.06] unit increase in log(BMI)	(Loos *et al.*, 2008)
18q21.32	*MC4R*	rs17782313-C	0.18	5 x 10^{-15} (children)	1.22 [1.05-1.40]	(Meyre *et al.*, 2009)
19q13.11	*KCTD15, CHST8*	rs29941-C	0.69	7 x 10^{-12}	4.18 [2.98-5.38] % SD	(Thorleifsson *et al.*, 2009)
19q13.11	*KCTD15*	rs11084753-G	0.67	2 x 10^{-8}	.06 [-0.01-0.13] kg/m2 increase	(Willer *et al.*, 2009)
20p12.3	*BMP2*	rs2145270-T	0.65	6 x 10^{-6}	NR	(Willer *et al.*, 2009)

NR = not reported, SD=standard deviation, OR=odds ratio, CI=confidence intervals

2.2.2 Lipids and lipoprotein metabolism

The term lipid is an umbrella for a diverse range of molecules. They are relatively water-insoluble or non-polar compounds of biological origin. They include waxes, fatty acids, phospholipids, sphingolipids and glycolipids. Lipids come in diverse shapes. Some are linear aliphatic molecules and others have ring structures. Some are aromatic, while others are not. This produces a huge range of structural properties. Generally lipids are largely non-polar with some additional polar chemical groups. For example, in cholesterol the polar group is one hydroxyl group (-OH). The non-polar part does not like to interact with polar solvents such as water.

Lipids and their structural components are obtained from diet. In a simplified model ingested fat (including triglycerides, phospholipids and cholesterol) passes through the stomach and through to small intestine where it is emulsified by bile. Fatty acids form micelles and are taken into intestinal cells where they are used in chylomicron synthesis. Lipoprotein is a biochemical term describing an assembly that contains both lipids and proteins, which may be bound covalently or non-covalently. Immature chylomicrons enter the blood stream where high density lipoprotein (HDL) particles donate apolipoproteins (CII and E) to the immature chylomicrons transforming them to the mature form. In the blood vessel apolipoprotein CII

activates lipoprotein lipase in the surface of endothelial cells allowing triglycerides in the lipoprotein to be broken down. Free fatty acids are taken into the adjacent tissue cell as a result and are either utilized or stored. Chylomicrons donate the ApoCII to HDL in the blood stream and become chylomicron remnants. Chylomicron remnants are then taken into the liver for endocytosis. The resulting components (triacylglycerol and cholesterol) are used for very low density lipoprotein (VLDL) synthesis. In this process VLDL particles gain ApoB-100 and enter blood stream. HDL particles in blood donate ApoCII and ApoE to produce mature VLDL particles. Again apolipoprotein CII activates lipoprotein lipase allowing triglycerides in the lipoprotein to be broken down. Free fatty acids are taken into the adjacent tissue cell as a result and are either utilized or stored. VLDL donates ApoCII to HDL in the blood stream and in the process VLDL turns into intermediate density lipoprotein (IDL). As IDL loses triacylglycerols it becomes less dense and transforms to low density lipoprotein (LDL). LDL are then taken up and used as fuel for tissue. The diameter, density and fat content of different lipoprotein classes are summarized in Table 10 and the schematic representation of lipid metabolism is depicted in Figure 9.

Table 10. Lipoproteins are broadly classified by their density. Lipoproteins are larger and less dense if they consist of more fat than protein (modified from Biochemistry 2nd Ed. 1995 Garrett & Grisham)

Density (g/mL)	Class	Diameter (nm)	Protein (%)	Cholesterol (%)	Phospholipid (%)	Triacylglycerol (%)
>1.063	HDL	5-15	33	30	29	4
1.019-1.063	LDL	18-28	25	50	21	8
1.006-1.019	IDL	25-50	18	29	22	31
0.95-1.006	VLDL	30-80	10	22	18	50
<0.95	Chylomicrons	100-1000	<2	8	7	84

HDL = high density lipoprotein, LDL = low density lipoprotein, IDL = intermediate density lipoprotein, VLDL = very low density lipoprotein

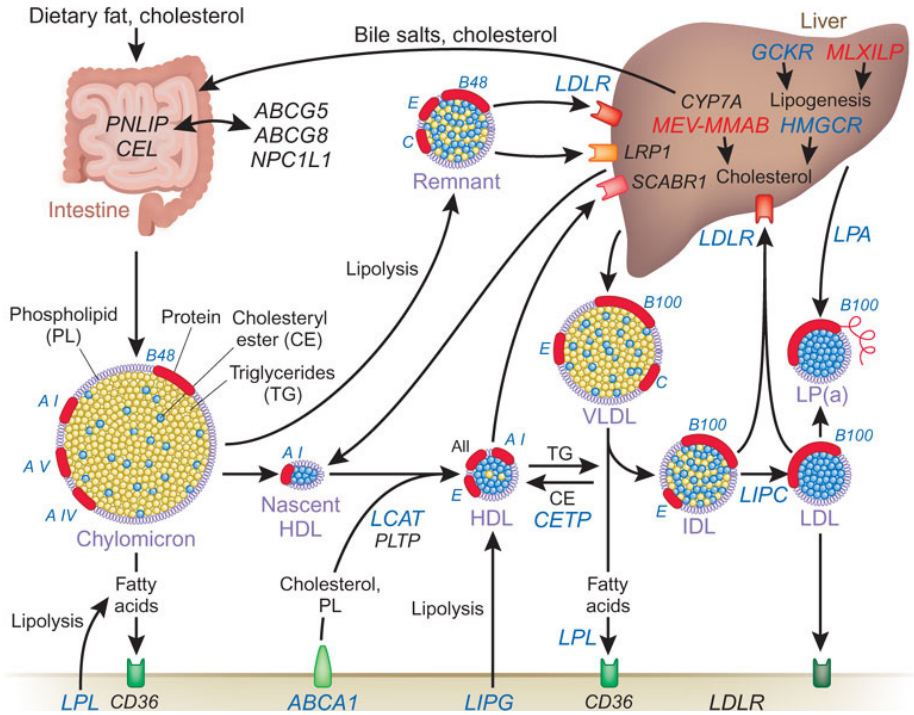


Figure 9. Schematic figure of lipid metabolism and the genes involved (modified from: (Lusis & Pajukanta, 2008))

2.2.3 Lipoproteins, obesity and insulin resistance

Dyslipidemia is a disruption of the normal blood lipid levels. Abnormalities in lipoprotein levels are strongly associated with central obesity (the accumulation of abdominal fat). Centrally obese individuals have an increased proportion of visceral fat packed inside the perinoteal cavity. Visceral fat promotes insulin resistance (Enzi *et al.*, 1986). Centrally obese individuals generally have increased triglyceride levels, low HDL levels, high ApoB levels and high (small and dense particles in particular) LDL levels. All of these metabolic traits are inter-correlated and are associated with insulin resistance. Insulin resistance plays a role in dyslipidemia through VLDL metabolism. Insulin resistance leads to increased hepatic glucose production, increased release of free fatty acids from adipose tissue and decreased muscle glucose uptake and oxidation (Howard *et al.*, 2003). This results in an increased intake of free fatty acids and glucose to the liver, which leads to an increased production of VLDL and triglycerides. VLDL is lipolysed by hepatic lipase, which results in small and dense LDL particles and a decrease in blood HDL cholesterol levels.

2.2.4 Lipoproteins and cardiovascular disease

Dyslipidemia has several functional mechanisms through which it affects the risk for a cardiovascular event. Elevated triglyceride levels promote atherogenic subspecies of LDL and HDL. It causes endothelial dysfunction and increases thrombotic risk. HDL particles are anti-atherogenetic and anti-inflammatory. They remove cholesterol from arterial wall in reverse cholesterol transport and take it back to the liver. They also inhibit LDL oxidation. Small and dense LDL particles have a better ability to enter the arterial wall. Small and dense LDL particles are more prone to oxidation and glycosylation. Inside the vascular wall these particles are engulfed by macrophages that turn into foam cells, the start of atherosclerosis. ApoB-containing lipoprotein particles are all atherogenic and it has been shown that ApoB levels may actually better predict a future cardiovascular event than LDL measures (Walldius *et al.*, 2001, Sniderman *et al.*, 2001). Inflammation is an important state in the atherosclerotic progress. To date, the most important inflammatory cells include macrophages and T cells. Early atherosclerotic lesions consist mainly of these cells. Mathieu and colleagues conclude in their obesity, inflammation and cardiovascular risk review paper that there is clear evidence that visceral fat is a source of inflammatory responses in obesity and cardiovascular disease (Mathieu *et al.*, 2010).

2.3 Transcriptomics

The transcriptome refers to the expression profile of active genes in a given cell, tissue or even whole organism. The most abundant of the known functional elements in the human genome are protein-coding genes. A protein coding gene consists of exons and introns. When a gene is producing protein, both exons and introns are copied into messenger ribonucleic acid (mRNA) in the transcription process. Exons contain the code of the final protein product, because introns are removed in the splicing process. After splicing and further modification, the code mRNA is translated to peptides by ribosomes. Transcription and translation are illustrated in . The transcriptome also includes transfer RNAs (tRNA), ribosomal RNAs (rRNA), and other non-coding RNAs with very diverse functions. The function of tRNA is to transfer amino acids to the polypeptide chain in protein synthesis. rRNA is the ribosomal RNA component that interacts with both tRNA and mRNA in translation.

The trascriptome, or expression profile, is studied using high throughput techniques. These techniques measure the levels of mRNA in the study sample using commercially available expression arrays, which contain probes for most of the expressed genes. The probe sets differ between manufacturers and vary depending on which gene annotation was used in the probe set creation. New sequencing technologies also allow expression detection and are not limited to given probe sets, but the ensuing data handling is very laborious.

The mRNA contains a sequence of codons, comprised of combinations of three nucleic acids, that code for one amino acid. The newly translated protein goes through further modifications and folds into its final functional structure.

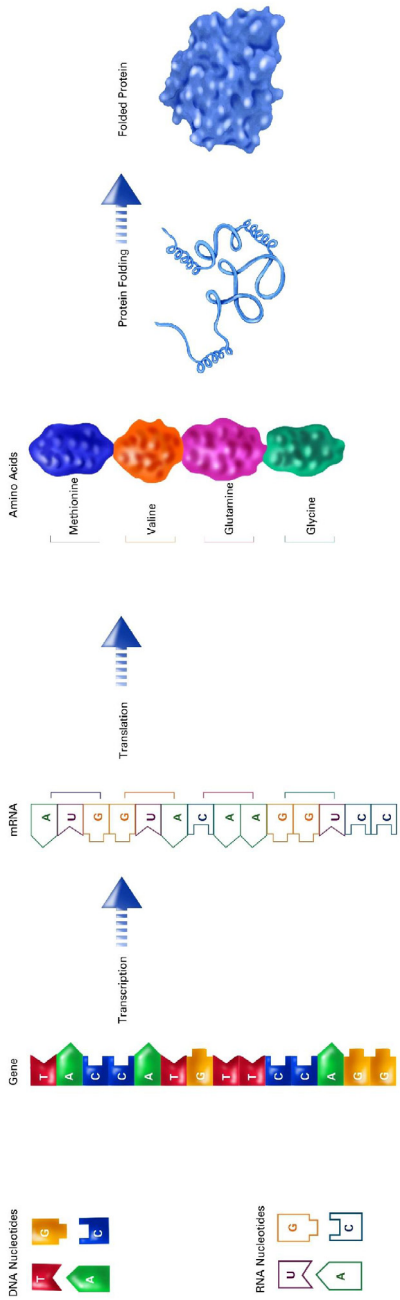


Figure 10. Transcription and translation (http://images.nigms.nih.gov/)

2.3.1 Transcription regulation

Gene expression is modified by transcription factors. These are proteins that bind, either alone or in a complex, to DNA at specific sequence motifs (transcription factor binding site, TFBS). The motif binding may promote expression (enhancer motif) by recruiting RNA polymerase or it may block RNA polymerase binding (silencer motif). The regulation of gene expression is a very complex process and involves a host of different proteins (Rockman & Kruglyak, 2006).

Mutations in the TFBSs can affect the factor binding. Mutation can even prevent the binding altogether. This is the case in lactase persistence (OMIM 223100). A mutation (SNP rs4988235) in the lactase-gene silencer binding site prevents the down regulation of the lactase enzyme in childhood (Enattah *et al.*, 2002). The mutation was found by Enattah *et al* in Finnish families and its effect has been shown *in vivo* (Rasinpera *et al.*, 2005). It enables the expression of the enzyme in intestinal epithelial cells and allows milk sugar lactose to be digested in adulthood. If the mutation occurs in the coding region of the gene, it can lead to a synonymous, non-synonymous or frameshift mutation. A synonymous mutation changes the nucleotide, but since the amino acids are coded with several combinations of codons, it may happen that the codon still codes for the same amino acid. For example, a point mutation in a codon coding for aspargine is TTA may change it to TTG which also codes for aspargine. Non-synonymous mutations change the amino acid in the protein (missense mutation) or introduce a premature stop codon (nonsense mutation) leading to a truncated, probably non-functional protein. Frameshift mutations are caused by insertions or deletions and they change the reading frame of the RNA after the mutation, given that the insertion/deletion is not divisible by three (in which case they do not change frame but insert or remove amino acids from protein).

Another heritable force that changes gene expression is epigenetics. Epigenetics refers to DNA methylation and chromatin remodeling and epigenetic patterns can be passed to offspring in what is called imprinting (Cooney *et al.*, 2002). Methylation generally occurs at CpG sites where cytosine is converted to 5-methylcytosine. In general, highly methylated regions tend to be less active, although the mechanism is not fully understood. Chromatin is the complex of DNA and histone proteins with which it associates. If the association between DNA and histones change, gene expression can change as well. Histone modifications occur throughout the entire DNA sequence. The N-termini of histones (called histone tails) are particularly prone to modification. These modifications include acetylation, methylation, phosphorylation, ubiquitylation and sumoylation. The transciptional effect of

modification is exemplified by acetylation, which is generally correlated with transcriptional competence. Epigenetic changes have been shown to be prone to environmental exposure. This is demostrated in work by Cooney and colleagues where they show that methyl dietary supplements alter the DNA methylation in mice offspring (Cooney *et al.*, 2002).

The environment where the human cell resides also affects its expression profile. This has been shown by Choi and Kim in their study of the expression profiles of monozygotic twins (Choi & Kim, 2007). While they are genetically identical, they display remarkable variation in phenotype. Some of this variation may be result of by epigenetic changes between monozygotic twins. Fraga and colleagues showed that the changes were larger in older twins, which underscores the responsiveness of epigenetics to environment (Fraga *et al.*, 2005).

During the last decade, several studies have been performed to assess the effect of genetic variation on gene expression due to technical advances in array technologies. These human expression quantitative trait loci (eQTL) have been mapped in cell lines and tissues using whole genome SNP platforms (Dimas *et al.*, 2009, Emilsson *et al.*, 2008, Stranger *et al.*, 2005). Dimas and colleagues made the observation that there are cell type-specific eQTLs as they studied lymphoblastoid cell lines, T-cells and fibroblasts (Dimas *et al.*, 2009). Work by Chen and colleagues showed that the expression of single genes could be grouped into networks and the summary profile of the whole network (pathway) could be correlated with a phenotype (Chen *et al.*, 2008).

2.4 Metabolomics

Metabolomics refers to the study of metabolic profiles of a cell, tissue and even a whole organism. Metabolites are the end products of cellular processes. Horning and Horning (Horning & Horning, 1970) were the first to introduce the term "metabolic profiles", defining the patterns of biochemically related metabolites. The modern technical era of metabolite measurement started in 1966 when Dalgliesh and colleagues demonstrated that it was possible to obtain multi-component gas-chromatographic analyses of the derivatives from a variety of trace organic compounds present in urine and tissue extracts (Dalgliesh *et al.*, 1966).

Currently, two methods are routinely used in profiling metabolites from human blood samples: nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) (Novotny *et al.*, 2008). MS is a very powerful method for screening metabolites. It is based in mass-to-charge ratio of ions fromed from molecules. It requires laborious cleaning and separation steps prior analysis and is therefore usually coupled with high performance liquid chromatograph or gas chromatograph for the preceding analytic separation step. NMR can detect a large number of metabolites: NMR detects metabolites that contain a nucleus with a nuclear spin (^{1}H, ^{13}C, ^{19}F, ^{31}P, etc). Nuclear spin is a phenomenon where a nucleus rotates around its axis. The spinning is caused by odd pairs of protons and neutrons in the nucleus that have opposite nuclear spins. Protons (^{1}H) are mostly used in magnetic spectroscopy due to their high abundance in organic compounds.

Metabolic profiling gives a range of intermediate phenotypes that may be closely related to a disease mechanism. It also subdivides commonly used intermediate phenotypes and gives more detailed information about them. For example ^{1}H NMR subdivides HDL particles into different density fractions that are not inter-correlated and may have different antiatherogenic functions (data presented in study III).

The GWAS era has extended to metabolic profiles. Illig and colleagues published a GWAS on the ratios of 163 MS measured blood metabolite phenotypes in 1029 German individuals (Illig *et al.*, 2009). They built on a previous underpowered genome scan of 284 males that did not reveal any significant loci (Gieger *et al.*, 2008). They hypothesized that there would be an enzyme that accounts for the transition from one metabolite to another. All significant loci from the study by Illig *et al* are listed in Table 11.

Table 11. Findings of the GWAS for the ratios of metabolites from study by Illig and colleagues (Illig *et al.*, 2009).

Cytoband	Gene	Strongest SNP-Risk Allele	Allele Frequency	P-value (trait transition)	Variance explained (%)
1p31.1	*ACADM*	rs211718-C	0.31	1×10^{-63} (C12/C10)	14.6
2q34	*ACADL*	rs2286963-T	0.37	3×10^{-60} (C9/C10:2)	13.8
4q32.1	*ETFDH*	rs8396-T	0.30	4×10^{-24} (C14:1-OH/C10)	5.6
11q12.2	*FADS1*	rs174547-T	0.30	7×10^{-179} (PC aa C36:3/PC aa C36:4)	36.3
12q24.31	*ACADS*	rs2014355-T	0.28	5×10^{-96} (C3/C4)	21.5

C3 = Propionylcarnitine, C4 = Butyrylcarnitine, C9 = Nonaylcarnitine, C10 = Decanoylcarnitine, C10:2 = Decadienylcarnitine, C12 = Dodecanoylcarnitine, C14:1-OH = Hydroxytetradecenoylcarnitine, PC aa C36:3 = Phosphatidylcholine diacyl C36:3, PC aa C36:4 = Phosphatidylcholine diacyl C36:4

Chasman and colleagues published a metabolomic GWAS which only included lipoproteins and their subclasses, which had a total of 22 measures (Chasman *et al.*, 2009). They had large sample size of 12 489 fasting samples and a total of 17 296 indviduals including replication samples, all of which were females. They identified 43 loci which associated significantly with metabolic traits including seven novel loci listed in Table 12. Although both of these studies report a hit in 12q24.31, they seem to be independent signals since they are 3 Mb apart.

Table 12. Novel loci and candidate genes with genome-wide significant associations ($p < 5*10^{-8}$) for 22 lipoprotein measures

locus	Whole sample	Fasting sample	candidate gene(s)
3q22.3	HDL:S	-	*PCCB, STAG1*
6p21.32	TG:N, VLDL:L	TG:N, VLDL:L	*BTNL2,HLA-DRA, HLA-DRB5*
7q32.2	HDL:Z, LDL:T, LDL:S, TG	-	*COPG2, KLF14, TSGA13*
12q23.2	HDL:T, HDL:N	-	*intergenic ASCL1, PAH*
12q24.31.B	HDL:L,HDL:Z, LDL:T, LDL:S, LDL:Z, TG	HDL:L	*CCDC92, DNAH10, ZNF664*
17q24.2	HDL:M	-	*PRKAR1A, WIPI1*
8p23.1	-	VLDL:Z	*intergenic PPP1R3B*

LDL = low density lipoprotein, HDL = high density lipoprotein, VLDL = very low density lipoprotein, TG = triglycerides, X:L = large particles, X:M = medium particles, X:S = small particles, X:Z = mean particle size, X:T total particles

3 AIMS OF THE STUDY

The aim of this study was to identify genetic factors correlated with variation in human obesity-related quantitative traits using following methods:

1) To investigate the possible cause for positive selection in the *lactase* region by genotyping the causative SNP for lactase persistence in large population cohorts. To test for a difference in BMI and height between cases and controls.

2) To combine European genome-wide linkage screens to perform a combined linkage scan for BMI in Australian, Danish, Dutch, Finnish, Swedish and United Kingdom families. To test for the effect of reduced environmental variation by limiting the analyses to dizygotic twin pairs.

3) To identify networks from gene expression data and test for correlation with metabolic profiles using genetic markers. To perform eQTL analyses and metabolite GWAS for the complete incorporation with pathway analyses.

4 MATERIALS AND METHODS

4.1 Study subjects

More detailed description of the study samples can be found in the original publications (I-III) and the references therein.

4.1.1 Population samples in the *lactase* study (I)

Study I consisted of cross-sectional population samples from Finland. Three of the studies (FINRISK, YF and HEALTH 2000) were collected across Finland. NFBC1966 was collected in the northernmost provinces of Finland. ATBC was collected in Southwest Finland. More detailed descriptions of these four cohorts can be found at www.nationalbiobanks.fi and http://vanha.med.utu.fi/cardio/youngfinnsstudy/. In brief: ATBC was collected for cancer study included males only; FINRISK was collected for the study of chronic disease, coronary risk factors and health behaviour; HEALTH 2000 was a general health examination and interview survey collected in year 2000; NFBC 1966 was collected in 1966 from the Oulu region as a birth cohort and follow up study, the NFBC 1966 subjects have been followed up since the first antenatal contact (10-16th week of pregnancy); YF is an ongoing follow-up study to assess the influence of childhood lifestyle and biological and psychological measures to the risk of cardiovascular diseases in adulthood. The European replication cohorts included KORA from Germany, two samples from the Netherlands (ERF, Rotterdam) and BWWHS from UK. The Rupchen Family (ERF) study is an extended-pedigree study that consists of a single pedigree from a population isolate in the Netherlands. The study's main focus areas are quantitative trait loci related to neuropsychiatric, cardiovascular, endocrinologic, ophthalmologic and musculoskeletal disorders. The Rotterdam study is an ongoing follow-up study. The aim of the study was to reveal the incidence and determinants of chronic disabling diseases that occur in the study subjects during follow up. The KORA (Cooperative health research in the Region of Augsburg, Southern Germany) study was designed to survey the development and course of chronic diseases. Its main focus traits are myocardial infarction and diabetes mellitus along with the associated risk factors. The British Women's Heart and Health Study (BWHHS) is a prospective cohort study of heart disease. Its aim is to provide information about existing patterns of treatment of heart disease and to better the understanding of risk factors and disease prevention. The samples included in the *lactase* study are summarized in Table 13.

Table 13. The sample demographics of the cohorts used in the lactase study

	N (M/F)	BMI (kg/m^2)	BMI SD	Age	Age SD	LP prevalence (%)
NFBC 1966	5498 (2636/2862)	24.6	3.9	31	0	85
ATBC	2126 (2126/0)	26.9	4.2	64	5	84
FINRISK	2265 (1555/710)	28.1	5.3	58	10	80
Health2000	5320 (2437/2883)	26.8	4.5	53	15	82
YF	2165 (985/1180)	25.9	5.0	38	5	83
BWHHS	3109 (0/3109)	27.3	4.5	69	5	94
ERF	2104 (909/1195)	26.9	4.8	50	15	90
Rotterdam	5689 (3320/795)	26.0	3.5	70	9	91
KORA S3	1578 (783/795)	27.2	3.8	53	10	87
KORA S4	1755 (859/896)	27.4	3.9	54	9	87

N = number of individuals, SD = standard deviation, NA = Not available, M = males, F = females,

LP = lactase persistence

4.1.2 The GenomEUtwin sample (II)

The samples in study II were collected from twin cohorts from seven European countries (Denmark, Finland, Italy, The Netherlands, Norway, Sweden and United Kingdom) and Australia. The sample consisted from 4401 families with twins and was collected by the GenomEUtwin consortium (www.genomeutwin.org). We were able to include 10535 individuals with genotype and phenotype information to this study. The sample demographics are presented in Table 14.

Table 14. Demographics of the GenomeUtwin sample

	Mean BMI	SD	Mean Age	N
All				
Males	25.3	3.4	50.2	3667
Females	24.7	4.5	48.9	6868
Australia				
Males	25.4	3.7	44.5	1214
Females	24.7	4.6	44.4	1876
Denmark				
Males	25.2	3.3	53.7	247
Females	23.5	3.7	61	377
Finland				
Males	25.5	3.7	52.1	518
Females	24.7	5.2	60.4	339
The Netherlands				
Males	25.2	3.3	45.1	1160
Females	24.5	4.1	43.6	1535
Sweden				
Males	25.5	2.7	74.6	528
Females	25.2	3.7	75	525
UK				
Females	25.1	4.7	47.3	2216

N= number of individuals, SD = standard deviation

4.1.3 The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (III)

The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study consist of 5025 individuals who took part in the larger FINRISK 2007 collection. The DILGOM sample was collected in order to study in greater detail the components affecting obesity and metabolic syndrome. It has five major components: 1) investigation of diet, physical activity, psychosocial factors, markers of obesity and glucose metabolism in 5025 men and women aged between 25 and 74 years; 2) investigation of the roles of psychosocial factors on diet, physical activity, obesity and metabolic syndrome; 3) assessment of the influence of diet on selected endocrinological factors, cytokines and other biomarkers, and their relationship to weight and glucose metabolism; 4) the profiling of abdominal obesity

and metabolic syndrome against whole genome SNP data and the testing of the effect of identified markers on weight gain and changes in glucose metabolism in a prospective manner. Funding for a five year follow-up is currently being applied for. This study included 303 females and 287 males of age between 25 and 74 years (average = 52, standard deviation = 13.6) from Helsinki region.

4.1.4 Laboratory measurements

4.1.4.1 Metabolome

The ^{1}H NMR technique is based on proton resonance in a magnetic field. The resonance is dependent on the structure of covalent bonds surrounding the proton of interest. The surrounding chemical composition gives distinct resonance spectra that can be utilized in the quantitation of the compound of interest. In study III we used ^{1}H NMR to quantitatively identify 134 serum metabolites. Measurement is performed in three separate steps or windows. The first window contains low molecular weight metabolites (LMWM, Figure 11). The LIPO window (Figure 12) contains information about lipoproteins and subclasses. The LIPID window (Figure 13) is analyzed using lipid extract and contains more detailed molecular information on various serum lipid constituents like free and esterified cholesterol, sphingomyelin, polyunsaturation and ω-3 fatty acids. Metabolites for the 518 individuals used in study III were quantified using a Bruker AVANCE III spectrometer operating at 500.36 MHz (^{1}H observation frequency; 11.74 T). Temperature was stabilized by using an A BTO-2000 thermocouple at the level of approximately 0.01 °C in the sample.

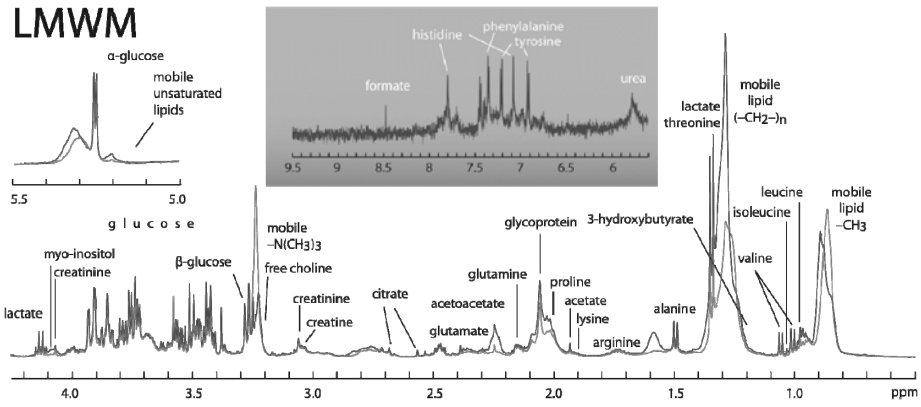


Figure 11. The LMWM window of the ^{1}H NMR (produced by Mika Ala-Korpela, reproduced with permission)

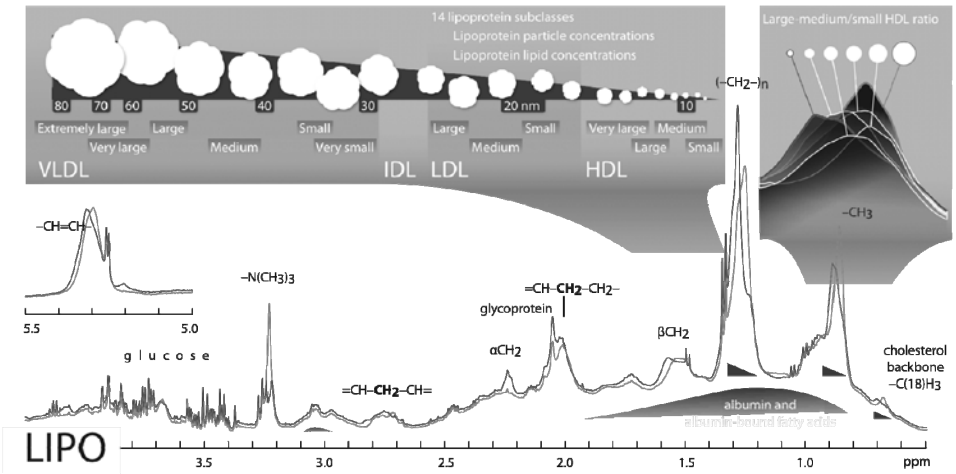


Figure 12. The LIPO window of the ^{1}H NMR (produced by Mika Ala-Korpela, reproduced with permission)

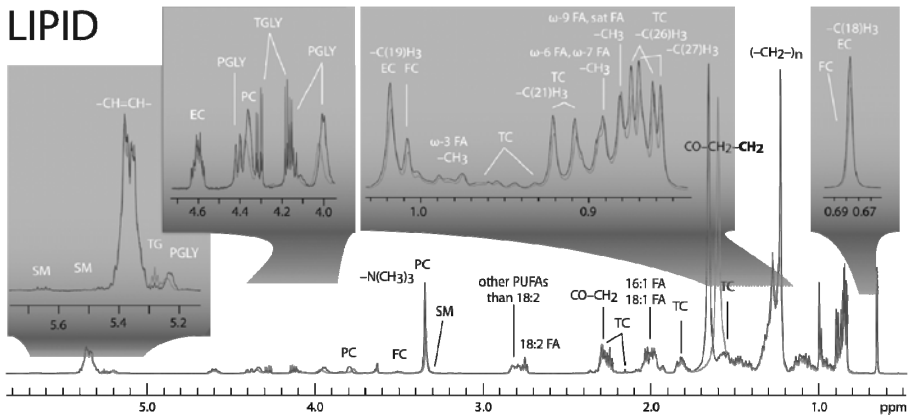


Figure 13. The LIPID window of the ^{1}H NMR (produced by Mika Ala-Korpela, reproduced with permission)

4.1.4.2 Transcriptome

To obtain stabilized total RNA in study III, the PAXgene Blood RNA System (PreAnalytiX GMbH, Hombrechtikon, Switzerland) was used. The protocol includes the collection of 2.5 ml peripheral blood into PAXgene Blood RNA Tubes (Becton Dickinson and Co., Franklin Lakes, NJ, USA) and the total RNA extraction using the PAXgene Blood RNA Kit (Qiagen GmbH, Hilden, Germany). We used the s standard manufacturer's protocol. The fragmentation of the RNA sample was measured with the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). We produced biotinylated cRNA from 200 ng of total RNA with the Ambion Illumina TotalPrep RNA Amplification Kit (Applied Biosystems, Foster City, CA, USA), following the manufacturer's protocol. Biotinylated cRNA (750 ng) was hybridized onto Illumina HumanHT-12 Expression BeadChips (Illumina Inc., San Diego, CA, USA), according to the manufacturer's protocol.

4.1.4.3 Genomic markers

The SNP *LCT* C/T-13910 was genotyped using the iPlex assay on the MassARRAY System (Sequenom, San Diego, CA, USA) using standard protocols in 15 209 Finnish individuals. The YF cohort LCT C/T-13910 genotyping was performed by using the 50-nuclease assay and fluorogenic, allele-specific, TaqMan probes and primers (Applied Biosystems, Foster City, CA, USA). Reactions were run using the ABI Prism 7000 sequence detection system (Applied Biosystems). The genotype frequencies were in Hardy–Weinberg equilibrium (HWE) in all five cohorts (YF, P = 0.814; ATBC, P = 0.84; FR, P = 0.77; NFBC, P = 0.93; Health2000, P = 0.54) and the call rate was greater than 95% in all samples. BWHHS genotypes for LCT C/T-13910 (rs4988235) were generated using KASPar chemistry, which is a competitive allele-specific PCR SNP genotyping system using FRET quencher cassette oligos following standard protocols (KBiosciences, Essex, UK).

Genome-wide SNP data was used in studies I and III. Samples were successfully genotyped after excluding chip failures and poor quality samples (as determined by visual inspection of a 0.75% agarose gel or poor Sequenom call rate). Low quality samples were removed from further analysis when the call rate was less than 95%. A sequenom fingerprint was performed prior to the SNP array. Twenty SNPs were genotyped for quality control purposes. An individual was removed if it failed to match Sequenom genotype fingerprinting (concordance < 0.90 for at least 10 genotypes). If two individuals were closely related or a sample was present in duplicate samples (pairwise identity by state pi-hat > 0.10), the sample with the smaller call rate was removed. SNPs failing to meet the following quality thresholds were also removed from further analysis: call rate > 0.95, minor allele frequency > 0.01 and Hardy-Weinberg equilibrium P value > 1.0×10^{-6}. In study I we had genome-wide SNP data (Illumina HumanHap 370, Illumina Inc., San Diego, CA, USA) available for 5 555 individuals from the NFBC 1966 cohort. Similarly, we had genome-wide SNP data (llumina HumanHap 610-Quad SNP array) available for a subset of the Health2000 cohort (2145 individuals) ascertained for a matched case–control study of metabolic syndrome. In study III the genotyping was performed with the Illumina 610 chip using standard protocols.

The genome-wide microsatellite markers used in study II were combined from genotype data provided by each registry as the participating twin registries released their data for the combined analysis. The participating registries were: the Australian Twin Registry (Hopper, 2002), the Danish Twin Registry (Skytthe *et al.*, 2002), the Finnish Twin Cohort (Kaprio & Koskenvuo, 2002), the Netherlands Twin Register (Boomsma *et al.*, 2002), the Swedish Twin Registry (Pedersen *et al.*, 2002) and the TwinsUK Adult Twin Registry (Spector & MacGregor, 2002).

4.2 Statistical methods

4.2.1 Quality control

4.2.1.1 Familial relationships

In the context of linkage mapping one must be certain that all familial relationships are correctly specified. The expected genetic covariance in linkage analysis is derived from the declared familial relationships. The validity of twin zygosity and other familial relationships were tested in study II with the program Graphical Representation of Relationships (Abecasis *et al.*, 2001).

In contrast, the population is expected to be unrelated in association studies. One cannot test for relatedness when one only has a few markers, such as in study I where only one genotype was studied. However, when a large number of polymorphic markers (at least roughly 500 SNPs or 100 microsatellites) spread evenly across the genome are available, one can detect close relatives using the average identity-by-state (IBS) sharing over the whole genome. This approach was used to remove closely related individuals in study III.

4.2.1.2 Genotype quality controls

The first quality control step usually performed after genotyping is assessing missing genotype rates or the success of genotyping. If the missing genotype rate is large (over 20 %), one can assume that the successful genotypes are unreliable and the marker should be discarded. This is a common threshold in studies with few markers since the genotypes are commonly called manually. In genome-wide SNP panels it would be too laborious to manually curate the genotypes and therefore a computational algorithm is applied to make the calls. In our studies we used the Illuminus-algorithm (Teo *et al.*, 2007). In the case of computationally curated genotypes, we have used a more stringent missing genotype rate threshold (over 5%).

The calling algorithm is sensitive to the genotype cluster sizes as it attempts to position the clusters. If the frequency of the minor allele is low (less than 1 %) the positioning becomes unreliable using the Illuminus algorithm, and therefore are removed from the analyses.

In family studies one can confirm that the markers segregate by the laws of Mendel. Mendelian inconsistencies were removed in study II using the program PedCheck (O'Connell & Weeks, 1998). Sometimes the genotyping error is not obvious and the surrounding haplotype structure provides evidence that the genotype is probably erroneous. The non-Mendelian inconsistencies were evaluated with the linkage analysis program Merlin and they were removed using the option pedwipe (Abecasis *et al.*, 2002).

Hardy-Weinberg equilibrium (HWE) states that allele and genotype frequencies are in equilibrium, or in other words that the frequencies of the genotype classes are the products of allele frequencies. While the HWE assumes, for example, that the population size is infinite, mating is random, there are no new mutations, there is no selection and there is no migration. It is robust enough to function in populations where these expectations are not met. Deviation from HWE can be tested with a χ^2 test, which we performed in all studies. There was no deviation from HWE in study I. We used $p > 0.001$ as a threshold in study II and removed all markers falling below the threshold. In studies I and III we used $p > 1*10^{-6}$ as a threshold since we had significantly more markers. One would expect to find a p-value as small as $2*10^{-6}$ just by chance on a 610k chip and as small as $3*10^{-6}$ on 370k chip.

4.2.1.3 Phenotype quality control, transformations and corrections

All statistical tests used in studies I-III assume that the trait is normally distributed. Outliers deviating far from the mean have larger weight in the test statistics. Therefore, in study II we tested for the skewness and kurtosis of BMI in the total sample. The values were 1.14 and 2.69, respectively, and thus a log base10 transformation was used. Normalization changed the distribution and the skewness and kurtosis of logBMI were 0.47 and 0.82, respectively. Sex, age and country of origin were used as covariates in the combined sex analysis since they correlated significantly with logBMI. Outliers (n = 393) were excluded from the analyses. Outliers were defined as individuals differing by more than 3 SD from the population mean. In study I the BMI was not transformed since after outlier removal (3 SD) the distribution was approximately normal (both skewness and kurtosis were between -1 and 1). There were a large number of phenotypes, with non-normal distribution in study III. Box-Cox power transformation was applied to all phenotypes, and values deviating more than 4 SD were removed. The variation in phenotype caused by non-genetic factors was removed prior analysis (studies I and III). Covariate correction was applied during the analysis (study I) for the same reason.

Table 15. Corrections for phenotypes used in study I

	Sex	Age	Area of origin	PCA
NFBC 1966	x			
ATBC		x		
FINRISK	x	x	x	
Health2000	x	x	x	
YF	x	x	x	
BWHHS		x	x	
ERF	x	x		
Rotterdam	x	x		x
KORA S3	x	x		x
KORA S4	x	x		x

PCA: Principal components analysis of genome wide SNP data was utilized to control for possible population stratification using the first ten principal components

In study I phenotypes were corrected, where applicable, for area of origin, gender, age and the ten first principal components. Corrections are summarized in Table 15. Sex, age and country of origin correlate significantly with BMI in study II and they were used as covariates in the analysis. We also performed sex stratified analyses in study II where sex was not a covariate. The phenotypes in study III were corrected for age, gender, and the first 10 principal components by taking the standardized residuals from a multiple linear regression with the above as covariates.

4.2.1.4 Quality control in expression arrays

Background corrected probe signal intensities and bead counts of the scanned data were taken from the Illumina's BeadStudio software for further processing in study III. Quantile normalization was used to force probe intensities for all samples on all arrays to be the same. Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient were used to measure the correlation of normalized technical replicate pairs for all samples. The MA plot (application of a Bland–Altman plot) was used to visualize the intensity-dependent ratio of microarray data. M (y-axis) is the intensity ratio and A (x-axis) is the average intensity for a dot in the plot. The MA plot gives a quick overview of the data. Each MA plot was manually curated for curvature or deviation from M = 0 axis, none of which were found. A sample was removed from further analysis if its P was <0.94

or ρ was <0.60 (9 samples failed). Raw signal intensities from corresponding replicates were combined. The number of beads that contributed to each signal were weighted and summed to obtain one measure of signal intensity for each sample on each probe. Probes that did not meet certain criteria were excluded from further analysis. Probes were removed if they were non-autosomal, were complementary to cDNA from erythrocyte globin components or mapped to several genomic positions. We use the expression markers defined in Whitney and colleagues' paper to correct for the relative cell type numbers (Whitney *et al.*, 2003). The corrected cell type proportions included B cells, cytotoxic T lymphocytes/natural killer cells, erythrocytes, lymphocytes, Myc-regulated cells, neutrophils and reticulocytesmyeloid cells (profiles for the time of day were also included). We were not able to correct for T cells (uncovered on HT-12 array), mast cells (not assessed in Whitney *et al.* 2003) and basophils (not assessed in Whitney *et al.* 2003).

4.2.1.5 Association analysis with expression data (III)

All univariate statistical tests and permutations were performed with PopGenomix, a custom C++ package. Using genome-wide SNP genotypes on the same individuals, we investigated the genetic effects on expression for each gene in the LL module and for the LL module as a whole; 35 419 $\log_2$ normalized expression probes and 541 654 SNPs (2 061 516 SNPs after imputation) were taken forward for further analyses. For SNPs in *cis*, within 1 Mb of the expression probe midpoint, a linear regression was performed. In order to determine significance, a permutation procedure was implemented (Stranger *et al.*, 2005). For *trans* SNPs, greater than 5 Mb away or on a different chromosome, the non-parametric Spearman rank correlation was used (Stranger *et al.*, 2005). The Spearman rank correlation offers a more robust test of association since permutation across the whole genome is computationally prohibitive. To determine the significance of the nominal Spearman P value, a threshold of $5.0*10^{-7}$ was implemented. 10 000 permutations were performed in order to evaluate the level of significance. We then assessed how gene expression correlated with metabolic measures by linear regression to obtain a list of metabolite-correlating genes for network analysis.

4.2.2 Statistical analyses

4.2.2.1 Linkage

Marker maps from all of the cohorts in study II were combined using the Cartographer program (Sammalisto *et al.*, 2005). We performed the variance components linkage analysis for logBMI in 4401 families (10535 individuals) using Merlin (Abecasis *et al.*, 2002). Merlin utilizes the Lander-Green algorithm (Lander & Green, 1987) for the IBD estimation and the standard variance components framework for linkage analyses. In total, there were 3356 dizygotic twin (DZ) pairs (706 male pairs, 2040 female pairs and 610 opposite-sex pairs). The linkage analysis program Merlin estimates the heritability of a given trait along with the linkage analyses. We performed country-specific analysis in extended families (where available) and DZ twins only. We performed an analysis of combined samples across countries for both the combined DZ twin sample and the extended families sample. The extended families sample included all available individuals. We also performed sex-stratified analysis in the combined sample to examine whether there was a sex-specific contribution to the linkage signal at a given locus.

4.2.2.2 Power calculation

We estimated the decrease of power when the incorrect analysis method is used. An additive model is a common compromise to avoid multiple testing when running GWAS for quantitative traits. The additive model may hinder the power to detect an association when true mode of inheritance is not additive. The effect is usually more pronounced when the true effect is seen between the minor allele homozygotes and the major allele carriers, as is the case in study I. The test of power reduction was performed by the bootstrapping method, using the R package. We simulated 10 000 samples of 17 374 individuals and randomly assigned a normally distributed quantitative “phenotype” to them. We distributed a random genotype with a 0.39 minor allele frequency and assigned a phenotype decrease of 0.08 trait value for minor allele homozygotes. Each of the 10 000 samples were subsequently analyzed with both additive and recessive models and p-values were collected.

4.2.2.3 Association testing and dominance deviation

Association testing was performed in studies I and III. The minor allele homozygotes were tested against major allele carriers in all of the cohorts in study I. All samples except for ERF were analyzed using standard linear regression on the corrected phenotype. The ERF cohort, a family sample, was analyzed with QTDT (Abecasis *et al.*, 2000) which utilizes variance components methodology and transmission disequilibrium test. Since the project was collaborative, all of the cohorts performed their individual analyses which were subsequently combined in meta-analysis. The population cohorts were analyzed using the following statistical packages: NFBC 1966 and Health2000 were analyzed with PLINK 1.04 (Purcell *et al.*, 2007); FINRISK, KORA and ATBC with R (R Development Core Team, 2007); and BWHHS, YF and Rotterdam with SPSS (SPSS, 1999). In addition to a regular test for association, we performed a formal test for rejecting the additive model in study I. The dominance deviation tests whether the heterozygous genotypes deviate from the additive model. We also performed sex stratified analyses to test if there was a gender-specific contribution to the signal. The metabolite GWAS in study III was performed using PLINK with linear regression and an additive model for all 134 transformed and corrected metabolic phenotypes (Purcell *et al.*, 2007).

4.2.2.4 Imputation

Imputation is a method used to fill in missing data with respect to a reference set (Figure 14). One can fill in missing genotypes in a haplotype if an available reference set has denser marker map.A commonly used reference is the HapMap Central European ancestry trio sample. It contains over 2 million phased SNPs. The NFBC 1966, Health2000 subset, Rotterdam, ERF and KORA samples in study I were imputed using the MACH program (Li *et al.*, 2009) and the HapMap (The International HapMap Consortium, 2003) CEU reference.

4.2.2.5 Principal components analysis

Principal components analysis is a mathematical method where the variation of a dataset of correlating variables is captured in a smaller number of uncorrelated variables. Principal components analysis was utilized in studies I and III. The genetic correlation between individuals was reduced to principal components to detect population structure as described in a paper by Jakkula and colleagues (Jakkula *et al.*, 2008). The genotype information should be available from no less than 50 000 SNP markers evenly spaced across the genome in order to obtain reliable information on the population structure. The EIGENSOFT program (Price *et al.*, 2006), which utilizes the genotypes of individuals, was used to reduce the genetic variation into principal components in study I. The first component is fitted so that it captures the largest amount possible of variation. Then the next principal components are fitted orthogonally to the previous. In study III, a different but analogous approach was used to detect population structure. Multidimensional scaling, utilized in the PLINK program (Purcell *et al.*, 2007), uses the relationship matrix to calculate the principal components. Relationship matrix contains mean IBS sharing across the whole genome for each pair of individuals in the data. Since the expression of the gene pathway is correlated, we used principal components analysis in study III to reduce the expression of the whole pathway into one vector that could be used in the correlation test with metabolites and genetic markers.

Figure 14. 1) The reference panel contains haplotypes of alleles 0 and 1. 2) The genotyped sample contains allele counts 0 and 2 represent homozygotes, 1 heterozygotes and ? missing data. 3) The missing values are filled in according to the haplotypes of the reference and the surrounding markers from data.

1. Reference set of haplotypes

0	0	1	1	1	1	?	1	1	0	0	0	?	1	1
0	0	0	0	0	0	?	1	0	1	1	1	?	0	1
1	1	1	1	1	1	?	0	1	0	0	0	?	0	0
1	0	1	1	1	1	?	1	1	1	1	1	?	0	1

2. Sample with missing genotypes respect to reference

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	?	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

3. Imputing the missing data

1	1	2	2	2	0	0	1	2	0	0	0	1	1	1
1	1	1	1	1	0	0	1	2	1	0	0	0	0	0
0	0	1	1	1	1	1	2	1	1	1	1	0	0	1
1	2	2	2	1	0	0	1	2	0	0	0	1	1	1
2	1	2	2	2	0	0	0	2	2	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	0	0	1
0	0	2	2	2	0	0	2	2	2	2	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

4.2.2.6 Meta-analysis, heterogeneity and interaction using summary statistics

All study cohorts were analyzed separately in study I. Therefore the test statistics were combined using meta-analysis to obtain the test statistics for the combined sample and sex-stratified analyses. In meta-analysis, the effect size estimates (β) of linear regression is weighted with their respective standard errors and combined together. Standard error will correct for the sample size differences. The response variable must be uniformly treated prior to linear regression because the effect estimate is a measure of change in standard deviation units. If one study has standardized (SD = 1, mean = 0) the trait values and another uses raw trait values, for example BMI (SD=4, mean =25), the effect estimate will be in different standard deviation units and combination of these is not obvious. Therefore, the phenotype was standardized in all cohorts in study I. Heterogeneity analysis tests if some of the effect estimates are very different from others. This is similarly performed using the effect estimates' standard errors.

We used the summary statistics in study I to test for gender interaction. In brief, we had the sex-specific analyses performed in all cohorts and those were meta-analyzed to produce a combined effect for both genders. We then tested if there was a significant difference between the effect size estimates between genders (Hardy, 1993).

4.2.2.7 Network analysis in gene expression

In biological pathways, many genes tend to co-express. Therefore, it is natural to incorporate these correlations into a network-based framework. Within this framework, pair-wise correlations between genes are used to describe the connectedness of the network, and clusters of tightly correlated genes (or modules) can define pathways. Network analysis was performed using the WGCNA R package (Langfelder & Horvath, 2008). To construct a co-expression network that characterizes metabolic traits, the method of Horvath and Dong (Horvath & Dong, 2008, Langfelder & Horvath, 2008) was used to assess the top 10% of expression signals for all metabolites (3,520 unique signals). The correlation matrix was constructed by obtaining all against all Pearson correlation coefficients and the adjacency matrix was calculated with a soft threshold power of seven. The initial modules were determined with the dynamic tree cut function in WGCNA where the minimum module size of 10 genes was used. Summary profiles were obtained from each individual module by singular value decomposition (first vector of PCA). Highly correlated modules were identified by clustering the summary module profiles. The modules with dendogram height less than 0.20 were merged.

4.2.2.8 Network orientation and putative causality

Genetic markers are used as causal anchors to direct the edges of a network (Chen *et al.*, 2007a, Thomas & Conti, 2004, Schadt *et al.*, 2005, Li *et al.*, 2006, Zhu *et al.*, 2007, Kulp & Jagalur, 2006, Chen *et al.*, 2007b, Sieberts & Schadt, 2007). Genetic markers are thought to be ideal in causality tests due to the underlying randomization by Medelian laws, and are thus less susceptible to confounding effects (Kulp & Jagalur, 2006, Greenland, 1990, Katan, 1986, Clayton & McKeigue, 2001, Zhu *et al.*, 2004, Thompson *et al.*, 2005). If a trait is associated with a genetic marker, the genetic marker must cause the variation in trait:

Marker → Trait A (Schadt *et al.*, 2005)

This is then extended to the graph where causation is referred as:

Marker → Trait A → Trait B

The SNP causes variation in Trait A which in turn causes variation in Trait B. Conditional independencies of the variables can be determined by the graphical property of d-separation, an algorithm used to compute all the conditional independence relations implicated by their graphs (Pearl, 1988, Pearl, 2000, Shipley, 2000). We used the R package NEO (Aten *et al.*, 2008) in our attempt to infer causality and directedness of the network. We included metabolites in edge orienting that were associated with Lipid-Leukocyte (LL) module expression and had at least one SNP associated with $P < 5*10^{-7}$. The LL module was the strongest network of genes that showed association with metabolomic phenotypes in study III. We were left with 29 metabolites in the directed network analysis along with the LL module genes. We selected SNPs as causal anchors by two criteria; either they were associated with a metabolite (P value $< 5*10^{-7}$) or the gene expression in *cis* of at least one of the LL module genes (P value $< 5*10^{-3}$). The automatic SNP selection approach implemented in NEO was used to assess the causality of a give SNP. The selection uses both greedy and forward-stepwise regression. The causality of an oriented edge was determined by (i) its NEO.NB.OCA score being greater than 0.30 (ii) its causal model *P* value being greater than 0.10 and (iii) the edge's A → B path coefficient had a Z score statistic smaller than -1.96 or greater than 1.96.

4.2.2.9 Connectedness of the network

We attempted to assess if the LL module's core co-expression changes with metabolite levels. The $\log_2$-normalized expression values were partitioned into quintiles for each module gene. We leveled the number of individuals in each quintile. We calculated a correlation matrix of the gene expression using Spearman rank correlation in each quintile and fitted a linear model across all co-expression pairs. We assessed from the slope of the curve if the expression of the module was dependent on the metabolite concentrations.

5 RESULTS AND DISCUSSION

5.1 Lactase persistence association with body mass index (I)

5.1.1 Association analyses and meta-analyses

We used five Finnish population cohorts to assess the effect of LP, none of which had been ascertained for height or BMI. The cohorts were The North Finland Birth Cohort 1966 (NFBC 1966), The Health 2000 Health Examination Survey (Health2000), The Cardiovascular Risk of Young Finns Study (YF), the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) study and the FINRISK. FINRISK is a cross-sectional population survey targeting coronary risk factors and is collected every 5 years. We performed additional analyses in European populations: the British population using the British Women's Heart and Health Study (BWHHS), the German using the KORA S3 and KORA S4 study samples, and the Dutch using the Erasmus Rupchen Family (ERF) study with extended pedigrees and the Rotterdam study. Each population cohort was analyzed using linear regression to test if the lactase persistent C_{-13910} allele homozygotes differed from the lactase non-persistent T_{-13910}-allele carriers. The effect estimate and its standard error in the ERF family sample were obtained from QTDT. **Table 16** summarizes the results from individual cohorts.

Table 16. Results summary of all cohorts

	n	β	se	p-value
NFBC 1966	5498	-0.11	0.04	0.002
ATBC	2126	-0.12	0.06	0.04
FINRISK	2265	-0.09	0.05	0.09
Health2000	5320	-0.06	0.03	0.1
YF	2165	-0.04	0.06	0.51
BWHHS	3109	-0.06	0.07	0.4
ERF	2104	-0.08	0.06	0.16
Rotterdam	5689	-0.02	0.05	0.69
KORA S3	1578	-0.001	0.07	0.99
KORA S4	1755	0.08	0.06	0.24

β = effect estimate, n = number of individuals, se = standard error

We combined the effect estimates and their standard errors in the meta-analysis of 17 374 Finns (β = -0.08, P = $1.5*10^{-5}$) and found that the CC genotype was associated with decreased BMI. We tested for heterogeneity in the meta-analysis. There was some evidence of heterogeneity (I^2 = 10), but the Q statistic was not significant (P = 0.35). The LP allele carriers had 0.3 kg/m^2 higher BMI than lactase non-persistent individuals. This corresponds to ~1 kg difference for an average person. LP explained 0.2% of the BMI variance in the NFBC 1966 cohort. We had the FTO rs9939609 genotypes available in NFBC 1966 to compare the proportion of variance explained. The proportions of variance explained by the FTO variant was found to be equal (0.2%) in the NFBC 1966. In sex-stratified analyses we found that males (n = 9739, β= -0.09, P = $1.6*10^{-4}$) had a somewhat higher effect estimate in the model than females (n = 7635, b = -0.06, P = 0.02). The difference was not statistically significant (P = 0.6). We added the effect sizes of the additional four European cohorts to the meta-analysis for a total of nine cohorts and 31 720 individuals. The association with BMI remained robust (β = -0.06, P = $7.9*10^{-5}$) in the combined meta-analysis of all samples. The results from linear regression and the meta-analyses are summarized as a forest plot (Figure 15).

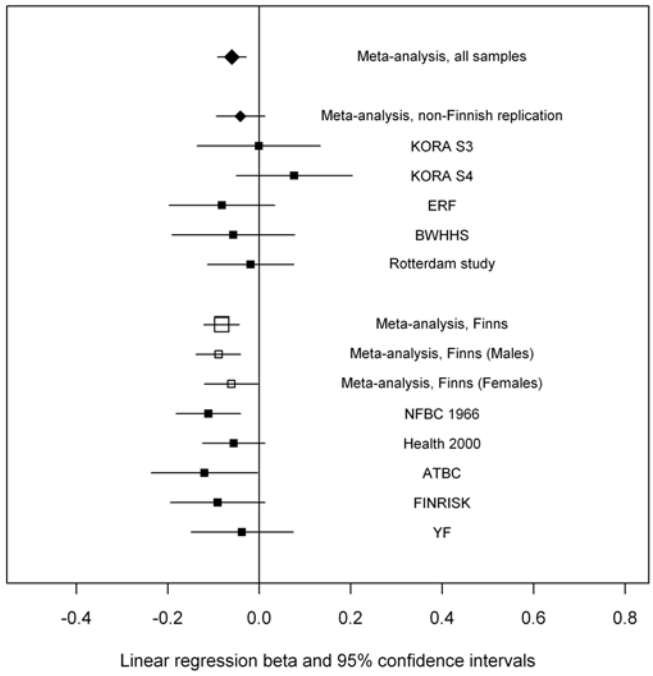


Figure 15 . Forest plot of the individual cohort analyses and the meta-analyses

5.1.2 Addressing stratification

We had genome wide SNP data available in two of the cohorts, NFBC 1966 (n = 4911) and Health2000 (n = 2145, metabolic syndrome case–control sample). The first 14 eigenvectors explained the largest proportion of the variance in the principal components analysis of NFBC 1966 genome-wide SNP data. We tested for the correlation between BMI and the first fourteen eigenvectors in NFBC 1966 and Health2000. Figure 16 is a Scree-plot of the eigenvalues. Figure 17 shows the first two principal components of the NFBC 1966 data, black dots indicate lactase persistent and gray dots controls. It has been previously shown by Lahti-Koski and colleagues that there is no correlation between geographic location and BMI within Finland, which might have lead to spurious association due to stratification (Lahti-Koski *et al.*, 2008).

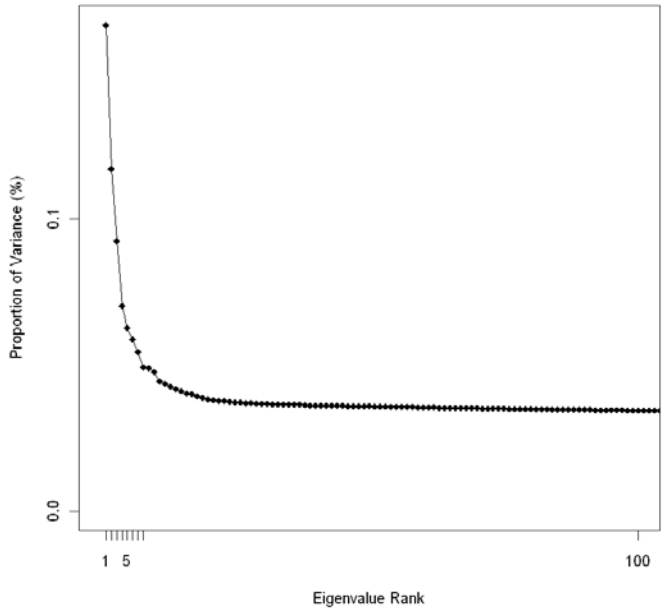


Figure 16. The Scree plot of the first 100 eigenvectors where the first eigenvector explains the largest proportion of variance

Linear regression was utilized to evaluate evidence of correlation between the eigenvectors and BMI. We did not find correlation in either NFBC 1966 or Health2000 ($P > 0.05$). We used logistic regression to test for the correlation between LP and eigenvectors. The association did not change after adding all significantly correlating eigenvectors as covariates in the NFBC 1966 data or in the Health2000 subsample. The Health2000 GWAS subsample was ascertained as a case-control cohort for metabolic syndrome and hence we tested for the effect of LP in both cases and controls separately. We found that there was no difference in the linear regression β between the groups.

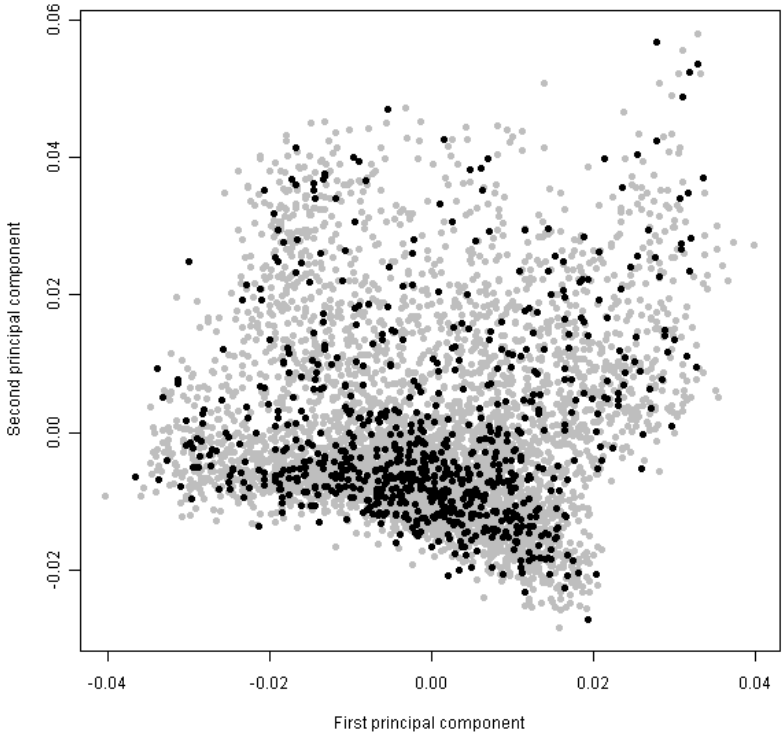


Figure 17. The first two principal components from the NFBC 1966 data; lactase persistent (black) and lactase non-persistent (gray)

5.1.3 Imputation

We compared the imputed genotypes of the rs4988235 SNP with the Sequenom genotyped genotypes for individuals in the Health2000 subsample and NFBC 1966. The imputation quality score was 0.91 and r^2 was 0.81, which suggest that the imputation was successful. After comparing the genotypes we found that the LP phenotype defining C/C genotype had a 23% difference between the genotyped and imputed genotypes. The imputation had not called 245 C/C genotypes that were confirmed by the 1060 C/C genotypes in the genotyped data. There were 15 cases where the imputation had called a C/C genotype and genotyping had called C/T. When counting all genotypes, the amount of discordance between imputation and genotyping was 11%.

5.1.4 Power reduction and model in analyses

Since the initial GWASs have not identified association in the LCT region, and lactase persistence shows a dominant pattern of inheritance, we used the dominance deviation test from additivity to see if the effect was additive or dominant. We analyzed each cohort separately and combined the β-estimates with their standard errors in meta-analysis. The dominance term showed evidence that we could reject the additive model ($P = 0.001$, $\beta = 0.047$). We used the additive model in our analyses with the Finnish data instead of the correct dominant model and combined the results with meta-analysis. The additive model showed a decrease of association ($P = 0.001$) compared with the dominant model ($P = 1.5*10^{-5}$). We used bootstrapping to simulate the reduction in power when using the wrong model. The result of this power reduction is summarized in Table 17.

Table 17. The results of power calculation for using the additive model instead of correct dominant model where sample size (n = 17 374), effect size (β = 0.082) and MAF (0.39) are the same as the LCT variant had in Finnish sample of this study

Significance level	Power (D/A, %)
0.05	98/83
0.01	91/63
0.001	73/35
$1*10^{-4}$	51/16
$1*10^{-5}$	30/7
$1*10^{-6}$	16/2
$1*10^{-7}$	8/1
$5*10^{-8}$	6/1

D = dominant model, A = additive model

5.1.5 Discussion

In the study I we present a novel association with genetically defined lactase persistence and BMI ($P = 7.9*10^{-5}$). The *LCT* region has been shown to cause spurious associations due to its frequency gradient by Cambell and colleagues (Campbell *et al.*, 2005). We were able to show that the association was not due to population stratification by using principle components analysis for population stratification detection. Additionally, the example made by Cambell and colleagues concentrated on stature, where we did not find any evidence of association (P = 0.41). The BMI association was mediated by the body weight of individuals ($P = 1.4*10^{-5}$).

We critically addressed two aspects that influence the success of GWAS probably not only in *LCT* region. All of the cohorts used in the meta-analyses have been imputed with the same HapMap 2 reference because the functional variant of the lactase gene is commonly not included in SNP panels and the LD to the nearest variant on the panel is insufficient. In addition to false genotypes, the wrong analysis model has been utilized. We showed that the wrong analysis model decreases power substantially when true effect is present among minor allele homozygotes only. The false genotypes and the inappropriate analytical model combined have a detrimental effect on the power of association analysis in this specific region even in a very large meta-analysis. This region shows strong LD structure and if a common haplotype is by chance missing from the reference set, the result may be false

negative finding. This may play a large role in why the Rotterdam and KORA cohorts did not replicate the finding. The ERF sample was not susceptible to the imputation issue, because the imputation reference was genotyped from within the family and no external reference set was used. Additionally, the family structure of ERF makes a large difference in the imputation. Therefore, the LD structure and problems with imputation may somewhat explain why the non-Finnish population samples did not replicate the association significantly (P = 0.24).

We cannot exclude the possibility that cultural influences on milk consumption override the rather minor discomfort consequent to milk ingestion by non-persistent individuals. It has been shown by Smith *et al* that the association of LP with milk consumption varies between populations. (Smith *et al.*, 2008) They showed that it was stronger in intermediate allele frequency regions than in those where one allele is rare. The consumption of dairy products is large in Finland and this may have inflated the effect size because in Finland there is a correlation with liquid dairy product consumption and LP (Lehtimaki *et al.*, 2006, Enattah *et al.*, 2004).

It remains unclear how LP affects body composition. It may be due to the dietary restriction caused by the trait or perhaps by the negative symptoms such as diarrhoea. The *lactase* locus has been under strong positive selection and this novel association with BMI may provide some explanation to the positive selection. The *lactase* association with BMI also gives an excellent example for James Neel's the thrifty gene hypothesis where a 'thrifty' genotype would have been advantageous during times of food scarcity and in today's obesogenic environment becomes detrimental (Neel, 1962). We have shown here that the recent GWASs have not yet been able to reveal all of the common variants affecting BMI. In summary, we show evidence that the European lactase persistence variant, and perhaps other regulatory variants in the lactase region, plays a role in the human adult BMI development.

5.2 Genome-wide linkage scan for body mass index in European twin cohorts (II)

The heritability of BMI was 54 % in the extended families sample and 73% in the DZ twin sample. We performed sex-stratified analyses in the combined samples of both DZ twin and extended families, but not in the country specific samples. All linkage results per chromosome in the extended family sample, DZ twins only sample and sex stratified analyses are shown. The linkage results of chromosomes with either a linkage peak reported in this study or a published GWAS hit can be viewed in Figure 18 and Figure 19. The DZ twin data provided evidence for more loci than the extended family sample. The DZ linkage loci of multipoint logarithm of odds score (MLOD) > 1 included 1p32 (MLOD = 1.3), 3q27 (MLOD = 2.5), 3q29 (MLOD = 2.6), 7q36 (MLOD = 2.4), 16p13 (MLOD = 1.3), 18q12 (MLOD = 1.6), 20p12 (MLOD = 1.1) and 22q13 (MLOD = 1.9). The extended families' loci of MLOD > 1 included only 16p13 (MLOD = 1.3) and 20q13 (MLOD = 1.7). Since the individual cohorts have smaller sample sizes, we only report the loci of MLOD > 2. The cohort-specific DZ linkage loci of MLOD > 2 included 2p24 (MLOD = 3.4, Dutch cohort), 6q26 (MLOD = 2.6, Dutch cohort), 7q36 (MLOD = 2.6, Australian cohort), 7q36 (MLOD = 2.4, Danish cohort) and 18q12 (MLOD = 2.2, Swedish cohort). The extended families' linkage loci of MLOD > 2 included 3q26 (MLOD = 2.0, Australian cohort), 10q22 (MLOD = 2.6, Finnish cohort), 16q23 (MLOD = 3.7, Dutch cohort), 17p13 (MLOD = 2.3, Finnish cohort) and 20q13 (MLOD = 3.2, Finnish cohort).

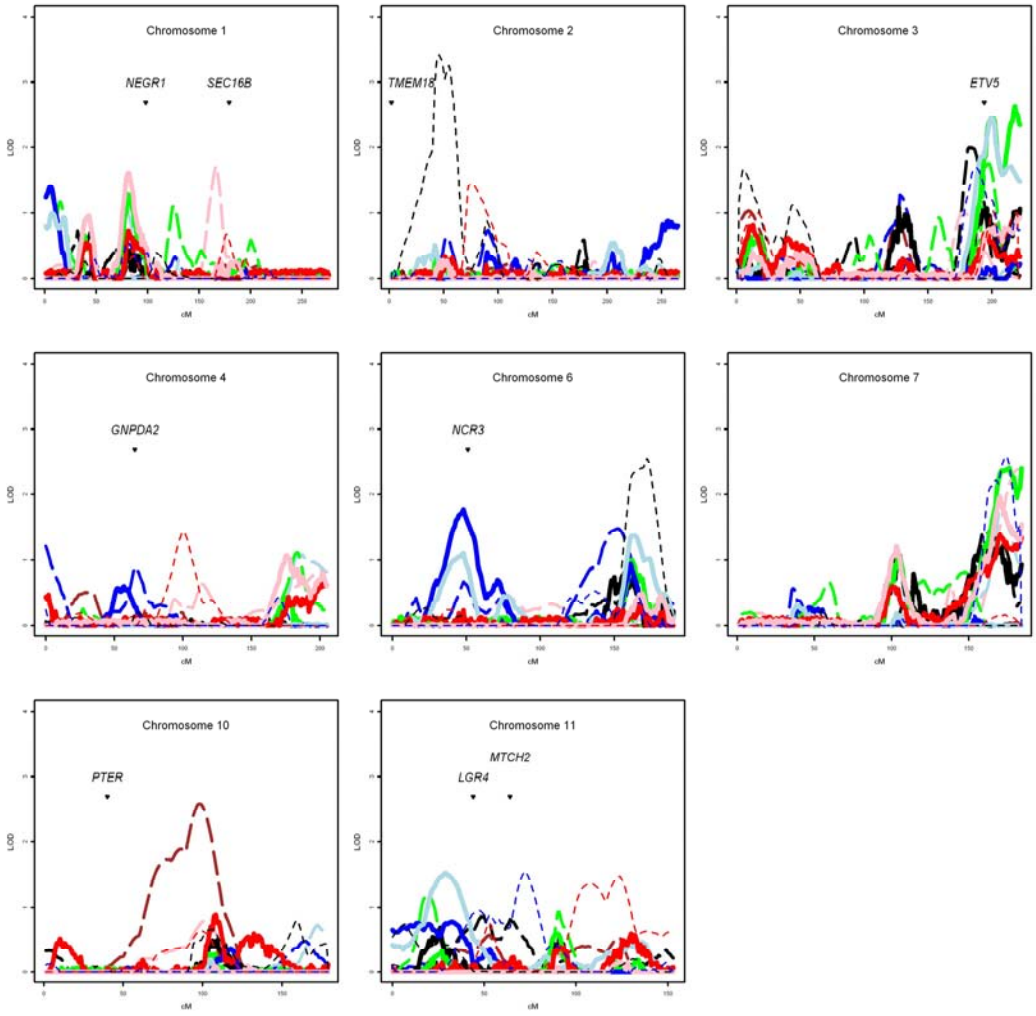


Figure 18. The multipoint variance components linkage results for chromosomes 1 to 11 that have either reported linkage in this study or are published genome-wide association loci

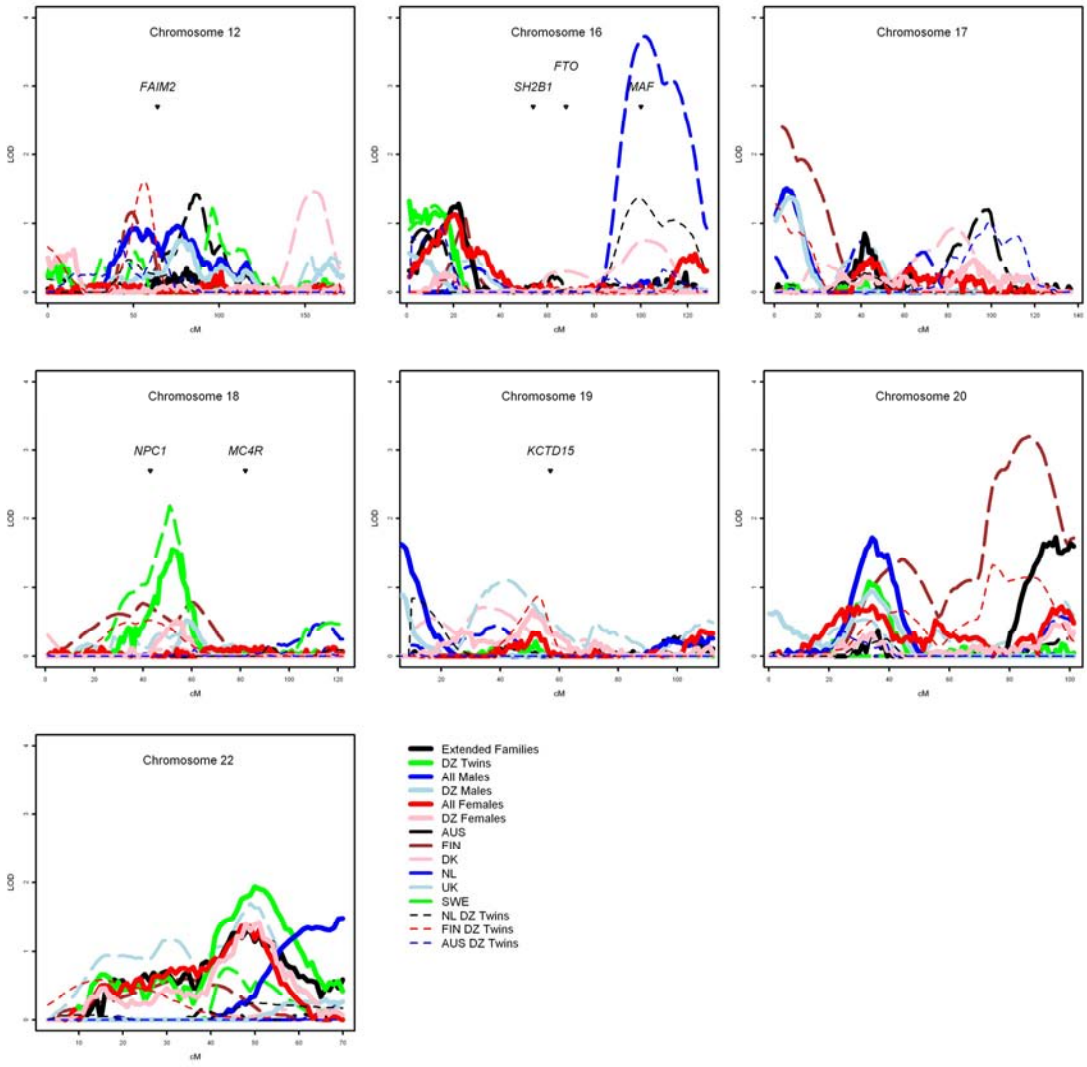


Figure 19. The multipoint variance components linkage results for chromosomes 12 to 22 that have either reported linkage in this study or are published genome-wide association loci

5.2.1 Discussion

Linkage studies have been thus far unable to reliably pinpoint regions in the human genome that contribute to human obesity. The human obesity gene map catalogue contains published linkage and candidate gene association findings in every chromosome, except Y, for obesity-related traits (Rankinen *et al.*, 2006). Perhaps this speaks of the very complex nature of the trait. In study II we provided further evidence for two previously replicated loci in the largest linkage scan to date. We found that the loci 3q29 and 7q36 (MLOD = 2.6 and MLOD = 2.4, respectively) showed suggestive evidence for linkage. These regions have been extensively linked to obesity-related traits. Francke showed evidence for linkage with coronary heart disease and myocardial infarction in the 3q29 region in a North Indian-originating population (MLOD = 2.1, families n = 99, individuals n =535) (Francke *et al.*, 2001). They studied. Kissebah *et al* have linked 3q29 with metabolic syndrome component traits in a Caucasian population (MLOD = 2.4-3.5, families n = 507, individuals n = 2209) (Kissebah *et al.*, 2000). Luke *et al* have found very strong evidence for linkage in an African American population with BMI and in the 3q29 chromosome region (MLOD = 4.3, families n = 329, individuals n = 1163) (Luke *et al.*, 2003). Vionnet *et al* have shown strong binomial likelihood in the 3q29 region in the French population for early onset type 2 diabetes (Binomial likelihood = 4.6, families n = 143, individuals n = 637) (Vionnet *et al.*, 2000). Walder and colleagues has shown linkage with BMI and the 3q29 region in Pima Indians (MLOD = 1.4, families n = 239, individuals n = 770) (Walder *et al.*, 2000). Wu *et al* show significant evidence of linkage for BMI in 3q29 in a large study of mixed European American, African American and Mexican American populations (Wu *et al.*, 2002). The 7q36 region has also been replicated several times. Feitosa *et al* show significant evidence of linkage for BMI in an European American population (MLOD = 4.9, families n = 536, individuals n = 3407) (Feitosa *et al.*, 2002). Hsueh and colleagues have shown evidence of linkage for BMI-adjusted leptin levels in an Amish population (MLOD = 1.8, families n = 28, individuals n = 672) (Hsueh *et al.*, 2001). Pérusse and colleagues have shown evidence for linkage in Caucasion population for abdominal subcutaneous fat (MLOD = 2, families n = 156, individuals n = 521) (Perusse *et al.*, 2001). Sex specific analyses provided evidence that the 7q36 locus may be female driven, which is supported by previous evidence provided by Sammalisto *et al* (MLOD = 2.9, families n = 3032, individuals n = 5788) (Sammalisto *et al.*, 2009). Our study strengthens the evidence that these loci do indeed harbour genetic variants affecting human adult obesity.

DZ twins are of the same age, have shared the same prenatal and family rearing environment and school experiences more closely than full siblings. As BMI

changes with age, matching on age is an important advantage of the twin sample. The DZ twin sample showed higher heritability estimates, which means there is a reduced total variance in the sample when compared to the sample where additional family members were included. This is something that could be tested by using a linear mixed effects model in a large sample of families with twins and siblings if the variance is smaller within twin pairs than twin-sib pairs or sib-sib pairs. On the other hand, this may be due to the problematic modelling of the shared environment in small pedigrees compared to large pedigrees. The shared environment mimics the effect of shared genes in small pedigrees. Similarly, the twins showed stronger linkage and replicated previous findings, which suggests that the sampling strategy was successful. Interestingly, the *FTO* gene region does not show evidence of linkage in our study even though it is the strongest common variant found in GWAS (Frayling *et al.*, 2007). Four of the 13 published GWA loci show evidence for linkage (MLOD > 1) in the combined sample including *NPC1*, *LGR4*, *ETV5* and *NCR3* as shown in Figure 18 and Figure 19. These regions may harbour both common small effect, and rare, strong effect variants. There are very good candidate genes residing under the two strongest linkage peaks. *APOD* gene is right under the 3q29 peak. *APOD* is primarily localized in HDL (~65%) but its function remains largely unclear. There are two excellent candidate genes under the 7q36 region linkage peak, namely *LEP* and *INSIG1*. The leptin gene has been shown to cause severe obesity both in mice and humans [OMIM 164160]. *INSIG1* [OMIM 602055] has been shown to regulate the cholesterol concentration in cells. All three genes may harbor strong effect variants, which are shared in the six European populations and are good candidates for resequencing studies.

In the light of this evidence, the next logical step would be to study these GWAS-replicated loci in more family samples and utilize the linkage method. In this way, one might screen the loci that harbor the high effect variants, and thus identify areas for selective resequencing. For example, the *FTO* locus does not show evidence for linkage suggesting that there are no high effect variants and it may not be a good target for resequencing. On the other hand, *MC4R* has been shown to harbour variants [OMIM 155541] that cause severe obesity and indeed there is a linkage peak 20 cM from the gene that may be caused by non-syndromic but high effect variants. These variants will explain some of the missing heritability, but more importantly they have a large effect on the individuals carrying them. Therefore, it is essential to pinpoint these mutations.

5.3 Integration of three omics in examination of obesity-related components in Finnish population cohort (III)

5.3.1 Network analysis and module association to metabolic phenotypes

The network analysis revealed 21 independent modules. Summary profiles (first principal component) of the module expression were correlated with metabolic traits using Spearman correlation. The strongest associations were identified for module A, the Lipid-Leukocyte (LL) module genes are summarized in Table 18 and significant associations in Table 19.

Table 18. Lipid-Leukocyte -module gene composition

Gene
C1ORF186
CPA3
ENPP3
FCER1A
GATA2
HDC
HS.132563
MS4A2
SLC45A3
SPRYD5
CACNG6

Table 19. Significant Lipid-Lleukocyte module associations with serum metabolites

Metabolite	Effect direction	P-value
Concentration of chylomicrons and extremely large VLDL particles	-	$4.46*10^{-29}$
Triglycerides in medium VLDL	-	$3.46*10^{-26}$
Concentration of large VLDL particles	-	$4.51*10^{-26}$
Free cholesterol in large VLDL	-	$4.89*10^{-26}$
Triglycerides in large VLDL	-	$5.20*10^{-26}$
Triglycerides in VLDL	-	$7.25*10^{-26}$
Concentration of very large VLDL particles	-	$8.88*10^{-26}$
Total cholesterol in large VLDL	-	$1.00*10^{-25}$
Triglycerides in very large VLDL	-	$1.09*10^{-25}$
Phospholipids in large VLDL	-	$1.43*10^{-25}$
Total lipids in large VLDL	-	$1.98*10^{-25}$
Triglycerides in VLDL (Lipido)	-	$3.77*10^{-25}$
Concentration of medium VLDL particles	-	$8.80*10^{-25}$
Cholesterol esters in large VLDL	-	$9.68*10^{-25}$
Total lipids in very large VLDL	-	$1.32*10^{-24}$
Total lipids in medium VLDL	-	$1.59*10^{-24}$
Serum total triglycerides	-	$2.53*10^{-24}$
Phospholipids in medium VLDL	-	$6.17*10^{-24}$
Triglycerides in small VLDL	-	$1.27*10^{-23}$
Total triglycerides	-	$2.87*10^{-23}$
Mobile lipids -CH2-	-	$3.25*10^{-23}$
Free cholesterol in medium VLDL	-	$3.26*10^{-23}$
Phospholipids in very large VLDL	-	$6.36*10^{-23}$
Ratio of triglycerides to phosphoglycerides	-	$2.14*10^{-22}$
Total cholesterol in medium VLDL	-	$7.78*10^{-22}$
Concentration of small VLDL particles	-	$1.18*10^{-20}$
Cholesterol esters in medium VLDL	-	$1.36*10^{-19}$
Phospholipids in chylomicrons and extremely large VLDL	-	$2.56*10^{-19}$
Total lipids in chylomicrons and extremely large VLDL	-	$7.91*10^{-19}$
Triglycerides in small HDL	-	$1.00*10^{-18}$
Total lipids in small VLDL	-	$1.18*10^{-18}$
Triglycerides in chylomicrons and extremely large VLDL	-	$2.03*10^{-18}$

Phospholipids in small VLDL	-	$1.70*10^{-17}$
Mobile lipids -CH3	-	$2.11*10^{-16}$
Isoleucine	-	$4.03*10^{-16}$
Unsaturated lipids	-	$9.41*10^{-16}$
Free cholesterol in small VLDL	-	$1.43*10^{-15}$
Triglycerides in very small VLDL	-	$2.97*10^{-15}$
Omega-9 and saturated fatty acids	-	$1.61*10^{-14}$
Total fatty acids	-	$1.66*10^{-13}$
Total cholesterol in IDL (Lipido)	-	$9.13*10^{-13}$
Apolipoprotein B by apolipoprotein A-I	-	$3.90*10^{-12}$
Total cholesterol in small VLDL	-	$1.18*10^{-10}$
Free cholesterol in large HDL	-	$2.59*10^{-10}$
Omega-6 and -7 fatty acids	-	$5.02*10^{-10}$
Apolipoprotein B	-	$1.17*10^{-9}$
Average number of methylene groups per a double bond	-	$1.24*10^{-9}$
Ratio of bisallylic groups to total fatty acids	+	$1.29*10^{-9}$
Ratio of bisallylic groups to double bonds	+	$4.16*10^{-9}$
Total cholesterol in large HDL	+	$5.15*10^{-9}$
Total lipids in large HDL	+	$1.57*10^{-8}$
Phospholipids in large HDL	+	$1.77*10^{-8}$
Glycoproteins	-	$1.83*10^{-8}$
Cholesterol esters in large HDL	+	$2.26*10^{-8}$
Concentration of large HDL particles	+	$2.87*10^{-8}$
Average number of double bonds in a fatty acid chain	+	$3.39*10^{-8}$
Leucine	-	$5.58*10^{-8}$
Total cholesterol in medium HDL	+	$5.94*10^{-8}$
Ratio of omega-9 and saturated fatty acids to total fatty acids	-	$6.82*10^{-8}$
Phospholipids in very large HDL	+	$4.00*10^{-7}$
Total cholesterol in large HDL	-	$1.30*10^{-6}$
Concentration of very large HDL particles	+	$2.56*10^{-6}$
Triglycerides in IDL	-	$5.24*10^{-6}$
Total lipids in very large HDL	+	$5.49*10^{-6}$
Ratio of omega-6/7 fatty acids to total fatty acids	+	$8.94*10^{-6}$
Total cholesterol in HDL	+	$1.17*10^{-5}$
Free cholesterol in very large HDL	+	$1.96*10^{-5}$

Concentration of very small VLDL particles	-	$3.23*10^{-5}$
Cholesterol esters in very large HDL	+	$5.08*10^{-5}$
3-hydroxybutyrate	+	$9.48*10^{-5}$
Concentration of small LDL particles	-	$1.02*10^{-4}$
Total cholesterol in very large HDL	+	$1.25*10^{-4}$
18:2, linoleic acid	-	$4.48*10^{-4}$
Concentration of small HDL particles	-	$4.48*10^{-4}$
Total phosphoglycerides	-	$5.78*10^{-4}$
Total lipids in small LDL	-	$7.87*10^{-4}$
Other polyunsaturated fatty acids than 18:2	-	$8.30*10^{-4}$
Creatine	+	$9.02*10^{-4}$
Total lipids in very small VLDL	-	$9.13*10^{-4}$
Description of average fatty acid chain length (not carbon number)	+	$1.26*10^{-3}$
Phospholipids in medium LDL	-	$1.31*10^{-3}$
Triglycerides in very large HDL	-	$1.33*10^{-3}$

HDL = high density lipoprotein, LDL = low density lipoprotein, IDL = intermediate density lipoprotein, VLDL = very low density lipoprotein

5.3.2 Genetic factors affecting LL module expression

Only two SNPs (rs12569123 and rs12569261, *SLC45A3*) associated significantly with gene expression in the LL module in *cis*, but they did not affect the LL module expression significantly (Table 20). It was of note that rs2251746, an experimentally-verified eQTL of *FCER1A* and the strongest signal in a recent GWAS for serum immunoglobulin E (IgE) levels, nominally influenced *FCER1A* expression (nominal $P = 1.83*10^{-4}$), but showed strong evidence in of association with total LL module expression ($P = 4.28*10^{-6}$). For *trans* SNPs, only three significant associations and all were observed between *MS4A3* expression and a haploblock on chromosome 6 containing *PNRC1* and *SRrp35*. These SNPs also strongly predicted LL module expression (Table 21).

Table 20. *Cis* expression quantitative trait loci in the Lipid-Leukocyte module

Gene	Chr	SNP	SNP position	P value	β	Permuted significance	LL association P
SLC45A3	1	rs12569123	204794597	$9.44*10^{-5}$	-0.29	0.05	$5.67*10^{-2}$
SLC45A3	1	rs12569261	204793883	$9.70*10^{-5}$	-0.29	0.05	$1.76*10^{-1}$
FCER1A	1	rs2251746	157538684	$1.83*10^{-4}$	-0.17	-	$4.28*10^{-6}$

β = effect estimate, LL = Lipid-Leukocyte module

Table 21. *Trans* expression quantitative trait loci in the Lipid-Leukocyte module

Gene	Probe chr	SNP	SNP chr	SNP position	P value	Rho	LL association P
MS4A3	11	rs10455501	6	89856456	$3.90*10^{-7}$	0.22	$8.77*10^{-3}$
MS4A3	11	rs6938490	6	89862949	$4.21*10^{-7}$	-0.22	$3.89*10^{-3}$
MS4A3	11	rs765798	6	89837202	$4.34*10^{-7}$	0.22	$8.46*10^{-4}$

Rho = Spearman's rank correlation coefficient, LL = Lipid-Leukocyte module

5.3.3 Integrity testing

We tested for the integrity of the network by dividing the sample into quintiles according to the metabolic phenotype of interest. The co-expression of all core LL module gene pairs was measured via Spearman rank correlation. For 63 metabolites there was significant association ($P < 2.4*10^{-3}$) with LL module expression and additionally a significant linear fit ($P < 0.05$). All associations displayed an inverse relationship between the direction of association for the metabolite concentration vs. LL module expression and metabolite quintile vs. co-expression stability exemplified by Figure 20.

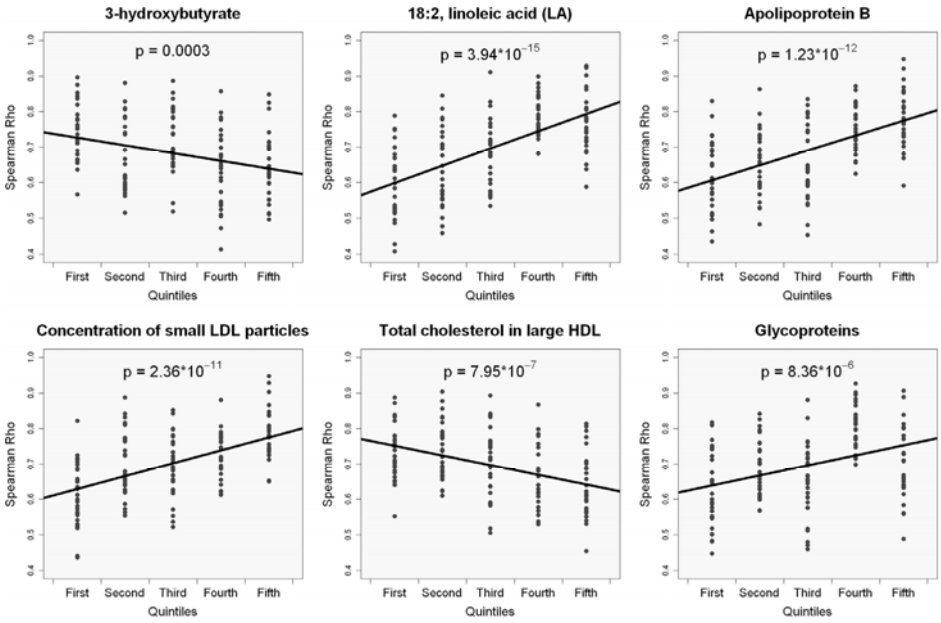


Figure 20. Lipid-Leukocyte module expression (y-axis) partitioned into quintiles (x-axis) based on metabolite concentration rank; a linear model is fitted to test for the expression dependence of the metabolite concentration.

5.3.4 Inferring causality

We used Network Edge Orientation (NEO) conditional correlation analysis to infer a Bayesian network of core LL module gene expression with all metabolites. The metabolites were selected so that they had significant association to the LL module expression as well as a genetic component for the edge orientation (at least one SNP associated with $P<5.0*10^{-7}$). The network consisted of 36 nodes (7 genes from the LL module and 29 metabolites) and 137 causal edges. The Bayesian network inferred a reactive role for the core LL module to fatty acids and all lipoprotein fractions. However, a contradicting trend was observed for free cholesterol and phospholipids in large VLDL, which appeared to be driven by expression of three of the module genes (*CPA3*, *HDC* and *SPRYD5*). We observed an interesting trend in the triglyceride levels of small HDL particles. It appeared to be driven by the concentration of large and medium HDL subclasses. Interestingly, the level of triglycerides present in the smallest *HDL* particles (those which showed the strongest negative relationship with larger *HDL* subclasses) was strongly driven by the concentration of large/medium *HDL*.

5.3.5 Discussion

In study III we presented the first evaluation of metabonomic, transcriptomic and genomic variation in a large population-based cohort. Gene networks were created to portray biological pathways. In networks, pair-wise correlations between genes were used to describe the connectedness of the network, and clusters of tightly correlated genes (modules) were used to characterize the pathways. The genes were chosen in relation to their association with metabolic phenotypes. After network analysis, 21 independent modules were identified. The LL module expression was shown to correlate strongly with several peripheral blood metabolites. The LL module contains genes (*HDC*, *FCER1A*, *GATA2*, *CPA3*, *MS4A2*, *SPRYD5*, and *SLC45A3*) that are key components of inflammation and allergy. The expression of the module was associated with the SNP rs2251746, which is the strongest regulator of serum igE levels. The functions of *SLC45A3*, *GATA2*, and *CPA3* are largely unknown, but *FCER1A* and *HDC* are involved in inflammation and allergy. *HDC* is the catalyst for the conversion of histidine to histamine, a well-known initiator of inflammation and immune response to pathogens that is secreted by basophils and mast cells (BMCs). BMCs have been previously associated, as have lipids and lipoproteins, with atherosclerosis and myocardial infarction (McQueen *et al.*, 2008, Tanimoto *et al.*, 2006, Kaartinen *et al.*, 1994, Kovanen *et al.*, 1995). Previous evidence in mice has shown *HDC*'s involvement in hyperleptinemia, glucose tolerance, body weight and atherosclerosis (Jorgensen *et al.*, 2006, Fulop *et al.*, 2003, Tanimoto *et al.*, 2006). *FCER1A* plays a powerful role in the immune response and in histamine release as the encoded receptor subunit directly interacts with antigen-bound IgE to initiate cross-linking and BMC activation on the cell surface (Kraft & Kinet, 2007). Serum igE levels have been shown to associate with *FCER1A*, but more interestingly co-expression with *GATA2* was shown in the same study (Weidinger *et al.*, 2008). Since there was strong correlation between LL module genes, we hypothesized that *HDC*, *FCER1A*, *SLC45A3*, *GATA2*, and *CPA3* function as part of the same pathway.

One additional interesting observation was the opposite correlations of HDL subclasses both in terms of LL module association and over all. The smallest subclass of HDL behaved in the same way as VLDL subclasses and had a negative correlation to other HDL subclasses. This suggests that the clinical serum HDL measurement is not a coherent physiological function and its physiological meaning may be confounded by a high relative concentration of small HDL particles. The smallest HDL fraction may have its own pro-atherogenic role in the determination of triglyceride levels and the APOB cascade.

We utilized the Bayesian network to evaluate gene/gene or gene/metabolite causality. The expression of the LL module genes seemed largely reactive to metabolites, but free cholesterol and phospholipids in large VLDL seemed to be driven by the LL module expression. Additionally, the triglyceride content in smallest HDL particles was driven by medium/large HDL particles.

We were able to assess the co-expression dynamics of the LL module in different metabolic environments. There were 63 metabolites that showed change in the connectedness of the LL module expression. This raises the hypothesis that the function of the LL module is dampened, for example, in the face of large/medium HDL, 3-hydroxybutyrate stimulated by metabolites like triglycerides, glycoproteins and small HDL.

This is the first study of its kind and we provide both a proof of concept and roadmap for integrated analysis of variation at the genetic, transcriptional, and metabonomic levels. Perhaps the future studies using this approach will catalogue the expressional pathways in other tissues' correlation with various human traits. Adipose tissue would be the next obvious tissue to study due to its impact on lipid levels and energy metabolism. In this study, we elucidated the role of the LL module as a pathway linking metabolic compounds and immune response.

6 CONCLUSIONS

In the course of this thesis science has taken giant leaps forward. The thesis work started when whole genome linkage scans and subsequent candidate gene association tests were the way to do genetics. The array technologies quickly became the trend and they have delivered many success stories in pinpointing loci for complex traits. However, a large part of the genetic variance still remains hidden and more work is needed to unveil the rest of the genetic influence on complex traits. Family studies will most likely come back into fashion after a few years of neglectance. They still offer different information than GWAS by helping isolate high impact, rare variants. The combination of these two will probably reveal the regions suitable for subsequent targeted sequencing efforts. Common obesity will most likely not be cured due to these genetic findings. They provide a means to pinpoint individuals who carry a high load of risk variants and cases in which early intervention in terms of lifestyle factors is crucial. This kind of screening clearly elicits ethical issues, but they are beyond the scope of this thesis. The monogenic forms of obesity, however, are more prone to medical intervention. The problem is that they are so rare that medical companies will not turn a profit by producing the drug. But as the work goes on, this projection may change.

We are in the wake of the 1000 genomes project, which will shed light on the total variation and the amount of rare variants in human genome. The project will finish in the next few years. We will, most probably, soon have thousands of exomes and genomes for analysis. These will offer a glimpse into the future before all studies will have sequence data from all individuals instead of array genotypes.

Transcriptomic profiling has the major benefit of producing more detailed information of the current state of a tissue of an organism. The future will tell whether we will have human transcriptomic profiles of every tissue, from which it is ethically possible to take samples, in individuals at each state of development. Here, we have made the initial effort to show the benefit of this approach.

Sequence data, transcriptomic profiles at different time points, epigenetic patterns and metabolic profiles at different time points along with detailed phenotyping will produce vast amounts of data for investigation of the molecular mechanisms underlying metabolic processes and human complex disease. Hopefully the rate of participation in cohort collections will remain high since without the volunteering individuals, none of this work would be possible.

7 ACKNOWLEDGEMENTS

This study was mainly carried out in the Department of Molecular Medicine, Finland's National Public Health Institute, Helsinki, Finland, later, The National Institute for Health and Welfare, Helsinki, Finland during years 2005-2010. Significant amount was carried out in the Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kindom. I would like to thank the director of National Institute of Health and Welfare, Professor Pekka Puska, director of Institute for Molecular Medicine Finland, Professor Olli Kallioniemi and head of department of Public Health Unit, Adjunct professor Anu Jalanko for providing the excellent research facilities.

I greatly appreciate the financial support to the GenomEUtwin-project, Academy of Finland and Wellcome Trust Sanger Institute.

I would like to thank my supervisors Adjunct professor Markus Perola and late Academician of Science Leena Peltonen-Palotie who guided my baby steps in biomedical research. I would like to thank Leena for showing the drive, inspiration and passion to this work as well as providing the excellent tools to work with. For Markus, the support and guidance along with great humour makes you what you are, don't ever change.

I would like to thank Juni Palmgren for accepting the Opponent's role in my thesis defence. This thesis improved greatly by the excellent comments from the reviewers Harald Göring and Johanna Schleutker.

This thesis would certainly not have been possible without the great collaboration with Aarno Palotie, Veikko Salomaa, Mika Ala-Korpela, Jaakko Kaprio, Olli Raitakari and numerous international groups. I hope the collaboration will continue as lively as it has been during this work.

I thank the senior scientists in our department for the help during the years: Anu Jalanko, Vesa Olkkonen, Matti Jauhiainen, Tiina Paunio, Marjo Kestilä, Ismo Ulmanen, Samuli Ripatti, Janna Saarela, Anu Loukola, Kaisa Silander and Teppo Varilo.

Many thanks to the secretaries and administration Sari Kivikko, Huei-Yi Shen, Liisa Penttilä, Sanna Tossavainen, Mika Kivimäki and Tuija Koski for their assistance during the years. Hannu Turunen, Teemu Perheentupa and Jari Raikko are thanked for help for the IT-support during the years. Warm thanks go to Minna Suvela, Anne Nyberg, Elli Kempas, Liisa Arala, Sisko Lietola, Arja Terola and Anne Vikman for the help in the lab.

I want to express my gratitude to the people in Sanger: Verneri, Anja, Chloe, Stella, Kirsi, Kati, Katta, Eija, Henna, Katja, Marja-Liisa, Guillaume, Pablo, Gudrun, Kay, Yvette, Olli, Pekka, Helena, Zoltan, Carola, Heidi, Jonna, Virpi, Mike and Karola. Without you guys the two years in UK would not have been as fun as they were.

I greatly appreciate the company during the work and travel from Minttu, Marika, Emma P, Annina, Henna K, Kaisu, Antti, Heidi M, Will, Pia, Juho, Marine, Katri, Suvi, Anu K, Jussi, Pööpö, Jarkko, Tea, Hanna, Perttu, Mikko, Taru, Tarja, Emma N, Niina, Liisa T, Tintti, Annina L, Mirkka, Päivi, Krista, Ansku, Nora, Markus L, Tanja, Antti S, Emmi, Tiia, Mari, Annu, Peter, Mikko Taru and Tarja. You guys are woderful and endless source of fun. This list is endless so apologies if I have forgotten someone. Without your company the coffee breaks, social events and conferences would not have been as great as they were!

My warm thanks for the ladies in the old office: Outi, Kati, Mervi, Kirsi A. Thanks for the good atmosphere and great company during the first stages of my research and Elina C who helped me to get started in the first place. Sampo, Tero and Markus, I can't imagine a better team to have worked with. Every day has been a pleasure.

I would like to express my heartfelt thanks to my parents Seppo and Tuula. Without your support and care, the path leading to this thesis would not have been possible. I thank Arttu, Milla and little Eetu for the care and company.

My warmest thanks and my love goes to my lovely wife Sanna and my dear daughter Aada. Sanna, without your help and support none of this would have happened. Sanna and Aada, my family, you are my everything.

Helsinki, August 5th, 2010

Johannes Kettunen

8 REFERENCES

Abecasis, G.R., Cardon, L.R. & Cookson, W.O. (2000) A general test of association for quantitative traits in nuclear families. *American journal of human genetics,* 66**,** 279-92.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. (2001) GRR: graphical representation of relationship errors. *Bioinformatics (Oxford, England),* 17**,** 742-3.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics,* 30**,** 97-101.

Allison, D.B., Kaprio, J., Korkeila, M., Koskenvuo, M., Neale, M.C. & Hayakawa, K. (1996) The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int J Obes Relat Metab Disord,* 20**,** 501-6.

Aten, J.E., Fuller, T.F., Lusis, A.J. & Horvath, S. (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst Biol,* 2**,** 34.

Aulchenko, Y.S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I.M., Pramstaller, P.P., Penninx, B.W., Janssens, A.C., Wilson, J.F., Spector, T., Martin, N.G., Pedersen, N.L., Kyvik, K.O., Kaprio, J., Hofman, A., Freimer, N.B., Jarvelin, M.R., Gyllensten, U., Campbell, H., Rudan, I., Johansson, A., Marroni, F., Hayward, C., Vitart, V., Jonasson, I., Pattaro, C., Wright, A., Hastie, N., Pichler, I., Hicks, A.A., Falchi, M., Willemsen, G., Hottenga, J.J., De Geus, E.J., Montgomery, G.W., Whitfield, J., Magnusson, P., Saharinen, J., Perola, M., Silander, K., Isaacs, A., Sijbrands, E.J., Uitterlinden, A.G., Witteman, J.C., Oostra, B.A., Elliott, P., Ruokonen, A., Sabatti, C., Gieger, C., Meitinger, T., Kronenberg, F., Doring, A., Wichmann, H.E., Smit, J.H., Mccarthy, M.I., Van Duijn, C.M. & Peltonen, L. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics,* 41**,** 47-55.

Bell, C.G., Walley, A.J. & Froguel, P. (2005) The genetics of human obesity. *Nat Rev Genet,* 6**,** 221-34.

Bishop, D.T. & Williamson, J.A. (1990) The power of identity-by-state methods for linkage analysis. *American journal of human genetics,* 46**,** 254-65.

Bochukova, E.G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'rahilly, S., Hurles, M.E. & Farooqi, I.S. (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature,* 463**,** 666-70.

Boomsma, D.I., Vink, J.M., Van Beijsterveldt, T.C., De Geus, E.J., Beem, A.L., Mulder, E.J., Derks, E.M., Riese, H., Willemsen, G.A., Bartels, M., Van Den Berg, M., Kupper, N.H., Polderman, T.J., Posthuma, D., Rietveld, M.J., Stubbe, J.H., Knol, L.I., Stroet, T. & Van Baal, G.C. (2002) Netherlands Twin Register: a focus on longitudinal research. *Twin Res,* 5**,** 401-6.

Bouchard, C., Tremblay, A., Despres, J.P., Nadeau, A., Lupien, P.J., Theriault, G., Dussault, J., Moorjani, S., Pinault, S. & Fournier, G. (1990) The response to long-term overfeeding in identical twins. *N Engl J Med,* 322**,** 1477-82.

Bultman, S.J., Michaud, E.J. & Woychik, R.P. (1992) Molecular characterization of the mouse agouti locus. *Cell,* 71**,** 1195-204.

Burkhardt, R., Kenny, E.E., Lowe, J.K., Birkeland, A., Josowitz, R., Noel, M., Salit, J., Maller, J.B., Pe'er, I., Daly, M.J., Altshuler, D., Stoffel, M., Friedman, J.M. & Breslow, J.L. (2008) Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler Thromb Vasc Biol,* 28**,** 2078-84.

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G. & Hirschhorn, J.N. (2005) Demonstrating stratification in a European American population. *Nature genetics,* 37**,** 868-72.

Challis, B.G., Luan, J., Keogh, J., Wareham, N.J., Farooqi, I.S. & O'rahilly, S. (2004) Genetic variation in the corticotrophin-releasing factor receptors: identification of single-nucleotide polymorphisms and association studies with obesity in UK Caucasians. *Int J Obes Relat Metab Disord,* 28**,** 442-6.

Challis, B.G., Pritchard, L.E., Creemers, J.W., Delplanque, J., Keogh, J.M., Luan, J., Wareham, N.J., Yeo, G.S., Bhattacharyya, S., Froguel, P., White, A., Farooqi, I.S. & O'rahilly, S. (2002) A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. *Hum Mol Genet,* 11**,** 1997-2004.

Chambers, J.C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., Balding, D., Scott, J. & Kooner, J.S. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nature genetics,* 40**,** 716-8.

Chasman, D.I., Pare, G., Mora, S., Hopewell, J.C., Peloso, G., Clarke, R., Cupples, L.A., Hamsten, A., Kathiresan, S., Malarstig, A., Ordovas, J.M., Ripatti, S., Parker, A.N., Miletich, J.P. & Ridker, P.M. (2009) Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS genetics,* 5**,** e1000730.

Chasman, D.I., Pare, G., Zee, R.Y., Parker, A.N., Cook, N.R., Buring, J.E., Kwiatkowski, D.J., Rose, L.M., Smith, J.D., Williams, P.T., Rieder, M.J., Rotter, J.I., Nickerson, D.A., Krauss, R.M., Miletich, J.P. & Ridker, P.M. (2008) Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. *Circ Cardiovasc Genet,* 1**,** 21-30.

Chen, J., Xu, H., Aronow, B.J. & Jegga, A.G. (2007a) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics,* 8**,** 392.

Chen, L.S., Emmert-Streib, F. & Storey, J.D. (2007b) Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol,* 8**,** R219.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., Macneil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., Leonardson, A., Castellini, L.W., Wang, S., Champy, M.F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T.A., Lusis, A.J. & Schadt, E.E. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature,* 452**,** 429-35.

Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., Cha, S.H., Kim, J.W., Han, B.G., Min, H., Ahn, Y., Park, M.S., Han, H.R., Jang, H.Y., Cho, E.Y., Lee, J.E., Cho, N.H., Shin, C., Park, T., Park, J.W., Lee, J.K., Cardon, L., Clarke, G., Mccarthy, M.I., Lee, J.Y., Oh, B. & Kim, H.L. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics,* 41**,** 527-34.

Choi, J.K. & Kim, S.C. (2007) Environmental effects on gene expression phenotype have regional biases in the human genome. *Genetics,* 175**,** 1607-13.

Clayton, D. & Mckeigue, P.M. (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet,* 358**,** 1356-60.

Clement, K., Vaisse, C., Lahlou, N., Cabrol, S., Pelloux, V., Cassuto, D., Gourmelen, M., Dina, C., Chambaz, J., Lacorte, J.M., Basdevant, A., Bougneres, P., Lebouc, Y., Froguel, P. & Guy-Grand, B. (1998) A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature,* 392**,** 398-401.

Collins, F.S., Morgan, M. & Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science (New York, N.Y,* 300**,** 286-90.

Comuzzie, A.G., Blangero, J., Mahaney, M.C., Haffner, S.M., Mitchell, B.D., Stern, M.P. & Maccluer, J.W. (1996) Genetic and environmental correlations among hormone levels and measures of body fat accumulation and topography. *The Journal of clinical endocrinology and metabolism,* 81**,** 597-600.

Comuzzie, A.G., Hixson, J.E., Almasy, L., Mitchell, B.D., Mahaney, M.C., Dyer, T.D., Stern, M.P., Maccluer, J.W. & Blangero, J. (1997) A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. *Nature genetics,* 15**,** 273-6.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. & Hurles, M.E. (2010) Origins and functional impact of copy number variation in the human genome. *Nature,* 464**,** 704-12.

Cooney, C.A., Dave, A.A. & Wolff, G.L. (2002) Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *The Journal of nutrition,* 132**,** 2393S-2400S.

Cotsapas, C., Speliotes, E.K., Hatoum, I.J., Greenawalt, D.M., Dobrin, R., Lum, P.Y., Suver, C., Chudin, E., Kemp, D., Reitman, M., Voight, B.F., Neale, B.M., Schadt, E.E.,

Hirschhorn, J.N., Kaplan, L.M. & Daly, M.J. (2009) Common body mass index-associated variants confer risk of extreme obesity. *Hum Mol Genet,* 18**,** 3502-7.

Dalgliesh, C.E., Horning, E.C., Horning, M.G., Knox, K.L. & Yarger, K. (1966) A gas-liquid-chromatographic procedure for separating a wide range of metabolites occuring in urine or tissue extracts. *Biochem J,* 101**,** 792-810.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J., Deloukas, P., Dermitzakis, E.T. & Antonarakis, S.E. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y,* 325**,** 1246-50.

Duggirala, R., Stern, M.P., Mitchell, B.D., Reinhart, L.J., Shipman, P.A., Uresandi, O.C., Chung, W.K., Leibel, R.L., Hales, C.N., O'connell, P. & Blangero, J. (1996) Quantitative variation in obesity-related traits and insulin precursors linked to the OB gene region on human chromosome 7. *American journal of human genetics,* 59**,** 694-703.

Duggirala, R., Williams, J.T., Williams-Blangero, S. & Blangero, J. (1997) A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet Epidemiol,* 14**,** 987-92.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E. & Stefansson, K. (2008) Genetics of gene expression and its effect on disease. *Nature,* 452**,** 423-8.

Enattah, N., Valimaki, V.V., Valimaki, M.J., Loyttyniemi, E., Sahi, T. & Jarvela, I. (2004) Molecularly defined lactose malabsorption, peak bone mass and bone turnover rate in young finnish men. *Calcified tissue international,* 75**,** 488-93.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. & Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nature genetics,* 30**,** 233-7.

Enzi, G., Gasparo, M., Biondetti, P.R., Fiore, D., Semisa, M. & Zurlo, F. (1986) Subcutaneous and visceral fat distribution according to sex, age, and overweight, evaluated by computed tomography. *Am J Clin Nutr,* 44**,** 739-46.

Faivre, L., Cormier-Daire, V., Lapierre, J.M., Colleaux, L., Jacquemont, S., Genevieve, D., Saunier, P., Munnich, A., Turleau, C., Romana, S., Prieur, M., De Blois, M.C. & Vekemans, M. (2002) Deletion of the SIM1 gene (6q16.2) in a patient with a Prader-Willi-like phenotype. *J Med Genet,* 39**,** 594-6.

Farooqi, I.S., Keogh, J.M., Yeo, G.S., Lank, E.J., Cheetham, T. & O'rahilly, S. (2003) Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N Engl J Med,* 348**,** 1085-95.

Feitosa, M.F., Borecki, I.B., Rich, S.S., Arnett, D.K., Sholinsky, P., Myers, R.H., Leppert, M. & Province, M.A. (2002) Quantitative-trait loci influencing body-mass index reside on chromosomes 7 and 13: the National Heart, Lung, and Blood Institute Family Heart Study. *American journal of human genetics,* 70**,** 72-82.

Fisher, R.A. (1918) The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh,* 52**,** 399-433.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C. & Esteller, M. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 10604-9.

Francke, S., Manraj, M., Lacquemant, C., Lecoeur, C., Lepretre, F., Passa, P., Hebe, A., Corset, L., Yan, S.L., Lahmidi, S., Jankee, S., Gunness, T.K., Ramjuttun, U.S., Balgobin, V., Dina, C. & Froguel, P. (2001) A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. *Hum Mol Genet,* 10**,** 2751-65.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., Shields, B., Harries, L.W., Barrett, J.C., Ellard, S., Groves, C.J., Knight, B., Patch, A.M., Ness, A.R., Ebrahim, S., Lawlor, D.A., Ring, S.M., Ben-Shlomo, Y., Jarvelin, M.R., Sovio, U., Bennett, A.J., Melzer, D., Ferrucci, L., Loos, R.J., Barroso, I., Wareham, N.J., Karpe, F., Owen, K.R., Cardon, L.R., Walker, M., Hitman, G.A., Palmer, C.N., Doney, A.S., Morris, A.D., Smith, G.D., Hattersley, A.T. & Mccarthy, M.I. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.,* 316**,** 889-94.

Freathy, R.M., Mook-Kanamori, D.O., Sovio, U., Prokopenko, I., Timpson, N.J., Berry, D.J., Warrington, N.M., Widen, E., Hottenga, J.J., Kaakinen, M., Lange, L.A., Bradfield, J.P., Kerkhof, M., Marsh, J.A., Magi, R., Chen, C.M., Lyon, H.N., Kirin, M., Adair, L.S., Aulchenko, Y.S., Bennett, A.J., Borja, J.B., Bouatia-Naji, N., Charoen, P., Coin, L.J., Cousminer, D.L., De Geus, E.J., Deloukas, P., Elliott, P., Evans, D.M., Froguel, P., Glaser, B., Groves, C.J., Hartikainen, A.L., Hassanali, N., Hirschhorn, J.N., Hofman, A., Holly, J.M., Hypponen, E., Kanoni, S., Knight, B.A., Laitinen, J., Lindgren, C.M., Mcardle, W.L., O'reilly, P.F., Pennell, C.E., Postma, D.S., Pouta, A., Ramasamy, A., Rayner, N.W., Ring, S.M., Rivadeneira, F., Shields, B.M., Strachan, D.P., Surakka, I., Taanila, A., Tiesler, C., Uitterlinden, A.G., Van Duijn, C.M., Wijga, A.H., Willemsen, G., Zhang, H., Zhao, J., Wilson, J.F., Steegers, E.A., Hattersley, A.T., Eriksson, J.G., Peltonen, L., Mohlke, K.L., Grant, S.F., Hakonarson, H., Koppelman, G.H., Dedoussis, G.V., Heinrich, J., Gillman, M.W., Palmer, L.J., Frayling, T.M., Boomsma, D.I., Davey Smith, G., Power, C., Jaddoe, V.W., Jarvelin, M.R. & Mccarthy, M.I. (2010) Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nature genetics,* 42**,** 430-5.

Fulop, A.K., Foldes, A., Buzas, E., Hegyi, K., Miklos, I.H., Romics, L., Kleiber, M., Nagy, A., Falus, A. & Kovacs, K.J. (2003) Hyperleptinemia, visceral adiposity, and decreased glucose tolerance in mice with a targeted disruption of the histidine decarboxylase gene. *Endocrinology,* 144**,** 4306-14.

Galton, F. (1889) *Natural inheritance.* McMillan.

Galton, F. (1890) Kinship and correlation. *North American Review,* 150**,** 419-431.

Galton, F. (1897) The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society,* 61**,** 401-413.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. & Snyder, M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res,* 17**,** 669-81.

Gibson, W.T., Pissios, P., Trombly, D.J., Luan, J., Keogh, J., Wareham, N.J., Maratos-Flier, E., O'rahilly, S. & Farooqi, I.S. (2004) Melanin-concentrating hormone receptor mutations and human obesity: functional analysis. *Obesity research,* 12**,** 743-9.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabe De Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T. & Suhre, K. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics,* 4**,** e1000282.

Greenland, S. (1990) Randomization, statistics, and causal inference. *Epidemiology,* 1**,** 421-9.

Halpern, J. & Whittemore, A.S. (1999) Multipoint linkage analysis. A cautionary note. *Hum Hered,* 49**,** 194-6.

Hardy, M.A. (1993) *Regression With Dummy Variables (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-093).* Newbury Park, CA: Sage: SAGE Publications.

Haseman, J.K. & Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics,* 2**,** 3-19.

Heard-Costa, N.L., Zillikens, M.C., Monda, K.L., Johansson, A., Harris, T.B., Fu, M., Haritunians, T., Feitosa, M.F., Aspelund, T., Eiriksdottir, G., Garcia, M., Launer, L.J., Smith, A.V., Mitchell, B.D., Mcardle, P.F., Shuldiner, A.R., Bielinski, S.J., Boerwinkle, E., Brancati, F., Demerath, E.W., Pankow, J.S., Arnold, A.M., Chen, Y.D., Glazer, N.L., Mcknight, B., Psaty, B.M., Rotter, J.I., Amin, N., Campbell, H., Gyllensten, U., Pattaro, C., Pramstaller, P.P., Rudan, I., Struchalin, M., Vitart, V., Gao, X., Kraja, A., Province, M.A., Zhang, Q., Atwood, L.D., Dupuis, J., Hirschhorn, J.N., Jaquish, C.E., O'donnell, C.J., Vasan, R.S., White, C.C., Aulchenko, Y.S., Estrada, K., Hofman, A., Rivadeneira, F., Uitterlinden, A.G., Witteman, J.C., Oostra, B.A., Kaplan, R.C., Gudnason, V., O'connell, J.R., Borecki, I.B., Van Duijn, C.M., Cupples, L.A., Fox, C.S. & North, K.E. (2009) NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS genetics,* 5**,** e1000539.

Heid, I.M., Henneman, P., Hicks, A., Coassin, S., Winkler, T., Aulchenko, Y.S., Fuchsberger, C., Song, K., Hivert, M.F., Waterworth, D.M., Timpson, N.J., Richards, J.B., Perry, J.R., Tanaka, T., Amin, N., Kollerits, B., Pichler, I., Oostra, B.A., Thorand, B., Frants, R.R., Illig, T., Dupuis, J., Glaser, B., Spector, T., Guralnik, J., Egan, J.M., Florez, J.C., Evans, D.M., Soranzo, N., Bandinelli, S., Carlson, O.D., Frayling, T.M., Burling, K., Smith, G.D., Mooser, V., Ferrucci, L., Meigs, J.B., Vollenweider, P., Dijk, K.W., Pramstaller, P., Kronenberg, F. & Van Duijn, C.M. (2009) Clear detection of ADIPOQ locus as the major gene for plasma adiponectin: results of genome-wide association analyses including 4659 European individuals. *Atherosclerosis,* 208**,** 412-20.

Herbert, A., Gerry, N.P., Mcqueen, M.B., Heid, I.M., Pfeufer, A., Illig, T., Wichmann, H.E., Meitinger, T., Hunter, D., Hu, F.B., Colditz, G., Hinney, A., Hebebrand, J., Koberwitz, K., Zhu, X., Cooper, R., Ardlie, K., Lyon, H., Hirschhorn, J.N., Laird, N.M., Lenburg, M.E., Lange, C. & Christman, M.F. (2006) A common genetic variant is associated with adult and childhood obesity. *Science (New York, N.Y,* 312**,** 279-83.

Hicks, A.A., Pramstaller, P.P., Johansson, A., Vitart, V., Rudan, I., Ugocsai, P., Aulchenko, Y., Franklin, C.S., Liebisch, G., Erdmann, J., Jonasson, I., Zorkoltseva, I.V., Pattaro, C., Hayward, C., Isaacs, A., Hengstenberg, C., Campbell, S., Gnewuch, C., Janssens, A.C., Kirichenko, A.V., Konig, I.R., Marroni, F., Polasek, O., Demirkan, A., Kolcic, I., Schwienbacher, C., Igl, W., Biloglav, Z., Witteman, J.C., Pichler, I., Zaboli, G., Axenovich, T.I., Peters, A., Schreiber, S., Wichmann, H.E., Schunkert, H., Hastie, N., Oostra, B.A., Wild, S.H., Meitinger, T., Gyllensten, U., Van Duijn, C.M., Wilson, J.F., Wright, A., Schmitz, G. & Campbell, H. (2009) Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS genetics,* 5**,** e1000672.

Hinney, A., Nguyen, T.T., Scherag, A., Friedel, S., Bronner, G., Muller, T.D., Grallert, H., Illig, T., Wichmann, H.E., Rief, W., Schafer, H. & Hebebrand, J. (2007) Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS ONE,* 2**,** e1361.

Hiura, Y., Shen, C.S., Kokubo, Y., Okamura, T., Morisaki, T., Tomoike, H., Yoshida, T., Sakamoto, H., Goto, Y., Nonogi, H. & Iwai, N. (2009) Identification of genetic markers associated with high-density lipoprotein-cholesterol by genome-wide screening in a Japanese population: the Suita study. *Circ J,* 73**,** 1119-26.

Holder, J.L., Jr., Butte, N.F. & Zinn, A.R. (2000) Profound obesity associated with a balanced translocation that disrupts the SIM1 gene. *Hum Mol Genet,* 9**,** 101-8.

Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A.I. & Swallow, D.M. (2001) Lactase haplotype diversity in the Old World. *American journal of human genetics,* 68**,** 160-172.

Hopper, J.L. (2002) The Australian Twin Registry. *Twin Res,* 5**,** 329-36.

Horning, E.C. & Horning, M.G. (1970) Metabolic profiles: chromatographic methods for isolation and characterization of a variety of metabolites in man. *Methods Med Res,* 12**,** 369-71.

Horvath, S. & Dong, J. (2008) Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol,* 4**,** e1000117.

Howard, B.V., Ruotolo, G. & Robbins, D.C. (2003) Obesity and dyslipidemia. *Endocrinol Metab Clin North Am,* 32**,** 855-67.

Hsueh, W.C., Mitchell, B.D., Schneider, J.L., St Jean, P.L., Pollin, T.I., Ehm, M.G., Wagner, M.J., Burns, D.K., Sakul, H., Bell, C.J. & Shuldiner, A.R. (2001) Genome-wide scan of obesity in the Old Order Amish. *The Journal of clinical endocrinology and metabolism,* 86**,** 1199-205.

Huszar, D., Lynch, C.A., Fairchild-Huntress, V., Dunmore, J.H., Fang, Q., Berkemeier, L.R., Gu, W., Kesterson, R.A., Boston, B.A., Cone, R.D., Smith, F.J., Campfield, L.A., Burn, P. & Lee, F. (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell,* 88**,** 131-41.

Igl, W., Johansson, A., Wilson, J.F., Wild, S.H., Polasek, O., Hayward, C., Vitart, V., Hastie, N., Rudan, P., Gnewuch, C., Schmitz, G., Meitinger, T., Pramstaller, P.P., Hicks, A.A., Oostra, B.A., Van Duijn, C.M., Rudan, I., Wright, A., Campbell, H. & Gyllensten, U. (2010) Modeling of environmental effects in genome-wide association studies identifies SLC2A2 and HP as novel loci influencing serum cholesterol levels. *PLoS genetics,* 6**,** e1000798.

Ikonen, E., Baumann, M., Gron, K., Syvanen, A.C., Enomaa, N., Halila, R., Aula, P. & Peltonen, L. (1991) Aspartylglucosaminuria: cDNA encoding human aspartylglucosaminidase and the missense mutation causing the disease. *EMBO J,* 10**,** 51-8.

Illig, T., Gieger, C., Zhai, G., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmuller, G., Kato, B.S., Mewes, H.W., Meitinger, T., De Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D., Adamski, J. & Suhre, K. (2009) A genome-wide perspective of genetic variation in human metabolism. *Nature genetics,* 42**,** 137-41.

Jackson, R.S., Creemers, J.W., Farooqi, I.S., Raffin-Sanson, M.L., Varro, A., Dockray, G.J., Holst, J.J., Brubaker, P.L., Corvol, P., Polonsky, K.S., Ostrega, D., Becker, K.L., Bertagna, X., Hutton, J.C., White, A., Dattani, M.T., Hussain, K., Middleton, S.J., Nicole, T.M., Milla, P.J., Lindley, K.J. & O'rahilly, S. (2003) Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. *The Journal of clinical investigation,* 112**,** 1550-60.

Jackson, R.S., Creemers, J.W., Ohagi, S., Raffin-Sanson, M.L., Sanders, L., Montague, C.T., Hutton, J.C. & O'rahilly, S. (1997) Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nature genetics,* 16**,** 303-6.

Jakkula, E., Rehnstrom, K., Varilo, T., Pietilainen, O.P., Paunio, T., Pedersen, N.L., Defaire, U., Jarvelin, M.R., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M.J., Palotie, A. & Peltonen, L. (2008) The genome-wide patterns of variation expose

significant substructure in a founder population. *American journal of human genetics,* 83**,** 787-94.

Johansson, A., Marroni, F., Hayward, C., Franklin, C.S., Kirichenko, A.V., Jonasson, I., Hicks, A.A., Vitart, V., Isaacs, A., Axenovich, T., Campbell, S., Floyd, J., Hastie, N., Knott, S., Lauc, G., Pichler, I., Rotim, K., Wild, S.H., Zorkoltseva, I.V., Wilson, J.F., Rudan, I., Campbell, H., Pattaro, C., Pramstaller, P., Oostra, B.A., Wright, A.F., Van Duijn, C.M., Aulchenko, Y.S. & Gyllensten, U. (2009) Linkage and genome-wide association analysis of obesity-related phenotypes: association of weight with the MGAT1 gene. *Obesity (Silver Spring, Md,* 18**,** 803-8.

Jorgensen, E.A., Vogelsang, T.W., Knigge, U., Watanabe, T., Warberg, J. & Kjaer, A. (2006) Increased susceptibility to diet-induced obesity in histamine-deficient mice. *Neuroendocrinology,* 83**,** 289-94.

Kaariainen, H., Ryoppy, S. & Norio, R. (1989) RAPADILINO syndrome with radial and patellar aplasia/hypoplasia as main manifestations. *Am J Med Genet,* 33**,** 346-51.

Kaartinen, M., Penttila, A. & Kovanen, P.T. (1994) Accumulation of activated mast cells in the shoulder region of human coronary atheroma, the predilection site of atheromatous rupture. *Circulation,* 90**,** 1669-78.

Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y. & Kamatani, N. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature genetics,* 42**,** 210-5.

Kang, S.J., Chiang, C.W., Palmer, C.D., Tayo, B.O., Lettre, G., Butler, J.L., Hackett, R., Adeyemo, A.A., Guiducci, C., Berzins, I., Nguyen, T.T., Feng, T., Luke, A., Shriner, D., Ardlie, K., Rotimi, C., Wilks, R., Forrester, T., Mckenzie, C.A., Lyon, H.N., Cooper, R.S., Zhu, X. & Hirschhorn, J.N. (2010) Genome-wide association of anthropometric traits in African- and African-derived populations. *Hum Mol Genet,* 19**,** 2725-38.

Kaprio, J. & Koskenvuo, M. (2002) Genetic and environmental factors in complex diseases: the older Finnish Twin Cohort. *Twin Res,* 5**,** 358-65.

Katan, M.B. (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet,* 1**,** 507-8.

Kathiresan, S., Manning, A.K., Demissie, S., D'agostino, R.B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N.P., Melander, O., Orho-Melander, M., Arnett, D.K., Peloso, G.M., Ordovas, J.M. & Cupples, L.A. (2007) A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet,* 8 Suppl 1**,** S17.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E.S., Ordovas, J.M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M.R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D.M. & Orho-Melander, M. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics,* 40**,** 189-97.

Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B.F., Bonnycastle, L.L., Jackson, A.U., Crawford, G., Surti, A., Guiducci, C., Burtt, N.P., Parish, S., Clarke, R., Zelenika, D., Kubalanza, K.A., Morken, M.A., Scott, L.J., Stringham, H.M., Galan, P., Swift, A.J., Kuusisto, J., Bergman, R.N., Sundvall, J., Laakso, M., Ferrucci, L., Scheet, P., Sanna, S., Uda, M., Yang, Q., Lunetta, K.L., Dupuis, J., De Bakker, P.I., O'donnell, C.J., Chambers, J.C., Kooner, J.S., Hercberg, S., Meneton, P., Lakatta, E.G., Scuteri, A., Schlessinger, D., Tuomilehto, J., Collins, F.S., Groop, L., Altshuler, D., Collins, R., Lathrop, G.M., Melander, O., Salomaa, V., Peltonen, L., Orho-Melander, M., Ordovas, J.M., Boehnke, M., Abecasis, G.R., Mohlke, K.L. & Cupples, L.A. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics,* 41**,** 56-65.

Kiel, D.P., Demissie, S., Dupuis, J., Lunetta, K.L., Murabito, J.M. & Karasik, D. (2007) Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med Genet,* 8 Suppl 1**,** S14.

Kissebah, A.H., Sonnenberg, G.E., Myklebust, J., Goldstein, M., Broman, K., James, R.G., Marks, J.A., Krakower, G.R., Jacob, H.J., Weber, J., Martin, L., Blangero, J. & Comuzzie, A.G. (2000) Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 14478-83.

Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gomez Perez, F.J., Frazer, K.A., Elliott, P., Scott, J., Milos, P.M., Cox, D.R. & Thompson, J.F. (2008) Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nature genetics,* 40**,** 149-51.

Kovanen, P.T., Kaartinen, M. & Paavonen, T. (1995) Infiltrates of activated mast cells at the site of coronary atheromatous erosion or rupture in myocardial infarction. *Circulation,* 92**,** 1084-8.

Kraft, S. & Kinet, J.P. (2007) New developments in FcepsilonRI regulation, function and inhibition. *Nat Rev Immunol,* 7**,** 365-78.

Krude, H., Biebermann, H., Luck, W., Horn, R., Brabant, G. & Gruters, A. (1998) Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature genetics,* 19**,** 155-7.

Kulp, D.C. & Jagalur, M. (2006) Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics,* 7**,** 125.

Kyttala, M., Tallila, J., Salonen, R., Kopra, O., Kohlschmidt, N., Paavola-Sakki, P., Peltonen, L. & Kestila, M. (2006) MKS1, encoding a component of the flagellar apparatus basal body proteome, is mutated in Meckel syndrome. *Nature genetics,* 38**,** 155-7.

Lahti-Koski, M., Taskinen, O., Simila, M., Mannisto, S., Laatikainen, T., Knekt, P. & Valsta, L.M. (2008) Mapping geographical variation in obesity in Finland. *European journal of public health,* 18**,** 637-643.

Lander, E.S. & Green, P. (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America,* 84**,** 2363-7.

Langfelder, P. & Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9**,** 559.

Lee, Y.S., Poh, L.K. & Loke, K.Y. (2002) A novel melanocortin 3 receptor gene (MC3R) mutation associated with severe obesity. *The Journal of clinical endocrinology and metabolism,* 87**,** 1423-6.

Lehtimaki, T., Hemminki, J., Rontu, R., Mikkila, V., Rasanen, L., Laaksonen, M., Hutri-Kahonen, N., Kahonen, M., Viikari, J. & Raitakari, O. (2006) The effects of adult-type hypolactasia on body height growth and dietary calcium intake from childhood into young adulthood: a 21-year follow-up study--the Cardiovascular Risk in Young Finns Study. *Pediatrics,* 118**,** 1553-9.

Li, R., Tsaih, S.W., Shockley, K., Stylianou, I.M., Wergedal, J., Paigen, B. & Churchill, G.A. (2006) Structural model analysis of multiple quantitative traits. *PLoS genetics,* 2**,** e114.

Li, Y., Willer, C., Sanna, S. & Abecasis, G. (2009) Genotype imputation. *Annu Rev Genomics Hum Genet,* 10**,** 387-406.

Lindgren, C.M., Heid, I.M., Randall, J.C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E.K., Thorleifsson, G., Willer, C.J., Herrera, B.M., Jackson, A.U., Lim, N., Scheet, P., Soranzo, N., Amin, N., Aulchenko, Y.S., Chambers, J.C., Drong, A., Luan, J., Lyon, H.N., Rivadeneira, F., Sanna, S., Timpson, N.J., Zillikens, M.C., Zhao, J.H., Almgren, P., Bandinelli, S., Bennett, A.J., Bergman, R.N., Bonnycastle, L.L., Bumpstead, S.J., Chanock, S.J., Cherkas, L., Chines, P., Coin, L., Cooper, C., Crawford, G., Doering, A., Dominiczak, A., Doney, A.S., Ebrahim, S., Elliott, P., Erdos, M.R., Estrada, K., Ferrucci, L., Fischer, G., Forouhi, N.G., Gieger, C., Grallert, H., Groves, C.J., Grundy, S., Guiducci, C., Hadley, D., Hamsten, A., Havulinna, A.S., Hofman, A., Holle, R., Holloway, J.W., Illig, T., Isomaa, B., Jacobs, L.C., Jameson, K., Jousilahti, P., Karpe, F., Kuusisto, J., Laitinen, J., Lathrop, G.M., Lawlor, D.A., Mangino, M., Mcardle, W.L., Meitinger, T., Morken, M.A., Morris, A.P., Munroe, P., Narisu, N., Nordstrom, A., Nordstrom, P., Oostra, B.A., Palmer, C.N., Payne, F., Peden, J.F., Prokopenko, I., Renstrom, F., Ruokonen, A., Salomaa, V., Sandhu, M.S., Scott, L.J., Scuteri, A., Silander, K., Song, K., Yuan, X., Stringham, H.M., Swift, A.J., Tuomi, T., Uda, M., Vollenweider, P., Waeber, G., Wallace, C., Walters, G.B., Weedon, M.N., et al. (2009) Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS genetics,* 5**,** e1000508.

Ling, H., Waterworth, D.M., Stirnadel, H.A., Pollin, T.I., Barter, P.J., Kesaniemi, Y.A., Mahley, R.W., Mcpherson, R., Waeber, G., Bersot, T.P., Cohen, J.C., Grundy, S.M., Mooser, V.E. & Mitchell, B.D. (2009) Genome-wide linkage and association analyses to identify genes influencing adiponectin levels: the GEMS Study. *Obesity (Silver Spring, Md,* 17**,** 737-44.

Liu, J.Z., Medland, S.E., Wright, M.J., Henders, A.K., Heath, A.C., Madden, P.A., Duncan, A., Montgomery, G.W., Martin, N.G. & Mcrae, A.F. (2010) Genome-wide association study of height and body mass index in Australian twin families. *Twin Res Hum Genet,* 13**,** 179-93.

Liu, X.G., Tan, L.J., Lei, S.F., Liu, Y.J., Shen, H., Wang, L., Yan, H., Guo, Y.F., Xiong, D.H., Chen, X.D., Pan, F., Yang, T.L., Zhang, Y.P., Guo, Y., Tang, N.L., Zhu, X.Z., Deng, H.Y., Levy, S., Recker, R.R., Papasian, C.J. & Deng, H.W. (2009a) Genome-wide association and replication studies identified TRHR as an important gene for lean body mass. *American journal of human genetics,* 84**,** 418-23.

Liu, Y.J., Liu, X.G., Wang, L., Dina, C., Yan, H., Liu, J.F., Levy, S., Papasian, C.J., Drees, B.M., Hamilton, J.J., Meyre, D., Delplanque, J., Pei, Y.F., Zhang, L., Recker, R.R., Froguel, P. & Deng, H.W. (2008) Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. *Hum Mol Genet,* 17**,** 1803-13.

Liu, Y.Z., Pei, Y.F., Liu, J.F., Yang, F., Guo, Y., Zhang, L., Liu, X.G., Yan, H., Wang, L., Zhang, Y.P., Levy, S., Recker, R.R. & Deng, H.W. (2009b) Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males. *PLoS ONE,* 4**,** e6827.

Loos, R.J., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M., Freathy, R.M., Attwood, A.P., Beckmann, J.S., Berndt, S.I., Jacobs, K.B., Chanock, S.J., Hayes, R.B., Bergmann, S., Bennett, A.J., Bingham, S.A., Bochud, M., Brown, M., Cauchi, S., Connell, J.M., Cooper, C., Smith, G.D., Day, I., Dina, C., De, S., Dermitzakis, E.T., Doney, A.S., Elliott, K.S., Elliott, P., Evans, D.M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C.J., Gwilliam, R., Hadley, D., Hall, A.S., Hattersley, A.T., Hebebrand, J., Heid, I.M., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H.E., Herrera, B., Hinney, A., Hunt, S.E., Jarvelin, M.R., Johnson, T., Jolley, J.D., Karpe, F., Keniry, A., Khaw, K.T., Luben, R.N., Mangino, M., Marchini, J., Mcardle, W.L., Mcginnis, R., Meyre, D., Munroe, P.B., Morris, A.D., Ness, A.R., Neville, M.J., Nica, A.C., Ong, K.K., O'rahilly, S., Owen, K.R., Palmer, C.N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J.C., Rayner, N.W., Ring, S.M., Sandhu, M.S., Scherag, A., Sims, M.A., Song, K., Soranzo, N., Speliotes, E.K., Syddall, H.E., Teichmann, S.A., Timpson, N.J., Tobias, J.H., Uda, M., Vogel, C.I., Wallace, C., Waterworth, D.M., Weedon, M.N., Willer, C.J., Wraight, Yuan, X., Zeggini, E., Hirschhorn, J.N., Strachan, D.P., Ouwehand, W.H., Caulfield, M.J., et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics,* 40**,** 768-75.

Lowe, J.K., Maller, J.B., Pe'er, I., Neale, B.M., Salit, J., Kenny, E.E., Shea, J.L., Burkhardt, R., Smith, J.G., Ji, W., Noel, M., Foo, J.N., Blundell, M.L., Skilling, V., Garcia, L., Sullivan, M.L., Lee, H.E., Labek, A., Ferdowsian, H., Auerbach, S.B., Lifton, R.P., Newton-Cheh, C., Breslow, J.L., Stoffel, M., Daly, M.J., Altshuler, D.M. & Friedman, J.M. (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS genetics,* 5**,** e1000365.

Lubrano-Berthelier, C., Durand, E., Dubern, B., Shapiro, A., Dazin, P., Weill, J., Ferron, C., Froguel, P. & Vaisse, C. (2003) Intracellular retention is a common characteristic of childhood obesity-associated MC4R mutations. *Hum Mol Genet,* 12**,** 145-53.

Luke, A., Wu, X., Zhu, X., Kan, D., Su, Y. & Cooper, R. (2003) Linkage for BMI at 3q27 region confirmed in an African-American population. *Diabetes,* 52**,** 1284-7.

Lusis, A.J. & Pajukanta, P. (2008) A treasure trove for lipoprotein biology. *Nature genetics,* 40**,** 129-30.

Ma, L., Yang, J., Runesha, H.B., Tanaka, T., Ferrucci, L., Bandinelli, S. & Da, Y. (2010) Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. *BMC Med Genet,* 11**,** 55.

Maes, H.H., Neale, M.C. & Eaves, L.J. (1997) Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics,* 27**,** 325-51.

Mathieu, P., Lemieux, I. & Despres, J.P. (2010) Obesity, inflammation, and cardiovascular risk. *Clin Pharmacol Ther,* 87**,** 407-16.

Mcqueen, M.J., Hawken, S., Wang, X., Ounpuu, S., Sniderman, A., Probstfield, J., Steyn, K., Sanderson, J.E., Hasani, M., Volkova, E., Kazmi, K. & Yusuf, S. (2008) Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study. *Lancet,* 372**,** 224-33.

Meyre, D., Delplanque, J., Chevre, J.C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F., Proenca, C., Gaget, S., Korner, A., Kovacs, P., Kiess, W., Tichet, J., Marre, M., Hartikainen, A.L., Horber, F., Potoczna, N., Hercberg, S., Levy-Marchal, C., Pattou, F., Heude, B., Tauber, M., Mccarthy, M.I., Blakemore, A.I., Montpetit, A., Polychronakos, C., Weill, J., Coin, L.J., Asher, J., Elliott, P., Jarvelin, M.R., Visvikis-Siest, S., Balkau, B., Sladek, R., Balding, D., Walley, A., Dina, C. & Froguel, P. (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature genetics,* 41**,** 157-9.

Montague, C.T., Farooqi, I.S., Whitehead, J.P., Soos, M.A., Rau, H., Wareham, N.J., Sewter, C.P., Digby, J.E., Mohammed, S.N., Hurst, J.A., Cheetham, C.H., Earley, A.R., Barnett, A.H., Prins, J.B. & O'rahilly, S. (1997) Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature,* 387**,** 903-8.

Nachman, M.W. & Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics,* 156**,** 297-304.

Naggert, J.K., Fricker, L.D., Varlamov, O., Nishina, P.M., Rouille, Y., Steiner, D.F., Carroll, R.J., Paigen, B.J. & Leiter, E.H. (1995) Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nature genetics,* 10**,** 135-42.

Neel, J.V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics,* 14**,** 353-62.

Norris, J.M., Langefeld, C.D., Talbert, M.E., Wing, M.R., Haritunians, T., Fingerlin, T.E., Hanley, A.J., Ziegler, J.T., Taylor, K.D., Haffner, S.M., Chen, Y.D., Bowden, D.W. & Wagenknecht, L.E. (2009) Genome-wide association study and follow-up analysis of adiposity traits in Hispanic Americans: the IRAS Family Study. *Obesity (Silver Spring, Md,* 17**,** 1932-41.

Novotny, M.V., Soini, H.A. & Mechref, Y. (2008) Biochemical individuality reflected in chromatographic, electrophoretic and mass-spectrometric profiles. *J Chromatogr B Analyt Technol Biomed Life Sci,* 866**,** 26-47.

O'connell, J.R. & Weeks, D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American journal of human genetics,* 63**,** 259-66.

Ollmann, M.M., Wilson, B.D., Yang, Y.K., Kerns, J.A., Chen, Y., Gantz, I. & Barsh, G.S. (1997) Antagonism of central melanocortin receptors in vitro and in vivo by agouti-related protein. *Science (New York, N.Y,* 278**,** 135-8.

Ozata, M., Ozdemir, I.C. & Licinio, J. (1999) Human leptin deficiency caused by a missense mutation: multiple endocrine defects, decreased sympathetic tone, and immune system dysfunction indicate new targets for leptin action, greater central than peripheral resistance to the effects of leptin, and spontaneous correction of leptin-mediated defects. *The Journal of clinical endocrinology and metabolism,* 84**,** 3686-95.

Pearl, J. (1988) *Probabilistic reasoning in Intelligent Systems 2nd edition.* San Fransisco, CA: Morgan Kaufmann Publishers, Inc.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference.* Camridge, UK: Cambridge University Press.

Pedersen, N.L., Lichtenstein, P. & Svedberg, P. (2002) The Swedish Twin Registry in the third millennium. *Twin Res,* 5**,** 427-32.

Penrose (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecied parentage. *Ann. Eugen.,* 8**,** 133-138.

Perusse, L., Rice, T., Chagnon, Y.C., Despres, J.P., Lemieux, S., Roy, S., Lacaille, M., Ho-Kim, M.A., Chagnon, M., Province, M.A., Rao, D.C. & Bouchard, C. (2001) A genome-wide scan for abdominal fat assessed by computed tomography in the Quebec Family Study. *Diabetes,* 50**,** 614-21.

Polasek, O., Marusic, A., Rotim, K., Hayward, C., Vitart, V., Huffman, J., Campbell, S., Jankovic, S., Boban, M., Biloglav, Z., Kolcic, I., Krzelj, V., Terzic, J., Matec, L., Tometic, G., Nonkovic, D., Nincevic, J., Pehlic, M., Zedelj, J., Velagic, V., Juricic, D., Kirac, I., Belak Kovacevic, S., Wright, A.F., Campbell, H. & Rudan, I. (2009) Genome-wide association study of anthropometric traits in Korcula Island, Croatia. *Croat Med J,* 50**,** 7-16.

Pollin, T.I., Damcott, C.M., Shen, H., Ott, S.H., Shelton, J., Horenstein, R.B., Post, W., Mclenithan, J.C., Bielak, L.F., Peyser, P.A., Mitchell, B.D., Miller, M., O'connell, J.R. & Shuldiner, A.R. (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science (New York, N.Y,* 322**,** 1702-5.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics,* 38**,** 904-9.

Price, R.A. & Gottesman, Ii (1991) Body fat in identical twins reared apart: roles for genes and environment. *Behavior genetics,* 21**,** 1-7.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics,* 81**,** 559-75.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Rankinen, T., Zuberi, A., Chagnon, Y.C., Weisnagel, S.J., Argyropoulos, G., Walts, B., Perusse, L. & Bouchard, C. (2006) The human obesity gene map: the 2005 update. *Obesity (Silver Spring, Md,* 14**,** 529-644.

Rasinpera, H., Kuokkanen, M., Kolho, K.L., Lindahl, H., Enattah, N.S., Savilahti, E., Orpana, A. & Jarvela, I. (2005) Transcriptional downregulation of the lactase (LCT) gene during childhood. *Gut,* 54**,** 1660-1.

Rasmussen, F., Johansson, M. & Hansen, H.O. (1999) Trends in overweight and obesity among 18-year-old males in Sweden between 1971 and 1995. *Acta Paediatr,* 88**,** 431-7.

Rice, T., Perusse, L., Bouchard, C. & Rao, D.C. (1999) Familial aggregation of body mass index and subcutaneous fat measures in the longitudinal Quebec family study. *Genet Epidemiol,* 16**,** 316-34.

Richards, J.B., Waterworth, D., O'rahilly, S., Hivert, M.F., Loos, R.J., Perry, J.R., Tanaka, T., Timpson, N.J., Semple, R.K., Soranzo, N., Song, K., Rocha, N., Grundberg, E., Dupuis, J., Florez, J.C., Langenberg, C., Prokopenko, I., Saxena, R., Sladek, R., Aulchenko, Y., Evans, D., Waeber, G., Erdmann, J., Burnett, M.S., Sattar, N., Devaney, J., Willenborg, C., Hingorani, A., Witteman, J.C., Vollenweider, P., Glaser, B., Hengstenberg, C., Ferrucci, L., Melzer, D., Stark, K., Deanfield, J., Winogradow, J., Grassl, M., Hall, A.S., Egan, J.M., Thompson, J.R., Ricketts, S.L., Konig, I.R., Reinhard, W., Grundy, S., Wichmann, H.E., Barter, P., Mahley, R., Kesaniemi, Y.A., Rader, D.J., Reilly, M.P., Epstein, S.E., Stewart, A.F., Van Duijn, C.M., Schunkert, H., Burling, K., Deloukas, P., Pastinen, T., Samani, N.J., Mcpherson, R., Davey Smith, G., Frayling, T.M., Wareham, N.J., Meigs, J.B., Mooser, V. & Spector, T.D. (2009) A genome-wide association study reveals variants in ARL15 that influence adiponectin levels. *PLoS genetics,* 5**,** e1000768.

Ridker, P.M., Pare, G., Parker, A.N., Zee, R.Y., Miletich, J.P. & Chasman, D.I. (2009) Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study. *Circ Cardiovasc Genet,* 2**,** 26-33.

Ripatti, S., Becker, T., Bickeboller, H., Dominicus, A., Fischer, C., Humphreys, K., Jonasdottir, G., Moreau, Y., Olsson, M., Ploner, A., Sheehan, N., Van Steen, K., Baur, M., Van

Duijn, C. & Palmgren, J. (2009) GENESTAT: an information portal for design and analysis of genetic association studies. *Eur J Hum Genet,* 17**,** 533-6.

Rockman, M.V. & Kruglyak, L. (2006) Genetics of global gene expression. *Nat Rev Genet,* 7**,** 862-72.

Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C.G., Zaitlen, N.A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., Mccarthy, M.I., Daly, M.J., Jarvelin, M.R., Freimer, N.B. & Peltonen, L. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics,* 41**,** 35-46.

Sammalisto, S., Hiekkalinna, T., Schwander, K., Kardia, S., Weder, A.B., Rodriguez, B.L., Doria, A., Kelly, J.A., Bruner, G.R., Harley, J.B., Redline, S., Larkin, E.K., Patel, S.R., Ewan, A.J., Weber, J.L., Perola, M. & Peltonen, L. (2009) Genome-wide linkage screen for stature and body mass index in 3.032 families: evidence for sex- and population-specific genetic effects. *Eur J Hum Genet,* 17**,** 258-66.

Sammalisto, S., Hiekkalinna, T., Suviolahti, E., Sood, K., Metzidis, A., Pajukanta, P., Lilja, H.E., Soro-Paavonen, A., Taskinen, M.R., Tuomi, T., Almgren, P., Orho-Melander, M., Groop, L., Peltonen, L. & Perola, M. (2005) A male-specific quantitative trait locus on 1p21 controlling human stature. *J Med Genet,* 42**,** 932-9.

Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., Song, K., Yuan, X., Johnson, T., Ashford, S., Inouye, M., Luben, R., Sims, M., Hadley, D., Mcardle, W., Barter, P., Kesaniemi, Y.A., Mahley, R.W., Mcpherson, R., Grundy, S.M., Bingham, S.A., Khaw, K.T., Loos, R.J., Waeber, G., Barroso, I., Strachan, D.P., Deloukas, P., Vollenweider, P., Wareham, N.J. & Mooser, V. (2008) LDL-cholesterol concentrations: a genome-wide association study. *Lancet,* 371**,** 483-91.

Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., De Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., Groop, L., Altshuler, D., Almgren, P., Florez, J.C., Meyer, J., Ardlie, K., Bengtsson Bostrom, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H.N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E.K., Taskinen, M.R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S.B., Chirn, G.W., Ma, Q., Parikh, H., Richardson, D., Ricke, D. & Purcell, S. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y,* 316**,** 1331-6.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A. & Lusis, A.J. (2005) An

integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics,* 37**,** 710-7.

Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C.I., Muller, T.D., Grallert, H., Wichmann, H.E., Balkau, B., Heude, B., Jarvelin, M.R., Hartikainen, A.L., Levy-Marchal, C., Weill, J., Delplanque, J., Korner, A., Kiess, W., Kovacs, P., Rayner, N.W., Prokopenko, I., Mccarthy, M.I., Schafer, H., Jarick, I., Boeing, H., Fisher, E., Reinehr, T., Heinrich, J., Rzehak, P., Berdel, D., Borte, M., Biebermann, H., Krude, H., Rosskopf, D., Rimmbach, C., Rief, W., Fromme, T., Klingenspor, M., Schurmann, A., Schulz, N., Nothen, M.M., Muhleisen, T.W., Erbel, R., Jockel, K.H., Moebus, S., Boes, T., Illig, T., Froguel, P., Hebebrand, J. & Meyre, D. (2010) Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS genetics,* 6**,** e1000916.

Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orru, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G.B., Fink, A.A., Weder, A.B., Cooper, R.S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E. & Abecasis, G.R. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS genetics,* 3**,** e115.

Service, S., Deyoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A., Heutink, P., Aulchenko, Y., Oostra, B., Van Duijn, C., Jarvelin, M.R., Varilo, T., Peddle, L., Rahman, P., Piras, G., Monne, M., Murray, S., Galver, L., Peltonen, L., Sabatti, C., Collins, A. & Freimer, N. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature genetics,* 38**,** 556-60.

Shipley, B. (2000) *Cause and Correlation in Biology 2nd edition.* Cambridge, UK: Cambridge University Press.

Sieberts, S.K. & Schadt, E.E. (2007) Moving toward a system genetics view of disease. *Mamm Genome,* 18**,** 389-401.

Skytthe, A., Kyvik, K., Holm, N.V., Vaupel, J.W. & Christensen, K. (2002) The Danish Twin Registry: 127 birth cohorts of twins. *Twin Res,* 5**,** 352-7.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. & Froguel, P. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature,* 445**,** 881-5.

Smith, G.D., Lawlor, D.A., Timpson, N.J., Baban, J., Kiessling, M., Day, I.N. & Ebrahim, S. (2008) Lactase persistence-related genetic variant: population substructure and health outcomes. *Eur. J. Hum. Genet.,* 17**,** 357-367.

Sniderman, A.D., Scantlebury, T. & Cianflone, K. (2001) Hypertriglyceridemic hyperapob: the unappreciated atherogenic dyslipoproteinemia in type 2 diabetes mellitus. *Ann Intern Med,* 135**,** 447-59.

Sorensen, T.I., Holst, C. & Stunkard, A.J. (1992a) Childhood body mass index--genetic and familial environmental influences assessed in a longitudinal adoption study. *Int J Obes Relat Metab Disord,* 16**,** 705-14.

Sorensen, T.I., Holst, C., Stunkard, A.J. & Skovgaard, L.T. (1992b) Correlations of body mass index of adult adoptees and their biological and adoptive relatives. *Int J Obes Relat Metab Disord,* 16**,** 227-36.

Spector, T.D. & Macgregor, A.J. (2002) The St. Thomas' UK Adult Twin Registry. *Twin Res,* 5**,** 440-3.

Spss, I. (1999) *SPSS Inc. (1999). SPSS Base 10.0 for Windows User's Guide. SPSS Inc., Chicago IL.*

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S., Deloukas, P. & Dermitzakis, E.T. (2005) Genome-wide associations of gene expression variation in humans. *PLoS genetics,* 1**,** e78.

Strobel, A., Issad, T., Camoin, L., Ozata, M. & Strosberg, A.D. (1998) A leptin missense mutation associated with hypogonadism and morbid obesity. *Nature genetics,* 18**,** 213-5.

Stunkard, A.J., Harris, J.R., Pedersen, N.L. & Mcclearn, G.E. (1990) The body-mass index of twins who have been reared apart. *N Engl J Med,* 322**,** 1483-7.

Stunkard, A.J., Sorensen, T.I., Hanis, C., Teasdale, T.W., Chakraborty, R., Schull, W.J. & Schulsinger, F. (1986) An adoption study of human obesity. *N Engl J Med,* 314**,** 193-8.

Sun, Q., Cornelis, M.C., Kraft, P., Qi, L., Van Dam, R.M., Girman, C.J., Laurie, C.C., Mirel, D.B., Gong, H., Sheu, C.C., Christiani, D.C., Hunter, D.J., Mantzoros, C.S. & Hu, F.B. Genome-wide association study identifies polymorphisms in LEPR as determinants of plasma soluble leptin receptor levels. *Hum Mol Genet,* 19**,** 1846-55.

Tanimoto, A., Sasaguri, Y. & Ohtsu, H. (2006) Histamine network in atherosclerosis. *Trends Cardiovasc Med,* 16**,** 280-4.

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P. & Clark, T.G. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics (Oxford, England),* 23**,** 2741-6.

The International Hapmap Consortium (2003) The International HapMap Project. *Nature,* 426**,** 789-96.

Thomas, D.C. & Conti, D.V. (2004) Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol,* 33**,** 21-5.

Thompson, J.R., Minelli, C., Abrams, K.R., Tobin, M.D. & Riley, R.D. (2005) Meta-analysis of genetic studies using Mendelian randomization--a multivariate approach. *Stat Med,* 24**,** 2241-54.

Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E.J., Olafsdottir, G.H., Jonsson, T., Jonsson, F., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Lauritzen, T., Aben, K.K., Verbeek, A.L., Roeleveld, N., Kampman, E., Yanek, L.R., Becker, L.C., Tryggvadottir, L., Rafnar, T., Becker, D.M., Gulcher, J., Kiemeney, L.A., Pedersen, O., Kong, A., Thorsteinsdottir, U. & Stefansson, K. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature genetics,* 41**,** 18-24.

Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. (1998) A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nature genetics,* 20**,** 113-4.

Walder, K., Hanson, R.L., Kobes, S., Knowler, W.C. & Ravussin, E. (2000) An autosomal genomic scan for loci linked to plasma leptin concentration in Pima Indians. *Int J Obes Relat Metab Disord,* 24**,** 559-65.

Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R.J., Marcano, A.C., Hajat, C., Burton, P., Deloukas, P., Brown, M., Connell, J.M., Dominiczak, A., Lathrop, G.M., Webster, J., Farrall, M., Spector, T., Samani, N.J., Caulfield, M.J. & Munroe, P.B. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American journal of human genetics,* 82**,** 139-49.

Walldius, G., Jungner, I., Holme, I., Aastveit, A.H., Kolar, W. & Steiner, E. (2001) High apolipoprotein B, low apolipoprotein A-I, and improvement in the prediction of fatal myocardial infarction (AMORIS study): a prospective study. *Lancet,* 358**,** 2026-33.

Weidinger, S., Gieger, C., Rodriguez, E., Baurecht, H., Mempel, M., Klopp, N., Gohlke, H., Wagenpfeil, S., Ollert, M., Ring, J., Behrendt, H., Heinrich, J., Novak, N., Bieber, T., Kramer, U., Berdel, D., Von Berg, A., Bauer, C.P., Herbarth, O., Koletzko, S., Prokisch, H., Mehta, D., Meitinger, T., Depner, M., Von Mutius, E., Liang, L., Moffatt, M., Cookson, W., Kabesch, M., Wichmann, H.E. & Illig, T. (2008) Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. *PLoS genetics,* 4**,** e1000166.

Whitney, A.R., Diehn, M., Popper, S.J., Alizadeh, A.A., Boldrick, J.C., Relman, D.A. & Brown, P.O. (2003) Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 1896-901.

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G. & Sibly, R.M. (2003) Likelihood-based estimation of microsatellite mutation rates. *Genetics,* 164**,** 781-7.

Who (2004) Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet,* 363**,** 157-63.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., Strait, J., Duren, W.L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A.J., Morken, M.A., Narisu, N., Bennett, D.,

Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W.M., Li, Y., Scott, L.J., Scheet, P.A., Sundvall, J., Watanabe, R.M., Nagaraja, R., Ebrahim, S., Lawlor, D.A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A.R., Collins, R., Bergman, R.N., Uda, M., Tuomilehto, J., Cao, A., Collins, F.S., Lakatta, E., Lathrop, G.M., Boehnke, M., Schlessinger, D., Mohlke, K.L. & Abecasis, G.R. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics,* 40**,** 161-9.

Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., Lettre, G., Lim, N., Lyon, H.N., Mccarroll, S.A., Papadakis, K., Qi, L., Randall, J.C., Roccasecca, R.M., Sanna, S., Scheet, P., Weedon, M.N., Wheeler, E., Zhao, J.H., Jacobs, L.C., Prokopenko, I., Soranzo, N., Tanaka, T., Timpson, N.J., Almgren, P., Bennett, A., Bergman, R.N., Bingham, S.A., Bonnycastle, L.L., Brown, M., Burtt, N.P., Chines, P., Coin, L., Collins, F.S., Connell, J.M., Cooper, C., Smith, G.D., Dennison, E.M., Deodhar, P., Elliott, P., Erdos, M.R., Estrada, K., Evans, D.M., Gianniny, L., Gieger, C., Gillson, C.J., Guiducci, C., Hackett, R., Hadley, D., Hall, A.S., Havulinna, A.S., Hebebrand, J., Hofman, A., Isomaa, B., Jacobs, K.B., Johnson, T., Jousilahti, P., Jovanovic, Z., Khaw, K.T., Kraft, P., Kuokkanen, M., Kuusisto, J., Laitinen, J., Lakatta, E.G., Luan, J., Luben, R.N., Mangino, M., Mcardle, W.L., Meitinger, T., Mulas, A., Munroe, P.B., Narisu, N., Ness, A.R., Northstone, K., O'rahilly, S., Purmann, C., Rees, M.G., Ridderstrale, M., Ring, S.M., Rivadeneira, F., Ruokonen, A., Sandhu, M.S., Saramies, J., Scott, L.J., Scuteri, A., Silander, K., Sims, M.A., Song, K., Stephens, J., Stevens, S., Stringham, H.M., Tung, Y.C., Valle, T.T., Van Duijn, C.M., Vimaleswaran, K.S., Vollenweider, P., et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics,* 41**,** 25-34.

Vionnet, N., Hani, E.H., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Lepretre, F., Lecoeur, C., Gallina, P., Zekiri, L., Dina, C. & Froguel, P. (2000) Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *American journal of human genetics,* 67**,** 1470-80.

Visscher, P.M., Medland, S.E., Ferreira, M.A., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W. & Martin, N.G. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics,* 2**,** e41.

Vogler, G.P., Sorensen, T.I., Stunkard, A.J., Srinivasan, M.R. & Rao, D.C. (1995) Influences of genes and shared family environment on adult body mass index assessed in an adoption study by a comprehensive path model. *Int J Obes Relat Metab Disord,* 19**,** 40-5.

Wu, X., Cooper, R.S., Borecki, I., Hanis, C., Bray, M., Lewis, C.E., Zhu, X., Kan, D., Luke, A. & Curb, D. (2002) A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *American journal of human genetics,* 70**,** 1247-56.

Yeo, G.S., Farooqi, I.S., Aminian, S., Halsall, D.J., Stanhope, R.G. & O'rahilly, S. (1998) A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature genetics,* 20**,** 111-2.

Zemunik, T., Boban, M., Lauc, G., Jankovic, S., Rotim, K., Vatavuk, Z., Bencic, G., Dogas, Z., Boraska, V., Torlak, V., Susac, J., Zobic, I., Rudan, D., Pulanic, D., Modun, D., Mudnic, I., Gunjaca, G., Budimir, D., Hayward, C., Vitart, V., Wright, A.F., Campbell, H. & Rudan, I. (2009) Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat Med J,* 50**,** 23-33.

Zhu, J., Lum, P.Y., Lamb, J., Guhathakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. & Schadt, E.E. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res,* 105**,** 363-74.

Zhu, J., Wiener, M.C., Zhang, C., Fridman, A., Minch, E., Lum, P.Y., Sachs, J.R. & Schadt, E.E. (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol,* 3**,** e69.