Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
"BIOLOGIA EVOLUZIONISTICA E AMBIENTALE"

CICLO XXVI

COORDINATORE Prof. Guido Barbujani

Dissection of pleiotropic effects in genome-wide association studies of phenotypes related to cardiometabolic health

Settore Scientifico Disciplinare BIO/18

Dottorando
Dott.ssa Marullo Letizia

Tutore
Prof.ssa Chiara Scapoli

Cotutore
Dott.ssa Inga Prokopenko

Anni 2011/2013

1 Index

2 Literature Review

2.1 Introduction

2.1.1 Mendelian and Complex phenotypes, causal variants and Genome Wide Association Studies (GWASs)

One of the most important challenges in human genetics is the identification of polymorphisms and variants in the DNA sequence, related to phenotypic traits and/or lead to an increased risk of developing diseases. In this context, the multifaceted goals of genetics can be summarised as describing, understanding and utilising the relationship between genotypes and phenotypes, or the genotype–phenotype map (GPM)[1].

Generally, human hereditary phenotypes are classified into two primary groups: Mendelian and non-Mendelian or multifactorial complex phenotypes.
Mendelian phenotypes have, as we can derive from the name, a hereditary modality which is ascribable to a Mendelian model. Commonly, they are rare, with a frequency in the population less than 0.05%.
In the case of Mendelian diseases, as for example sickle-cell anaemia or cystic fibrosis, the genetic association is with a single gene, therefore the genotype-phenotype relationship is easily interpretable.
There exist six main different schemes of heredity for Mendelian characters:

- Autosomal dominant,
- Autosomal recessive,
- X-linked dominant,
- X-linked recessive,
- Y-linked,
- Mitochondrial.

Non-Mendelian phenotypes represent a more serious hazard for public health as they can assume a population frequency more than 1%. In fact, the most common human diseases - such as Type 2 Diabetes (T2D), obesity, Cardio-vascular Diseases (CVD) and schizophrenia - reside in this group.
From this, it is easy to deduce that understanding risk factors and etiological processes involved in the development of complex traits and disorders, in particular of common complex human diseases, is one of the biggest challenges of human genetics. A common characteristic of complex phenotypes is that they present an increased familiarity without recognising a clear Mendelian model of inheritance: for example, generally, in the same family several individuals are affected by the same pathology, but this is not attributable to either a dominant model, nor to a recessive one, nor to a sex-linked heredity.
Non-Mendelian complex phenotypes are caused by multiple genetic, but also environmental, factors; for this reason, they are labelled as "multi-factorial" phenotypes. One set of traits that are

particularly difficult to deal with are those that exhibit continuous or metrical variation. For these traits multiple genetic and non-genetic factors contribute to their population-level variability. Therefore, the genetic dissection of complex traits and diseases may require study designs and research protocols that are various and sophisticated[2].

Actually, Mendelian and non-Mendelian characteristics can be imagined as the two extremes of a shade of intermediate situations where we can find, for example, genetic heterogeneity (polymorphisms in different genes can cause similar clinical profile), clinical heterogeneity (the same phenotype, with same genotype, can have different features), incomplete penetrance (when the effect of a variant in the DNA is not always manifested) or oligogenic phenotypes (a handful of genes are involved).

The combination of the effects of genetic and environmental factors which augment and diminish a quantitative phenotype or the risk of developing a disease determines a curve of distribution of the phenotype that has a Gaussian trend (figure 2.1). Central values of the distribution represent the most common values for a quantitative trait or a population risk of developing a disease. The left tail of the distribution represents extreme lowest values for the quantitative trait or a lower risk of developing a disease compared with the normal population or, in other words, a situation of protection from the disease.

On the other hand, if one considers the right tail of the distribution, it contains extreme highest values for a quantitative characteristic or an increased risk of developing the disease, therefore it is possible to define a threshold beyond which the disease occurs.

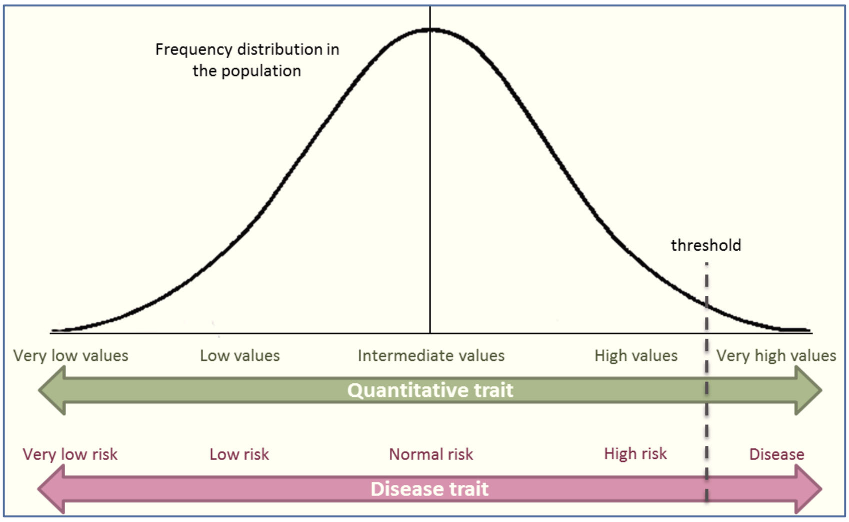


Figure 2.1: *Gaussian distribution of a complex phenotype determined by all influencing (risk and protective) factors. On the green line there are definitions referred to a quantitative phenotype; on the pink line there are definitions based on the evaluation of a disease risk.*

The combinations of factors, genetic and non-genetic, determinant for particular multifactorial phenotypes, can be represented as complex interactive networks, as shown in figure 2.2. It is possible to identify some genes directly connected with the influenced phenotype (B and D and F for Phenotype 1; F and H for Phenotype 2; L for Phenotype 3), as well as gene-gene interactions (A, B, C, D and E, F for Phenotype 1,; E, F, G, H, I for Phenotype 2; I, J, K, L for Phenotype 3). Some environmental factors have a direct influence on the trait (x on Phenotype 1) while others influence phenotypes through gene-environment interactions (z with C and F, y with I). A gene can be involved in more than one phenotype (F for Phenotype 1 and 2), and a phenotype can have an effect on other traits or diseases (Phenotype 2 on Phenotype 3).

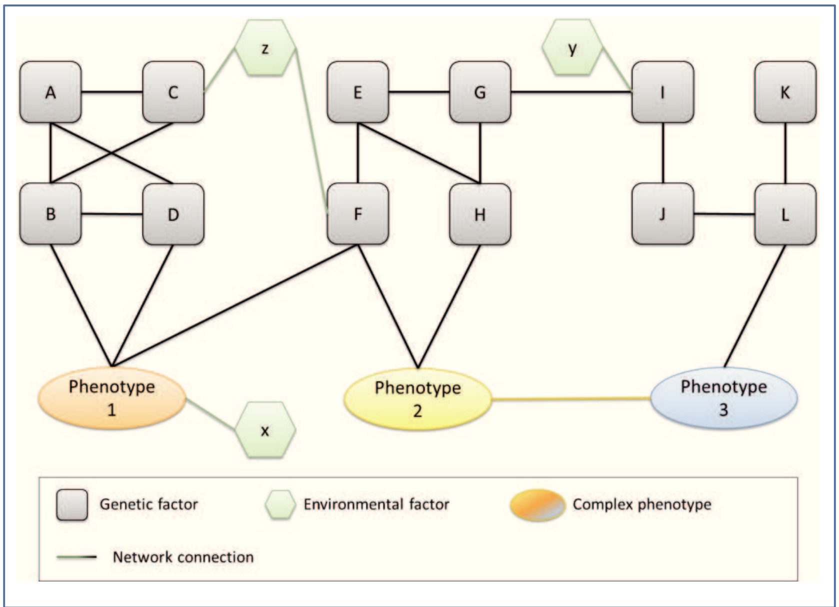


Figure 2.2: *Complex interactive networks of genetic (grey squares) and environmental factors (green hexagons) involved in the determination of three different complex non-Mendelian phenotypes (coloured ovals). For a detailed description, see the paragraph in the main text.*

If we consider the only genetic component of susceptibility for a non-Mendelian disease, it has already a notable complexity by itself. The allele frequency of variants that contribute to cause a common disease and the magnitude of their contribution is subject of debate[3]. In particular, the two main hypotheses proposed in literature are:

- Common Disease/Common Variant (CDCV) hypothesis: on the basis of this theory, the genetic component of a complex disease is constituted by a number of variants with low penetrance, any of which is rather frequent in the control population (minor allele frequency, MAF ≥ 5%). Simultaneous combination of multiple risk alleles at these variants leads to a greater susceptibility to the disease.
- Common Disease/Rare Variant (CDRV) hypothesis: this theory says that a complex disease is genetically determined by several low frequency (MAF < 5%) variants with bigger effects compared to those of common variants.

There is evidence from the literature of studies in favour of both theories; however none of these studies clarified what is the exact allele-frequency spectrum of risk variants involved, the effect size at any disease gene, and hence the total number of risk alleles[3].
Nowadays, different approaches for genetic studies demonstrated that complex diseases cannot be explained by a small number of rare variants with large effects, but neither by a limited number of common variants of moderate effect. Thus, the most accepted hypothesis is a “unifying” one, where variants with all combinations of allele frequency and strength of genetic effect, as represented in figure 2.3, contribute to the genetic susceptibility of a particular phenotype[4].

Defining the genetic architecture of a trait or disease means to define its biological and physiological characterisation of effects of single genes, of functional gene-gene interactions, and of possible influence of environmental factors. An articulate genetic architecture is peculiar for complex common phenotypes and its resolution, reconstructing the molecular and physiological mechanisms involved, has as the final aim the translation of the findings into clinical practice, for achieving better diagnosis and prevention and for the development of more specific treatments. There are two main ways through which such translation might be undertaken: in the first, identification of novel causal pathways might lead to the characterisation of novel therapeutic targets and/or novel therapeutic agents for treatment and prevention. Another positive outcome is the discovery of biomarkers, allowing improved disease prediction and monitoring of disease progression and treatment response[5]. The second translational route considers using the knowledge of individual patterns of disease predisposition (for example, through genetic profiling) to develop more specialised approaches to disease treatment[5].

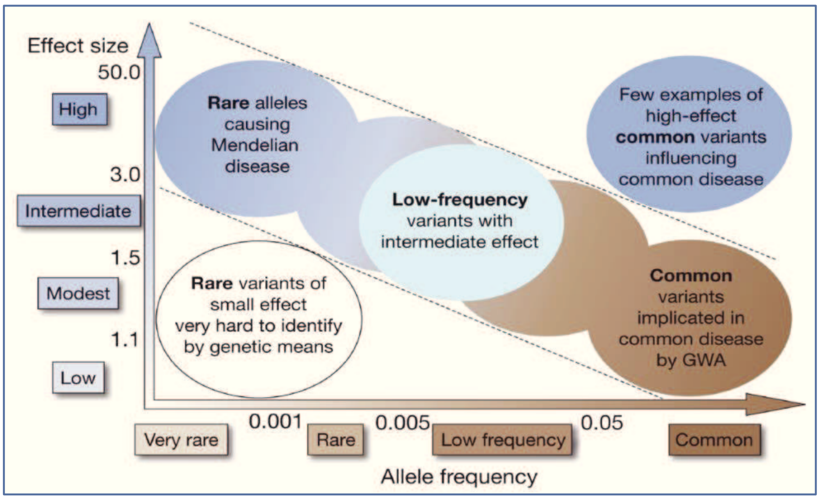


***Figure 2.3:** Possible combinations of frequency and genetic effect for genetic variants of susceptibility. The majority of published genetic studies for human traits aims in identifying associations with the characteristics shown within diagonal dotted lines. From Manolio et al. 2009*[8].

One widely discussed point is the exact definition of "causal variant" for a disease or a trait. Mutations that directly contribute to a particular quantitative trait, or to an increased or decreased risk of developing a disease, are associated with other variants in the genome through linkage disequilibrium (LD): LD is the non-random association between alleles at different positions in the DNA sequence, and is created by evolutionary forces such as mutation, drift, and selection, but it is broken down by recombination[3]. Therefore, it is possible that a genotyped variant, robustly associated with a disease in multiple samples, is not directly causative in risk predisposition, but instead, it is just a mutation lying sufficiently near the causal variant and in LD with it. A ''causal variant'', in fact, is a variant that has a direct functional effect on disease risk, rather than a variant that is associated with disease risk through LD; hence, it is the variant that causes the observed association signal[3]. It is important to keep in mind this concept when researching genetic variants in association with phenotypes, and to remember that, when a polymorphism is detected as significantly associated with a disease, it can be just a "tag" of a causal mutation, and not the causal mutation itself.

Gene mapping through linkage analysis relies on the co-segregation of causal variants with tag polymorphisms (also called "markers") within pedigrees. Because the number of recombination events per meiosis is relatively small, tagging a causal variant requires only a few genetic markers per chromosome[3]. The use of linkage analysis to map genomic loci -specific locations in the DNA of genes, groups of genes, or specific sequences on chromosomes - that have an effect on diseases, or on other traits, have been ubiquitous in the last two decades, and they have been extremely successful for Mendelian phenotypes, but much less so for common diseases and, in particular, in the identification of the underlying causative mutations.

The most widely used method for studying the genetic component of complex traits and diseases is association analysis, which aims to identify genetic variants that are statistically correlated with a phenotype in a population-based sample, without distinguishing between real causal variants and those in LD with the causative ones.

In particular, in the context of association analysis, the genome-wide association study (GWAS) approach has been an important advance compared to "candidate gene" studies, in which sample sizes are generally smaller, and the assayed variants are limited to a selected few, often based on imperfect understanding of biological pathways, and often yielding associations that are difficult to replicate.

Genome-Wide Association Studies (GWASs) are based upon the principle of LD at the population level: thanks to the ability of accurately genotyping hundreds of thousands of single-nucleotide polymorphisms (SNPs) in an automated and affordable manner and to the knowledge of the correlation (LD) structure of those markers in the human genome, these studies enable the analysis of a list of tag SNPs that capture most of the common genomic variation in a number of human populations in association with phenotypes of interest, avoiding the bias of prior biological knowledge (or prior beliefs), and of knowledge of genomic location.

Commercial companies produce dense SNP arrays or "SNP chips" that could genotype many markers in a single assay, capturing most, although not all, common variation in the genome. The technological advances together with bio-banks of either population cohorts or case-control samples, have facilitated the ability to conduct GWASs[3].

The underlying rationale for GWAS is the CDCV hypothesis: in fact SNPs that lie on the majority of SNP chips have been selected to be common (most of them have a minor allele frequency > 5%).

During the past seven years, GWASs have identified more than 8,500 confirmed associations with more than 350 human complex traits and diseases[6]. Published GWASs can be found at the National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI)'s catalog (http://www.genome.gov/gwastudies/, figure 2.4)[7]. These findings have considerably surpassed early expectations, reproducibly identifying hundreds of variants associated with many dozens of traits, and providing valuable insights into the genetic architecture of complex human disease.

Despite the great success of GWASs, we are still far from full comprehension of all the mechanisms behind most common human phenotypes and several challenges underlie limitations of this kind of studies taken alone:

- Follow-up studies are not always able to replicate the discoveries of a previous GWAS.
- For most of the studied phenotypes, discovered variants explain only a fraction of observed familial aggregation.
- The patterns of association observed in GWAS at individual risk-loci are highly variable.
- Allelic heterogeneity is often observed for associations within and between phenotypes.
- GWAS discoveries minimally help in clarification of biological and pathophysiological mechanisms underlying particular phenotypes.

Figure 2.4: *Published Genome-Wide Associations through 12/2012 at p-value of significance ≤ 5x10^{-8} for each chromosome, for 17 trait categories as they are reported in the legend below. From NHGRI GWA Catalog[7]: www.genome.gov/GWAStudies.*

Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait

However, at the present, there is little consensus about the best approaches and priorities for the research of these “dark matter” aspects of GWASs[8].

Manolio and colleagues proposed a list of steps to help GWASs in clarifying different aspects of this “dark matter”[8]:

- Carefully plan the samples to use for the analyses: ensure the ancestry and other possible forms of population structure; choose carefully the individuals for follow-up analyses.
- Increase sample size, for instance thorough meta- and mega-analyses of comparable data: in association studies, in fact, the number of discovered variants is strongly correlated with experimental sample size, and an ever-increasing discovery sample size is expected to increase the number of discovered variants.
- Possibly expand the studies to non-European samples.
- Enhance the investigation of the X and Y chromosome.
- Expand the study to rare variants and copy number variants (CNVs): much of the speculation about missing heritability from GWAS has been attributed to the contribution of variants of low minor allele frequency, defined as roughly 0.5% < MAF <5%, or rare variants (MAF < 0.5%); on the other hand structural variation, including CNVs may contribute to the genetic basis of human traits and disorders[8].
- Investigate gene-gene and gene-environment interactions, including dominance and epistasis, since the detection or characterization of any one of the relevant genetic factors might be obscured or confounded by the influence of others.
- Improve phenotypes by expanding to subtypes, or to more quantitative ones, or to more precise ones.
- Explore multi-phenotype effects: there are thousands of quantitative or qualitative traits in a complex organism, such as the human, but the number of genes is limited (in the human genome it is only around 30,000) and therefore a single gene can simultaneously influence multiple characteristics. Considering this may help in the detection of processes that elucidate part of the missing heritability because some loci may be detectable only by analysing combination of multiple effects on combined phenotypes.

The study of multiple phenotypes simultaneously is becoming more and more relevant: the concept of "omics" is in fact gaining enormous importance. Both, clinical and molecular, phenotypes can be measured and analysed as part of metabolome, transcriptome, proteome, or other groups of "omics" phenotypes (phenome) for a wide spectrum of diseases and quantitative traits. Furthermore, systematic and "phenome-wide" association studies (PheWASs), in which a SNP with an established association with a phenotype is tested for association with hundreds of other phenotypes, have just been published[6]. An example of such an effort is the Population Architecture using Genomics and Epidemiology (PAGE) network, a large-scale collaboration that started in 2011 for harmonizing phenotypes characterisation and for conducting PheWASs on replicated GWAS hits across eight epidemiological studies and five ethnic groups[9]. Other efforts aim to analyse a broad range of phenotypes extracted from electronic medical records.

These new sampling and analysis strategies create a need for appropriate methodology for the identification of associations between genetic markers and combinations of multiple traits and diseases which denote causal relationships between them, and ultimately help in elucidating the underlying biological processes[10].

2.1.2 Cross-Phenotype association and definition of pleiotropy

As cited above, GWASs have identified hundreds of variants associated with a wide variety of complex human phenotypes. Interestingly, many genetic loci appear to harbour variants that are associated with multiple, sometimes seemingly distinct, traits or disorders. We will term such associations as "Cross-Phenotype (CP) associations" as proposed by Nadia Solovieff and colleagues in their review[6], or as "multi-phenotype effects". Evidence of CP associations also comes from less recent discoveries for genetic studies, described below.

The most striking examples of CP effects are in monogenic syndromes. For example, the Pallister–Hall syndrome is caused by a mutation in a single gene encoding the transcription factor Glioma-Associated Oncogene Family Zinc Finger 3 (GLI3), but it manifests with a wide range of symptoms that include extra digits, webbing between digits, shortened limbs, structural abnormalities in the central nervous system, and kidney abnormalities[11]. This is because GLI3 acts as a transcription factor in several organ systems during foetal development.

Twin and family studies have also provided evidence for genetic correlations among diseases[6]. For example a bivariate twin analysis conducted by Kendler and colleagues in 1992 revealed that genetic factors were completely shared between major depression and generalized anxiety disorder[12]. Another example was reported by Criswell et al. on behalf of the multiple autoimmune disease genetics consortium (MADGC) in 2005; by studying 265 families, they discovered that a functional SNP in the intracellular tyrosine phosphatase gene (*PTPN22*) confers risk of four separate autoimmune disorders: type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, and Hashimoto thyroiditis[13].

Association studies and, especially, GWASs have suggested numerous CP effects. For example a SNP on chromosome 8q24 demonstrated association with both prostate[14] and colorectal cancer[15]. Other examples are not only for single SNPs, but also for gene regions; this is the case of the fat mass and obesity associated (*FTO)* locus, where different variants have been associated with body mass index (BMI)[16], melanoma[17], fasting insulin[18] and T2D[19].
A recent evaluation of genome-wide-significant SNPs listed in the National Human Genome Research Institute (NHGRI)'s catalog found that 4.6% of SNPs, and 16.9% of genes known to be associated to physiological or disease traits, have CP effects[20]. These numbers can be underestimated because of many reasons: for example, many human phenotypes have not been extensively studied yet and therefore their associated variants and genes are not known; then, not all genes involved in the determination of studied phenotypes are known; in addition, it can happen that several SNPs distinctly identified as associated with different traits or diseases may underlie a common causal variant that is shared between phenotypes.
GWASs and other genomic analyses have also identified rare structural variants, such as rare CNVs, with CP effects. For example, multiple CNVs across the genome have been demonstrated to be associated with a variety of neurodevelopmental disorders[21].

CP associations highlight that phenotypes share common underlying genetic pathways. However it is important to be cautious with the inference of their causes and to not wrongly label them as "pleiotropic" effects. In fact, we define that a CP association occurs when a genetic locus is associated with more than one trait, regardless of the underlying cause for the observed association. Pleiotropy, instead, underlies a specific mechanism that leads to multiple effects.

There are several potential genetic mechanisms that can explain loci showing overlapping associations for multiple traits, and pleiotropy is just one of the possibilities[6]; we distinguish three main mechanisms of CP effects (figure 2.5):

- **Pleiotropy** occurs when the same genetic causal element affects more than one phenotype. It can appear at the single variant level, where a single causal variant is related to multiple phenotypes (figure 2.5a or 2.5d), or at a locus level, that is when multiple variants in the same gene or locus are associated with different phenotypes by affecting the same functional element (figure 2.5g)[6]. The functional mechanism behind pleiotropy can be related to a gene product that is used by different tissues or cell types, or that is targeted to different signalling receptors. In general, we will refer to pleiotropy as occurring when a genetic variant or a set of variants in LD that constitute a functional unit, are independently associated with more than one phenotype, upon reciprocal conditioning on each phenotype in single-trait (or disease) analyses preserves the association signal at the other. Therefore we can say that multiple associations occur "in parallel"[22].
- **Mediation or mediated pleiotropy** occurs when a genetic variant is directly associated with a phenotype and that phenotype is itself causal for a second phenotype (figure 2.5b) or more phenotypes (figure 2.5e)[6]. In other words, the multi- trait association is "in series". This mechanism includes also cases of pathophysiological change from healthy variation to disease.
- **Multi-phenotype Allelic Heterogeneity** is a phenomenon which involves independent uncorrelated variants at the same locus which cause changes in multiple phenotypes. It can be determined by two causal variants lying in different adjacent genes (figure 2.5c) or around the same gene locus, but affecting the two phenotypes through independent paths underlied by distinct functional elements of the same locus (figure 2.5f).

Other phenomena can bias multi-phenotype analyses, leading to an erroneous identification of CP associations. For example, an ascertainment bias can occur when the recruitment of individuals with one phenotype increases the prevalence of a second, unrelated phenotype in the cohort[6], and this is common in clinical samples, as patients suffering from two conditions are likely to seek treatment more often than those suffering from only one. Since unaffected control individuals are often shared across multiple studies, a biased CP association could also occur if an artefact (such as population stratification or batch effects) is present in the shared controls. Furthermore false CP effects can also be identified when subjects with a particular phenotype are systematically misclassified with a different one, as occurs for some behavioural disorders: for example patients with schizophrenia are sometimes misdiagnosed as affected by bipolar disorder and vice versa[6].

The interpretation of CP effects is not simple, but understanding the real mechanisms behind a CP effect is important, since the identification and characterisation of real pleiotropic mechanisms is crucial for a comprehensive biological understanding of complex traits and disease states, enabling better reconstruction of GPM[6].

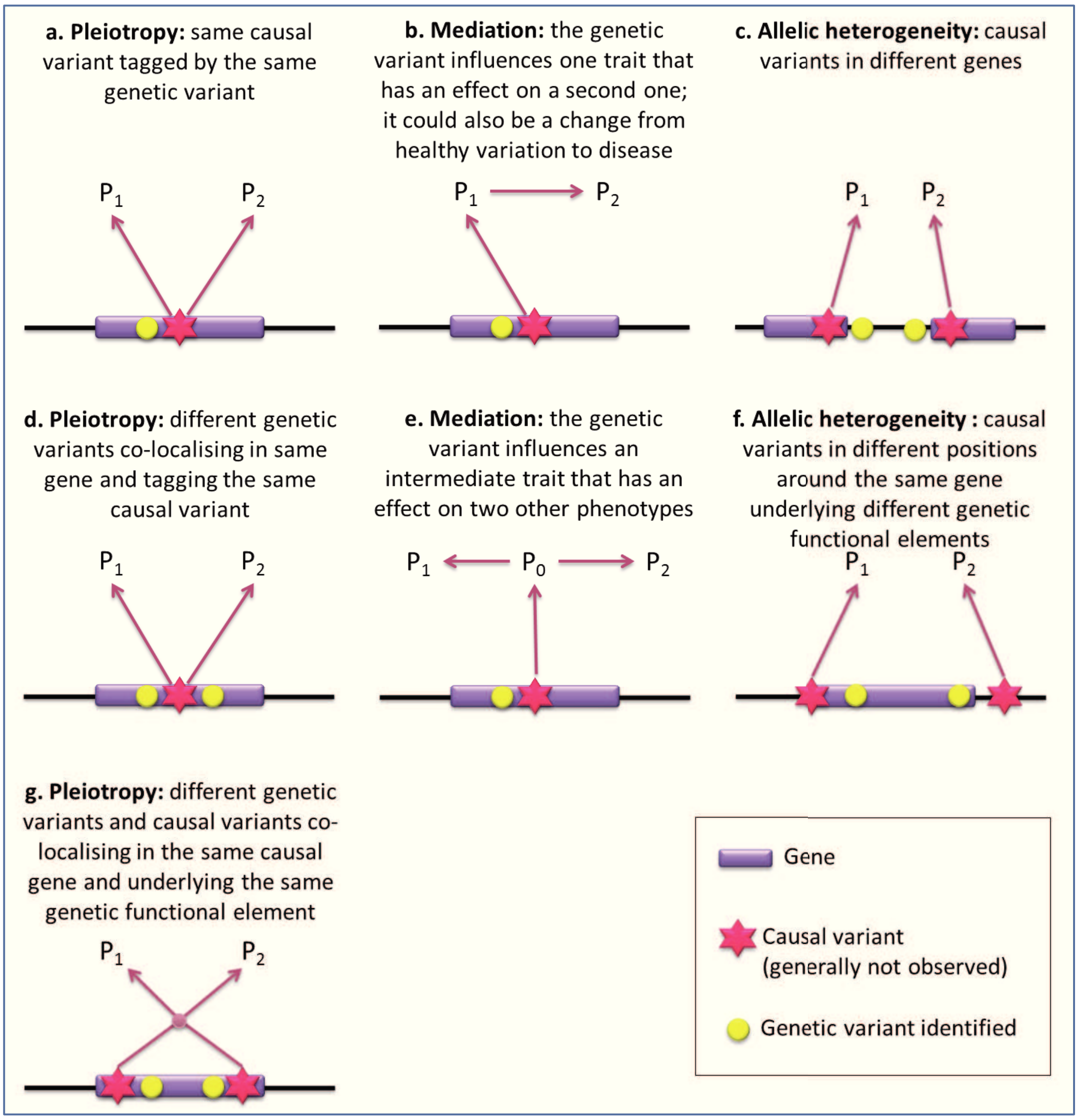


Figure 2.5: *Different mechanisms which determine overlapping associations for multiple traits. a., d. and g. represent real pleiotropic effects; b. and e. represent mediated effects; finally, c. and f. represent multi-trait allelic heterogeneity.*

The impact of genetic studies of pleiotropy for common complex human diseases has been widely recognised and described. However, until now, it has not received sufficient attention, and few multi-phenotype analyses of empirical datasets have been undertaken. Recently, the idea of extending observations of CP effects, by considering a wider range of phenotypes (as described in the chapter above), is emerging. These multi-phenotype analysis approaches will improve our understanding of the extent of shared genetics between traits and diseases, and our global understanding of phenotypes as a range of inter-related manifestations of biological mechanisms, and not as isolated events[6,20].

An understanding of pleiotropic effects is of key importance for drug development too: for example, statins inhibit 3-Hydroxy-3-Methylglutaryl-CoA (HMG-CoA) reductase, but they also have multiple other molecular actions with effects beyond cholesterol reduction, and this has been proposed as the cause for their efficacy in the reduction of cardiovascular outcomes. If a gene has opposing effects on different common diseases, this is likely to greatly complicate drug development and marketing. However, at the same time, knowledge of pleiotropic associations could help to improve drug efficacy and predict side effects.

Furthermore, gaining insight into the level of genetic connectivity between different phenotypes will provide an opportunity to rethink current classification/categorisation of diseases by considering distinctions based on different genetic determinants or whether genetic similarities traverse current divisions[20].

These issues are likely to gain in importance as the full extent of pleiotropy in the genome becomes clear.

2.1.3 History of Pleiotropy definition

Pleiotropy is a concept that has evolved over time, also following the advent of ever more modern techniques for the study of DNA sequences, and pathological and physiological molecular mechanisms; in this section we will retrace the chronological history of definitions and concepts related to pleiotropy (for a synthesis see figure 2.6), placing them in the context of historical discoveries in genetics and molecular biology. In the following section (2.1.4_Insights into the definition of pleiotropy) we will deepen the concepts cited in this chapter.

The first time that the term "pleiotropy" was used in a published manuscript was in 1910, when the German geneticist Ludwig Plate used it to indicate some distinct phenotypes that were explicable only through the mechanism of a single gene:

> *"I call a unit of inheritance pleiotropic if several characteristics are dependent upon it; these characteristics will then always appear together and may thus appear correlated".*
> *"The more research into Mendelian factors advances, the more examples become known which can be explained only under the assumption of pleiotropy"*[23].

In 1925, Haecker described the same mechanism under the name "polyphean", but "pleiotropy" had received more attention and became established in the literature[24,25].

Fisher, in 1930, proposed the idea of "universal pleiotropy", which asserts that a mutation at any locus has the potential to affect almost all traits[26]. This idea was then reclaimed and upgraded by Wright and Mayr[27] in 1963.

Gruneberg published an article in 1938 about the study of rat developmental genetics and, in particular, about skeletal abnormality. From his experiments, he firstly deduced a theory on the mechanism of pleiotropy: he designed the division of pleiotropy into "genuine" and "spurious". Genuine pleiotropy was characterised by two distinct primary products, each arising from a single locus. Spurious pleiotropy indicated, instead, two possible mechanisms: a single primary product that was utilised in different ways, or a primary product that initiated a cascade of events with different consequences for the phenotype[25]. This was the first definition of what we call today "type I" and "type II" pleiotropy[1]; but in 1941, Beadle and Tatum proposed their idea of "one gene/one enzyme", that is a single gene codes for a single protein[28], leaving no room for mechanisms of genuine pleiotropy[25].

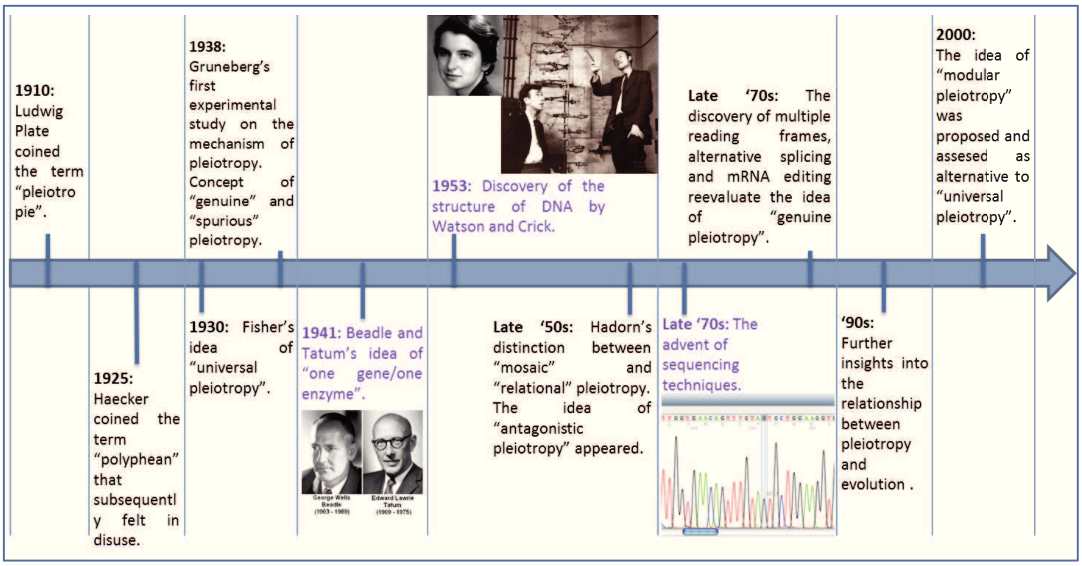


Figure 2.6: *Historical salient steps which contributed to the study and to the modern concept of pleiotropy.*

In the late '50s, after the discovery of the structure of DNA by Watson and Crick, other classifications of pleiotropy were defined based on insights in the ways a single gene product could have multiple uses. Richard Ernst Hadorn made a useful distinction between two types of pleiotropy that were defined as "mosaic" and "relational": mosaic pleiotropy denotes instances where a single locus directly affects two phenotypes; relational pleiotropy describes the action of a single locus that initiates a cascade of events impacting multiple independent phenotypes[29]. These two definitions better describe the two possible mechanisms of spurious pleiotropy hypothesised by Gruneberg.
Additionally, it was in those years that the idea of "antagonistic pleiotropy" started to be viewed as a well-known application of pleiotropy in evolution and medicine. In particular, Williams suggested that genes with antagonistic effects at different life stages could contribute to aging in a way that natural selection could not alter: genes that are beneficial prior to reproduction, but negative after

reproduction, would be favoured by natural selection over those that increase longevity, but which are less favourable to reproduction and survival to reproductive age[30]. This concept will be explained better below.

The advent of sequencing techniques in the late '70s demonstrated that a single locus can produce different primary products at all levels of gene expression and protein processing, for example due to multiple or overlapping reading frames (a strand could be read starting at different points producing different mRNAs and, thus, different proteins from the same single locus)[31], or due to alternative splicing and alternate start/stop codons[32], or to mRNA editing in different tissues and with differential expression[33]. These discoveries gave plausibility back to Gruneberg's idea of "genuine pleiotropy" (or type I) as a possible molecular pleiotropic mechanism.

After the stabilisation of the "antagonistic pleiotropy" concept, the relationship between pleiotropy and evolution was further explored in the '90s. In particular, Waxman and Peck (1998) proposed a theory about the maintenance of pleiotropy, which asserts that pleiotropic traits under stabilising selection are more likely to reach an optimum genetic sequence. This suggests that pleiotropic phenotypes are more likely to be favoured by natural selection[34,35].

In 2000, departing from Fisher's concept of universal pleiotropy, Orr elaborated the "cost of complexity" theory[36], but also the contrasting view about the extent of pleiotropy, and its consequent implications in evolution, has emerged more recently. Following on from Welch and Waxman's idea, organisms can be broken up into modules, and pleiotropy is restricted to the action within these modules[37]. Several recent studies have tried to asses if pleiotropy is more universal or more modular, and their conclusion is that modular pleiotropy is more likely to represent reality[38-42].

2.1.4 Insights into the definition of pleiotropy

2.1.4.1 <u>Other types of pleiotropy</u>

The above mentioned definitions of possible mechanisms of multi-phenotype effects which explain CP associations are not the only ones proposed. Several researchers or research groups have tried to order and define multi-phenotype genetic effects[1,25,35].

Hodgking for example, defined seven different types of "pleiotropic effects"[35]:

- Artefactual pleiotropy: when adjacent but functionally unrelated genes are affected by the same mutation;
- Secondary pleiotropy: when a simple primary biochemical disorder leads to a complex final phenotype (similar to "mediation");
- Adoptive pleiotropy: one gene product is used for quite different chemical purposes in different tissues;
- Parsimonious pleiotropy: one gene product is used for identical chemical purposes in

multiple pathways;

- Opportunistic pleiotropy: arises when one gene product plays a secondary role in addition to its main function;
- Combinatorial pleiotropy: when one gene product is employed in various ways, and with distinct properties, depending on its different protein partners;
- Unifying pleiotropy: one gene, or cluster of adjacent genes, encodes multiple chemical activities that support a common biological function[35].

This classification is rather complicated, and Hodgking's definitions are not always easily discernible from each other.

From the point of view of the molecular basis of a pleiotropic phenomenon, Hans Gruneberg had already, in 1938, distinguished two main mechanisms of pleiotropy: "genuine" and "spurious" pleiotropy (already defined above)[25].

Wagner and Zhang reconsidered Gruneberg's definitions by defining "type I" and "type II" pleiotropy. Type I pleiotropy occurs when a gene product has multiple molecular functions; an example is the human serum albumin that maintains osmotic pressure in body tissues, but is also a plasma carrier for hydrophobic steroid hormones, a transport protein for haemin and fatty acids, and participates to the oxidation of nitric oxide[1,43]. Type II pleiotropy is, instead, characterised by a singular molecular function with multiple consequences, for example glutamine amidotransferase in yeast, which acts through its function of removal of the ammonia group from a glutamine molecule in both histidine biosynthesis and purine nucleotide monophosphate biosynthesis[1,43].

From a study about the relationship between yeast gene pleiotropy and gene function, He and Zhang discovered that, at a genome-wide level, gene pleiotropy is generally represented by a singular molecular function in multiple biological processes, since part of gene products is distributed into multiple cellular components or contributes to multiple protein-protein interactions. This discovery has not to be taken as a rule because it was found only in yeast: in fact, yeast genes do not undergo alternative splicing, and therefore we do not know if this mechanism can importantly contribute to pleiotropy in species with prominent alternative splicing; similarly we cannot estimate the contribution of pleiotropy that arises from gene expression in multiple tissues of a multicellular organism. Anyway, it is important to highlight that this study found no correlation between pleiotropy and the number of different molecular functions[43].

2.1.4.2 The extent of pleiotropy and its relationship with evolutionary processes

Another important point of discussion is the extent of pleiotropy in relation to different phenotypic characters: due to its importance in biology, several mathematical models of pleiotropy have been developed, and important theoretical results have been derived from the analyses of these models.

The oldest hypothesis about the extent of pleiotropy is the "universal pleiotropy theory", proposed by Fisher as part of his geometric model: every mutation affects every trait, and the effect size of mutations on a trait is uniformly distributed[26].

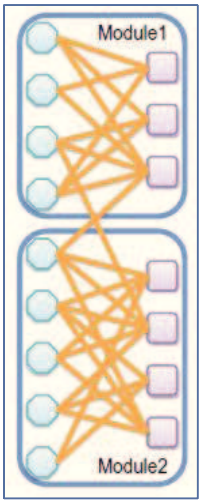


Figure 2.7: *Visual explanation of "modular pleiotropy" theory from Wang et al. 2010[42]. Genes are blue hexagons on the left; phenotypes are violet squares on the right.*

The main alternative hypothesis is that of "modular pleiotropy", which is equally important because of a number of theories about development and evolution derived from it[37]: gene–phenotype relationships can be represented by a bipartite network of genes and traits. where a link between gene nodes and phenotype nodes indicates that the gene affects the phenotype; modular pleiotropy is based on the definition of modules, which include limited number of genes and phenotypes, and refers to the phenomenon where links within a particular module are significantly more frequent than those across modules (figure 2.7)[42].
Another proposed thesis is that of "rare pleiotropy" - where pleiotropic effects are attributable only to a few genes, and affect a very limited number of traits or disorders- but it has found little support in the literature.

On the basis of Fisher's geometric model (FGM), and the assumption that the total effect size of a mutation is constant in different organisms, Orr derived the so-called "cost of complexity" hypothesis: if the "universal pleiotropy" theory is true, the more traits that are observed in an organism (more complexity), the more of its genes are pleiotropic (as every gene affects all traits); complex organisms then are inherently less evolvable or adaptable to changing environments than simple organisms, because their mutations are more likely to be subject to the action of purifying selection[36].
In other words, both the fixation probability of a beneficial mutation, and the fitness gain that is conferred by the fixation of the beneficial mutation, decrease with organismal complexity because there are more possibilities that that the beneficial mutation for a particular phenotype is deleterious in its effect on another phenotype[1,42].

> *"The conformity of these statistical requirements with common experience will be perceived by comparison with the mechanical adaptation of an instrument, such as the microscope, when adjusted for distinct vision. If we imagine a derangement of the system by moving a little each of the lenses, either longitudinally or transversely, or by twisting through an angle, by altering the refractive index and transparency of the different components, or the curvature, or the polish of the interfaces, it is sufficiently obvious that any large derangement will have a very small probability of improving the adjustment..."*[26]

The modularity reduces the probability that a random mutation is deleterious, because that mutation will affect just a set of related traits, rather than all traits. Moreover, Wang et al. found a greater per-trait effect size for pleiotropic mutations in more complex organisms with consequent greater probability of fixation, and a larger amount of fitness gain when a beneficial mutation occurs; through this mechanism, pleiotropy may promote the evolution of complexity. Together, these two reflections lead to the conclusion that organisms of intermediate levels of effective complexity have greater adaptation rates than organisms of lower levels, and explain why complex organisms could have evolved, despite the cost of complexity[42].

From the literature, we can say that the model of universal pleiotropy is not empirically supported.
For example, in 2008, Quantitative Trait Loci (QTLs) underlying a set of traits that represented all major subsystems of the bony skeleton were mapped in inbred mice with increased or reduced body size and, on a total of 102 QTLs identified for 70 traits, the median degree of pleiotropy was only six traits, or 8.6% of the traits examined[38].
Similar work on 54 body-shape traits in sticklebacks identified approximately an average number of 3.5 traits affected by single QTL[39].
Li and colleagues, in 2006, analysed the protein interaction networks of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*, addressing several aspects of network properties. They determined that each gene in the three genomes affects, on average, four or five proteins[41].
In 2010, Su, Zheng and Gu were able to estimate the number of traits affected by each gene in their sample of 321 genes from eight vertebrate species by using comparative data from protein sequence and microarray analysis, in conjunction with mathematical modelling: they found that the number of traits affected per gene was about six to seven[40].
Additionally, in a genome-wide analysis of pleiotropy in yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), and mouse (*Mus musculus*), Wang and colleagues robustly revealed a generally low level of pleiotropy for most genes, and a pleiotropic structure that is highly modular, with an average of 4.6 trait associations per gene, and larger per trait phenotypic effects of those genes affecting more traits[42].
Therefore, the conclusion from current studies is that pleiotropic effects per gene involve a limited number of phenotypes. Consequently, previous estimates from evolutionary theory of the cost of complexity are flawed, since their basic assumptions are not empirically supported[1].

It is largely thought that pleiotropy causes compromises among adaptations of different phenotypes, on the basis that a genetic change beneficial to one phenotype may also be deleterious to another. This property should underlie many fundamental principles and phenomena in biology, including senescence, trade-off, and cooperation[43].
The most popular form to express this idea is the antagonistic pleiotropy theory of senescence: it asserts that mutant genes, advantageous to development and reproduction, are deleterious after the reproductive age and cause senescence; this may explain why all species have a limited life span (Williams 1957). An example that supports this theory is represented by an experiment conducted on social amoeba *Dictyostelium discoideum*: this organism can aggregate during starvation where some cells die to form a stalk that holds the other cells aloft as reproductive spores; deleting the gene *dimA* in *D. discoideum* allows cells to avoid death, but leads to a great reduction in spore production and, therefore, in reproduction[39]. Hence, *dimA* has a pleiotropic effect that stabilises the cooperation among amoeba[43].

2.1.4.3 Features of pleiotropic genes

We have already cited a review by Sivakumaran and colleagues who found that pleiotropy is a property of only 17% of genes and 5% of SNPs that are known to be associated with diseases or disease-related traits in humans, and that these are likely to be lower-bound estimates[20] .

It has also been demonstrated that pleiotropic genes are longer than non-pleiotropic ones: an effect that might be caused by: firstly, the fact that longer genes might encode an increased number of protein structural domains which might give rise to multiple functions; and secondly, longer genes usually contain more variants with a concomitant rise in the opportunity for some to be involved in different functions[20].

Moreover, it seems more probable that pleiotropic SNPs are mostly exonic and structurally functional than non-pleiotropic SNPs. As yet, no data support the hypothesis that pleiotropic SNPs would be more likely to be present in regulatory regions[20].

From an evolutionary perspective, highly pleiotropic genes are expected to be under stronger stabilising selection because they affect multiple traits, and thus are less likely to experience beneficial mutations as a result of the interwoven web of genetic and physiological interactions that are involved in development and function[35]. To this end, the genome-wide study by He and Zhang, published in 2006, found that, testing 21 different phenotypes, the 39.5 ± 0.8% of no-effect yeast genes have homologs in the fruit fly *D. melanogaster*, and that this proportion increases if we consider pleiotropic genes: 49.2 ± 2.3% of genes with effects on one or two phenotypes, and 54.7 ± 3.6% of high pleiotropic genes (with multiple effects on more than two phenotypes) have fruit fly homologs. Similarly, 52.6±2.7% of pleiotropic yeast genes have detectable homologs in the nematode *C. elegans*, in comparison to 38.3±1.1% of no-effect genes. In the same study, when the fungus *S. pombe* is compared, 71.7±3.3% of pleiotropic yeast genes have detectable homologs, in comparison to 58.4±1.3% of no-effect ones. These findings were all significant, and supported the idea that pleiotropy leads to the evolutionary conservation of genes and gene sequences[43].

2.2 State of the art in the study of pleiotropic effects

2.2.1 General introduction

One of the major limitations of association studies and GWASs is that they have tended to focus on single phenotypes through "univariate" analyses.
The complexity in the overlap of associations for different phenotypes observed within univariate analyses might be due to several underlying factors: (i) the power of genetic analyses can change based on the differences in the magnitude of the observed effects for common signals and differences in sample sizes; (ii) on the other hand, heterogeneity increases when larger number of studies is included to maximise the sample size of the meta-analysis, and this has a detrimental effect on power; (iii) sometimes there is a non-genetic component of observed phenotypic correlations, for instance due to epigenetic effects or environmental impact; (iv) moreover, a limited knowledge of the functional physiological role of associated loci, with an impact on different groups of phenotypes, may lead to a misunderstanding of the relationships between traits and diseases.
Over the last few years, it became clear that it is important to dissect the majority of the phenotypic and, to this aim, sampled cohorts have been surveyed with a large number of traits, hundreds of clinical phenotypes, and genome-wide profiling of gene expression, many of which are correlated[10].
The inability to properly dissect this kind of data, due to the absence of appropriate methodology, extensively complicates its analysis and interpretation.
There are two main challenges: the first is to obtain the greatest knowledge from the past and future univariate GWASs, by developing strategies to join together single-phenotype analyses to identify common determinants not yet discovered; the second is to explore methods to analyse a large number of variables at the same time through multivariate analysis. The projects that have been developed during my PhD programme, and that will be described in following sections, concern both these two challenges.

The analysis of multiple phenotypes enhances the ability to estimate both, the number of loci contributing to risk of multiple traits and diseases, and the spectrum of phenotypes that each locus influences, thus clarifying genetic relationships between them.
The biological advantages of performing joint analysis of multiple phenotypes include the ability to address the issue of pleiotropy vs. tight linkage or mediation, and the ability to investigate intermediate endophenotypes, e.g., serum metabolites, as a step toward understanding how biochemical pathways relate to complex traits and disorders[44].
A variety of different approaches have been proposed in the last few years to test the relationship of genes with multiple phenotypes. These approaches are based on different statistics, some of which were applied to linkage studies, and others to case-control studies (see table 2.1 for a summary of reviewed methods).
Based on the reasoning described above, these methods can be broadly classified into two main groups: univariate analyses and multivariate analyses.

Within all proposed approaches, It is not possible to define a uniformly most powerful test, because the most suitable method depends on the circumstances and on the available data[6].

Another important aspect to highlight is that the majority of proposed methods are able to detect co-association with multiple traits, that is CP effects, but this does not mean that they represent real pleiotropy.

In some cases, in fact, the same variants show association with multiple traits, but in other cases, although the same overall region is implicated, distinct nearby markers show signals of association with different traits: in this situation, it becomes fundamental to be able to distinguish the associations that represent genuinely shared effects of single variants from those that represent the effects of co-localising, but independent variants (multi-trait allelic heterogeneity, see figure 2.5)[6].

Equally important, although more difficult, it is to distinguish real pleiotropy from mediation (where the association of a genetic locus with more than one trait is due to a real association with only one of them and then to an influence of the gene-associated phenotype on the others).

An important issue to deal with, when you begin to plan a multi-phenotype association analysis, is whether the effects of a gene on correlated traits can be counted as independent contributions to the degree of pleiotropy. In other words, it is the problem of identifying the basic building blocks of the phenotype.

Just to give an example, a question can be: “are the depth and the width of a bird beak really two different characters?”[1]. Maybe the beak depth and width are two different measures of the same thing, and any mutation that affects both really has only one effect.

In fact, different phenotypes can be substantially correlated, and some correlation might be due to shared genetic covariance. A detected genetic association for one phenotype might reflect associations with other correlated phenotypes, in the sense that some genetic effects are partly or totally explained through an association with the other phenotype[45].

In addition, as a gene variant might be truly associated with two or more different correlated phenotypes, other genes could also have clear pleiotropic effects on phenotypes that are apparently clinically uncorrelated[45].

In general, ignoring phenotype correlations and relationships leads to an upward bias in estimates of pleiotropy.

This problem can be addressed, for example, by calculating and evaluating the degree of correlation between traits, and by detecting an “effective” number of phenotypes before running the analyses.

Another empirical approach to estimate if two phenotypes are independent is to evaluate the presence of mutations that dissociates them, meaning for instance that a mutation affects one phenotype but not the other, or that a mutation has same directional effects on the two phenotypes and another has opposite effects[1].

Method	Linkage or association studies	Based on p-values or effects	Allows for effect heterogeneity	Types of phenotype	Accomodate overlapping subjects	Identification of subsets of associated phenotypes	Variants or region identification	Reference
Multiple univariate analyses								
Simple comparison	Both	Both, primarly p-value	Yes	Any	Yes	Two traits per time	Both	46
Fisher's omnibus test	Both	P-value	Yes	Any	No	No	Variants	49
CPMA	Association	P-value	Yes	Any	No	No	Variants	50
Fixed-effects MA	Association	Effect	No	Same	No	No	Variants	45,48
Random-effects MA	Association	Effect	Yes, not opposite effects	Same	No	No	Variants	45,48
Subset-based MA	Association	Effect	Yes	Same	Offer extension to do it	Yes	Variants	51
Extensions to O'Brien	Both	Effect	Yes	Any	Yes, only	No	Variants	52,53
TATES	Association	P-value	Yes	Any	Yes, only	No	Variants	54
PRIMe	Association	P-value	Yes	Any	Yes	No	Regions	55
Dimension reduction techniques								
Decomposition of covariance matrix	Both	A priori transformation	Yes	Any	Yes, only	Yes	Variants	56,57
PCA	Both	A priori transformation	Yes	Any	Yes, only	Yes	Variants	58
CCA	Both	A priori transformation	Yes	Any	Yes, only	Yes	Variants	61,62
Multivariate analyses								
Multivariate linear regression	Both	Raw data	Yes	Quatitative	Yes, only	Should test different models	Variants	47,63-67
Multivariate logistic regression	Both	Raw data	Yes	Discrete	Yes, only	Should test different models	Variants	44
Log-linear regression	Both	Raw data	Yes	Discrete	Yes, only	Should test different models	Variants	68
Bayesian model search	Association	Raw data	Yes	Discrete	Yes, only	Yes	Variants	69,70
Variance-components method for multipoint linkage	Linkage	Raw data	Yes	Any	Yes, only	Should test different models	Variants	71
Variations of GEE	Both	Raw data	Yes	Any	Yes, only	Should test different models	Variants	72-74
EGEE	Association	Raw data	Yes	Any	Yes, only	Should test different models	Variants	75
Multiphen	Association	Raw data	Yes	Any	Yes, only	Yes	Variants	62
Non-parametric tests	Association	Raw data	Yes	Any	Yes, only	Should test different models	Variants	76
Graphical multivariate approaches								
Graph-based methods	Association	Raw data	Yes	Any	Yes, only	Yes	Variants	10
Tree-based methods	Association	Raw data	Yes	Any	Yes, only	Yes	Variants	77
Bayesian network methods	Association	Raw or summary data	Yes	Quantitative	Yes, only	Yes	Variants	78
Polygenic approaches								
Polygenic score	Association	Effect	Yes	Same	No	Two traits per time	None	79
Genetic correlation	Both	Effect	Yes	Same	No	Two traits per time	None	81

Table 2.1: *Summary of proposed approaches for the study of the relationship of genes with multiple phenotypes.*

2.2.2 Methods for studying cross-phenotype effects

2.2.2.1 Multiple univariate analyses

A possible strategy to detect and study CP effects is to combine results from standard univariate analyses, such as linkage analyses or association analyses (for example GWASs), across various phenotypes, to identify those variants that are associated with multiple traits[6].

In the standard univariate approach, when considering a quantitative phenotype, a linear regression is usually performed for phenotype, Y, on genotype, X. $Y_i = [Y_{i1}, \ldots, Y_{iK}]$ denotes the phenotype data corresponding to K phenotypes for an individual i and $X_i = [X_{i1}, \ldots, X_{iG}]$ denotes their genotype data at G SNPs, where, under an additive model, $X_{ig} \in [0,1,2]$. The regression performed at a SNP, g, and a phenotype, k, to test for association between the SNP genotype and the phenotype, is thus modelled as:

$$Y_{ik} = \alpha_k + \beta_{gk}X_{ig} + \varepsilon_{igk}$$

where ε_{igk} is the residual error assumed to be normally distributed.

The null hypothesis of no association between genotype and phenotype ($\beta_{gk} = 0$) can be tested by performing a t-test or an ANOVA.

Studies that used univariate approaches on different phenotypes, not necessarily measured on the same individuals, may be combined together as described below; therefore it is clear that they are well suited to analysing existing GWAS results, including those already conducted by consortia that, moreover, can be organised into cross-disease groups. These methods are especially important for rare diseases, which are less likely to be ascertained simultaneously in the same cohort studies.

Another advantage of univariate approaches is that, unlike multivariate approaches, most of them are based on summary statistics, which do not divulge individual-level data.

Below, several univariate approaches for the detection of CP effects are reported (see also table 2.1). There is not a single most powerful approach, but the appropriate statistical test should be chosen based on study design, the type of phenotypes to be analysed, assumptions on effect heterogeneity (do we expect that the effects have different direction and different sizes on different phenotypes, or not? Can we define a “prior” of the directionality and of the extent of multiple effects?), and other factors[6].

Simple comparison of univariate analysis results

The simplest strategy to analyse genetic relationships with multiple phenotypes is to run a separate linkage or association analysis for each phenotype of interest, and to compare the results. Alternatively, the set of genome-wide significant SNPs for one phenotype can be tested for association with other phenotypes; in this case, the advantage is that the significance level for multiple testing is adjusted only for the number of tested SNPs, rather than for all SNPs genome-wide.

An example of a similar approach comes from an “expression quantitative trait loci” (eQTL) association study in mice by Chen et al. where the authors assembled a co-expression network and then applied a clustering algorithm to this network for the identification of subgroups of expressed genes whose members participate in the same molecular pathway or biological process. After that, within each subgroup of expressed genes, a univariate eQTL analysis was performed between genotypes and expression data: if the majority of expressed genes in each subgroup were mapped to a same locus in the genome, that locus was considered to be significantly associated with the subgroup[46].

This kind of approach has two main disadvantages: first, it does not take into account the multivariate structure of the data; and second, testing of multiple phenotypes increases the type I error rate (experiment-wise false-positive rate), if not properly accounted for in the analysis[47]. Moreover, robust discovery is required as a starting point because these approaches are fairly underpowered if we think that known associations are probably only a subset of the possible true associations[6].

Simple meta-analytical approaches

A meta-analysis is a statistical method for the combination of summary statistics obtained from different studies to provide an overall summary result, with the aim of statistically increasing power, reducing false-positive findings, and eventually identifying new, previously unsuspected, associated loci[48].

Traditional meta-analysis approaches combine evidence for association with the same phenotype across numerous studies. Variations on meta-analysis have also been adapted for CP effect detection: in these, meta-analytical approaches aggregate summary statistics from individual studies of multiple phenotypes into one statistic, and can be applied genome-wide, or on a pre-specified set of SNPs.

These methods can be divided into those which combine p-value and those which combine effect estimates. Methods based on association p-values ignore allelic effect direction (positive versus negative) and effect heterogeneity (different effect directions and sizes) across phenotypes. Methods, based on the effect statistics, instead, are sensitive to allelic effect direction and magnitude.

In GWASs, methods that combine p-values test the null hypothesis of no association in any of the combined data sets. The alternative hypothesis is that there is association in at least one data set. These methods are easy to compute and have adequate power[45]; for this reason they were widely used until the 1980s, but then they became unpopular, and were almost abandoned in biomedical sciences, because of several limitations, such as an inability to provide a summary effect, difficulties in addressing heterogeneity issues, and dependence on normality assumptions[48].

The simplest meta-analytical approach aggregates p-values across phenotypes in different studies to test the null hypothesis that the genetic variant is not associated with any phenotype.

An example is Fisher's method[49] for combining *N* p-values (*p)* in a cumulative association statistic S_{cum} trough the Fisher's Omnibus test:

$$S_{cum} = -2 \times \sum_{i=1}^{N} lnp_i$$

S_{cum} follows the χ^2 distribution with *2N* degrees of freedom (df)[50].

This approach does not explicitly test for CP effects, and a significant association could be driven just by one phenotype, as well as by two or more phenotypes[6]. We will better discuss these aspects in the sections below, where we applied Fisher's method as primary simple meta-analysis of our data.

Cross-phenotype meta-analysis (CPMA) method

The cross-phenotype meta-analysis (CPMA) was proposed by Cotsapas at al. to investigate the genetic commonality in immune-mediated inflammatory and autoimmune diseases[50].

The CPMA uses p-values from univariate analyses for single traits and diseases and assesses association across multiple phenotypes by testing whether the observed p-values deviate from an expected distribution.

The expected distribution of association p-values for each SNP across diseases represents the null hypothesis of no additional associations beyond those already known: deviations from it are indicative of multiple associations. The alternative hypothesis is thus that each independent SNP has multiple phenotypic associations.

The alternative hypothesis includes only models in which two or more of the phenotypes, but not necessarily all of them, are associated with the SNP, with the result that this approach explicitly tests for CP effects, although it ignores the direction of effect in each disease.

Under the null hypothesis of no additional associations beyond those already known, we expect association values to be uniformly distributed, and hence *-ln(p)* to decay exponentially with a decrease rate $\lambda = 1$.

The likelihood of the observed ($\hat{\lambda}$) and expected (1) values of λ is calculated and expressed as a likelihood ratio test:

$$CPMA = -2 \times \frac{P[Data|\lambda = 1]}{P[Data|\lambda = \hat{\lambda}]}$$

Because only a single parameter is estimated (the deviation in p-value behaviour), rather than performing a meta-analysis, which would detect association with all phenotypes, or test all combinations of phenotypes increasing the multiple testing burden, this test is asymptotically distributed as a χ^2 with df = 1. This gives more statistical power to reject the null hypothesis than relying on strategies based on combining association statistics that have multiple degrees of freedom. This power comes at the price of not knowing which phenotypes the marker is associated with[50].

Moreover, CPMA assumes that the p-values used for the individual traits and diseases come from different non-overlapping cohorts; as such, it cannot be applied in the case of large consortia that investigate many phenotypes but usually share the same control samples. Modest overlap of the control samples (< 50%) is tolerable, but larger overlaps erode the power of this statistic[48].

Meta-analyses of the effects of genetic variants on multiple phenotypes

Standard meta-analysis based on effect estimates is commonly used to combine evidence of association across multiple GWASs for the same phenotype, and has also been adapted to combine evidence across multiple phenotypes.

Effect size meta-analysis methods use information on the effect sizes of the variants, and calculate summary effect sizes that can be meaningfully translated; they also allow the between-study heterogeneity to be estimated and tested[45]. The widely used approaches are described below.

<u>Fixed-effects meta-analysis</u> is the most popular approach for synthesising GWAS data and the most powerful approach for prioritising and discovering phenotype-associated SNPs.

It assumes that the genetic variant has the same effect on each phenotype, in other words, that the true underlying genetic effect in all data sets is the same, and that the observed differences are due to chance alone; this assumption is tenuous because of potential differences in phenotype definitions, linkage disequilibrium structure, and many other sources of variation, but has the major advantage of maximizing discovery power compared to other methods[45,48].

For fixed-effects models, the inverse variance weighting method is widely used. The weighted average of the effect sizes can be calculated as:

$$\hat{\theta}_F = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i}$$

and the variance is:

$$var(\hat{\theta}_F) = \frac{1}{\sum_i w_i}$$

where $\hat{\theta}_i$ is the i^{th} study normalised effect (for example, logarithm of odds ratio or β-coefficient for a logistic regression of a binary phenotype, or mean difference or standardised mean difference for a continuous phenotype), and w_i is the reciprocal of the estimated variance of the effect from that study[48].

<u>Random-effects meta-analysis</u> allows the genetic effect to differ across phenotypes. This model assumes that each data set has its own underlying effect within a population of possible underlying effects.

Random-effects are not typically used in discovery efforts owing to their limited power compared to fixed effects models; however, they are more appropriate when the aim is to consider the generalizability of the observed association, and estimate the average effect size of the associated variant and its uncertainty across different populations: for example, for predictive purposes[45,48].

The most popular method for estimating the between-study variance in random-effects meta-analysis is the DerSimonian and Laird method, but more sophisticated methods also exist.

The random effects model incorporates the between-study variance of heterogeneity, and therefore the weight for the random-effects model is calculated as:

$$w_i^R = \frac{1}{\left[\frac{1}{w_i} + \hat{\tau}^2\right]}$$

where:

$$\tau^2 = \frac{(Q - (k-1))}{\left[\sum_i w_i - \left[\frac{\sum_i w_i^2}{\sum_i w_i}\right]\right]}$$

and Q is Cochran's Q statistic, which is given by:

$Q = \sum_i w_i(\hat{\theta}_i - \hat{\theta}_F)$.

Although random-effects meta-analysis incorporates a moderate level of effect heterogeneity, it is not well suited for situations in which the genetic variant has opposite effects on different phenotypes. In addition, both fixed-effects and random-effects models will have lower power when only a subset of analysed phenotypes is associated[6].

<u>Subset-based meta-analysis</u> extends standard fixed-effects meta-analysis to an agnostic approach that allows for opposite effects and to include situations in which association is observed with only a

subset of traits, offering an improved power, and more interpretable results when compared to traditional methods for the analysis of heterogeneous phenotypes[51].

This method exhaustively evaluates all possible combinations of all possible subsets of ‘‘non-null’’ studies to identify the strongest association signal, and then evaluates the significance of the signal while accounting for the multiple tests required by the subset search. An efficient approximation is used for rapid evaluation of p-values, bypassing computational problems of multiple testing. A two-sided extension of the test allows for effects with opposite directions.

Subset-based meta-analysis was firstly proposed by Bhattacharjee and colleagues in 2012. In their paper, they evaluated the evidence of the association for a SNP for any given subset *S* of the *I* studies on the basis of the Z statistic:

$$\mathrm{Z(S)} = \sum_{i \in S} \sqrt{\pi_i(S)} Z_i,$$

in which $\pi(S) = n_i / \sum_{i \in S} n_i$ denotes the sample size for the *i*th study relative to the total sample size for the given subset *S*. The overall evidence for the association of the SNP is then evaluated on the basis of the maximum (in absolute value) of the subset specific Z statistics over the class *S* of all possible $2^i - 1$ subsets of the *I* studies.

The authors evaluated the method through simulations and application to real data, comparing it with classical alternative meta-analysis approaches. They demonstrated how subset-based meta-analysis gains substantial power—sometimes approaching between 100% and 500%—over some of the alternatives, and also performs well in distinguishing the subsets of associated phenotypes for a specific variant[51].

At present, this is the only method that identifies which phenotypes are influenced by a variant, although this advantage comes with a multiple testing price: the number of possible non-null combinations to be adjusted for increases exponentially with the number of traits selected, so that detection power decreases for even moderate phenotype counts[6].

O’Brien’s linear combination test and its extensions

The <u>O’Brien’s linear combination</u> method was proposed by O’Brien in 1984 and consists of a simple approach to combine test statistics, from linkage or association studies, of correlated individual phenotypes[52].

With *K* correlated phenotypes, $T = T^1, T^2, \dots, T^K$ is the vector of *K* statistics from association analyses; *T* follows a multivariate normal distribution with mean $\beta = (\beta_1, \beta_2, \dots, \beta_K)^T$ and covariance matrix $\boldsymbol{V}$.

The test uses a weighted sum of the univariate test statistics that is a linear combination of *T* with weight *e*:

$U = e^T \boldsymbol{V}^{-1} T.$

Under the null hypothesis H_0 there is no association: $\beta = 0$ and *U* follows the normal distribution with variance $e^T \boldsymbol{V}^{-1} e$; the alternative hypothesis instead is that at least one $\beta_k \neq 0$.

This approach can be readily used to combine univariate GWAS test statistics to create a test of pleiotropic effects; for each SNP, *U* is obtained as a test for the SNP affecting at least one of the phenotypes[52,53]. This approach is very useful for analysing multiple phenotypes of any type (continuous, dichotomous), obtained on unrelated individuals or families; however the power of this method may be less optimal when the βs are heterogeneous.

To overcome this problem, several groups have proposed extensions to the linear combination test, and in particular Yang and colleagues in 2010 proposed two extensions of O'Brien's approach that allow the weights to differ by phenotype, but which mainly differ in how they arrive at those weights: a sample splitting method and a cross-validation method. The sample spitting method first splits the sample into two subsets, one for estimating weights, and the other for constructing the final test statistic. The test statistic obtained using the estimation set is $\hat{T}w$ and using the testing set is T. Thus the final statistic becomes:

$$S = \hat{T}_w{}^T \boldsymbol{V}^{-1} T$$

which is approximately normally distributed with variance $= \hat{T}_w{}^T \boldsymbol{V}^{-1} T_w$.
The cross-validation method is a repeated random sample splitting method: it randomly divides the data set into training and testing data of a fixed size, the partition is repeated multiple times and the resulting statistics from all splits are averaged[53].
These two extended approaches can be easily applied to data consisting of unrelated individuals or families, and to individual phenotypes that are not of the same type.
Yang, using simulation studies, demonstrated that O'Brien's method provides the highest power when the means of individual test statistics are homogeneous. However, on the other hand, newly proposed approaches outperform O'Brien's method when the effects are very heterogeneous. When the effects are in different directions, O'Brien's method may have a very low power, whilst the new methods (sample splitting method and cross-validation method) gain additional power[53].

TATES
Similar to O'Brien's test, the "Trait-based Association Test that uses Extended Simes" (TATES) procedure was developed to detect associations across correlated phenotypes, but uses the p-value for each association instead of the effect[6].
TATES combines the p-values obtained in standard univariate GWASs carried out on each phenotype to arrive at a minimum global phenotype-based p-value P_T, correcting, at the same time, for the number of phenotypes and the observed correlation structure between them[54].
With m phenotypes measured in the same individual, this test aims to analyse the association between all m phenotypes and all n genotyped genetic variants (SNPs); TATES combines within each SNP the m phenotype-specific p-values (*p (1), …, p(m)*) to obtain one overall trait-based p-value P_T as follows:

$$P_T = min\left(\frac{m_e p_j}{m_{ej}}\right)$$

where m_e denotes the effective number of independent p-values of all m phenotypes for a given SNP, and m_{ej} the effective number of p-values among the top j p-values, where j runs from 1 to m; pj denotes the jth p-value in the list of ordered p-values. P_T is thus the smallest weighted p-value, associated with the null hypothesis that none of the phenotypes is associated with the SNP, and the alternative hypothesis that at least one of the phenotypes is associated with the SNP[54].
An estimate of the effective number of p-values m_{ej} is derived through a correction based on eigenvalue decomposition of the $m \times m$ correlation matrix between the p-values associated with the

m phenotypes. From this derivation, it is clear that if the *j* phenotypes are all uncorrelated, then all *j* eigenvalues equal 1, and $m_{ej} = j\text{-}0 = j$; in contrast, if the *j* phenotypes are perfectly correlated, then the first eigenvalue equals *j*, and the other eigenvalues equal 0, rendering $m_{ej} = j\text{-}(j\text{-}1) = 1$. In practice, phenotypes show inter-correlations of variable magnitude, so the effective number of p-values m_{ej} will usually be smaller than *j*, but greater than 1. m_e results equal to m_{ej} for the case that $j = m$, that is when the selection of top phenotypes covers all phenotypes. Note that the $m{\times}m$ correlation matrix between the p-values is accurately approximated through the $m{\times}m$ correlation matrix between the phenotypes[54].

PRIMe

The methods reported above only consider single nucleotide polymorphism (SNP) level but not region-level pleiotropy.

The "Pleiotropy Regional Identification Method" (PRIMe) searches for regions of the genome that contain genetic variants associated with multiple traits, but does not require the same genetic variant to be associated with multiple phenotypes[55].

Firstly, with P_S being the threshold for association significance of SNPs, and *r* being the correlation coefficient between a SNP pair, measured as the square root of the linkage disequilibrium (LD) measure r^2, PRIMe iteratively finds SNPs with the lowest association p-value among all traits and defines them as *drivers*; it then searches for SNPs whose r^2 with the drivers is above the user-specified threshold (≥ 0.8 by default), and defines them as *passengers*. Once a SNP is designated as a passenger, it will not be considered again as a new driver or passenger. In this manner, PRIMe identifies genomic regions of interest out of the whole genome.

Subsequently, a pleiotropic index is calculated as the number of traits that have at least one SNP with a univariate p-value less than P_S at a particular genomic region.

The significance of the pleiotropic index is then assessed by comparison with its distribution under the null hypothesis of no genotype–phenotype association for any of the traits/diseases. For uncorrelated phenotypes this is a simple binomial distribution; for correlated phenotypes the expected distribution is approximately a multivariate normal distribution and it requires the correlation among phenotypes to be taken into account[55].

2.2.2.2 Dimension reduction techniques

Another class of approaches that allows for multiple phenotypes considers first performing dimension reduction on the phenotypes. These techniques include both principal components analysis and linear discriminant analysis, which seek to identify linear combinations of variables that explain the most variance in the data (for principal components analysis) or which discriminate between classes and disjoint subgroups of the data (for linear discriminant analysis)[44].

Decomposition of covariance matrix

Weller et al. (1996) proposed multiple analysis of univariate, uncorrelated eigentraits, derived by a canonical transformation that consists of eigen decomposition of the covariance matrix for the original traits/disorders, in order to avoid the complexity of a very large multivariate analysis[56].

More specifically, for a given set of phenotypes with known covariance matrix, a new set of phenotypes can be derived by multiplication of the vector of the original phenotypes by a matrix, whose columns are the eigenvectors of the phenotypic covariance matrix. This way it is possible to obtain linear functions of the original phenotypes that are called "canonical variables" and are phenotypically uncorrelated. Canonical variables with very low eigenvalues, relative to the sum of all eigenvalues, can be deleted from the analysis because they explain only a minuscule fraction of the variance of the original phenotypes. In doing so, marker-linked effects can then be tested on the reduced set of canonical variables, rather than on the original one, with the advantage of reducing the number of analysed variables. Moreover, since canonical variables are uncorrelated, it is possible to exclude the possibility that a significant marker association with two phenotypes is due to mediation or to correlation.

Once significant effects are detected for the canonical variables, the effects on the original phenotypes can be derived by the reverse transformation, that is by multiplication of the inverse of the eigenvector matrix by the vector of allele effects on the canonical variables[56].

This approach can be useful to increase power of detection, and to reduce the number of analyses, but anyway the final step consists in a comparison of results from multiple univariate analyses for different canonical variables.

In 2001, Korol et al. proposed a similar eigen decomposition of the phenotypic covariance matrix in order to reduce the multiple phenotypes into a single variable only[57].

The major limitation of approaches that decompose the covariance matrix is that it is not always possible to find a transformation which guarantees that all loci influence only one canonical phenotype[44].

Principal components analysis

The best known method for dimension reduction involves using one or more of the principal components of the phenotypes in place of the original phenotypes[58]. Principal components analysis (PCA) extracts linear combinations of multiple variables that can be used as phenotypes in a genetic association analysis[6]. This approach requires only one test, and can be based on a pre-set significance level instead of running *m* different tests and adjusting the significance level for this multiplicity. The disadvantage of this approach is that principal components (PCs) depend on the variance-covariance matrix of the data, and they are not genetically based; indeed, it is possible that PCs have a low heritability.

An efficient alternative approach is a method based on the principal component of heritability (PCH), which derives a trait based on the measured phenotypes to enhance the heritability. PCH is based on the notion of optimising the phenotypic variance explained by genetic variants: for each SNP the phenotypes are reduced to a single variable that has a higher heritability than any other linear combination of the phenotypes; the association between a SNP and the derived variable is often easier to detect than an association with any of the individual phenotypes or the PCs[58].

This approach can be applied in the context of pedigree studies. Ott and Rabinowitz developed it for family-based data, where available phenotypes are combined into scales: *Y* is the *p* dimensional vector of phenotypes composed for a family-specific component, *A*, and an individual-specific component, E, that are uncorrelated with each other:

$$Y = A + E.$$

From the variance-covariance matrices of A and E, it is possible to derive the heritability of a linear combination of phenotypes. The principal components of heritability are defined not as the scores with maximum variance, but instead, as the scores with maximum heritability, subject to being uncorrelated with each other. That is, the first PC has highest heritability, the second PC has highest heritability among all PCs uncorrelated with the first, the third has highest heritability among those uncorrelated with the first and second, and so on[59].

The notion of heritability attributable to a genetic variant should not be confused with the total genetic heritability of a phenotype: the latter is usually calculated using family data, without reference to any specific genetic variants, while the heritability attributable to a genetic variant can be calculated directly from a random sample from the same population. Using the heritability attributable to a genetic variant, PCH can be applied also to association studies[58] of unrelated subjects, but in this manner a drawback arises: because the linear combination differs for each genetic variant, it is necessary to estimate the PC that maximises the heritability over all phenotypes for each single SNP each time. To address this challenge, Klei and colleagues proposed an iterating method of sample splitting and cross-validation that uses one portion of the data as training set and the remainder of the sample as a testing set for population-based association analysis[58].

From simulation experiments, both on family-based and population-based samples, PCH resulted in substantial gains in power over standard PCA when the phenotypes are not primarily repeated measures of a single trait[58,59]. When several phenotypes are repeated measures, instead, a better approach is to replace them with a simple average of the measures[57,58].

If the number of phenotypes that have been measured is very large, and exceeds the number of individuals, as for example in a typical gene expression experiment, a ridge penalty can be added to prevent over fitting, as proposed by Wang and colleagues[60].

Another approach is the one proposed by Ferreira and Purcell in 2009, where they used a canonical correlation analysis (CCA), which is a multivariate generalization of the Pearson product-moment correlation. CCA extracts the linear combination of phenotypes that explains the largest possible amount of covariation between the marker and all phenotypes[61].

The method starts with a sample of n unrelated individuals, with data for two sets of variables, a bi-allelic marker (set 1) and k phenotypes (set 2), and aims to measure the association between these two sets. The analysis can also be extended to multiple markers by expanding the first set of variables to include more than one marker. The test is based on Wilk's lambda (λ) and approximates to an F-distribution:

$$\lambda = 1 - \hat{\rho}^2$$

Where ρ is the canonical correlation between the marker and k phenotypes, and the F approximation is:

$$F_{(k,n-k-1)} = [(1-\lambda)/\lambda] \cdot [(n-k-1)/k].$$

The test can also be extended to the analysis of family-based data: prior to CCA, it is necessary to partition each individual's genotype into the orthogonal between-families (B) and within-family (W) components; then CCA is performed using the *k* phenotypes and either the B (between-family association test), the W (within-family association test), or the B+W (total association test) genotype scores. An adaptive permutation procedure is then used to account for family structure.

From simulation studies, this method was both robust and powerful; although it is most appropriate for the analysis of normally distributed traits, it shows good performance, even when considering non-normally distributed phenotypes or disease outcomes[61]. A weak point of this method is that CCA also treats genotypes as normally distributed, instead of using a more appropriate ordinal model. Moreover, CCA inflates type 1 error rates when applied to non-normal continuous phenotypes or binary phenotypes at low frequency variants[62].

2.2.2.3 Multivariate approaches

Multivariate analyses jointly analyse more than one phenotype in a unified framework. Thus, they simultaneously test for the association of multiple phenotypes with a genetic variant, given a mathematical model of the relationships among phenotypes, which can be either correlations or conditional dependencies.

Numerous multivariate parametric and non-parametric approaches have been proposed for genetic association studies, particularly for correlated phenotypes. The choice of the most appropriate method depends on the types of phenotypes included in the analysis: continuous, categorical, binary or mixed[6].

Many methods for multivariate analysis in genetics were first employed in linkage analysis, but are easily adapted to genome-wide association data from human studies[44].

Multivariate approaches are generally more efficient than multiple univariate ones, in the presence of correlated phenotypes, and when phenotypes depend on different sets of independent variables and predictors. In addition, multivariate analysis can prevent problems arising from missing data and interpretation that may complicate multiple univariate analyses when different sets of individuals are included[44].

Most multivariate methods require that all phenotypes are measured on the same individuals; this can be a limitation because they are only well suited for studies in which subjects are phenotyped across various diseases (for example, large cohort studies or cross-sectional studies), and they are not well suited for diseases with a low prevalence.

Another major difficulty is that parameter estimation reflects different alternative hypotheses to be compared to the global null hypothesis of no association. Ideally, the analyst can specify one of these alternative hypotheses, *a priori*, but sometimes interest may be in more than one, or even all of the alternative hypotheses. This situation raises considerable uncertainty about how to appropriately correct for multiple comparisons[44].

Multivariate regression framework for continuous phenotypes

For continuous phenotypes, a multivariate regression framework can be used, but the approach requires that the phenotypes are approximately normally distributed.

A first example is represented by linear regression approaches based on the Haseman-Elston method and applied to linkage studies. This group of methods is based on a robust algorithm for detecting linkage developed by Haseman and Elston for data from sib pairs.

The extension to incorporate observations of multiple correlated phenotypes on each individual is justified by the idea that these kinds of linkage studies may be more powerful if they use the information from each of the phenotypes that are affected by a same gene.

The Haseman-Elston method consider y_{ij} the measure of a phenotype for the *i*th sib ($i = 1,2$) for the *j*th pair of sibs, with μ mean, *gij* major genetic effect and e_{ij} random independent effect. If the major gene has two possible alleles (A, a), the considered model is:

$$y_{ij} = \mu + g_{ij} + e_{ij}$$

and

$$g_{ij} = \begin{cases} a \; if \; the \; ijth \; individual \; has \; genotype \; AA \\ d \; if \; the \; ijth \; individual \; has \; genotype \; Aa \\ -a \; if \; the \; ijth \; individual \; has \; genotype \; aa \end{cases}$$

If π_{mj} is the proportion of genes (0, 1/2, or 1) that the *j*th sib pair shares identical by descent (ibd) at a marker locus, and f_{m1j} denotes the probability that the sib pair shares one gene ibd; if $Y_j = (y_{1j} - y_{2j})^2$, where y_{1j} and y_{2j} are the trait values for the two sibs composing the *j*th pair, I_{mj} is the observed ibd of the sib pair at the marker locus, and the variance of the difference in residuals for the pair is σ_2, then

$$E(Y_j | I_{mj}) = \alpha + \beta\pi_{mj} + \gamma f_{m1j}$$

The coefficient β is negative if ϑ (recombination fraction between the phenotype and marker loci) < 0.5 and the additive genetic variance is greater than zero ($\alpha > 0$)[63].

From this equation we understand that the variability between a pair of sibs can be linearly modelled as a function of the genetic component of variance and the recombination fraction between the phenotype and marker loci. The ratio of the estimate of β to its standard error is distributed as a standard normal variable: a one-sided test of linkage can be obtained by comparing this statistic with the appropriate t distribution[64].

Amos and colleagues proposed a two-step approach where Haseman and Elston's function can be extended to multiple variables to find a linear function having the strongest correlation to ibd at a marker among sib pairs. A conservative test of no significant regression of the identified linear function on the proportion of genes ibd can be obtained by calculating an F statistic, that is, the linear function is then analysed as a univariate phenotype. The calculated F statistic is then compared with the critical value from an F distribution with the appropriate df[64].

Amos and colleagues tested this method on a sample pedigree of 200 individuals, considering all the possible combinations of four lipid traits (high-density lipoprotein, low-density lipoprotein, apo AI, and apo B levels) in relation to the marker-locus haptoglobin, and demonstrated that testing multiple traits can lead to the identification of stronger relationships with ibd.

Allison and colleagues, in 1998, evaluated the method developed by Amos et al. (1990), through a series of simulations, confirming that such multivariate analysis can substantially increase the power of quantitative-trait locus (QTL)–mapping studies[47].

Multivariate linear regression approaches based on the Haseman-Elston method were also applied in studies of multipoint linkage, as for example in the study by Eaves at al[65].

Linear regression methods are used in population-based association studies using General Linear Models, but they require that the phenotypes are approximately normally distributed.
Suppose there are $N = 1,2,...,n$ continuous traits measured for z individuals; a general linear model function can be written as:

$$Y = XB + E$$

where $Y = (y_1, y_2, ..., y_n)$ is a data matrix of a series of n multivariate measurements, in our case a series of n normally distributed traits, $X = (x_1, x_2, ..., x_m)$ is a matrix of covariates, and functions of the genotype probabilities for the m markers being considered, $B = (\beta_1, \beta_2, ..., \beta_m)$ is a matrix containing estimated regression coefficients, and $E = (e_1, e_2, ..., e_m)$ is a matrix containing errors. The errors are usually assumed to follow a multivariate normal distribution.
The general linear model is a generalization of multiple linear regression model to the case of more than one dependent variable y and it incorporates several statistical models[66]: χ^2 statistic, likelihood estimation, MANOVA, F-test.
An example is represented by the study of Yang and colleagues, published in 2009, where they applied a linear regression model for the study of association of a single marker to two quantitative traits[67].

Multivariate methods for discrete phenotypes

To model multiple categorical phenotypes (for example, multiple binary diseases), a multivariate logistic regression framework can be applied. To simplify, we can consider a bivariate logistic regression, which analyses two binary dependent variables jointly as functions of possibly different sets of independent variables. If we use π to indicate the probability of a particular combination of two dichotomous phenotypes, the joint outcome follows a Bernoulli distribution: Bernoulli (π_{00}) Bernoulli (π_{10}) Bernoulli (π_{01}) Bernoulli (π_{11}), with the constraint that $\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$. These joint probabilities are modelled with three parameters: the marginal probability $P(y_1 = 1) = \pi_{10} + \pi_{11}$, the marginal probability $P(y_2 = 1) = \pi_{01} + \pi_{11}$, and the odds ratio that relates the two dependent variables $\pi_{00}\pi_{01}/\pi_{10}\pi_{11}$. The bivariate regression model also analyses two binary dependent variables jointly as functions of possibly different sets of independent variables. The joint outcomes are described by two latent continuous variables that follow the bivariate normal distribution[44].

Lee and colleagues proposed also a log-linear regression model to explicitly test and compare causal models for multiple diseases subtypes; their approach can be easily applied to multiple correlated diseases. Imagine we have a group of individuals all characterised for two diseases: some are unaffected (U), some are affected by disease1 (d_1), some are affected by disease2 (d_2), and the rest by both diseases (d_{12}). Under this scenario, a series of log-linear models are fitted that corresponding to different relationships: a null model represents a variant with no effect on risk of either disease1 or disease2, a disease1 model represents a variant with effect on disease1 risk and no effect on disease2 risk, a disease2 model vice versa, and a gradient model suggests that the variant increases risk for both diseases. Each model has four parameters to estimate, q_U, q_{d1}, q_{d2} and q_{d12}, that are the frequencies for the A allele in groups U, d_1, d_2 and d_{12}. In a group $i = U, d_1, d_2$ or d_{12}, the observed counts of alleles are labelled A_i and a_i , then the log likelihood L for a model is:

$$\Sigma_i = [U, d_1 d_2 d_{12}] A_i \log(q_i) + a_i \log(1 - q_i).$$

Maximum likelihood parameter estimation is used to obtain the allele frequency parameters, along with two commonly used information criteria, the Akaike information criterion (AIC) and Bayesian information criterion (BIC). For a model with k parameters, the AIC is $-2ln(L) + 2 k$; the BIC is $-2ln(L) + k\, ln(n)$ where n is the total number of observations. These information metrics can be used for selecting the causal model that best fits the observed data. Models with lower values for these metrics are to be preferred because they provide a good fit to the data without the need for large numbers of model parameters[68]. This approach could be easily extended to model genotype or haplotype frequencies instead of allele frequencies, or be re-parameterised in terms of genotypic relative risks , or include affected offspring/parent trio data[68].

<u>A Bayesian model search</u> is a flexible, robust, and computationally efficient alternative approach, which lends itself naturally to the creation of genetic risk classifiers. A Naive Bayesian Classifier (NBC) is a simple tool that can be used to capture the complex genetic basis of a multigenic phenotype and can predict a subject's phenotype based on the posterior probability of the phenotype itself, given their genetic profile[69]. Bayesian classifiers have been used before in GWAS because they can use a large number of genetic variants, but generally only one individual phenotype. Pleiotropic associations can be modelled via the construction of simple Bayesian networks, and these models can be applied to produce single or ensembles of Bayesian classifiers that leverage pleiotropy to improve genetic risk prediction[69].

A model approach based on Bayesian classifiers has been proposed for including multiple diseases: this method starts from a GWAS dataset with multiple known related disease phenotypes as input, for which it identifies relationships between SNPs and phenotypes, and uses these relationships to generate classifiers and ensembles of classifiers that can predict one or multiple phenotypes. The algorithm does this by operating in two distinct phases[70].

In phase I, SNPs are ranked by significance of association, and the most likely association model is determined for each SNP. More specifically, we indicate with s a single-SNP with 2 or 3 possible values, depending on the mode of inheritance being tested; in the recessive mode, s is like a Bernoulli random variable with two possible values: *1 = [AA | AB]* and *2 = [BB]*; in the dominant mode, s is coded as *1 = [AA]* and *2 = [AB | BB]*; in the allelic or additive mode, each allele is treated as a separate observation, with *1 = [A]* and *2 = [B]*; in the genotypic mode, s has three possible values*: 1 = [AA], 2 = [AB], and 3 = [BB]*. If d_1 and d_2 represent two diseases, n different SNPs are modelled as having distributions that are conditional on the phenotype classes, considering four possible equally likely relationships between d_1, d_2, and each s:

- M_0, the null model, in which the distribution of the SNP is independent of either phenotype;
- M_1 and M_2, the single-phenotype association models in which the genotype frequencies of s are associated with d_1 or d_2; and
- M_{12}, the pleiotropic model, in which the distribution of s is correlated to both d_1 and d_2.

Let S be the vector of observed genotypes for each SNP, s, in m samples, and D_1 and D_2 the vectors of observed phenotypes for the two diseases respectively. Firstly, single-phenotype Bayes factors are calculated for each phenotype (d_1 and d_2) to compare the likelihood of observed genotypes S, given observed phenotypes D_1 and D_2, under the model M_1 and M_2 respectively, with the likelihood under the null model (M_0):

$$BF_{1\ vs\ 0} = \frac{p[S|D_1,M_1]}{p[S|M_0]} \ \ and \ \ BF_{2\ vs\ 0} = \frac{p[S|D_2,M_2]}{p[S|M_0]} \text{ [70]}.$$

The calculations are carried out under the four different modes of inheritance and the model with the largest Bayes factor is selected for each SNP. Only *t* SNPs, whose Bayes factors satisfy a significance threshold of $ln(BF) > 1$ are then considered. Secondly, the pleiotropic model is tested for each of the *t* SNPs: if D_x is the chosen phenotype (with x=1 or 2), then if $p[S/M_{12}, D_1, D_2]>p[S/M_x,D_x]$ the model M_{12} would be selected for this SNP; otherwise, the first-pass model (M_1 or M_2) would be selected.

The SNPs are ranked based on the Bayes factor comparing their respective selected models against the corresponding null models, and nested SNP sets (classifiers) are defined. After that, three types of genetic risk prediction can be carried out: marginal, conditional and naïve. Marginal prediction is the prediction of the risk of only one of the phenotypes, using only the subject genotype; the other phenotype is assumed unknown for prediction, but the classification rule is trained on a discovery set that includes both phenotypes. Conditional prediction is for only one of the phenotype, but now we assume that both the subject genotype and the value of the other phenotype are known; once again, the classification rule is trained using both phenotypes. Naïve prediction is based on naïve Bayesian classifiers and the classification rule is trained using one phenotype alone, ignoring all data on the other phenotype[70].

In phase II, cross-validation is used to: (1) determine the optimal number of SNPs to use in the final classifier, (2) estimate various accuracy metrics of the classifiers, and/or (3) select alternative classification thresholds. Either 10-fold or leave-one-out cross-validation (LOOCV) can be used. The discovery dataset is split into training and test sets. For each training/test set, phase I model selection is repeated on the training set, and the corresponding test set is classified using the resultant nested SNP sets. The final number of SNPs to include in the model is determined by finding the set of SNPs that, in the cross-validation, achieves the highest area under the Receiver Operating Characteristic (ROC) curve[69,70].

Testing with simulated and real data demonstrated that these models may improve genetic risk prediction under numerous circumstances[70].

Multivariate methods for continuous and categorical phenotypes together

Several methods extend multivariate approaches to allow non-normally distributed phenotypes and/or a mixture of continuous and categorical phenotypes for linkage and association studies.

Williams and colleagues in 1999 proposed a variance-components method for multipoint linkage analysis that allows joint consideration of a discrete variable and a correlated continuous trait in pedigrees of arbitrary size and complexity. The continuous trait is assumed to be normally distributed[71].

In contrast to the situation where all phenotypic data are either quantitative or qualitative in nature, and the likelihood of the complete pedigree data can be specified without any special partitioning of the variables, when some phenotypes are continuous and others are discrete, it becomes convenient to partition the total likelihood into factors that are descriptive of each type of data, and to develop each factor accordingly. The joint likelihood of observing a particular configuration of continuous phenotype values and discrete-phenotype statuses within a pedigree can be factored as $L(x,y) = L(x)L(y|x)$, where $L(x)$ is the likelihood of observing the continuous data on the pedigree members and $L(y|x)$ is the conditional likelihood of observing liabilities consistent with the affection statuses of the pedigree members, given their values for the continuous phenotype. The two likelihoods are then multiplied to give the total joint likelihood of the discrete and continuous observations in a pedigree. The results of this approach, when applied to simulated data, showed that joint consideration of a discrete phenotype and a correlated quantitative trait can improve the estimation of genetic parameters and increase evidence for linkage of the phenotypes to a major gene, compared with univariate analysis of individual phenotypes[71].

An extended regression framework, for example, can be based on variations of generalized estimating equations (GEE) that represent a multivariate version of generalized linear model (GLM) and were introduced by Liang and Zeger in 1986[72].
Generalized estimating equation models do not rely on assumptions of standard parametric distributions such as multivariate normality; the user is only required to specify the mean function and the variance[44].
Under the GLM, if we have an individual phenotype indexed by *k,* and a tested marker indexed by *m*, the model relates phenotypes and genotypes by this function:

$$L_{ijk} = \beta_{0k} + \beta_{1k} g_{ij}$$

for each *j*th individual from the *i*th family. In a population-based cohort, the number of *j* is the same of *i*, as each individual belongs to a distinct family. L_{ijk} is the link function for μ_{ijk}, the expected value of *k*; β_{0k} and β_{1k} represent population mean and genotypic effects, respectively; and g_{ij} is the genotype score for *m*. The derivation of the log-likelihood with respect to β_{1k} yields the score:

$$S_1 = \sum_i \sum_j t_{ijk} g_{ij}$$

where $t_{ijk} = y_{ijk} - \mu_{ijk}$. Under the null hypothesis of no association ($\beta_{1k} = 0$), μ_{ijk} is identical in all subjects, that is, $\mu_{ijk} = \mu_k$[73].
For multivariate data with arbitrary distributions with *K* phenotypes, Liang and Zeger's GEEs estimate model parameters while accounting for correlations among variables. A multivariate score is:

$$S_2 = \sum_i \sum_j g_{ij}\, \Delta_{ij} Var(\boldsymbol{t}_{ij})^{-1} \boldsymbol{t}_{ij}$$

where $\boldsymbol{\Delta}_{ij}$ is a diagonal matrix depending on the underlying GEE model, and $\boldsymbol{t}_{ij} = (t_{ij1}, \ldots, t_{ijK})$ is a *K*-dimensional vector containing all the phenotypic information. Under the null hypothesis of no association, $\boldsymbol{\Delta}_{ij}$ and $Var(\boldsymbol{t}_{ij})^{-1}$ are identical for all subjects, therefore the resulting score is:

$$S = \sum_i \sum_j \boldsymbol{t}_{ij} g_{ij}$$

(S_1 is a special situation of *S* with only one phenotype *k*)[72].

Lange et al. used GEE scores to extend family-based association tests (FBAT) creating a FBAT- GEE test. FBAT-GEE is a valid multivariate test that does not require any distributional assumption for the phenotypes and can be applied directly to multiple dichotomous outcome variables, counts, continuous variables, and to combinations of different types of variables[73].

In *n* independent families, each consisting of parents and one offspring, the authors tested the null hypothesis that a marker locus is not linked to any disease-susceptibility locus for any of *m* selected phenotypes. The bi-allelic marker has alleles *A* and *B*, with g_i counting the number of transmitted *A* alleles in the offspring of the ith family, and p_{i1} and p_{i2} are the parental genotypes for that family. Under the assumption that the phenotype y_i, given g_i, can be modelled by a generalized linear model, the likelihood score is given by the statistic:

$$S_3 = \sum_i t_i g_i$$

S_3 can then directly be utilized to construct a FBAT χ^2:

$$\chi^2 = \frac{(S - E(S))^2}{V_S}$$

where $E(S) = \sum_i t_i E(g_i \,| p_{i1}, p_{i2})$ is the mean value of S, and its variance is $V_S = \sum_i t_i{}^2 Var(g_i \,| p_{i1}, p_{i2})$.

When there are multiple phenotypes, instead of just one, per offspring, χ^2 is integrated with a GEE model where t_{ij} is substituted by $\boldsymbol{t}_{ij}$. With no missing phenotypic data and no covariates, the variance matrix of $\boldsymbol{t}_i$ and $\boldsymbol{\Delta}_i$ are identical for all subjects under the null hypothesis, thus they vanish when the score test is constructed under the null-hypothesis:

$$\tilde{S} = \sum_i \boldsymbol{t}_i (g_i - E(g_i | p_{i1}, p_{i2}))$$

The variance matrix is $V_{\tilde{S}} = Var\left(\tilde{S}\right) = \sum_i \boldsymbol{t}_i \boldsymbol{t}_i{}^t Var(g_i \,| p_{i1}, p_{i2})$.

The multivariate extension of the univariate FBAT can be defined by:

$$\chi^2{}_{FBAT-GEE} = \tilde{S}^t V_{\tilde{S}}{}^{-1} \tilde{S}$$

With df given by the rank of the variance matrix V_S[73].

Simulation experiments involving quantitative traits show that the multivariate FBAT clearly outperforms permutation tests and univariate FBATs with corrections for multiple testing. Moreover it can be easily extended to multi-allelic markers or to linear transformations of dependent variables[73].

This calculation can be also easily extended to population-based association studies, by just changing:

$$\tilde{S} = \sum_j \boldsymbol{t}_j (g_j - E(g_j)$$

And the χ^2 test is:

$$\chi^2 = \tilde{S}^t V_{\tilde{S}}{}^{-1} \tilde{S}$$

Finally, this statistic has also been implemented for joint analysis of population- and family-based samples: the total sample is divided into two complement sets, *U* and *R*, where *U* contains *Nu* unrelated individuals, and *R* contains the remaining *N* - *Nu* related offspring in each family. For the *U* set, the population genotype mean and variance, denoted by $\bar{g}$ and *Var(g)* respectively, are estimated[74]. For each individual in the *R* set, the genotype mean and variance are estimated from its parents' genotypes. $\tilde{S}$ thus becomes:

$$\tilde{S} = \sum_{(i,j)\in U} \boldsymbol{t}_{ij}\left(g_{ij} - \bar{g}\right) + \sum_{(i,j)\in R} \boldsymbol{t}_{ij}\left(g_{ij} - \frac{g_{i1} + g_{i2}}{2}\right)$$

The proposed test χ^2 can then be regarded as the uniform integration of population- and family-based association tests. This method can be further integrated with principal component analysis to adjust for population stratification, and also to take account of multiple siblings and missing parents[74].

Extended generalized estimating equation (EGEE) methods have been proposed by Liu et al. as powerful approach to bivariate association analysis for candidate genes or GWA studies which incorporates both continuous and discrete phenotypes, and which can be also extended to multiple correlated phenotypes with complex distributions. The advantages are: 1) offering consistent estimations of regression and association parameters, 2) being efficient[75].
The authors used seemingly unrelated regression (SUR) model by which two generalized linear models (GLMs) with different link functions, as for example different phenotypic distributions, can be combined in a unique function. In their bivariate simplification, they used an identity link for continuous traits, and a logit link for binary phenotypes.
For *N* unrelated individuals, each having observations on two phenotypes (T_1 normally distributed trait and T_2 binary variable), this unique function of the relationship between the explanatory variables and the marginal means of the two phenotypes can be expressed as:

$$L(\mu_i) = X_i'\beta$$

Where μ_i is the mean vector of the two phenotypes, X_i' is a compound function vector of explanatory variables, including genetic markers and other covariates, and β is a vector of regression parameters for the two phenotypes (β_1 and β_2) to be estimated[75].
There are two additional parameters to estimate: the dispersion parameter ψ for each phenotype (ψ_1 for T_1 and ψ_2 for T_2), and the association parameter ξ. In the context of binary outcomes, there exists no over-dispersion, so that $\psi_2 \equiv 1$. ψ_1 is squared transformed to φ^2 and ξ and φ are combined in a single vector $\tilde{\alpha} = (\xi, \varphi)'$.
Within EGEE, the authors took two steps: an Estimation Step, where the regression vector β and the association vector $\tilde{\alpha}$ are estimated; and a Testing Step, where they employed a Wald χ^2 statistic to test the significance of β parameter, that is to test if the explanatory variable has an effect on either the continuous phenotype or the binary outcome, and of the ξ parameter[75].
Simulation experiments and GWA real data analyses demonstrated better performance of this method over univariate analyses, in terms of improved power with comparable false-positive rates, under almost all the scenarios simulated.

An interesting alternative to GEE methods is the ordinal regression analysis, where the genotype of a marker is used as outcome variable, and the set of multiple phenotypes as predictors in the model. This is the strategy behind MultiPhen software, which is based on the idea that modelling the association between linear combinations of phenotypes and the genotype at each SNP can uncover novel genetic associations not detectable in single phenotype GWASs and in those based on an *a priori* definition of a phenotype as a fixed function of variables. MultiPhen rapidly performs multi-

phenotype analysis by identifying the linear combination of phenotypes most associated with genotype at each SNP. This is achieved by reversing the regression, such that the K phenotypes under investigation become the predictor variables, and genotype is regressed on phenotypes, rather than phenotypes on genotype as in standard univariate approaches (as described in paragraph "2.2.2.1. Multiple univariate analyses"). The genotype data is an allele count and is therefore modelled using ordinal proportional odds logistic regression. Ordinal regression (proportional odds logistic regression) is applied without making any assumption on the distribution of phenotypes: binary, ordinal and continuous measurements can thus be accommodated. The test for association is then an omnibus likelihood ratio test for model fit to test whether all regression weights in the model are together significantly different from zero[62]; in other words, at each SNP $g = 1, \ldots, G$, a likelihood ratio test is used to test the null hypothesis $b_{g1} = \ldots = b_{gK} = 0$. The test does not assume Hardy-Weinberg Equilibrium[62].
MultiPhen was shown to outperform other multivariate methods when minor allele frequency (MAF) was low and the phenotypes were case-control status or non-normally distributed continuous variables[54]. Another key advantage of this method is its computational speed, as well as its applicability to directly genotyped or imputed SNPs or CNV data[62]. For all its advantages, we developed a strategy of ordinal regression analysis similar to that used in MultiPhen approach, and we applied it for the analysis on multiple cardiometabolic phenotypes. In-depth description of this analysis will be reported hereinafter.

Non-parametric tests for multiple phenotypes have been proposed, even if not extensively utilised and extended. An example is Zhang and colleagues' rank-based approach that uses the generalised Kendall's τ and corresponding non-parametric U-statistics for analysing differences between pairs of individuals, as more flexible forms of test statistics.
As the authors explained, D and M denote a causal locus of interest and a marker locus, respectively, and A_D and A_M denote the alleles at D and M, respectively. The coefficient of LD between D and M is $\delta = P(A_D, A_M) - P(A_D)P(A_M)$. The null hypothesis, $H0$, is that there is no linkage disequilibrium ($\delta = 0$) between the alleles at the marker and the causal locus of interest. In other words, the assumption under the null hypothesis is equivalent to the independence of the phenotype distribution and the marker distribution. The U-statistic is then used for testing this independence assumption[76].
Kendall's τ is a classic nonparametric measure of correlation between two variables based on the difference between the probability of observing the two variables in the same order in two observations, and the probability of observing the two variables in the opposite order. For a sample of n observations $(X_1, Y_1), \cdots, (X_n, Y_n)$, two observations (X_i, Y_i) and (X_j, Y_j) are concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. Then Kendall's τ is based on the difference between the numbers of concordant pairs and discordant pairs. Firstly a multiplicative kernel function is defined as following:

$$\phi\left((X_i, Y_i), (X_j, Y_j)\right) = \phi_1(X_i, X_j)\phi_2(Y_i, Y_j) =$$
$$= \mathrm{sign}[(X_i - X_j)(Y_i - Y_j)] \begin{cases} 1, & if\,(X_i - X_j)(Y_i - Y_j) > 0 \\ -1, & if\,(X_i - X_j)(Y_i - Y_j) < 0 \\ 0, & if\,(X_i - X_j)(Y_i - Y_j) = 0 \end{cases}$$

where $\phi_1(X_i,X_j)$and $\phi_2(Y_i,Y_j)$ measure the dissimilarity of (X_i,X_j) and (Y_i,Y_j), respectively; and the corresponding U-statistic is:

$U = \binom{n}{2}^{-1} \phi((X_i, Y_i), (X_j, Y_j));$

The Kendall's τ thus is:

$$\tau = \frac{U}{\sqrt{Var_0(U)}}$$

Where $Var_0(U)$ is the variance of U under the null hypothesis.

Suppose we observe a vector of measured or coded phenotypes $\boldsymbol{T} = (T^{(1)},\cdots, T^{(p)})'$, and a vector of markers $\boldsymbol{M} = (M^{(1)},\cdots, M^{(g)})'$, for each of n study subjects, which can be substituted for Y and X respectively. The U-statistic becomes:

$$\boldsymbol{U} = \binom{n}{2}^{-1} \sum_{i<j} \phi((\boldsymbol{M}_i, \boldsymbol{T}_i), (\boldsymbol{M}_j, \boldsymbol{T}_j))$$

And the association test statistic is

$$W = \boldsymbol{U}' Cov_0^{-1}(\boldsymbol{U}|\boldsymbol{T})\boldsymbol{U}$$

where $Cov_0(\boldsymbol{U}|\boldsymbol{T})$ is the co-variance matrix of **U** given trait $\boldsymbol{T}$ under the null hypothesis that there is no association between marker alleles and any T-phenotype linked locus.

This approach can handle mixed outcomes, but does not consider additional covariates beyond the genetic variant.

Simulation studies revealed an increased power of the method in detecting significant associations compared to the Lange and colleagues' FBAT- GEE test[76].

2.2.2.4 Graphical multivariate approaches

Graphical methods for jointly analysing multiple phenotypes have been recently developed based on network theory. The application of network theory to genetics has given rise to systems genetics, which is the study of networks of interactions between genes and phenotypes, as well as networks of interactions among phenotypes, ideally integrating functional data into the GPM[44].

Graph-based methods

Some methods envisage the consideration of the information provided by networks or graphs of phenotypes before applying a regression analysis. A graph is a set of nodes and edges; in multiple phenotype analysis, nodes represent phenotypes and edges represent the relationships between them.

There are many strategies to define whether an edge should be drawn between two phenotypes. For example, one could compute correlation coefficients for all pairs of phenotypes, and connect two nodes with an edge if the correlation coefficient is larger than some threshold value, or, in the absence of a threshold, all edges exist and weights can be assigned equal to the corresponding correlation coefficients.

A sophisticated example of the use of this type of approach is described by Kim and collaborators: they proposed to use a multivariate regression function that incorporates a quantitative-trait network as representation of the correlation structure between phenotypes, combining in this manner multiple traits in a single statistical framework and subsequently analysing them jointly to identify SNPs associated with subsets of tightly correlated traits, instead of combining results from multiple univariate analyses[10].

They started from a GWAS method called graph-guided fused lasso (GFlasso), which is a multivariate regression with the *L1* penalty, named “lasso”, which sets many of the regression coefficients for irrelevant markers to “0”.

As a starting point, the equation of single-trait association via linear regression model is:

$$y_k = \boldsymbol{X}\beta_k + \epsilon_k \qquad \forall k = 1, \dots, K$$

Where: $\boldsymbol{X}$ is an $N \times J$ matrix of genotypes for N individuals and J SNPs, and each element x_{ij} of $\boldsymbol{X}$ is assigned 0, 1 or 2 according to the number of minor alleles at the *j*th locus of the *i*th individual. $\boldsymbol{Y}$ is an $N \times K$ matrix of K quantitative trait measurements over the same set of individuals so that y_k denotes the k*th* column of $\boldsymbol{Y}$; β_k is a J-vector of regression coefficients for the *k*th trait that can be used in a statistical test to detect SNP markers with significant association, and ϵ_k is a vector of N independent error terms with mean 0 and a constant variance. The estimates of $\boldsymbol{B} = (\beta_1, \dots, \beta_K)$ are obtained by minimizing the residual sum of squares:

$$\widehat{\boldsymbol{B}} = argmin \sum_k (y_k - \boldsymbol{X}\beta_k)^T \cdot (y_k - \boldsymbol{X}\beta_k)$$

Using lasso, which penalises the residual sum of square with the *L1* norm of regression coefficients and has the property of setting regression coefficients with weak association markers exactly to 0, the estimate of the regression coefficients can be obtained as:

$$\widehat{\boldsymbol{B}}^{lasso} = argmin \sum_k (y_k - \boldsymbol{X}\beta_k)^T \cdot (y_k - \boldsymbol{X}\beta_k) + \lambda \sum_k \sum_j |\beta_{kj}|$$

where λ is a regularization parameter that controls the amount of penalization. This is equivalent to solving a set of K independent regressions for each trait with its own *L1* penalty, and does not provide combined information across multiple traits.

Kim and colleagues added an additional penalty to this equation, named “fusion penalty”, which uses weighted connectivity between phenotypes as a guide and combines regression coefficients across correlated phenotypes, based on the idea that if two traits are highly correlated, their variation across individuals might be explained by genetic variations at the same loci. The assumption of this modified graph-weighted fused lasso (G_wFlasso) is that a representation of the correlation structure over the set of K traits as an edge-weighted graph G is known.

For example, the authors proposed computing a pairwise Pearson’s correlation coefficient for all pairs of phenotypes, and then connect two nodes with an edge if their correlation coefficient is above a given threshold ρ. Considering E as a set of edges, the weight, representing the strength of correlation between the two nodes, of each edge $(m,l) \in E$, is fixed as the absolute value of correlation coefficient $|r_{m,l}|$.

Given the correlation graph of phenotypes, the G_wFlasso estimate of the regression coefficients is calculated as follows:

$$\widehat{\boldsymbol{B}}^{GW} = argmin \sum_{k} (y_k - \boldsymbol{X}\beta_k)^T \cdot (y_k - \boldsymbol{X}\beta_k) + \lambda \sum_{k} \sum_{j} |\beta_{kj}| + \gamma \sum_{(m,l)\epsilon E} f(r_{ml}) \sum_{j} |\beta_{jm} - sign(r_{ml})\, \beta_{jl}|$$

where: β_{jm} and β_{jl} are the two regression coefficients for the j*th* marker, fused together if traits *m* and *l* are connected in the graph, λ and γ are regularization parameters that determine the amount of penalization; and the last term of the equation is the fusion penalty, which counts both for the direction (*sign*(r_{ml})), and for the amount ($f(r_{ml}) = |r_{m,l}|$) of the correlation.
This method, compared with a univariate regression approach, and a multivariate one that doesn't account for any structural information in the phenotypes, through simulated and real data demonstrated an improvement of accuracy in detecting true associations[10].

Tree-based methods

Tree-based approaches include classification trees (CT) and regression trees (RT), which are both based on recursive partitioning of a sample into homogeneous disjointed subgroups. The optimal tree is created by both growing and pruning procedures. Tree-based association analysis is implemented by using genotype measurements such as allelic covariates, and related phenotype measurements, to construct binary trees. An allele shows association with the phenotype if its corresponding covariate is included in the optimal tree. Figure 2.8 illustrates this procedure: imagine we have 1000 sampled individuals and we want to test the association of fasting glucose with the derived allele at a SNP (A>a), accounting for three covariates: Body Mass Index (BMI), hypertension (HTN) and total triglyceride level (TG). Firstly, the total number of subjects is divided into two groups according to whether mean BMI is less than 25 or not; then subgroups are further subdivided according to their HTN status; finally the obtained subgroups are further divided based on TG.

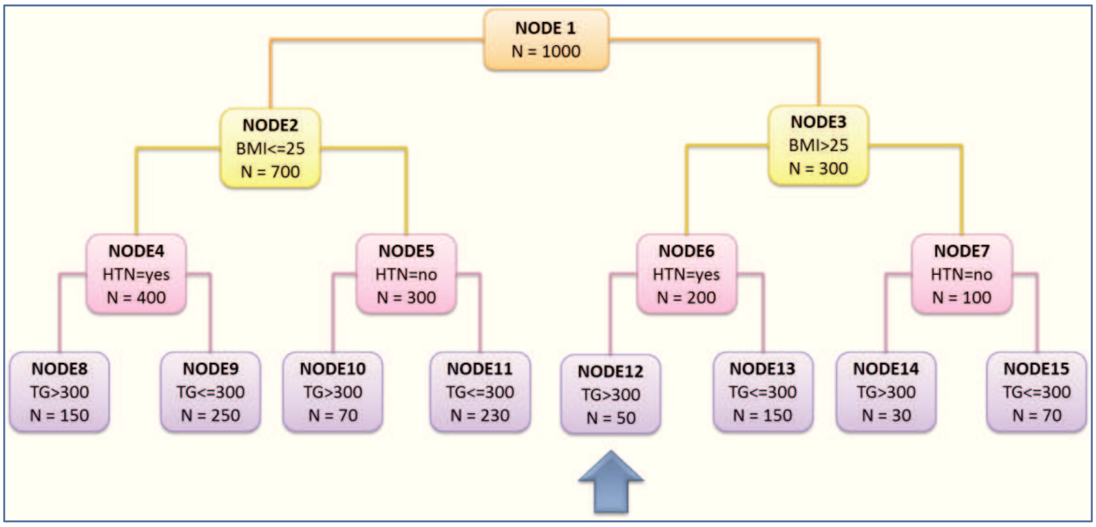


Figure 2.8: *Example of procedure for tree-based association analysis.*

The association with the genotype is then assessed for each final subgroup and, if a significant association is discovered, for example in NODE12 (indicated by the blue arrow in figure 2.8), it

means that the analysed genetic variant is associated with fasting glucose levels for those subjects with higher BMI (>25), HTN, and higher triglycerides[77]. An example application of this method is reported in a paper by Chen and colleagues[77].
In gene mapping, these approaches have been used more often with multiple independent variables than with multiple dependent variables[44].

Bayesian network methods

A Bayesian network is a directed acyclic graph in which the nodes represent random variables, and edges represent conditional dependencies between random variables, that is conditionally independent variables[44].
The Bayesian network analysis framework is based on model comparison, which effectively includes both standard univariate and multivariate association tests. Framing the association analysis as a model comparison problem, rather than as a testing problem focussed only on rejecting the null hypothesis, provides the interpretation of significant associations, and in particular by distinguishing which phenotypes are associated with each genetic variant. A collection of models is defined, each of which corresponds to a different association scenario, and the support for each model relative to the ''null'' scenario of no association is computed.
More specifically, consider assessing association between a single predictor variable *g* (a SNP genotype) and *d* related variables ***Y*,** each measured on *n* individuals randomly sampled from a population (so *g* is an *n*x*1* vector, and ***Y*** is an *nxd* matrix). *d* should be reasonably small, in the range of 2 to 10, and should include ''related'' variables in the sense that these variables either are significantly statistically correlated with one another, or are approximately uncorrelated but plausibly mechanistically linked, and so could be expected to share some genetic influences[78].

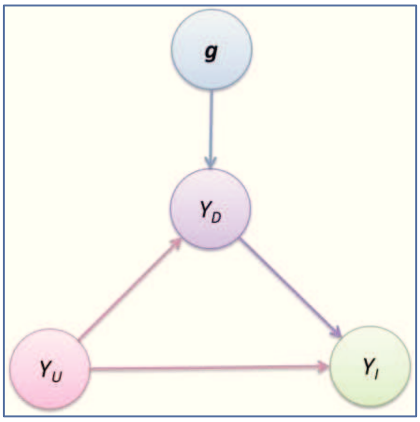


Figure 2.9: *A Bayesian network consisting of a marker g, Y_D phenotypes directly associated with g, Y_I phenotypes indirectly associated with g, and Y_U phenotypes not associated with g. Edges represent conditional dependencies, with the arrow pointing from the parent node to the child node.*

$\gamma = (U,D,I)$ denotes a partition of $\boldsymbol{Y}[1, \ldots, d]$ into disjoint subsets *U, D* and *I*, which represent, respectively, the variables that are not associated, directly associated and indirectly associated with *g*. Y_U, Y_D and Y_I are the corresponding columns of the matrix ***Y***. Each partition is then associated to a probability model $p_\gamma(\boldsymbol{Y}|g)$ that satisfies the following conditional independence relations: Y_U is independent of *g*; and Y_I is conditionally independent of *g* given Y_D, Y_U. These conditions imply that $p_\gamma(\boldsymbol{Y}|g)$ factorises as:

$$p_\gamma(\mathbf{Y}|g) = p_\gamma(Y_U)\, p_\gamma(Y_D|Y_U, g)\, p_\gamma(Y_I|Y_U, Y_D).$$

The relationships among Y_U, Y_D, Y_I and *g* can be visualised graphically as in the Bayesian network in figure 2.9.
Since γ identifies which coordinates of ***Y*** are associated with *g*, inferring γ can be viewed as the main goal. Inference for γ using Bayesian methods involves specifying a prior distribution, $p(\gamma)$, and computing the posterior distribution

using:

$$p(\gamma|\mathbf{Y}, g) \propto p(\gamma)p_\gamma(\mathbf{Y}|g).$$

Because each value of γ effectively defines a different statistical ''model'', performing inference for aspects of γ by summing over models is often referred to as ''Bayesian model averaging'' (BMA), which has the potential to answer questions about aspects of γ even when the actual ''true'' value of γ may be difficult to infer reliably.

Implementing this inference approach involves specifying a model, $p_\gamma(\mathbf{Y}|g)$, for each possible value of γ. The support for partition γ, relative to the global null hypothesis *H0*, is given by the likelihood ratio, or Bayes Factor (BF):

$$BF_\gamma = \frac{p_\gamma(\mathbf{Y}|g)}{p_0(\mathbf{Y})}$$

where large values of BF_γ indicate support for partition γ compared with the null.

The support for each partition γ corresponds to a test in which some subset of variables (Y_D) is treated as the response variables, another subset (Y_U) is controlled for, and the remaining subset (Y_I) is ignored. BF_γ is then:

$$BF_\gamma = \frac{p_1(Y_D|Y_U, g)}{p_0(Y_D|Y_U)}$$

for comparing a model where Y_D depends on g given Y_U with a model where Y_D is independent of g given Y_U. The overall evidence against the global null *H0* is summarised by an overall Bayes Factor. All possible values of γ represent a large number of models even if d is only moderate. A shortcut involves explicitly specifying only two models, and then deriving all other models from these; the two models that must be specified are those corresponding to the ''global null'', in which all variables are in U, and the ''full alternative'', in which all variables are in D. $p_0(\mathbf{Y})$ and $p_1(\mathbf{Y}|g)$ denote these two probability distributions. For multivariate normal outcomes, it is possible to use Bayesian Multivariate Regression (BMVR) to specify the null distribution $p_0(\mathbf{Y})$ and general alternative distribution $p_1(\mathbf{Y}|g)$.

This method for multivariate normal phenotypes is easily implemented, and can be applied genome-wide, requiring only summary data. However, implementing the framework for other phenotype distributions may be challenging. Another limitation is that the effect of genotype is assumed to affect only the mean, and not the variances or covariances, of phenotypes[78].

2.2.2.5 Polygenic approaches

As an alternative to the methods proposed above, or as a preliminary analysis that can be performed before searching for specific CP variants, is it possible to use a polygenic approach that analyses the information of all or of a large proportion of SNPs genome-wide, to evaluate the genetic overlap between two phenotypes[6].

This kind of approaches can use a polygenic score or a genetic correlation.

A polygenic score is based on risk alleles and their effect sizes estimated for single-nucleotide polymorphisms from independent genome-wide association studies and it aggregates the number of risk alleles that each subject carries weighted by the effect sizes of the alleles, for a particular

phenotype[6]. This scoring procedure aims to indirectly measure the collective effect of many weakly associated alleles that tend to show only very small allele frequency differences between cases and controls, but will nonetheless have higher average association test statistics and lower p-values than null loci[79].
Purcell and colleagues, on behalf of the International Schizophrenia Consortium, used this approach to directly test the polygenic inheritance theory to evaluate whether common variants have an important role, "en masse", on schizophrenia risk. Subsequently, they examined whether this component is shared with bipolar disorder[79].
The authors calculated the polygenic score using the PLINK software[80], as explained at the URL http://pngu.mgh.harvard.edu/~purcell/plink/profile.shtml, and then applied a logistic regression to test the association with the diseases.
As result of their study, the schizophrenia-derived score alleles were also associated with bipolar disorder (p-value = 7×10^{-9} and p-value = 1×10^{-12} in two independent samples), indicating a substantial, shared genetic component. However, they were largely not shared with several non-psychiatric diseases. The authors estimated also that common polygenic variation accounts for more than one-third of the total variation in schizophrenia risk[79].

Genetic correlation is the genome-wide aggregate effect of causal variants affecting two separate phenotypes. Traditionally, genetic correlations between complex phenotypes are estimated from pedigree studies, but such estimates can be biased by several factors, such as shared environmental exposures. Lee and colleagues proposed and validated a methods based on linear mixed models to obtain unbiased estimates of the genetic correlation between pairs of quantitative traits, or pairs of binary phenotypes, using population-based case–control studies with genome-wide SNP data[81].
They started from standard bivariate linear mixed models for two phenotypes:

$$y_1 = \mathbf{X}_1 b_1 + \mathbf{Z}_1 g_1 + e_1 \text{ for phenotype 1 and } y_2 = \mathbf{X}_2 b_2 + \mathbf{Z}_2 g_2 + e_2 \text{ for phenotype 2,}$$

where y is a vector of observations for trait, b_1 and b_2 are vectors of fixed effects, g_1 and g_2 are vectors of random polygenic effects for each individual, e_1 and e_2 are residuals for phenotypes 1 and 2, respectively, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices for the effects b and g, respectively. Based on this, the authors elaborated a linear approximation where the correlation between two diseases is the same on both the observed and liability scale.
Using this approach, Lee and colleagues demonstrated a significant genetic correlation between type 2 diabetes and hypertension (p-value = 0.023)[81].

It is important to notice that both approaches, polygenic score and genetic correlation, assess whether CP effects may exists between phenotypes but do not point to any particular DNA variant or genomic region[6].

2.2.2.6 Knock-out, knock-down and knock-in models

Experimentally, CP effects can be detected also by the observation of co-segregation of phenotypic differences through the use of knock-out or knock-down or knock-in genotypes in a homogenous

background using cultured cells *in vitro* or animal models.
An example is represented by a series of functional studies for an endogenous *β*-galactoside-binding protein galectin 3[6]: the knock-out mouse model of galectin 3 revealed that the deficiency of the protein leads to a concanavalin-A induced hepatitis in the liver[82], whereas inhibition of galectin 3 expression suppressed tumour growth in human breast carcinoma cells[83].
Dudley and collaborators applied this strategy on a yeast model[84]. A subsequent interesting step in the analysis proposed in Dudley's study, after common phenotype profiles are identified for several genes, consists of applying a clustering algorithm to group pleiotropic genes. Comparisons of these clusters to biological process classifications, synthetic lethal interactions, and protein complex data, support the hypothesis that this method can be used to genetically define cellular functions[84].
This knock-out method avoids the problem of closely linked genes, but it assesses only mutations that lead to the complete loss of gene activity, and therefore has to be taken as an upper limit of pleiotropy due to allele substitutions. Another limitation is that this approach applies only to knock-out genotypes that are not lethal[1]. To overcome these limitations, a knock-down strategy can be an alternative.

2.2.3 Distinguishing real pleiotropy from mediation and allelic heterogeneity

As explained in chapter "2.2.1. General introduction", the identification of a significant CP effect does not equate the identification of a pleiotropic effect: in fact, phenomena such as mediation and allelic heterogeneity may lead to a situation which can be easily confused with pleiotropy.
It is important to distinguish real pleiotropy from other forms of CP effects because they imply distinct molecular mechanisms, and have different implications for disease risk and pathogenesis[6].
In the following subsections we report a summary of principal methods to distinguish potential pleiotropy from mediation and allelic heterogeneity, this point will also be a central matter of the development of my study projects.

2.2.3.1 Identifying mediation

The definition of mediation is reported in chapter "2.1.2. Cross-Phenotype association and definition of pleiotropy": it is when a genetic variant is directly associated with a phenotype and that phenotype is causal for a second phenotype.
The association between the genetic variant and the second phenotype (also called "target phenotype") can be easily tested while adjusting or stratifying by the first phenotype ("intermediate phenotype"): if the association persists, the CP effect is probably not fully mediated. The disadvantage of this very simple approach is that it can be biased when the phenotypes share confounding factors (C)[6].

A popular framework for causal inference commonly used to test if the intermediate phenotype causally affects the target phenotype is Mendelian randomisation where the effect of a genetic variant can be taken as a proxy for the intermediate phenotype, and this is used to establish the causal relationship between the intermediate phenotype and the target phenotype[6].

Mendelian randomization refers to the random assortment of genes from parents to offspring that occurs during gamete formation and conception[85], and it was proposed for the first time by Martin Katan[86]. It is an instrumental variable analysis that uses a genetic variable (*G*, the instrumental variable), which is assumed to be randomly distributed within a population, and thus independent of confounders (*C*), to test whether an intermediate phenotype (P_A) causes another target phenotype (P_B) (see figure 2.10).

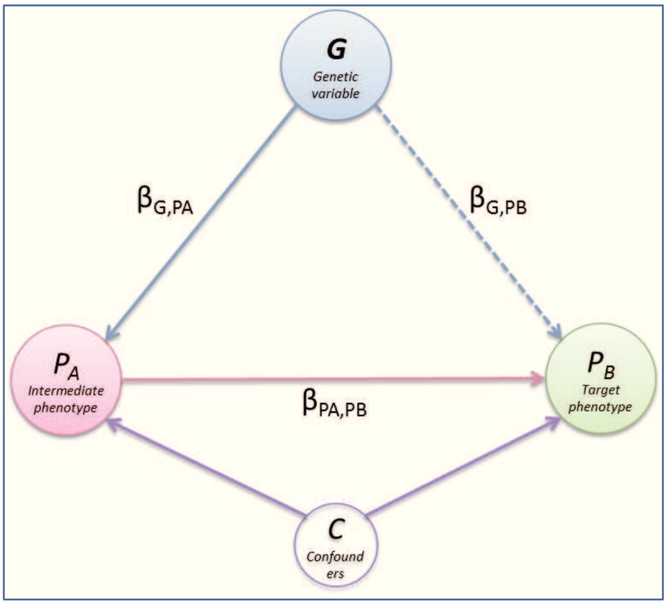


Figure 2.10: *Example of relationship model between a genetic marker (G) and two phenotypes (P_A and P_B), with the participation of some confounders (C).*

The tested hypothesis is that P_A causes P_B and the estimate of this relationship is $\beta_{PA,PB}$. To assess this, the magnitude of the estimated effects of a gene (*G*) on an intermediate phenotype (P_A), and on a target phenotype (P_B), can be combined to yield an estimate of the causal effect of P_A on P_B ($\beta_{PA,PB}$). In other words, if a causal pathway is correctly specified, then the causal effect $\beta_{PA,PB}$ can be estimated by the ratio of the regression coefficients from the association analyses of *G* on P_B, and of *G* on P_A:

$\beta_{PA,PB} = \beta_{G,PB}/\beta_{G,PA}$[87]

To conduct a valid Mendelian randomisation experiment, the following assumptions must be met:

- Assumption 1: *G* (which is a SNP or a combination of multiple SNPs) is robustly associated with P_A.
- Assumption 2: *G* is unrelated to *C*, which represents confounding factors that bias the relationship between P_A and P_B. In other words, there are no common causes of *G* and P_B.
- Assumption 3: *G* is related to P_B only through its association with P_A.

The assumptions of Mendelian randomisation are strong, and thus extreme care needs to be taken in the experimental design, in the selection of the instrumental variable (*G*), and in data interpretation[6].

Mendelian randomisation provides a potential research framework to assess causal links between phenotypes and, when correctly performed, provides insights into aetiological mechanisms and causality. Nevertheless, large sample sizes are needed, and gene-gene and gene-environment interactions could lead to false-positive or false-negative inferences, and population stratification can distort the results.

An example application of this approach is reported by Voight and colleagues in a paper in 2012, where they found that LDL levels causally affect myocardial infarction risk, whereas high-density lipoprotein (HDL) levels do not[88]. Another example is provided by the study of the relationship of the BMI-associated locus *FTO* and other metabolic and cardiometabolic related traits. Freathy et al., in 2008, through the use of the Mendelian randomisation approach, found that the *FTO* genotype is associated with metabolic syndrome and its components to an extent entirely consistent with its effect on BMI[89]. These results were replicated and extended in 2013 in a sample of ~150,000 individuals: this analysis demonstrated a causal relationship between adiposity and hypertension, adiposity and dyslipidemia, adiposity and heart failure, and adiposity and increased concentrations of the liver enzymes ALT and GGT[90].

2.2.3.2 Identifying allelic heterogeneity

Another important issue is the distinction of CP effects that are caused by proximal variants that actually represent independent association signals. This is defined as multi-phenotype allelic heterogeneity (see chapter "2.1.2. Cross-Phenotype association and definition of pleiotropy").

A preliminary approach to solve this problem can be the evaluation of LD between variants at a locus that is associated with multiple phenotypes. In this context, variants in very high LD can be considered as representative of a single underlying signal of association, whilst variants in very low or insignificant LD can be interpreted as uncorrelated or independent.

In addition, fine mapping of the region that surrounds a CP effect can help to discriminate allelic heterogeneity from real pleiotropy. Such mapping is used to more precisely locate the causal variant or variants that are responsible for a CP effect: if a single variant in the same gene, or variants in the same high LD block are discovered to be most probably causal for the diseases, this can be indicative of pleiotropy[6]. Notably, in many cases, establishing whether a variant is truly causal cannot be recognised just by fine mapping alone, and therefore this approach is approximate and not always useful to distinguish allelic heterogeneity.

A more precise and powerful method consists of performing association analyses for each phenotype, conditional on each most significantly associated SNP, within a specific locus. If the two analysed variants are not independent, and their signals overlap, the conditional analysis will show a decrease in the strength of the original signals of association. On the other hand, if the two variants are independent, thus representing allelic heterogeneity, the conditional analysis will show no change in the effects on the phenotypes that are independently associated within the same locus. Since this process is not hypothesis generating, but simply an evaluation of the architecture of multi-phenotype associations at these loci, multiple testing correction is not required.

2.2.3.3 Functional characterisation

The identification of the underlying mechanisms of multi-phenotype effects can be enriched by combining phenotypic and genetic data with functional data.

Several bioinformatics tools and databases are available for predicting the deleterious, potentially disease-causing biomolecular effects of mutations on the basis of the functional category, for example, PolyPhen[91] or SIFT[92]. However, most of these tools focus on the functional effects of either protein-coding or splice-site variants.

We know that mutations in non-protein-coding genes (such as microRNAs), or intergenic regulatory elements (such as enhancers), are also important and can result in the dysregulation of hundreds of target proteins, and thus could have a major role in phenotypic determination. Recently, the possibility of exploring and analysing several aspects for functional characterisation of coding/non-coding DNA elements arose thanks to the publication of data by the Encyclopedia of DNA Elements (ENCODE) project[93]. Since it is noteworthy that regulatory variants may confer tissue-specific effects on multiple genes, some of which reside on different chromosomes, and that single variants can thus have distinct effects on different tissues, tissue-specific investigations should be undertaken.

The examination of expression quantitative trait loci (eQTL) data in relevant tissue types can also help to identify the regulatory changes caused by mutations, as demonstrated in the Genotype-Tissue Expression (GTEx) eQTL Project[94].

The knowledge of biological processes or pathways that involve multi-phenotype associated variants can help in discerning the real nature of a cross-phenotype effect; in fact, systematic investigation of such complex biological networks would help to elucidate genetic and cellular mechanisms underlying various phenotypes, and consequently to prioritise candidate factors[95]. Multiple public resources of canonical pathways, biological functions, or protein–protein interaction data, can be used to compare and contrast diverse biological roles of gene products, as well as potential pathogenetic mechanisms underlying distinct disorders ([96] for a list of tools).

2.3 Overview of genetics of cardiometabolic phenotypes

2.3.1 Genetic discoveries for cardiometabolic phenotypes

2.3.1.1 General introduction

In our project about the study of pleiotropic effects for cardiometabolic phenotypes, we consider a series of diseases and quantitative traits related to cardiac and metabolic aspects of an organism. This is the list of considered phenotypes, grouped in categories based on the aspect of the metabolism they are related to:

- Glycaemic Phenotypes: 2 hour post-prandial glucose (2hGlu), 2 hour post-prandial insulin (2hIns), fasting glucose (FG), homeostasis model assessment for beta-cell function (HOMAB), fasting insulin (FI), homeostasis model assessment for insulin resistance (HOMAIR), fasting pro-insulin (PROINS), glycated haemoglobin (HbA1c), type 2 diabetes (T2D, disease outcome).
- Anthropometric and obesity-related traits: body max index (BMI), waist circumference (WC), hip circumference (HIP), waist-hip ratio (WHR), height, body fat percentage (PCBFAT).
- Lipids: high density lipoprotein (HDL) cholesterol, low density lipoprotein (LDL) cholesterol, total cholesterol (TC), triglycerides (TG).
- Blood Pressure-related phenotypes: diastolic blood pressure (DBP), systolic blood pressure (SBP), hypertension (HTN, disease outcome).

A detailed description of these groups and phenotypes, and of main genetic discoveries for them, is provided below.

We chose to analyse these variables for two main reasons: first, they describe in an exhaustive manner the different multifaceted physiological and pathophysiological aspects of human metabolism; second, for these variables publicly available data exists and their information is present in the majority of studied samples.

Why is it important to study cardiometabolic traits and diseases?

The rising prevalence of metabolic-related diseases indicates a crisis in global health. From a report of the World Health Organisation 2010, in 2004 over 112,000 deaths in the United States were attributed to increased cardiovascular disease (CVD), and in the same year, diabetes related complications were estimated to account for 5% of all global mortality. In 2006, more people died as a result of being overweight than underweight[97].

Therefore, it is evident that an improved understanding of pathophysiology of these diseases, achieved through genetic discovery, can provide new opportunities for treatment, diagnosis, and monitoring[98]. For this reason, numerous genetic analyses for cardiometabolic phenotypes have followed one another during the last 20 years.

In general, the discovery of causal genes for cardiometabolic traits and disorders has followed three main waves:

- The first wave consisted of family-based linkage analyses focused on candidate-genes. This approach especially permitted the identification of genes responsible for rare, monogenic extreme forms of diseases and phenotypes segregating as single-gene (Mendelian) disorders. Thanks to their high penetrance, in fact, the alleles responsible for these particular forms were relatively easy to identify[98].

- The second wave of discovery switched to tests of association for specific candidate variants or genes of interest. Most of these studies were seriously underpowered or focused on inappropriate candidates. Nevertheless, by accruing data over the course of multiple studies, some genuine susceptibility variants were identified.
- The third, and most successful, wave of discovery has been driven by systematic, large-scale surveys of association between common DNA sequence variants and phenotypes through genome-wide association studies (GWASs).

In the following sections, I will give an overview of the most important genetic discoveries for cardiometabolic phenotypes following these waves of studies.

In the past few years, genetic studies have identified hundreds of novel susceptibility loci for cardiometabolic diseases. In addition, GWASs have been undertaken on a number of related quantitative risk factors for these diseases. In fact, taken together, the inference from quantitative traits in terms of the (large) number of loci involved, the allelic frequency spectrum of associated variants, and the nature of the candidate genes, suggests that models arising from quantitative traits appropriately reflect the genetic architecture of related diseases, and reinforce the emerging evidence that it is the cumulative effect of many loci that underlies susceptibility to such pathologies.

The main relevance of the genetic discoveries achieved to date lies in potential insights into biological mechanisms underlying disease pathogenesis/progression and the potential for clinical translation through novel approaches to the diagnosis, prevention, treatment, and monitoring of cardiometabolic diseases, even though this step will take some time, because most GWAS discoveries were made in the last few years[3].

However, today, clinical translation is still limited: one of the fundamental obstacles for efforts to clinical translation, and thus to build efficient diagnostic and prognostic tools for more typical forms of cardiometabolic diseases, lies in difficulties defining the alleles and transcripts mediating association effects that have frustrated efforts to gain early biological insights. Moreover, the modest effect sizes of the common variants so far studied and discovered, and therefore the limited proportion of heritable variance which they explain has limited their value in guiding treatment of individual patients[99]. A third problem to the translation of the knowledge of risk variants implicated in multifactorial phenotypes relates to the concreteness with which risk-allele discovery has led to an improved understanding of their biological basis. The majority of associated variants, in fact, map to “noncoding” regions of the genome, making it more difficult to characterise their downstream consequences[98].

The growing power of techniques for genetic and functional evaluation is likely to catalyse further successes in characterising causal variants and connecting them to the genes, pathways and networks they modulate.

Most of the early GWASs involved individuals of European descent, but trans-ethnic fine mapping approaches, for example, particularly in samples of African origin, are growing and should help to localise the causal variants within common GWAS signals. Moreover, additional ongoing efforts to track causal variants through fine-mapping and resequencing, sequence based discovery of lower-frequency alleles, as well as functional characterisation of associated polymorphisms through the analysis of their interactions and participation in common pathways and, finally, the analysis of tissue-specific expression or regulation, should provide acceleration in the capacity for clinical translation[99].

2.3.1.2 Type 2 Diabetes

Type 2 diabetes (T2D) is a common, chronic, complex disease that accounts for more than the 95% of diabetes worldwide, and is characterised by concomitant defects in both insulin secretion from the β-cells in the pancreatic islets, and insulin action (insulin resistance) in fat, muscle, liver and elsewhere, the latter being typically associated with obesity (see figure 2.11). Although strong evidence for familial clustering highlights a strong contribution of genetic mechanisms to the disease aetiology, environmental and lifestyle factors are also of relevance[100].

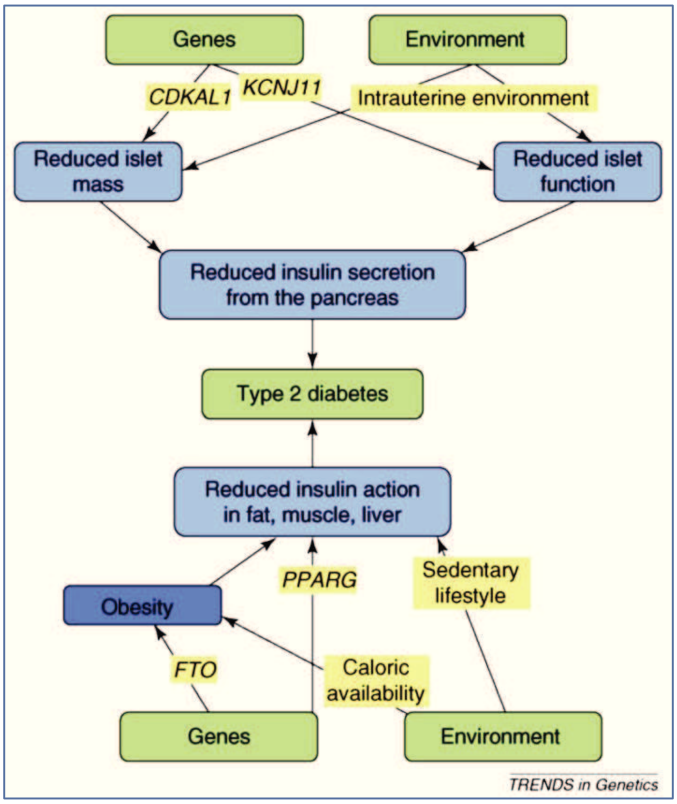


Figure 2.11: *Schema for the pathogenesis of T2D. T2D generally derives from concomitant defects in both insulin secretion and insulin action. Abnormalities in both β-cell mass and β-cell function contribute to the former, whereas obesity is a major cause of deficient insulin action. All processes involve contributions from both inherited and environmental effects. Examples of some of the genes and exposures implicated are shown in the yellow boxes. From Prokopenko et al. 2008*[100]*.*

T2D accounts for substantial morbidity and mortality from adverse effects on cardiovascular risk and disease-specific complications such as blindness and renal failure[98]. The global prevalence of T2D is of 220 million affected (figure 2.12), and this number is projected to rise to 366 million by 2030, according to the estimates of the World Health Organisation 2010[97,99].

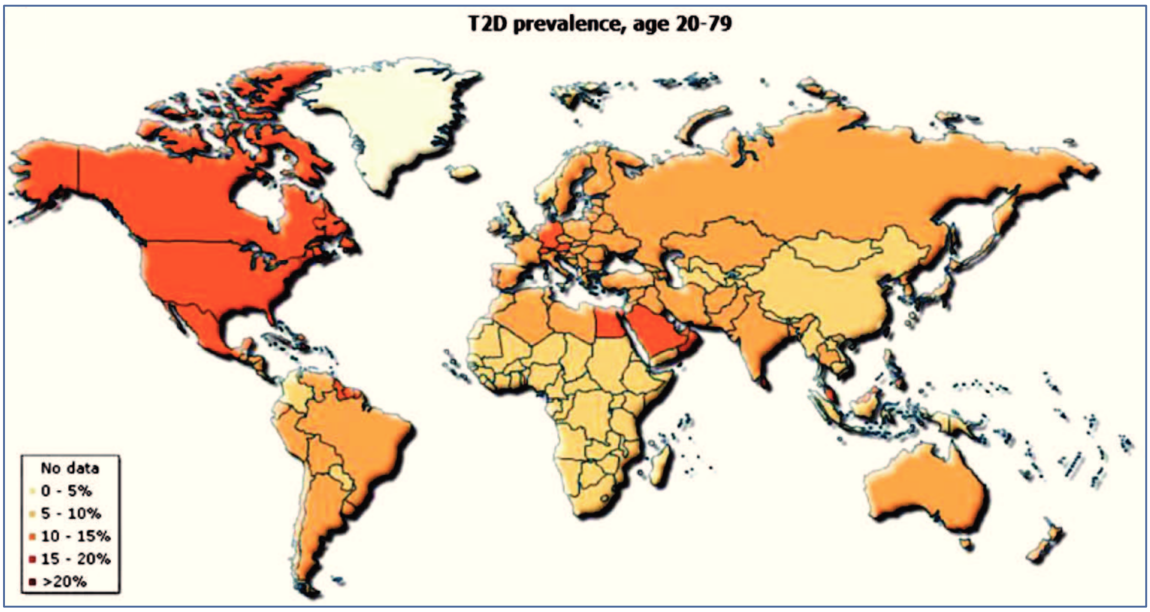


Figure 2.12: *Prevalence of type 2 diabetes by country. Colour intensity represents percentage of individuals aged 20–79 with diabetes (fasting plasma glucose > 7.0 mmol/L). From Travers et al. 2011*[99].

For T2D, the discovery of causal genes has followed the three main waves cited above. Linkage analysis was very successful in identifying the mutations responsible for monogenic and syndromic subtypes of T2D and has led to molecular classifications of disease with demonstrable prognostic and therapeutic relevance. For example, individuals with maturity onset diabetes of the young (MODY) due to mutations in *HNF1A* (Hepatic Nuclear Factor 1A) respond particularly well to treatment with sulfonylureas, whilst those with mutations in glucokinase (*GCK*) gene can often come off medication entirely because of their relatively benign prognosis. Infants with neonatal diabetes due to mutations in the *KCNJ11* (potassium inwardly-rectifying channel subfamily J member 11) gene, conventionally treated with insulin, typically showed substantial improvements when their treatment was changed to sulfonylureas[98].

However, family-based linkage studies and candidate gene association studies did not prove fruitful in revealing the variants of lower penetrance implicated in more common forms of the disease. Two of the many candidate-gene associations claimed for T2D have stood the test of time: the Pro12Ala variant in the peroxisome proliferator-activated receptor gamma (*PPARG*) gene, encoding the target for the thiazolidinedione class of drugs used to treat T2D, and the Glu23Lys variant in *KCNJ11*, which encodes part of the target for another class of diabetes drug, the sulphonylurease. These polymorphisms are both common and confirmed, in multiple studies, to influence risk of T2D. Their effect sizes are only modest: each copy of the susceptibility allele increases risk of disease by 15–20%[99].

Interestingly, rare mutations in both *KCNJ11* and *PPARG* loci are also known to be causal for certain rare monogenic syndromes characterized by severe metabolic disturbance of β-cell function and insulin resistance, respectively[100].

The number of loci for which there is convincing evidence that they confer susceptibility to T2D started to grow in early 2007 with the publication of the first GWAS[98]. Since then, more than 20 major GWASs for T2D have been published, and a cumulative number of around 80 genome-wide significant hits was discovered[3] (more than 60 loci; see figure 2.13 and Appendix table 1 for an overview).

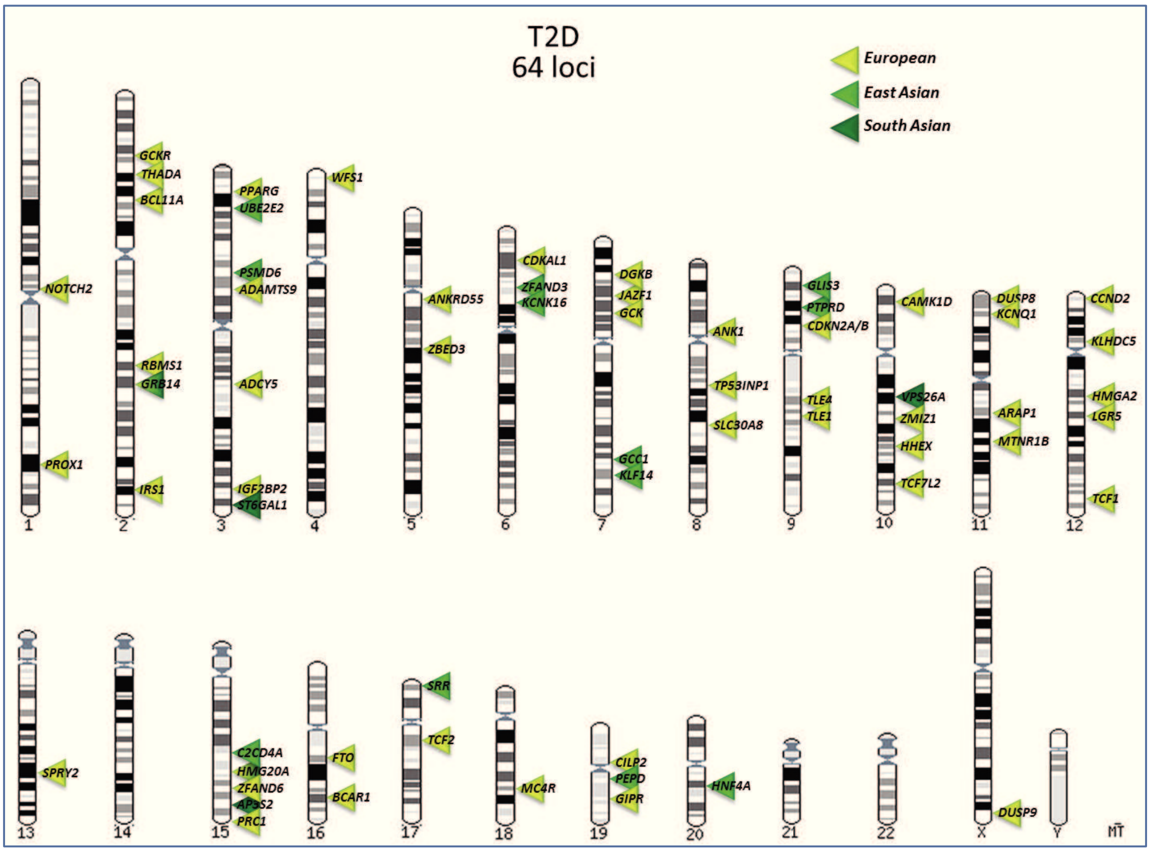


Figure 2.13: *Overview of genome-wide T2D-associated loci, through December 2012.*

The first wave of GWAS, in 2007, confirmed the already known loci *PPARG*, *KCNJ11* and *TCF7L2* (transcription factor 7-like 2), but added a further six novel loci including signals near *CDKAL1* (CDK5 regulatory subunit associated protein 1-like 1) and *CDKN2A/CDKN2B* (cyclin-dependent kinase inhibitor 2A/B), which encode putative or known regulators of cyclin-dependent kinases, *HHEX* (hematopoietically expressed homeobox) which transcribes a homeobox protein implicated in β-cell development, *SLC30A8* (solute carrier family 30 member 8), *IGF2BP2* (insulin-like growth factor 2 mRNA binding protein 2), and *FTO*[101-105]. Each copy of a susceptibility allele at one of these loci is associated with a 15 to 20% increased risk of diabetes[98].

Within successive rounds of GWA meta-analyses, the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, including more than 47,000 genome-widely characterised individuals and 94,000 samples for replication, firstly combined data from three published GWASs to reveal six novel loci[106] and subsequently aggregated data from additional five GWASs to capture a further 12 signals[19], bringing the count of confirmed common variant signals for T2D to more than 60.

DIAGRAM also coordinated a new run of meta-analysis of genetic variants genotyped on the Metabochip SNP array, including 34,840 cases and 114,981 controls of European descent. The Illumina CardioMetabochip (Metabochip) array for genotyping was published in 2012[107]: it is a custom array of 196,725 SNPs developed to support cost-effective, large-scale follow-up studies of putative association signals for a range of cardiovascular and metabolic traits and to fine map established loci. This analysis added another ten loci to the list of confirmed common variants associated with T2D[108].

Most published studies have considered individuals of European descent. More recently, equivalent studies have emerged from samples of East Asians[109-111], and South Asians[112], and large studies involving African Americans and other major ethnic groups are underway. Despite differences in allele frequency and LD patterns, most of the signals found in one ethnic group, in particular 40 European signals, showed some evidence of association in others, indicating that the common-variant signals identified by GWASs are likely to be the result of widely distributed causal alleles[3].

GWAS in East Asians also revealed several novel associations for T2D: for example variants in the potassium voltage-gated channel, KQT-like subfamily member 1 gene (*KCNQ1*), which have since been replicated in European ancestry populations.

The strongest common-variant association signal identified for T2D remains *TCF7L2,* detected just prior to the GWAS era, and subsequently confirmed by various GWASs; fine-mapping studies have converged upon the intronic SNP rs7903146 as the most compelling candidate variant in this region, with a per-risk allele odds ratio (OR) of around 1.35, and lifetime prevalence rates that, in persons carrying two copies of a risk allele, roughly double those seen in persons with none[98,99]. At this locus, ChIP-Seq (chromatin immunoprecipitation sequencing) studies have shown that rs7903146 maps within a region of islet-specific open chromatin, and the two alleles differ in their capacity to achieve or maintain this state[113]. *TCF7L2* mRNA levels in human pancreatic islets increase with the number of risk alleles, and over-expression of *TCF7L2* leads to reduced glucose-stimulated insulin secretion[114]. To date, pharmacogenetic studies in common forms of T2D have not offered dramatic applications. The only convincing result concerns the association of genotype at the *TCF7L2* with variation in response to sulfonylurea treatment. In a retrospective observational study, patients carrying two risk alleles at *TCF7L2* variant were almost twice as likely to fail treatment objectives than those carrying no risk alleles, with an intermediate effect for heterozygotes[115].

At other T2D-susceptibility loci, including *GCKR*, *PPRG* and *SLC30A8*, there is substantial statistical and biological evidence to support particular coding sequence variants as causal. Functional characterisation has shown that the T2D-risk allele alters fructose-6-phosphate-mediated regulation of the protein coded by *GCKR* (glucokinase regulatory protein), with consequences for glycolytic flux. *SLC30A8* encodes a zinc transporter, ZnT8, known to be expressed in the pancreatic islets and implicated in the proper function of β-cell insulin granules; in mice, β-cell-specific knock-outs of Znt8 are glucose intolerant, and display defects in insulin production, crystallisation, packaging and secretion, highlighting the importance of zinc as a modulator of islet function[99].

2.3.1.3 Glycaemic Traits

Studies of risk variants for T2D in healthy populations have shown that most of them act through perturbation of insulin secretion rather than insulin action, establishing inherited abnormalities of β-cell function or mass (or both) as critical components of the progression to T2D[98]. A role for other complex processes influencing other quantitative physiological T2D-related traits cannot be excluded and may have an action on the susceptibility in developing the disease.

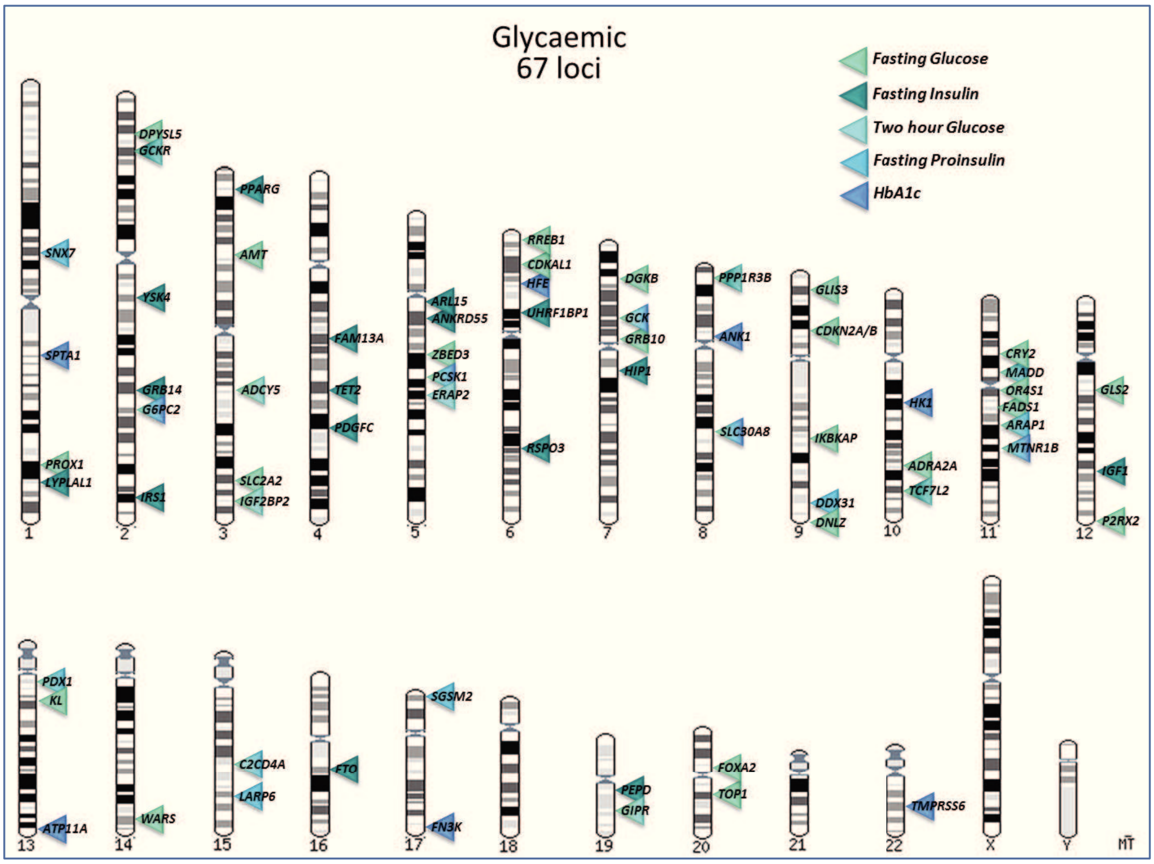


Figure 2.14: *Overview of genome-wide associated loci for glycaemic traits, through December 2012.*

With the aim of studying such processes, the Meta-Analysis of Glucose- and Insulin- Related Traits Consortium (MAGIC) investigators have been carrying out genetic analyses focused on the identification of variants influencing normal physiological variation in levels of continuous glycaemic traits in healthy non-diabetic individuals[18,116-119] (for a complete list of glycaemic-associated loci see figure 2.14 and Appendix table 2). Glycaemic trait large-scale GWAS meta-analyses so far comprised

FG, FI, PROINS, 2hGlu assay, and HbA1C levels.

Glucose is the major source of energy for most cells of the body, including those in the brain. It derives from carbohydrates that are found in fruit, cereal, bread, pasta, and rice, and which are quickly turned into glucose in the body, raising blood glucose levels. Hormones such as insulin and glucagon help control blood glucose levels. A blood fasting glucose (FG) test measures the amount of glucose in a sample of blood after having not eaten anything for at least 8 hours (fasting): a level between 70 and 100 milligrams per decilitre (mg/dL) is considered normal; while a level of 100-125 mg/dL means impaired fasting glucose, a condition of pre-diabetes, and a level of 126 mg/dL or higher most often means diabetes.

Another way to measure glucose tolerance is the oral glucose tolerance test (OGTT): after giving patients a liquid containing a certain amount of glucose (usually 75 grams) to drink, the glucose concentration in blood is measured after time intervals of 30 minutes. The measurement taken after two hours is called two hour post-prandial glucose level (2hGlu) and is considered normal if less than 140 mg/dL.

Glycated haemoglobin (HbA1c) is a form of haemoglobin that is measured to identify the average plasma glucose concentration over prolonged periods of time as it is influenced by average glycaemia over a 2- to 3-month period. It is formed in a non-enzymatic glycation pathway by haemoglobin's exposure to plasma glucose. The HbA1c test indicates the body’s long term control of blood sugar and is used to monitor and diagnose diabetes: a normal level is considered when HbA1c is less than 5.7% of total haemoglobin, whilst levels between 5.7% and 6.4% are indicative of pre-diabetes, and if they are 6.5% or higher indicate diabetes.

Insulin is a hormone secreted by the pancreas in response to eating carbohydrates that facilitates the transport of sugars from the bloodstream into the cells where they are used to make energy. Insulin resistance occurs when insulin does not work optimally to drive glucose into cells and tissues. Measuring FI in the blood is helpful in the diagnosis of insulin resistance and type 2 diabetes. Insulin excess is defined when levels are equal to or greater than 15 µIU/mL (micro International Units per millilitre).

Proinsulin is the pro-hormone precursor of mature insulin and C-peptide, made in the β-cells of the islets of Langerhans that are pancreatic specialised regions. Higher circulating levels of proinsulin are associated with impaired β-cell function, raised glucose levels, insulin resistance, and T2D, and seem to indicate an advanced stage of β-cell exhaustion. Consequently, fasting proinsulin might be used as marker detecting and for therapeutic decision in T2D. A normal proinsulin level is 2 to 6 pmol/L (picomoles per litre).

Part of this clinical information is taken from PubMed Health (https://www.ncbi.nlm.nih.gov/pubmedhealth/t/a/) and MedlinePlus (http://www.nlm.nih.gov/medlineplus/encyclopedia.html).

Prior to the GWAS era, the only compelling association signal for fasting glucose levels was known at *GCK* locus, coding for a glucokinase[3]. The first GWAS in European samples (about 46,000 individuals) expanded that number to 16 loci[117]. These variants explain around 10% of the inherited variation in fasting glucose levels. Only two signals, near *GCKR* and *IGF1* (insulin-like growth factor 1), were shown to influence fasting insulin levels in the same analysis.

Comparable analyses for two hour glucose (15,000 GWAS samples and up to 30,000 replication samples) identified further signals, including variants near the locus for GIP Receptor *GIPR*[118].

A genome-wide meta-analysis exploration for glycated haemoglobin HbA1c, equivalent to the one for fasting glucose and fasting insulin, identified ten loci that reached genome-wide significant association, including six new loci near *FN3K* (fructosamine 3 kinase), *HFE* (hemochromatosis), *TMPRSS6* (transmembrane protease, serine 6), *ANK1* (ankyrin 1), *SPTA1* (spectrin alpha 1) and *ATP11A*/*TUBGCP3* (ATPase class VI type 11A/tubulin gamma complex associated protein 3), and four known HbA1c/glycaemic/T2D loci: *HK1* (hexokinase type 1), *MTNR1B* (melatonin receptor 1B), *GCK* and *G6PC2*/*ABCB11* (G-6-phosphatase catalytic subunit 2/ATP-binding cassette, sub-family B member 11)[120]. Three of the ten signals (*GCK*, *G6PC2* and *MTNR1B*) of association with HbA1c are partly related to an association with hyperglycaemia. The remaining seven non-glycaemic loci accounted for a 0.19% HbA1c difference between the extreme 10% tails of the risk score.

Similarly, a GWAS analysis was conducted on proinsulin levels[121]. A meta-analysis for this trait resulted in nine SNPs at eight loci achieving genome-wide significant association (p-value < $5x10^{-8}$)[121]. Two loci (*LARP6* (La ribonucleoprotein domain family member 6) and *SGSM2* (small G protein signalling modulator 2)) were new, as previously unknown to be related to metabolic traits; one variant (near *MADD*, MAP-kinase activating death domain) had already been associated with fasting glucose, and another (*PCSK1*, protein convertase subtilisin/kexin type 1) with obesity; finally four SNPs (*TCF7L2*, *SLC30A8*, *VPS13C/C2CD4A/B* (vacuolar protein sorting 13 homolog C/C2 calcium-dependent domain containing 4A/B), and *ARAP1* (ArfGAP with RhoGAP domain ankyrin repeat and PH domain 1)) were already known as associated with increased T2D risk.

The proinsulin-raising allele of *ARAP1* was also associated with lower fasting glucose, improved β-cell function, and lower risk of T2D. Notably, this gene encodes the protein prohormone convertase 1/3, the first enzyme in the insulin processing pathway[121].

In 2012, the Illumina CardioMetabochip (Metabochip) array for genotyping was published[107]; a second run of GWAs for glycaemic traits genotyped with the Metabochip was thus conducted[18], resulting in discovery of 41 glycaemic associations not previously described: 20 for FG, 17 for FI, and four for 2hGlu.

This raised the number of associated loci to 36 for FG, 19 for FI, and 9 for 2hGlu, explaining 4.8%, 1.2%, and 1.7% of the variance in these traits, respectively.

Since obesity is an important determinant of insulin resistance, Manning and colleagues decided to carry out a joint meta-analysis (JMA) approach for genetic association to simultaneously test both the main genetic effects on glycaemic traits, on glycaemic traits adjusted for BMI (as index of obesity), and potential interaction between each genetic variant and BMI[119]. Six loci not previously known to be associated with fasting insulin levels were discovered, as well as seven additional loci associated with fasting glucose levels. Further, all previously reported associations for fasting glucose (16 loci) and fasting insulin (two loci) were replicated. The association of fasting insulin accounting for BMI with the genetic variant located at the *COBLL1/GRB14* (Cordon-Bleu WH2 Repeat Protein-Like 1/growth factor receptor-bound protein 14) locus is of particular interest since several studies suggested that *GRB14* is a tissue-specific negative regulator of insulin receptor signalling via the regulation of adipose tissue distribution. In addition, another suggested candidate locus was *PPP1R3B* (protein phosphatase 1 regulatory subunit 3B), which is likely to act via hepatic metabolism to influence fasting insulin and glucose levels, as well as the lipid profile and C-reactive protein levels[119].

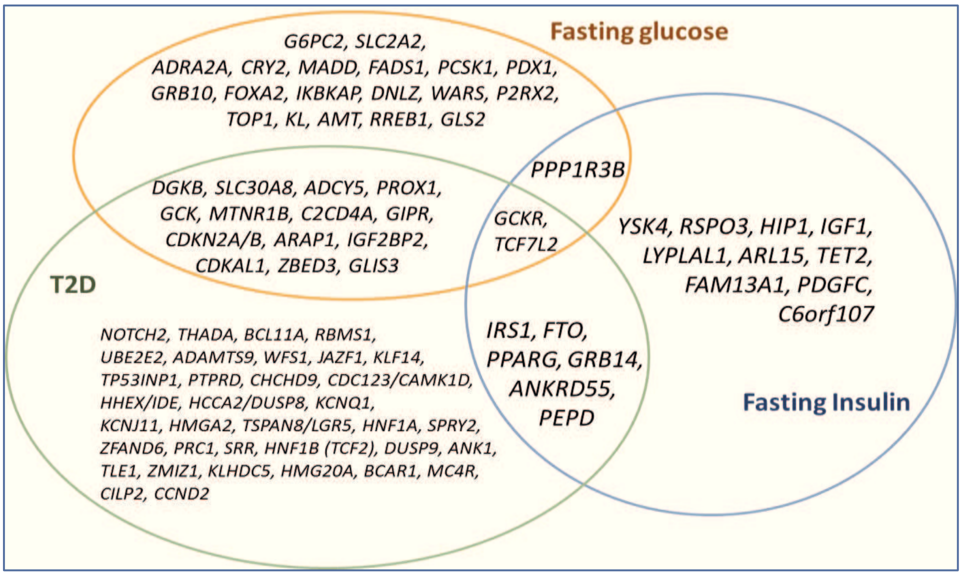


Figure 2.15: *Overlaps between T2D-associated loci and fasting glucose/ fasting insulin-associated loci.*

From the results described above and reported in figure 2.14, there is an incomplete overlap of T2D associated loci with those influencing physiological variation in glycaemic traits (figure 2.15). Some loci, for example *MTNR1B*, have a relatively large effect on both, whereas others, such as *G6PC2,* influence fasting glucose levels but have a minimal effect on T2D risk. On the other hand, *CDKN2A/B* has an impact on T2D but only modest effects on fasting glucose levels in healthy, non-diabetic individuals. The loci included in this last group appear to have their primary effect on the functionality of β-cells (rather than on insulin resistance) highlighting the importance of β-cell function with respect to normal and abnormal glucose homeostasis[3], supporting the idea that the mechanisms influencing physiological and pathophysiological variation in glycaemic homeostasis are only partially overlapping[99].

Physiological characterisation of some of the genetic loci influencing glycaemic traits demonstrated regulation activity by diverse pathways as reported in figure 2.16: the glucose-raising allele in *MADD* was related to abnormal insulin processing and higher proinsulin levels, but not to insulinogenic index. Defects in both insulin processing and insulin secretion, were seen in glucose-raising allele carriers at *TCF7L2*, *SCL30A8*, *GIPR*, and *C2CD4B*, while abnormalities in early insulin secretion only were suggested in glucose-raising allele carriers at *MTNR1B*, *GCK*, *FADS1* (fatty acid desaturase 1), *DGKB* (diacylglycerol kinase beta), and *PROX1* (prospero homeobox 1)[122]. *MTNR1B* is also associated with fasting glucose and T2D risk. From functional analyses *MTNR1B* expression results localised to the β-cells within human islets and showed altered expression in islets from type 2 diabetic donors, whilst the receptor, it encodes, mediates the inhibitory effect of melatonin on glucose-stimulated insulin response. Inhibition of this melatonin-ligand receptor system is therefore a potential therapeutic option for T2D[123].

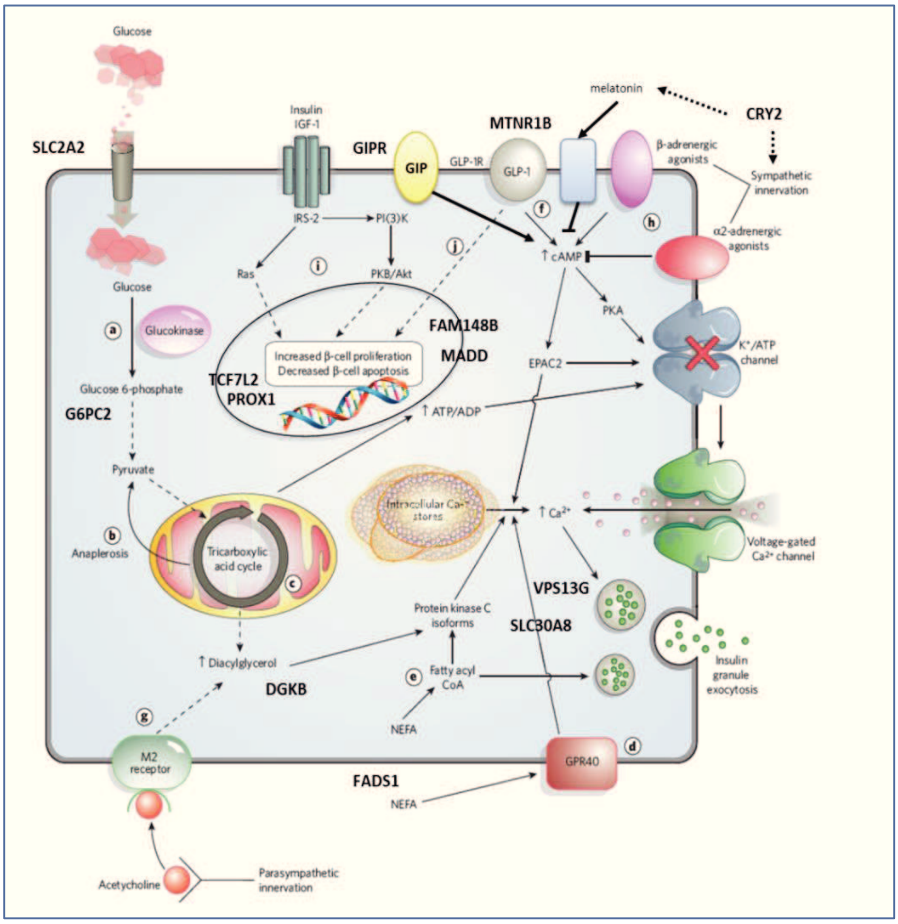


Figure 2.16: Suggestive mechanisms by which some of published genetic loci could influence glycaemic regulation. From Ingelsson et al. 2010[122].

2.3.1.4 <u>Obesity, obesity-related traits and Height</u>

Obesity is a rapidly growing health problem worldwide, conferring substantial excess risk for morbidity and mortality, especially from obesity-related complications, such as T2D and atherosclerotic cardiovascular disease (CVD)[124].

Obesity is a complex disorder, where genetic predisposition interacts with environmental exposures to produce a heterogeneous phenotype. Heritability of obesity is between 50 and 80%.

BMI is typically used as an indication of obesity status and it has consistently been associated with health outcomes[124]. It is a number calculated from a person's weight and height with the formula weight(kg)/[height(m)]^2 and is used as a screening tool to identify possible weight problems for adults. Worldwide, there are more than 400 million adults with a BMI exceeding 30 kg/m^2, the universally established threshold to define ‘‘obesity’’, and this number is projected to rise to 700 million by 2030 (see figure 2.17)[99].

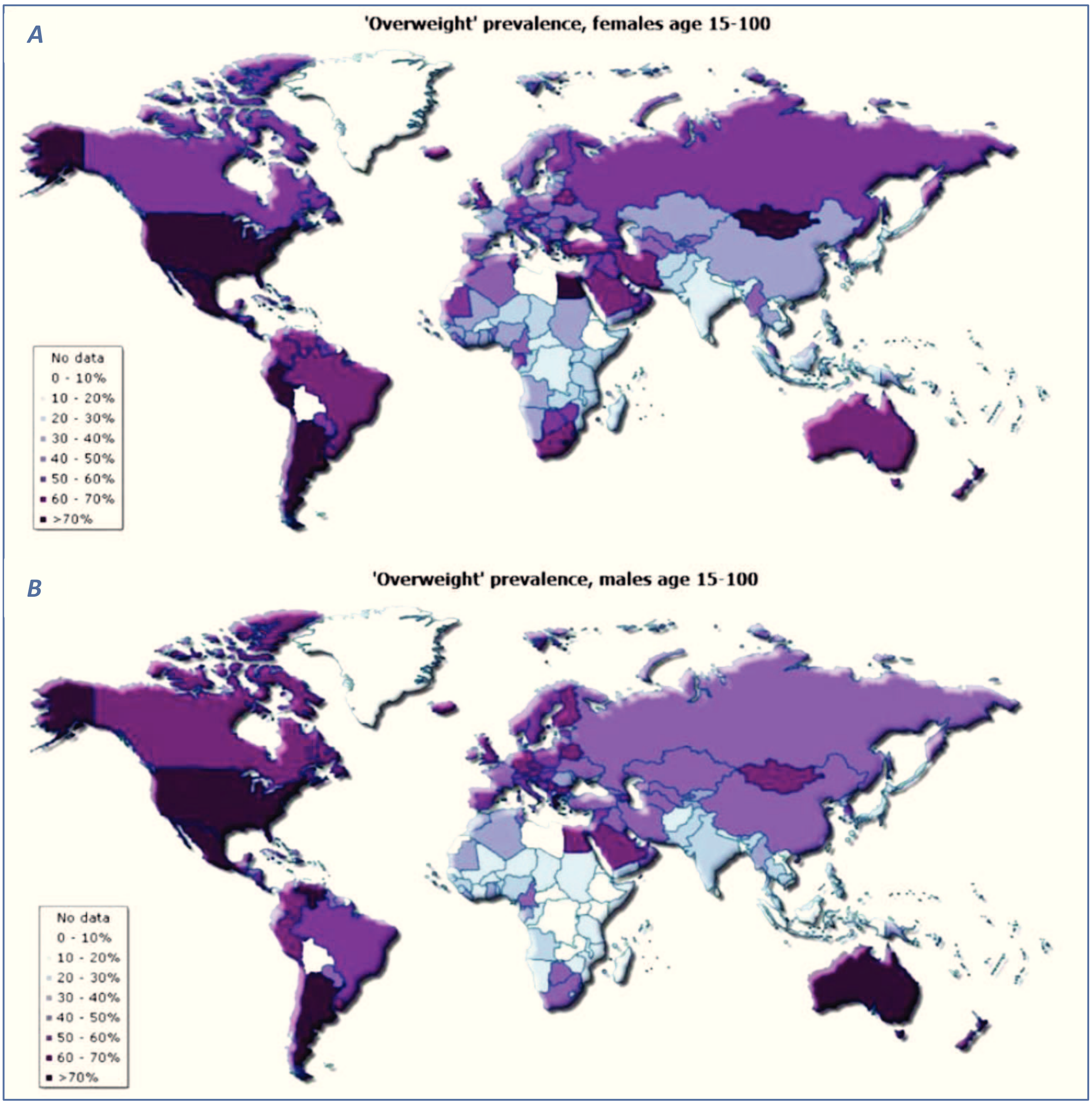


Figure 2.17: *Prevalence of obesity by country.* ***A:*** *colour intensity represents percentage of females aged 15–100 with BMI > 25 kg/m*^{2}*.* ***B:*** *colour intensity represents percentage of males aged 15–100 with BMI > 25 kg/m*^{2} *(''overweight'' data from World Health Organisation 2010: https://apps.who.int/infobase/). From Travers et al. 2011*[99]*.*

Other indices of fat distribution that can be used to monitor obesity are waist circumference (WC), as representative of central obesity, and hip circumference (HIP). Waist-to-hip ratio (WHR) is a measure of central obesity corrected by a peripheral mass index. Another index, finally, is body fat percentage (PCBFAT). The use of these alternative measures, instead of BMI, is prompted by the particularly deleterious health effects of visceral fat accumulation rather than of BMI.

From monogenic disease studies for extreme forms of obesity, identification of mutations in the leptin gene (*LEP*) causing severe early onset obesity resulted in the development of recombinant leptin therapy as a life-saving treatment for affected children[99].

Before the GWAS era, the only robust association between DNA sequence variation and either BMI or weight was observed from tests of association and involved low-frequency coding variants in *MC4R* gene, encoding the melanocortin-4 receptor. This variant explains approximately 2 to 3% of cases of severe obesity[3,98].

Genome-wide association studies of population-based samples for genetic variants influencing BMI and obesity have been more productive and have identified several loci influencing BMI and the risk of obesity. The strongest signal identified is the association with variants within *FTO* (the fat-mass and obesity–related gene). Successive rounds of GWA meta-analysis have brought the count of confirmed common variant signals for BMI and obesity to over 30[16,125-128].

Subsequently, the Genomic Investigation of Anthropometric Traits (GIANT) Consortium firstly combined data from 15 GWAS cohorts to reveal six new loci contributing to variation in BMI, as well as replicating the established common variant signals at *FTO* and *MC4R*[127]. Almost in parallel, the deCODE group reported ten new BMI-influencing loci[128]. The synthesis of these two efforts, involving genetic analysis of almost 250,000 individuals, confirmed 14 existing loci and revealed 18 novel signals for BMI and obesity[16].

The role of rare CNVs in obesity has not been well examined so far, but rare deletions at chromosome 16p11.2 have been shown to have high penetrance for obesity and mental retardation[99].

The largest signal for obesity-related traits remains that at *FTO*: its association signal accounts for less than 0.5% of the overall variance in BMI, equivalent to a difference of 2 to 3 kg between adults that are homozygous for the risk allele and those that are homozygous for the alternative allele. Consideration of all 32 currently known BMI-influencing loci increases this figure to only 1.45%[16].

The cited studies have tackled obesity through its cognate quantitative trait, BMI. As for T2D, case–control studies of extreme obesity have identified loci only partly overlapping with those associated with physiological variation of BMI (for example, *PCSK1*, *POMC* (proopiomelanocortin), *BDNF* (brain-derived neurotrophic factor), *MC4R*, and *SH2B1* (SH2B adaptor protein 1))[99].

GWAS of patterns of other indices of fat distribution, such as WC, HIP, WHR and body fat percentage, have characterized approximately 16 loci that are largely distinct from those influencing overall adiposity[126,129-132]; many of these signals display markedly stronger associations in women than in men.

For an overall view of the loci associated with obesity and body fat distribution, see figure 2.18 and Appendix table 3.

Additional studies in Indian Asians confirmed BMI-associated variants in *MC4R*[129], whilst other studies based on East Asian population revealed four new BMI-associated loci[133,134].

As for T2D and fasting glucose, most of the signals for obesity and fat distribution map to regulatory regions and the causal transcript is known for only a minority of the loci[3].

The BMI association signal near *FTO* is the most established signal and comprises a 47-kb LD block which may be involved in the regulation of the adjacent gene, *RPGRIP1L* (retinitis pigmentosa GTPase regulator interacting protein 1-like), as well as *FTO* itself. Although the region of association is clearly defined, and its effect is comparatively large, there is still some doubt as to whether *FTO* itself is responsible for the weight phenotype. Studies of mice demonstrated that disruption of *Fto* sequence influences adiposity with changes in body weight[135] thus being consistent with the hypothesis that *FTO* itself has a direct effect on BMI; studies of human *FTO* mutations instead are less clear-cut, as no direct evidence linking coding variants to body-weight variation has been demonstrated[3,98]. *RPGRIP1L* is expressed in the hypothalamus, with responses to alterations in nutritional and hormonal status that are similar to those of *FTO*.

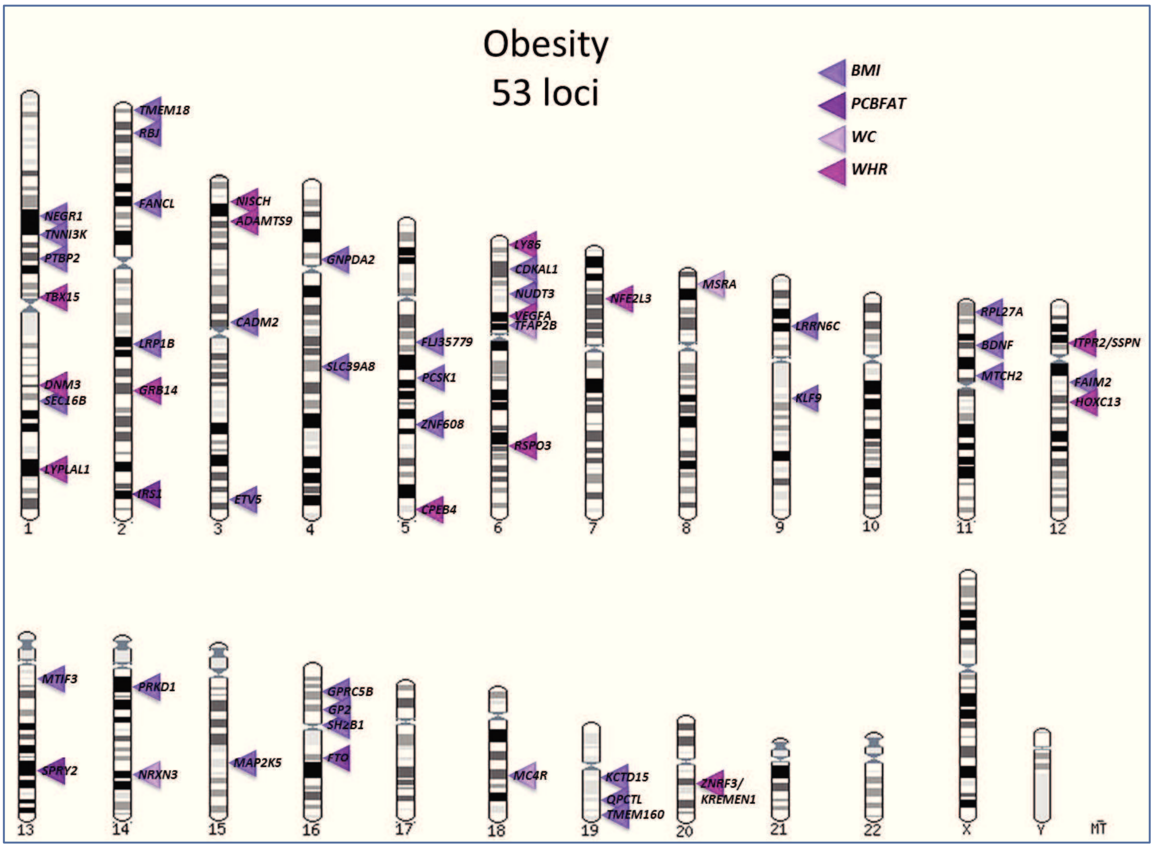


Figure 2.18: *Overview of genome-wide obesity and body fat distribution associated loci, through December 2012.*

The fact that *RPGRIP1L* and many of the other most obvious positional candidates at BMI and obesity-associated loci (*BDNF*, *SH2B1*, and *NEGR1* (neuronal growth regulator 1)) are all implicated in aspects of neuronal function is consistent with the known role of the hypothalamus in appetite regulation, and with the suspected role of other compartment of the central nervous system (CNS) in obesity. For example, BMI-associated *NEGR1* is involved in neuronal growth, whilst *SH2B1* is involved in hypothalamic leptin signalling: *Sh2b1* knockout mice are, in fact, obese and the phenotype can be rescued by targeted expression of *Sh2b1* in neurons[136].

These findings reinforce the view of common obesity as a behavioural, rather than a metabolic disorder, mediated through hypothalamic dysregulation. In contrast, equivalent studies of fat distribution, rather than overall adiposity, have highlighted candidate transcripts implicated in the regulation of adipocyte development and function[99].

Special consideration has to be given to height, an important anthropometric trait that should be taken into account when studying BMI and other obesity indices. For human adult height, a combined discovery and validation study on cohorts of about 180,000 samples identified 180 robustly associated loci, many non-randomly clustered in meaningful biological pathways, and enriched for genes that are involved in growth-related processes, that underlie syndromes of abnormal skeletal growth and that are directly relevant to growth-modulating therapies (*GH1* (growth hormone 1), *IGF1R* (insulin-like growth factor 1 receptor), *CYP19A1* (cytochrome P450 family 19 subfamily A polypeptide 1), *ESR1* (estrogen receptor 1))[137] (for a complete list of these loci see figure 2.19 and Appendix table 4).

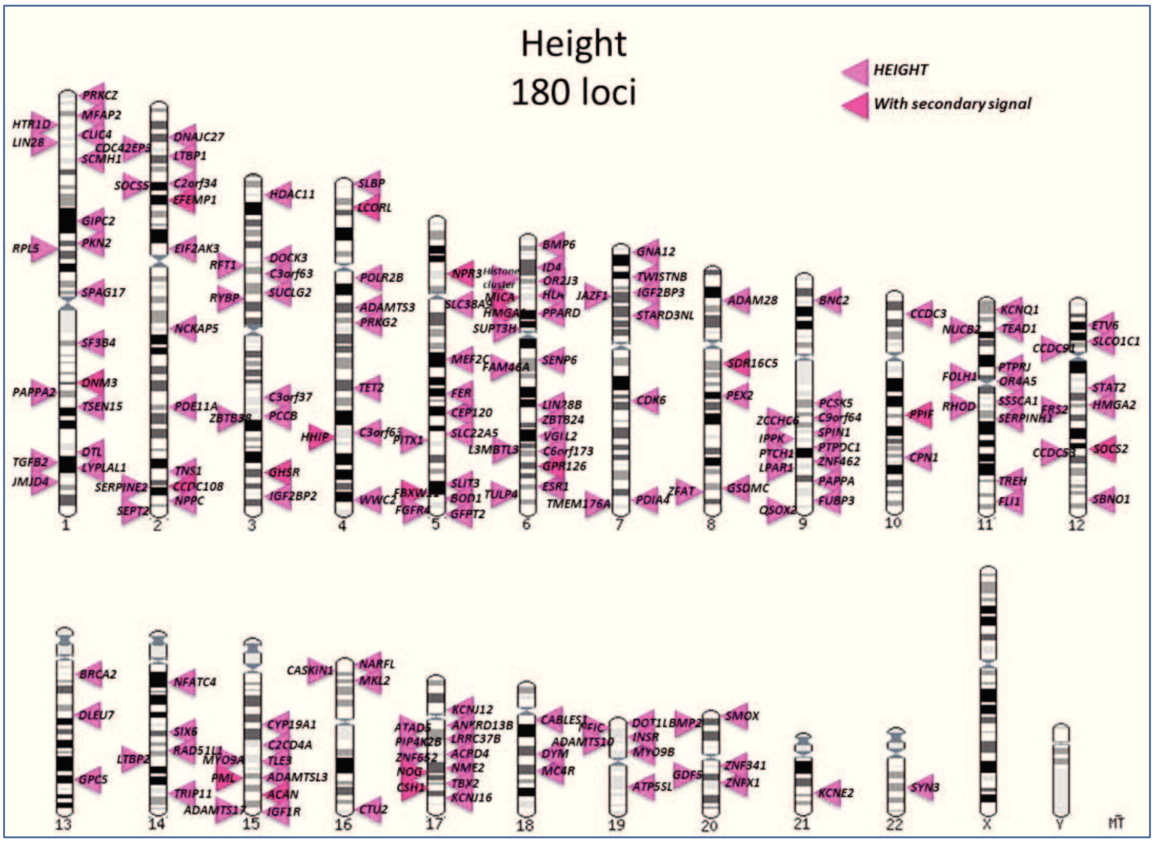


Figure 2.19: *Overview of genome-wide height-associated loci, through December 2012.*

For instance, genes such as *TGFB2* (transforming growth factor beta 2)and *LTBP1/3* (latent transforming growth factor beta binding protein 1/3) highlight a role for the TGF-β signalling pathway in regulating human height, consistent with the implication of this pathway in Marfan syndrome, a genetic disorder of the connective tissue. *Fgfr4*$^{-/-}$ *Fgfr3*$^{-/-}$ mice show severe growth retardation that is not seen in either single mutant, suggesting that the height-associated *FGFR4* (fibroblast growth factor receptor 4) variant might modify *FGFR3*-mediated skeletal dysplasias. Other genes, such as *NPPC* and *NPR3* (encoding the C-type natriuretic peptide and its receptor), influence skeletal growth in mice and likely influence also human growth[137].

Altogether, the discovered loci GW significantly associated with height explain approximately 12%–

14% of additive genetic variation (about the 10% of phenotypic variation).

2.3.1.5 Lipids

Plasma-lipid and lipoprotein levels, if high, are heritable risk factors for cardiovascular disease and targets for therapeutic intervention.

Total cholesterol (TC) represents all types of cholesterol in the blood. Clinically, a healthy level of TC is lower than 200 mg/dL; while levels equal or higher than 240 mg/dL are indicative of an elevated risk of cardiac dysfunctions.

Low-density lipoprotein cholesterol (LDL) is the fraction of TC which carries cholesterol, triglycerides, and other lipids in the blood to various parts of the body. LDL displays a positive association with atherogenesis. Atherosclerosis requires the build-up of LDL deposits in the arterial wall where they undergo oxidation and subsequent inflammatory response, leading to the formation of foam cells and further exacerbation of arterial LDL adhesion; this picture is compatible with cardiovascular disease status[138]. A healthy LDL level should be less than 100 mg/dL.

High-density lipoprotein cholesterol (HDL), instead, is a sub-group of cholesterol composed by a small, dense complex of phospholipids and apolipoproteins, including apolipoprotein A1 (APOA1), which is synthesized in the liver and which carries cholesterol, triglycerides, and other lipids in the blood from other parts of the body to the liver to be metabolised. It is negatively associated with atherogenesis, and thus helps to protect against heart disease[138]. Optimal HDL levels should be above 40 mg/dL in men and above 50 mg/dL in women.

Triglycerides (TG) are lipids composed by an ester derived from glycerol and three fatty acids, and in the blood they help the bidirectional transfer of adipose fat and blood glucose from the liver. The normal amount of triglycerides in the blood should be less than 150 mg/dL, while levels higher than 200 mg/dL are linked to atherosclerosis and heart disease.

Plasma concentrations of blood lipids are highly heritable: estimates range from 40% to 60% for total TC, LDL, HDL and TG, respectively[139], and numerous genetic studies have come, in succession, to discover heritable variants that influence their levels (for a complete list of discovered loci see figure 2.20 and Appendix table 5).

The first GWASs, involving up to 20,000 individuals of European ancestry, identified about 30 genetic loci contributing to inter-individual variation in plasma lipid concentrations[101,140-144]. Half of these loci harboured genes previously known to influence plasma lipid concentrations[145].

Among detected loci were *HMGCR* (3-hydroxy-3-methylglutaryl-CoA reductase), a well-established drug target of statins for the treatment of hyperlipidaemia; *LPA*, which encodes lipoprotein; *PLTP*, which encodes a phospholipid transfer protein; and *ANGPTL3* and *ANGPTL4* (angiopoietin-like 3 and 4), lipoprotein lipase inhibitors.

A meta-analysis for common variants associated with plasma lipids in more than 100,000 individuals of European ancestry, followed by an evaluation of mapped variants in other ethnic groups, detected a total of 95 loci significantly associated with plasma concentrations of cholesterol and triglycerides, with 59 showing genome-wide significant association with lipid traits for the first time[145].

The newly reported associations included SNPs near known lipid regulators (for example, *CYP7A1* (cholesterol 7-alpha-hydroxylase), *NPC1L1* (Niemann-Pick disease type c1 gene like 1) and *SCARB1* (scavenger receptor class B, member 1)), as well as in loci not previously implicated in lipoprotein metabolism[145]. The 95 loci contribute not only to normal variation in lipid traits, but also to extreme lipid phenotypes. Moreover, most of them also had an impact on lipid traits in three non-European populations: East Asians, South Asians and African Americans. These observations indicate that most (but probably not all) of these identified lipid loci contribute to the genetic architecture of lipid traits widely across global populations[145]. Overall variation at these 95 loci explains 10% - 12% of the total variance and 25% - 30% of the genetic variability in lipid phenotypes[139].

One of the discovered loci, *NPC1L1*, is a known drug target for the treatment of hyperlipidaemia (ezetimibe). Several other loci harbour genes that were already known to influence lipid metabolism, before this study: *SCARB1*, a HDL-receptor that mediates selective uptake of cholesteryl-ester; *CYP7A1*, which encodes cholesterol 7-alpha-hydroxylase; *STARD3* (StAR-related lipid transfer domain containing 3), a cholesterol transport gene; *LRP1* and *LRP4* (low density lipoprotein receptor-related protein 1 ad 4), members of the LDL receptor-related protein family; and *MYLIP* (myosin regulatory light chain interacting protein)*,* which protein product of is an ubiquitin ligase regulator of cellular LDL receptor levels[145].

Four novel lipid genes - *GALNT2*, *PPP1R3B*, *TTC39B* and *SORT1* (see description below) – have been validated in functionality and with experiments in mouse models.

GALNT2 (encoding UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyl transferase 2) is a member of a family of GalNAc-transferases, which transfer an N-acetyl galactosamine to the hydroxyl group of a serine/threonine residue in the first step of O-linked oligosaccharide biosynthesis. Liver-specific overexpression of mouse orthologue *Galnt2* resulted in significantly lower plasma HDL (24% compared to control mice); while reduction of the transcript level of about the 95% (knock-down) resulted in higher HDL.

Higher expression of *PPP1R3B* was related to lower plasma lipids by expression quantitative trait loci (eQTL) studies; consistently, overexpression of the mouse orthologue *Ppp1r3b* in mouse liver resulted in significantly lower plasma HDL levels.

The HDL-associated locus on chromosome 9p22, *TTC39B* (encoding tetratricopeptide repeat domain 39B), resulted in significantly higher plasma HDL levels when its orthologue (*Ttc39b*) expression were knocked-down in mice[145].

Finally, *SORT1* (sortilin 1) on chromosome 1p13 is another interesting locus, which is strongly associated with both, plasma LDL and myocardial infarction (MI) in humans. Associated variants at this locus have a minor allele frequency of about 30% in Europeans, and they are also common in other ethnicities (African Americans, Hispanics, Asian Indians and Chinese). This observation suggests that *SORT1* could be an important global genetic determinant of MI risk. Through a series of studies in human cohorts and human-derived hepatocytes, it was demonstrated that a lipid-associated common non-coding polymorphism at the 1p13 locus, rs12740374, creates a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site and alters the hepatic expression of *SORT1* gene. In mouse liver, *Sort1* alters plasma LDL and very low-density lipoprotein (VLDL) particle levels by modulating hepatic VLDL secretion: this observation provides functional evidence for a novel regulatory pathway of lipoprotein metabolism and suggests that modulation of this pathway may alter risk for MI in humans with a clinical difference of about 40% between alternative 1p13 homozygotes[146].

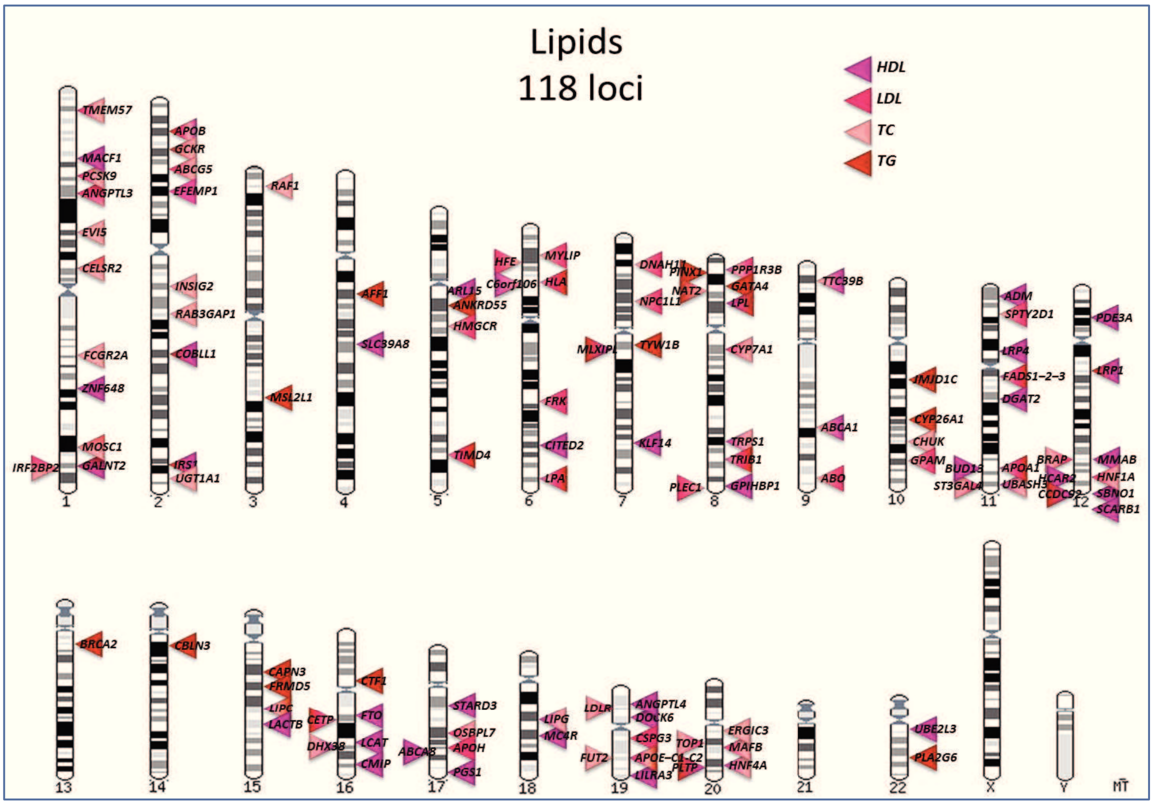


Figure 2.20: *Overview of genome-wide associated loci for lipids, through December 2012.*

Recently, to identify additional genetic associations underlying variation in plasma-lipid phenotypes, a large meta-analysis of 32 studies (comprising 66,240 individuals of European ancestry) was undertaken using a dense gene-centric approach: genotypes were in fact obtained using the candidate-gene HumanCVD BeadChip (Illumina), which is a custom gene-centric array that was designed to capture genetic diversity by using ~50,000 SNPs across ~2,000 gene regions selected, *a priori*, as primarily related to cardiovascular, inflammatory, and metabolic phenotypes[139]. Through this analysis, the authors confirmed a number of the previously reported associations and identified four, six, ten, and four unreported SNPs in established lipid genes for HDL, LDL, TC, and TGs, respectively. Several lipid-related SNPs in previously unreported genes were also identified: *DGAT2* (diacylglycerol O-acyltransferase 2), *HCAR2* (hydroxycarboxylic acid receptor 2), *GPIHBP1* (glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1), *PPARG* and *FTO* for HDL; *SOCS3* (suppressor of cytokine signalling 3), *APOH* (apolipoprotein H), *SPTY2D1* (Suppressor of Ty domain containing 1), *BRCA2* (breast cancer 2 gene) and *VLDLR* (very low density lipoprotein receptor) for LDL; *SOCS3*, *UGT1A1* (UDP glucuronosyltransferase 1 family polypeptide A1), *BRCA2*, *UBE3B* (ubiquitin protein ligase E3B), *FCGR2A* (Fc fragment of IgG low affinity IIa receptor), *CHUK* (conserved helix-loop-helix ubiquitous kinase) and *INSIG2* (insulin induced gene 2) for TC; and *SERPINF2* (serpin peptidase inhibitor clade F member 2), *C4B* (complement component 4B), *GCK*, *GATA4* (GATA binding protein 4), *INSR* (insulin receptor) and *LPAL2* (lipoprotein Lp(a)-like 2,pseudogene) for TG[139].

The most significantly associated locus for HDL in this study was *CETP* (cholesteryl ester transfer protein). CETP is a hydrophobic glycoprotein, secreted by the liver and bound mainly to HDL particles in the plasma. Its inhibition was significantly related to increased plasma HDL levels.

LDLR (low density lipoprotein receptor), the most associated locus for both LDL and TC, encodes the cell-surface LDL receptor, which removes circulating LDL via receptor-mediated endocytosis.

Finally, the locus most strongly associated with TG levels was *BUD13* (functional spliceosome-associated protein 71), located near the *APOA1-C3-A4-A5-ZNF259* cluster. In yeast, its homolog is an active spliceosome, but little is known about its function in humans. Variants in this gene have long been associated with clinical hypertriglyceridemia[139].

2.3.1.6 Blood pressure and Hypertension

Systemic blood pressure (BP) is the pressure exerted by circulating blood upon the walls of blood vessels, and is determined primarily by cardiac output and total peripheral resistance, which are controlled by a complex network of interacting pathways involving renal, neural, endocrine, vascular and environmental factors[147]. During each heartbeat, blood pressure varies between a maximum, the systolic blood pressure (SBP), and a minimum, the diastolic blood pressure (DBP).

SBP occurs near the end of the cardiac cycle when the ventricles contract; DBP, instead, occurs near the beginning of the cardiac cycle when the ventricles are filled with blood. Normal values of BP for a resting, healthy adult human are 120 mmHg SBP and 80 mmHg DBP (120/80 mmHg).

High blood pressure is defined as hypertension (HTN) and occurs when SBP is >= 140mmHg and/or DBP is >= 90mmHg. Over one billion people worldwide have hypertension and, in 2008, its prevalence was around 40% in adults aged 25 and over[148]; it is estimated that HTN contributes to 13.5 million deaths worldwide each year, and to about half the global risk for stroke and ischemic heart disease[149].

HTN is a major cardiovascular disease risk factor, but even small increments in blood pressure within the normal range are associated with an increased risk of cardiovascular damaging events and, thus, with effects on cardiovascular morbidity and mortality at the population level[149-151]: in fact, observational data indicate that a prolonged increase in DBP of 5 mmHg is associated with a 34% increase in risk for stroke, and a 21% increase in risk of coronary events[149], while 2 mm Hg lower SBP is estimated to translate into 6% less stroke and 5% less coronary heart disease[152].

Although lifestyle influences (excess salt and alcohol intake, and lack of exercise) are known to increase blood pressure and the risk of developing HTN, a substantial heritability of blood pressure, around 30–60%, has been documented, and has prompted extensive efforts to identify the contribution of genetic factors to overall disease pathogenesis[149] (for an overview of discovered loci see figure 2.21 and Appendix table 6).

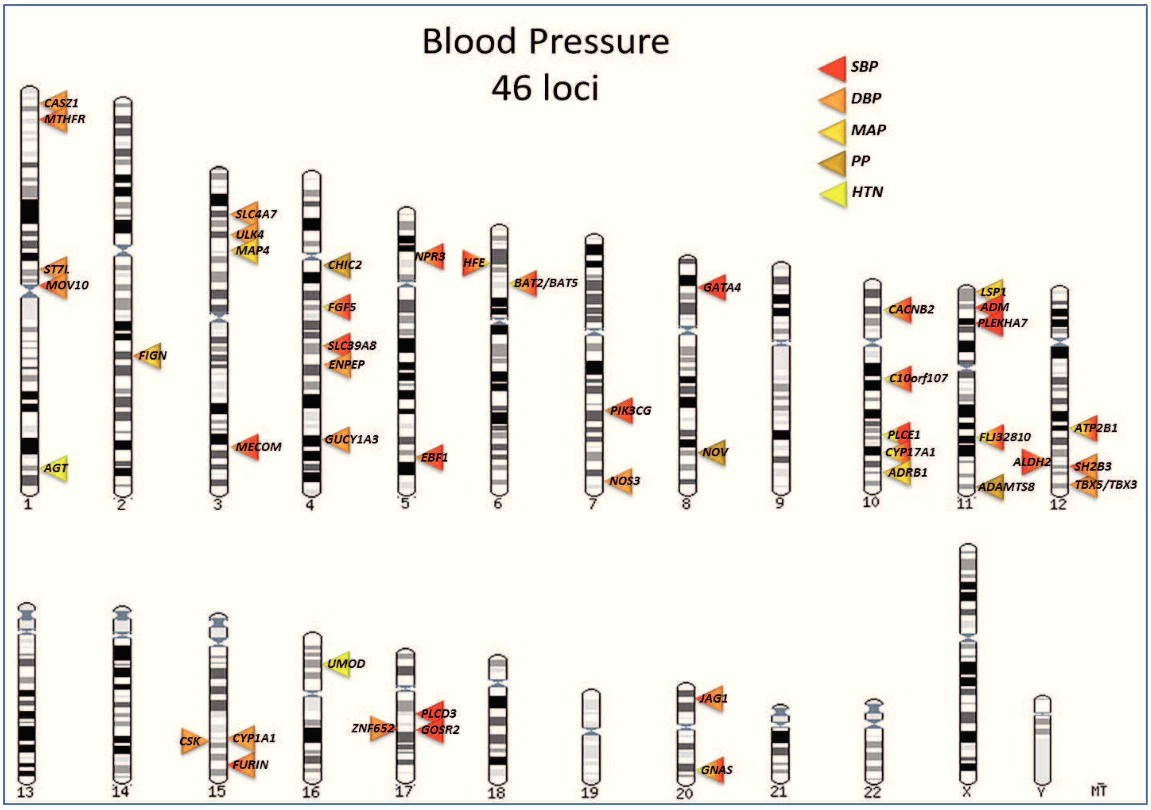


Figure 2.21: *Overview of genome-wide associated loci for blood pressure traits and hypertension, through December 2012.*

Despite considerable knowledge about pathways that are critical to blood pressure homeostasis, linkage and candidate gene studies provided limited consistent evidence of BP quantitative trait loci, identifying few variants associated with inter-individual blood pressure variation.
The study of families with rare Mendelian disorders of hypertension or of hypotension syndromes produced most notable progresses toward identifying mutations with gain or loss of function in about a dozen of genes, and other common variants with less strong effects in two additional genes, all influencing renal sodium regulation[149,152].

It was with GWASs that the majority of common genetic variation associated with BP was identified.
The first tranche of GW analyses consisted of two GWASs in European ancestry individuals within two major consortia: the Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium and the Global BPgen Consortium.
The first study identified four GW significant loci attained for SBP (*ATP2B1* (ATPase Ca++ transporting plasma membrane 1), *CYP17A1* (cytochrome P450 family 17 subfamily A polypeptide 1), *PLEKHA7* (pleckstrin homology domain containing family A member 7), *SH2B3* (SH2B adaptor protein 3), six for DBP (*ATP2B1*, *CACNB2* (calcium channel voltage-dependent beta 2 subunit), *CSK/ULK3* (c-src tyrosine kinase/unc-51 like kinase 3), *SH2B3*, *TBX3/TBX5* (T-box 3/5), *ULK4*), and one

for hypertension (*ATP2B1*). The top ten risk alleles for SBP and DBP were each associated with about a 1 and 0.5 mm Hg increase in SBP and DBP, respectively[149].

The second GWAS identified eight loci (*CYP17A1*, *CYP1A2* (cytochrome P450 family 1 subfamily A polypeptide 2), *FGF5* (fibroblast growth factor 5), *SH2B3*, *MTHFR* (methylenetetrahydrofolate reductase), *c10orf107* (chromosome 10 open reading frame 107), *ZNF652* (zinc finger protein 652) and *PLCD3* (phospholipase C delta 3)) showing genome-wide significant association with SBP or DBP, each of which was also associated with hypertension[152].

In total, the two studies recognised 13 loci associated with SBP, DBP and HTN, with a considerable concordance among top loci across all three phenotypes: for example *ATP2B1* and *CACNB2* showed significant association with SBP, DBP and HTN and *SH2B3* showed significant association with SBP and DBP.

ATP2B1 is a strong candidate gene: it encodes PMCA1, a plasma membrane calcium/calmodulin-dependent ATPase that is expressed in vascular endothelium and is involved in calcium pumping from the cytosol to the extracellular compartment. Another interesting locus is *CYP17A1*, which is also associated with a rare Mendelian form of hypertension[149].

The second tranche of GWAS for BP consisted of a multi-stage designed analysis in 200,000 individuals of European descent, which identified 29 independent SNPs at 28 loci significantly associated with SBP, DBP, or both[150]. Sixteen of the 29 SNPs were novel: six contain genes previously known or suspected to regulate blood pressure (*GUCY1A3/GUCY1B3* (guanylate cyclase 1 soluble alpha/beta 3), *NPR3/C5orf23*, *ADM* (adrenomedullin), *FURIN/FES* (furin/feline sarcoma oncogene), *GOSR2* (golgi SNAP receptor complex member 2), *GNAS/EDN3* (guanine nucleotide binding protein alpha stimulating activity polypeptide / endothelin 3), whilst the other ten provide new clues to blood pressure physiology. Of the 13 previously reported associations, only the association at *PLCD3* was not supported by the new results. Eight loci contained non-synonymous coding SNPs.

Some of the discovered signals were also replicated in individuals of different ancestry: nine SNPs were replicated in East Asians, and six in South Asians[150].

Among the discovered loci, *NPPA* and *NPPB* at the *MTHFR/NPPB* locus are particularly interesting as they encode precursors for atrial- and B-type natriuretic peptides (ANP, BNP). Three other loci harbour genes involved in natriuretic peptide and related nitric oxide signalling pathways: *NPR3*, *GUCY1A3*, and *ADM*. Two loci then have plausible connections to blood pressure via genes implicated in renal physiology or kidney disease: *SLC4A7* (solute carrier family 4 sodium bicarbonate cotransporter member 7) and *PLCE1* (phospholipase C epsilon 1). Finally, missense variants in two genes involved in metal ion transport also resulted associated: *HFE* and *SLC39A8* (solute carrier family 39 member 8)[150].

A GWAS of blood pressure extremes (extreme case-control design) identified an additional variant on chromosome 16 in the region of uromodulin (*UMOD*), where each copy of the minor G allele was associated with a lower risk of HTN, reduced urinary uromodulin excretion, better renal function, and with a 7.7% reduction in risk of CVD events. The putative role of this variant in HTN may be due to an effect on sodium homeostasis: the *UMOD* gene encodes for the Tamm Horsfall protein (THP)/uromodulin, a glycosylphosphatidylinositol (GPI) anchored glycoprotein that is the most abundant tubular protein in the urine, and is expressed primarily in the thick ascending limb of the loop of Henle (TAL), with negligible expression elsewhere[147].

Two further blood pressure phenotypes that can be studied to find genetic determinants of cardiovascular disease risk are pulse pressure (PP) and mean arterial pressure (MAP). PP is the difference between SBP and DBP and represents a measure of stiffness of the main arteries; MAP is a weighted average of SBP and DBP. Both PP and MAP are predictive for hypertension and cardiovascular disease[151]. A GW study for these two phenotypes discovered four new PP loci (*CHIC2* (cysteine-rich hydrophobic domain 2), *PIK3CG* (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma), *NOV* (nephroblastoma overexpressed) and *ADAMTS8* (ADAM metallopeptidase with thrombospondin type 1 motif 8)), two new MAP (microtubule-associated protein) loci (*MAP4* and *ADRB1*), and one locus associated with both of these traits (*FIGN*, fidgetin) that was also associated with SBP in East Asians. For three of the new PP loci, the estimated effect for SBP was opposite of that for DBP, in contrast with the majority of common SBP- and DBP-associated variants, which show concordant effects on both traits; this fact suggests the need of further investigations[151].

In 2011, using the HumanCVD BeadChip (Illumina), genotypes were tested for association with four continuous BP traits, SBP, DBP, MAP and PP, and also for association with HTN[148]. Discovery and follow-up analyses identified eight independent genetic variants associated with BP, confirming some signals at previously known loci (*LSP1/TNNT3* (lymphocyte-specific protein 1/troponin T type 3), *MTHFR/NPPB*, *AGT* (angiotensinogen) and *ATP2B1*), but also contributing to the discovery of four new loci (*NPR3*, *HFE*, *NOS3* (nitric oxide synthase 3), and *SOX6* (sex determining region Y-box 6))[148].

Further genetic studies for BP phenotypes in other ethnic groups have been undertaken. A meta-analysis of GWASs for SBP and DBP in East Asian ancestry subjects confirmed seven loci previously identified in populations of European descent, and also identified new loci (*ST7L/CAPZA1* (suppression of tumorigenicity 7 like/capping protein muscle Z-line alpha 1), *FIGN/GRB14*, *ENPEP* (glutamyl aminopeptidase) and *NPR3*) and a newly discovered variant near *TBX3*. Significant replication in an independent sample was observed for all of these loci, with the exception of *NPR3*. Additionally, an associated variant near *ALDH2* (aldehyde dehydrogenase 2 family) showed ethnic specificity, as it is not polymorphic in Europeans[153].

An extensive replication study in Japanese subjects replicated significant associations for seven loci, *CASZ1* (castor zinc finger 1), *MTHFR*, *ITGA9* (integrin alpha 9), *FGF5*, *CYP17A1*, *ATP2B1*, and *CSK/ULK3*, with any or all of the phenotypes SBP, DBP and HTN[154]. In this study the strongest association was observed for *FGF5,* a promising candidate because it encodes a member of the fibroblast growth factor family, the protein fibroblast growth factor, which is known for its effects in promoting angiogenesis in the heart.

2.3.2 Evidence of CP effects in cardiometabolic phenotypes

Findings from genetic studies, and in particular from GWASs, highlighted multiple loci that are associated with more than one cardiometabolic phenotype, suggesting shared molecular pathways. In some cases, the same variant shows association with more than one phenotype; in other cases, distinct nearby markers have indicated a multi-phenotype association pattern for a genomic region. The patterns of such multiple associations often do not follow epidemiological expectations, underscoring the importance of focused investigations about the role of pleiotropy in cardiometabolic diseases[20]. Below I refer to some examples of multiple cardiometabolic associations reported in the literature.

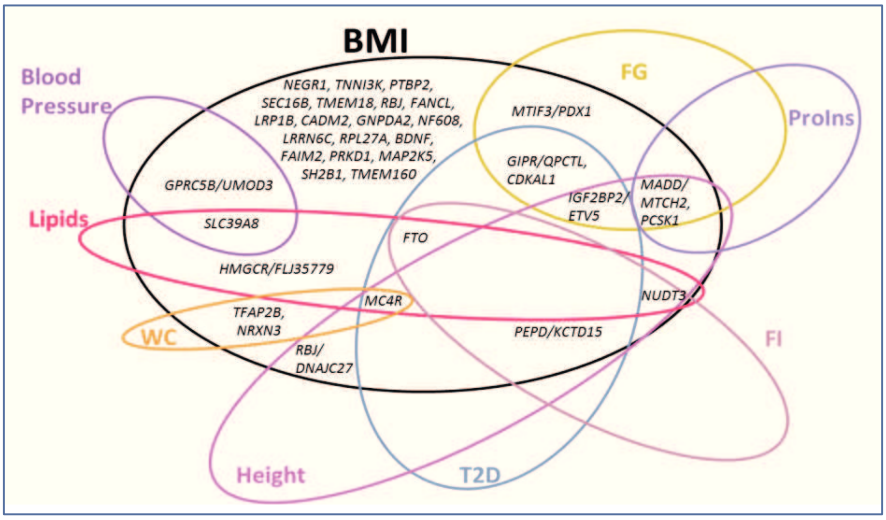


Figure 2.22: *Overlapping association signals between BMI and other metabolic traits.*

Obesity-related traits have been widely studied, and a substantial number of identified genetic associations are shared with other cardiometabolic phenotypes, in particular BMI shares 16 signals (figure 2.22) and WHR seven (figure 2.23). This is expected if we consider the biological causes and consequences of obesity[155].

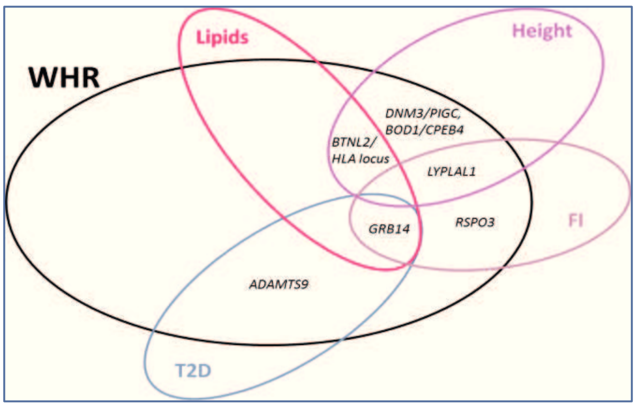


Figure 2.23: *Overlapping association signals between WHR and other metabolic traits.*

Particularly interesting is the connection between obesity and glycaemic phenotypes, especially FI. Obesity is a consequence of human conserved adaptive traits with maladaptive effects in the modern "obesogenic" environment, characterised by a chronic imbalance between caloric intake and energy expenditure, resulting in the storage of excess nutrients in white adipose tissue. With chronic over-nutrition, the storage capacity of professional metabolic tissues (white adipose tissue, liver, and skeletal muscle) is eventually exceeded, leading to cell-intrinsic and -extrinsic dysfunctions. Obesity-induced cellular dysfunction activates a diverse range of stress-responsive and counter-regulatory signalling pathways (including activation of Jun N-terminal kinases (JNK) and inhibitor of nuclear factor kB kinase subunit b (IKKb)). These pathways interact to produce two metabolically important effects: first, it converges on and inhibits insulin signalling pathways, primarily through serine phosphorylation of IRS (insulin receptor substrate) proteins; second, it initiates, supports, and augments an inflammatory response within metabolic tissues (figure 2.24)[155].

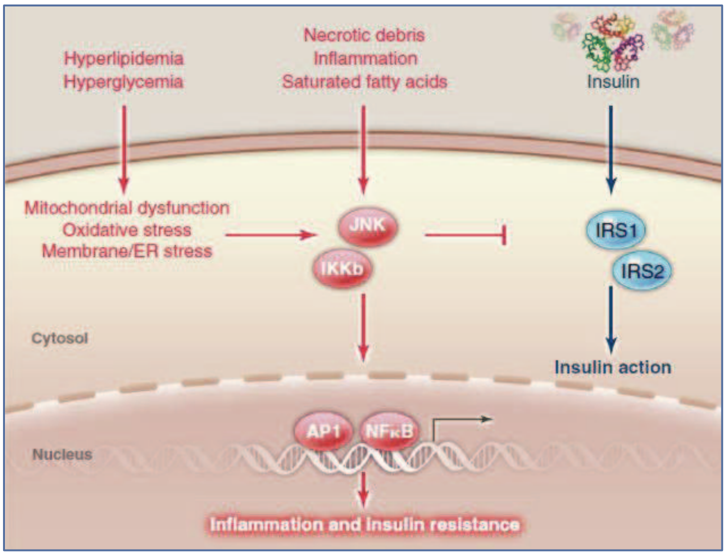


Figure 2.24: *Nutrient excess consequences through inflammatory signalling pathway and link with insulin resistance. Insulin's presence at the cell surface is transduced to cytoplasmic and nuclear responses by tyrosine phosphorylation of IRIS1 and IRIS2. Serine phosphorylation of these same proteins by JNK and IKKB, which in turn are activated by exceeded nutrient storage, however, potently inhibits insulin signalling and activates inflammatory response. From Odegaard et al. 2013*[155].

An example of shared association which relates obesity with insulin resistance is represented by the *GRB14* locus, which is associated with both WHR and FI. The protein coded by this gene is the Growth Factor Receptor-Bound Protein 14, which regulates adipose tissue distribution and consequently insulin receptor signalling in a tissue-specific negative manner[119].

Another interesting example is represented by the BMI-locus *FTO*: as I have already reported above, this locus showed significant signals for BMI, and also for T2D, lipids, FI and, as secondary effect, on the risk of coronary artery disease. Actually, *FTO* was firstly characterised as a T2D-associated locus, and only subsequently it demonstrated that association with T2D was predicated entirely by case–control differences in adiposity[99]. The exact physiological function of *FTO* is unknown, but it is believed to be involved in the regulation of food intake and to affect lipolysis in adipose tissue[139].

T2D is associated with obesity and other metabolic dysfunctions, such as cardiovascular disease. This relationship with other cardiometabolic phenotypes is also represented by a corresponding overlap of association signals (see figure 2.25).

An example is the *cis*-acting expression quantitative (eQTL) *KLF14* (Kruppel-like factor 14) locus with its association with HDL and T2D; *KLF14* is a trans-regulator of adipose gene expression, correlated with levels of several metabolic traits[19]. Another example is a pool of common genetic variants that were found to underlie T2D and hypertension in a linear mixed-effect model[81].

Some multi-phenotype associations are explained by changes of phenotype from variability within the physiological range to pathological values: this can be the case that explains the relationship and, thus, the many shared associated variants between T2D and glycaemic traits (as represented in figure 2.15). T2D and glycaemic traits GWAS meta-analyses showed directional consistency, with increased FG levels for most identified T2D signals; however, we already discussed the fact that the mechanisms responsible for the pathogenesis of T2D and those influencing physiological glucose homeostasis do not completely overlap. To this end, it is noteworthy that most of T2D risk variants are related to decreased β-cell function, while variants at only a few loci (*PPARG*, *FTO*, and *IRS1* (insulin receptor substrate 1)) are associated with increased insulin resistance. *IRS1* locus is in fact associated with increased risk of T2D, insulin resistance, along with decreased HDL, increased TG, and increased risk of cardiovascular disease[19].

T2D-susceptibility loci were also associated with other phenotypes apart from cardiometabolic traits: variants at *HNF1B* and *JAZF1* (juxtaposed with another zinc finger gene 1) showed clear effects on susceptibility to prostate cancer, while *CDKAL1* is also a susceptibility locus for Crohn's disease[100].

Lipids lipoprotein levels in plasma are related to cardiovascular disease; in fact, several HDL and TG loci are also associated with cardiovascular disease (*IRS1*, *C6orf106*, *KLF14* and *NAT2* (N-acetyltransferase 2)), suggesting that there may be selective mechanisms by which HDL or TG can be altered in ways that also modulate heart disease[19].

However, lipid traits share associated variants also with other cardiometabolic phenotypes, as represented in figure 2.26. For example, at the glucokinase regulator gene *GCKR* one common variant allele increases TG levels, but also lowers glucose levels, effects that run counter to epidemiological correlations[156].

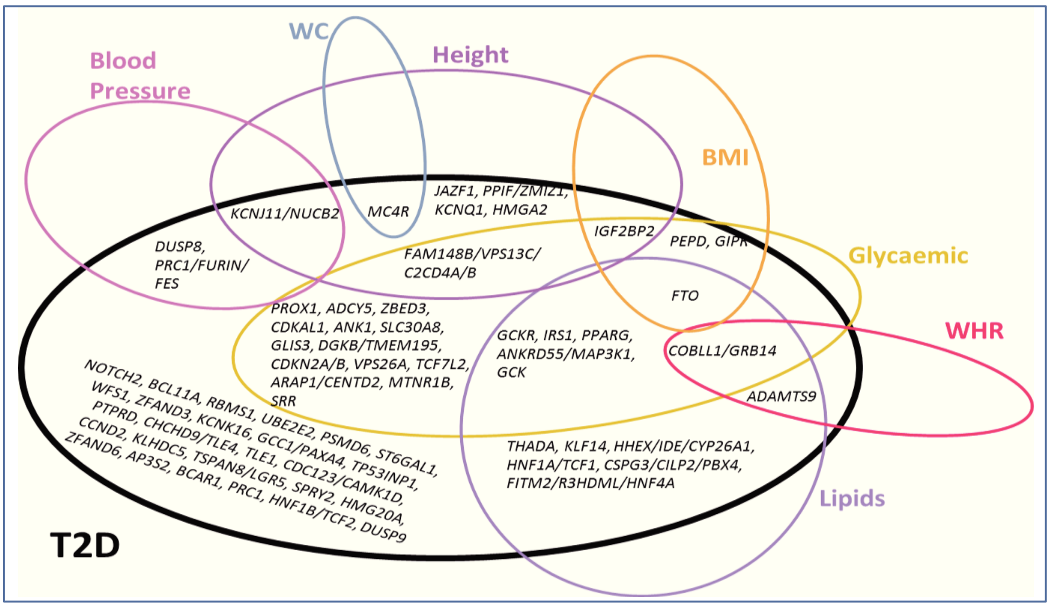


Figure 2.25: *Overlapping association signals between T2D and other metabolic traits.*

2.3.3 Relationships between cardiometabolic phenotypes

2.3.3.1 Proposed models: Metabolic Syndrome

Clinically and epidemiologically, metabolic, anthropometric and cardiovascular phenotypes are highly correlated, and are thought to be etiologically connected[157]. On one hand, quantitative metabolic traits underlie risk for several complex diseases, and are used as diagnostic criteria to define disease outcomes: this is the case of T2D, diagnosed and monitored through FG/FI levels; but also of hypertension. On the other hand, it is common to observe a concurrence of some cardiometabolic phenotypes that cluster together, in particular: increased risk of T2D, obesity, high blood pressure (BP), high triglycerides, low HDL-cholesterol levels (HDL) and insulin resistance (IR)[157].This cluster of related phenotypes is usually epidemiologically described, and it has been clinically defined as Metabolic syndrome (MetS)[158].
MetS has an estimated prevalence of 20-25% among adults around the globe. Cardiovascular disease and T2D represent the primary clinical outcome of MetS; just to give an example, in the Framingham cohort, MetS alone predicted the 25% of all new-onset cardiovascular diseases, and it is also highly predictive for new-onset diabetes. Beyond these two main outcomes, MetS individuals have been reported to be susceptible to other conditions, such as polycystic ovary syndrome, fatty liver, cholesterol gallstones, asthma, sleep disturbances, and some forms of cancer[158]. This last relationship is confirmed by the observation of some common genetic determinants for both T2D and prostate cancer[100].

In 2004, the National Cholesterol Education Program's Adult Treatment Panel III report (ATP III) identified six main components of metabolic syndrome:

- Abdominal obesity (1);
- Atherogenic dyslipidaemia: further partitioned into
    - low HDL (2),
    - high TG (3);
- Raised blood pressure (4);
- Insulin resistance with or without glucose intolerance (5);
- Pro-inflammatory state: elevations of C-reactive protein (CRP);
- Pro-thrombotic state: characterised by increased plasma plasminogen activator inhibitor (PAI)-1 and fibrinogen.

When at least three characteristics of (1-5) are present, a diagnosis of MetS can be made[158].
Recently, the link between inflammatory response and metabolism has been the subject of intense research, and two companion studies demonstrated an enrichment of immune pathways in MetS by integrating genomic and transcriptional variation[138].

The metabolic syndrome seems to have three potential etiological categories: obesity and disorders of adipose tissue that have been considered as mainly responsible for the rising prevalence of MetS and are the primary target of therapeutic intervention; insulin resistance, on which many investigators place a greater priority than on obesity; and a constellation of other factors, each of which subject to its own regulation through both genetic and acquired factors.

The genetic association of the *FTO* locus to cardiometabolic phenotypes supports the idea that obesity is one of the major risk factors for MetS: several studies have, in fact demonstrated that *FTO* genotypes are associated with MetS components to an extent entirely consistent with the *FTO* effect on BMI, and consequently that adiposity has a causal relationship on hypertension, dyslipidemia, and heart failure[89,90].

On the other hand, *KLF14* locus is a good example that strongly supports a major role of insulin resistance in MetS. It is, in fact, associated with T2D through a primary effect on insulin action, which is not driven by obesity, as well as with dyslipidaemia and heart diseases[19].

2.3.3.2 Alternative models and methods of study

Despite the great number of clinical observations, and numerous studies in the literature, abundant controversy exists about the extent of MetS and its capacity in explaining the relationships between cardiometabolic phenotypes.

In fact, the pair-wise genetic correlations between the MetS components showed large variability, and clinical exceptions to the definition of MetS have been recognised. An example is represented by metabolically healthy obesity and metabolically unhealthy leanness phenotypes. Ruderman and other researchers described metabolically obese normal-weight individuals who, despite having a normal-weight BMI, demonstrate metabolic disturbances that are typical of MetS individuals, including insulin resistance, increased levels of central adiposity, low levels of high-density lipoprotein-cholesterol (HDL) and elevated levels of triglycerides, impaired fasting glucose, and hypertension[159]. Some data suggest that this phenotype is reasonably common, with a prevalence of 3–28%[124].

Metabolically healthy obese individuals have also been described: despite having BMI > 30 kg/m^2, these subjects do not present any metabolic disease (T2D, HTN or other cardiovascular diseases), they are insulin sensitive, and lack most of the metabolic abnormalities typical of MetS[160]. Also this phenotype appears to be reasonably common, with a prevalence of 11–28%[124].

These two particular multi-phenotype conditions are interesting because they separate obesity from its usual metabolic consequences, and describe heterogeneity in the metabolic risk status of individuals with normal weight, overweight, or obesity, suggesting new pathways in cardiometabolic phenotype regulation that explains risks, independent of overall obesity, or risks associated with obesity that are independent of adiposity's intermediate metabolic abnormalities[124].

Even the complexity of the genetic association signals for metabolic phenotypes underlines an important feature of discontinuity and little consistency in the patterns of overlap, compared to that expected by common epidemiology. Overall, many genetic loci show effects on multiple phenotypes, but few of them cluster in a way consistent with a common genetic basis of MetS.

An example is the *GCKR* locus, already cited above[156].

Another unexpected pattern for cardiometabolic phenotypes was observed by Voight and colleagues: using a Mendelian randomisation approach, they found that LDL levels causally affect myocardial infarction risk, whereas high-density lipoprotein (HDL) levels do not. This counter-intuitive result can be explained by the facts that low HDL may be a consequence, rather than a cause, of myocardial infarction risk, thus contradicting the established view that increasing the levels of HDL cholesterol will uniformly lower the risk of myocardial infarction and cardiovascular disease[88].

The examples described above are explicative of the fact that MetS is just one combination of complex phenotypes and that alternatives exist.
In general this is consistent with the idea that uncovered alternative and/or combined pathways are involved in the determination of complex phenotypes, and in the relationships between them. Clarifying those pathways and relationships will shed light on the underlying cellular processes and biological mechanisms that determine diseases and physiological traits, with enormous advantages for the clinical translation into prevention, diagnosis and treatment. The study of genetic associated determinants, especially accounting for combined effects on multiple phenotypes aims to contribute to this clarification.

3 PhD Project

3.1 Preliminary data and General aim

3.1.1 Preliminary analysis: multi-phenotype effects of glycaemic loci and evidence of directional consistency

3.1.1.1 Introduction

My PhD project was envisaged after our initial observation of the overlap between the glycaemic and other cardiometabolic trait and disease loci, within the results of meta-analyses for identifying new loci influencing glycaemic traits[18].

The study combined previous discovery meta-analyses with newly available samples of European ancestry, including those genotyped using the Metabochip SNP genotyping array, for a total of up to 133,000 individuals. A follow-up meta-analysis of all included samples for 66,000 SNPs was performed, discovering 41 new glycaemic associations: 20 for fasting glucose concentration, 17 for fasting insulin concentration, and four for 2hGlu[18].
In this study we performed a series of additional analyses by testing for overlaps of significant associations and directional consistency of the effects with other metabolic phenotypes; in particular, we implemented:

- a graphical comparison of significance and direction of effects of newly discovered glycaemic SNPs in five other phenotypes (T2D, TG, HDL, BMI and WHR adjusted for BMI (WHRadjBMI));
- a binomial analysis of directional consistency of associations in follow-up results for glycaemic traits for those variants reported in Metabochip as associated with other cardiometabolic phenotypes.

The results of these analyses led us to the hypothesis about pleiotropic effects on cardiometabolic phenotypes.

3.1.1.2 Materials and Methods

False Discovery Rate analysis

When pursuing multiple inferences, researchers tend to select the most significant ones for emphasis, discussion and support of conclusions, but such a reporting usually results in a greatly increased false positive rate.
As a new point of view on the problem of multiplicity, the number of erroneous rejections (type I errors) should be taken into account in addition to the question about the number of errors made. The rate of erroneous rejections is inversely related to the number of hypotheses rejected.
A desirable error rate to control is the expected proportion of errors among the rejected hypotheses, defined as False Discovery Rate (FDR)[161].

When we test, simultaneously, m null hypotheses H0, m_0 are the true ones, and R is the number of rejected ones as represented in table 3.1.

	H0 declared non-significant	H0 declared significant	Total
True H0	U	V	m_0
False H0	T	S	$m-m_0$
Total	m-R	R	m

Table 3.1: *Number of erroneous and correct classifications when testing m null hypotheses.*

The m hypotheses are assumed to be known in advance; R is an observable variable; U, V, S and T are unobservable variables.

The proportion of errors committed by falsely rejecting null hypotheses can be viewed as the random variable $Q = V/(V+S)$, that is the proportion of erroneously rejected null hypotheses. When no error of false rejection is committed, $V+S = 0$ and therefore $Q = 0$.

FDR $E(Q)$ is the expectation of Q:

$$FDR = E(Q) = E\left[\frac{V}{V+S}\right] = E\left(\frac{V}{R}\right).$$

Considering this equation as a function of the significance level α at which the individual testing is done, FDR formula becomes:

$$FDR(\alpha) = q - value = \frac{\alpha m_0}{R(\alpha)}.$$

We applied FDR calculation to all the results in our analysis.

Graphical visualisation of associations of glycaemic trait variants with other cardiometabolic traits

For those SNPs that we identified as associated at genome-wide significance (p-value < $5x10^{-8}$) to one of the following glycaemic traits in the meta-analysis of more than 133,000 individuals - fasting glucose (FG), fasting insulin (FI), fasting insulin adjusted for BMI (FIadjBMI), two hour glucose

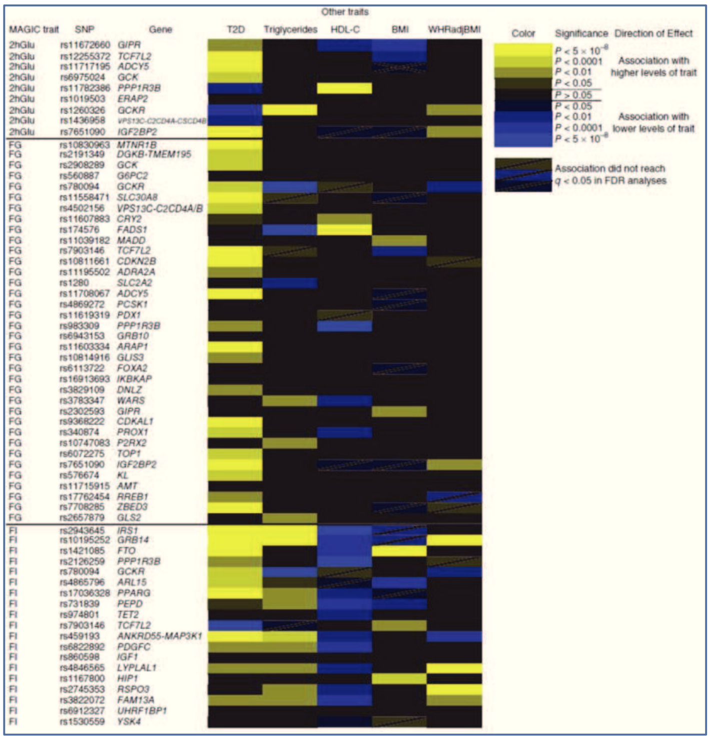


Figure 3.1: *Heat map of associations between glycaemic loci and T2D, HDL and TG concentrations, BMI and WHR. Loci associated with these phenotypes (P < 0.05) are highlighted. Those with positively correlated effect directions are shown in yellow, and those with negative correlations are shown in blue. Those which did not reach q-value < 0.05 in FDR analyses are indicated by a diagonal line through the corresponding rectangle. From Scott et al. 2012[18].*

(2hGlu) - we also investigated their association with other metabolic phenotypes and disease outcomes.

We looked-up the meta-analysis of association results for such SNPs in the latest DIAGRAM Metabochip analyses[108] for T2D and examined associations of these SNPs in publicly available data from previous studies of lipid traits from the Global Lipids Genetics Consortium (GLGC)[145] –TG, HDL and LDL cholesterol - as well as BMI and WHR from GIANT Consortium[16,126]. From these data, we extracted p-values of association and the directions of effect aligned to glycaemic trait-raising alleles. We highlighted associations with other phenotypes at p-value < 0.05, and displayed their directions using a colour code from bright yellow (very significant p-value < $5x10^{-8}$, positive association) to bright blue (very significant p-value < $5x10^{-8}$, negative association), with an intermediate black colour for non-significant associations (p-value > 0.05, figure 3.1). We also performed a false discovery rate (FDR) analysis for each trait, separately.

Analyses of directional consistency of cardiometabolic trait associations between discovery and follow-up studies

The Illumina CardioMetabochip (Metabochip) is a custom Illumina iSELECT array of 196,725 SNPs designed to support efficient large-scale follow-up analyses of putative associations for glycaemic and other metabolic and cardiovascular phenotypes (as represented in figure 3.2) and to enable the fine mapping of established loci.

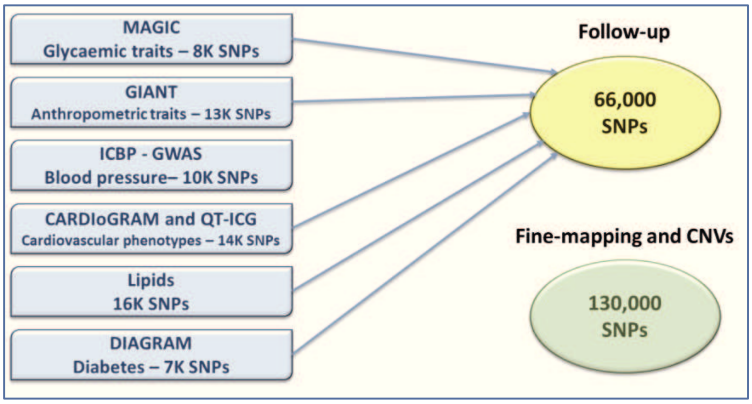


__Figure 3.2:__ Metabochip design: consortia and the number of SNPs they submitted to be followed-up using the Metabochip genotyping array. From Scott et al. 2012[18].

We investigated whether the Metabochip follow-up SNPs were likely to contain further true associations, in addition to those SNPs that reached genome-wide significance and whether more SNPs than expected by chance (50%) had a consistent direction of effect on glycaemic traits in follow-up analyses with that observed in the discovery analyses.
To do so, we performed two separate meta-analyses: the first one is of those studies involved in the original discovery analyses, comprising 42,078 individuals for fasting glucose, 34,230 for fasting insulin and 15,252 for 2hGlu; and the second one is a separately performed meta-analysis of all studies that were newly available to follow-up, comprising 85,710 individuals for fasting glucose, 69,240 for fasting insulin, and 27,602 for 2hGlu.
SNPs were filtered by LD ($r^2 < 0.01$) to identify independent variants. All SNPs in LD ($r^2 \geq 0.01$), and

those associated with glycaemic traits (FG, FI, 2hGlu, HbA1c and proinsulin) at genome-wide levels of significance (including SNPs identified in the present study), were excluded.
For each trait (FG, FI, FIadjBMI and 2hGlu), we identified all SNPs that had a nominally significant association (p-value < 0.05) in the follow-up studies alone and, for these SNPs, we performed a two-sided binomial test to test whether more SNPs than those expected by chance (50%) had a consistent direction of effect in the follow-up results with that observed in the discovery analyses. These analyses were initially performed for all 66,000 SNPs together, and then we were also able to compare across SNPs submitted to the Metabochip by different consortia (see figure 3.2), and for SNPs submitted for particular phenotypes from these consortia (table 3.2).

The results of each of these tests were plotted, overall, within SNPs from each consortium, and within SNPs submitted for follow-up of each trait (figure 3.3). We supplemented these results with FDR analyses, and noted the q-value at a p-value = 0.05 in the follow-up studies to identify the likelihood of true positives among these nominally significant SNPs.

3.1.1.3 Results

From the graphical visualisation of associations between significant glycaemic loci and T2D, HDL, TG, BMI, and WHR (figure 3.1), we observed that, in general, there is a significant effect of glycaemic loci on T2D risk: usually the increasing glycaemic trait level allele is significantly associated with increased risk of the disease. Exceptions are loci *TCF7L2* for FI, and *GCKR*, *PPP1R3B* and *VPS13C* for 2hGlu.
FI-associated loci showed also marked effects on TG levels with same directions, and opposite significant effects on HDL. FDR analysis was non-significant (q-value > 0.05) in a few cases: FI-associated variant rs7903146 in the *TCF7L2* locus for TG, and FI-associated variants in *GCKR* and *ARL15* for HDL.

For the overall follow-up study of each glycaemic trait, evaluation of the 66,000 Metabochip follow-up SNPs revealed a significant excess of SNPs showing directionally consistent associations (p-value < 0.05) compared to that expected by chance (table 3.2): FG $\text{p-value}_{binomial} = 5.01 \times 10^{-12}$, FI $\text{p-value}_{binomial} = 7.58 \times 10^{-13}$; FI adjusted for BMI $\text{p-value}_{binomial} = 9.76 \times 10^{-9}$; 2hGlu $\text{p-value}_{binomial} = 2.37 \times 10^{-6}$. FDR analyses suggested that a number of these nominal associations in the follow-up studies are true positives for fasting glucose and fasting insulin in particular (23% for FG; 24% for FI).

Notably, when we evaluated consistency of association with FI between discovery and follow-up stages among SNPs submitted to the Metabochip by other consortia, SNPs submitted by GIANT Consortium to be associated with anthropometric traits ($\text{p-value}_{binomial} = 1.52 \times 10^{-8}$), and by GLGC for lipid traits ($\text{p-value}_{binomial} = 1.15 \times 10^{-6}$), showed a marked excess of directional consistency, for BMI and triglycerides in particular (table 3.2, figure 3.3B). When we performed the same test for fasting insulin concentration adjusted for BMI, the observed enrichment among SNPs submitted by GIANT and GLGC was attenuated (table 3.2, figure 3.3C), although SNPs nominated for follow up on

TG associations remained the most significant (p-value = 3.18×10^{-7}). Of the 3,353 SNPs submitted for follow-up of TG associations, 158 SNPs showed nominal significance (p-value < 0.05) in follow-up studies and consistent direction of association with FI (adjusted for BMI) in both discovery and follow-up stages (data not shown). In 139 (88%) of these SNPs, the insulin-raising alleles were associated with higher levels of triglycerides, consistent with the positive correlations previously described between fasting insulin and triglyceride associations observed among the genome-wide significant loci for fasting insulin concentration (figure 3.1).

3.1.1.4 Discussion

From our results, the number of glycaemic loci associated with other metabolic phenotypes (q-value < 0.05; 34 of 53), also at genome-wide levels of significance (p-value < 5×10^{-8}; 14 of 53) (figure 3.1), is of particular note. Fasting insulin loci showed directionally consistent association with lipid levels (HDL and triglycerides); that is, the insulin-raising allele was associated with lower HDL and higher triglyceride levels, a hallmark combination in insulin-resistant individuals.

Further support for this notion comes from the analysis of loci nominated for the Metabochip by other consortia, and their associations with glycaemic traits. Effectively, comparing the consistency of the direction of associations for glycaemic traits between discovery and follow-up studies, we observed more directionally consistent associations than expected by chance among Metabochip follow-up SNPs; and this is particularly true when analysing FI association with those SNPs selected for BMI and TG. The significance for triglycerides SNPs remained also after BMI adjustment of FI, indicating that this association was not driven by obesity. Moreover, for 88% of triglyceride SNPs which showed consistency in directions of effects with fasting insulin, the insulin-raising alleles were associated also with higher levels of triglycerides.

These primary observations highlighted the fact that unexpected CP effects within cardiometabolic phenotypes may exist, and suggested to us the idea of deepening this outcome and developing research about the study of pleiotropy in cardiometabolic traits and diseases.

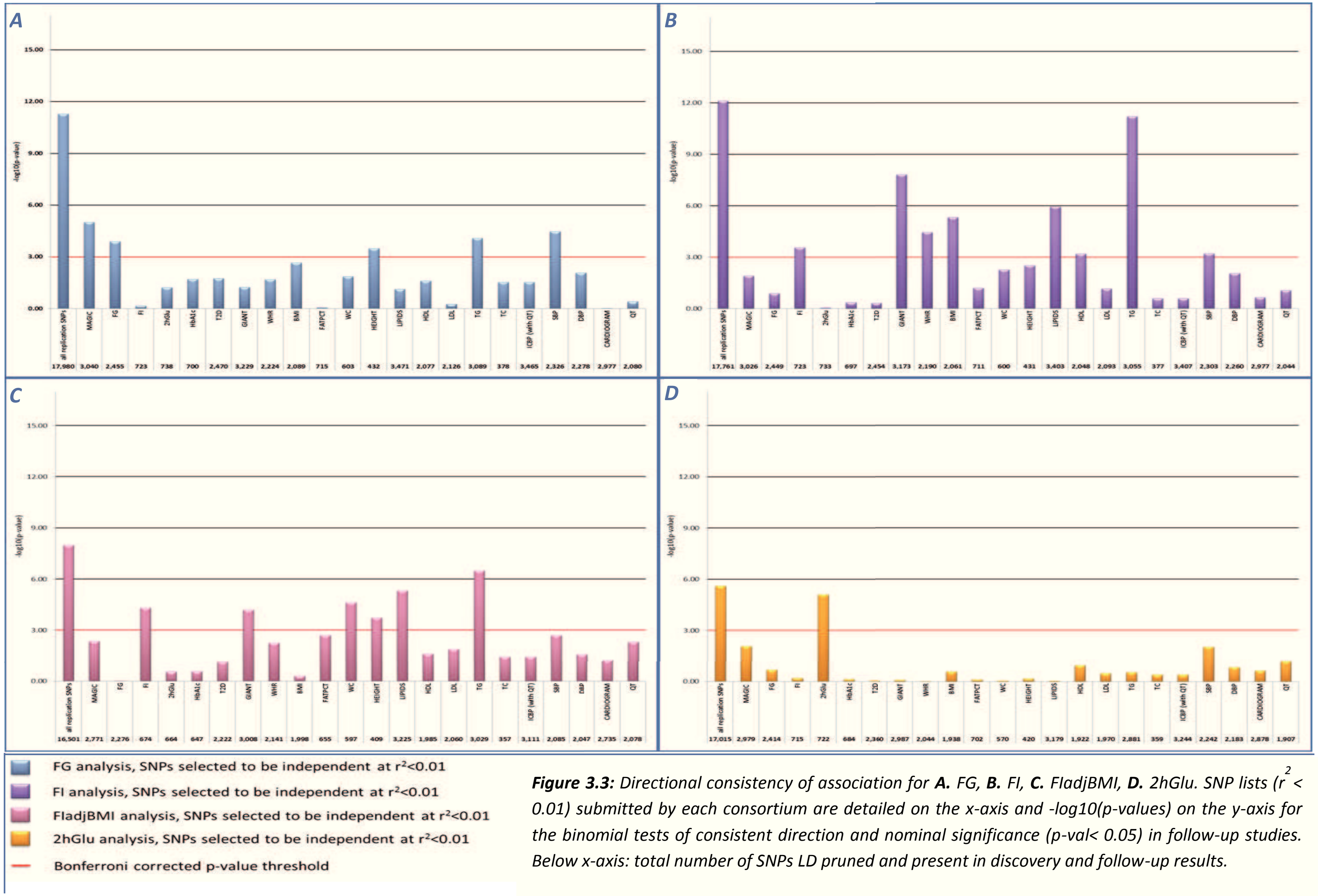


Figure 3.3: *Directional consistency of association for* ***A.*** *FG,* ***B.*** *FI,* ***C.*** *FIadjBMI,* ***D.*** *2hGlu. SNP lists (r^2 < 0.01) submitted by each consortium are detailed on the x-axis and -log10(p-values) on the y-axis for the binomial tests of consistent direction and nominal significance (p-val< 0.05) in follow-up studies. Below x-axis: total number of SNPs LD pruned and present in discovery and follow-up results.*

		Overall Follow-up SNPs	MAGIC					DIAGRAM	GIANT						GLGC					ICBP (with QT)				CARDIOGRAM
			All	FG	FI	2hGlu	HbA1c	T2D	All	WHR	BMI	FATPCT	WC	HEIGHT	All	HDL	LDL	TG	TC	All	SBP	DBP	QT interval	CAD
	Total Follow-up SNPs on METABOCHIP	65,345	8,473	5,055	1,046	1,081	1,082	5,270	13,454	5,268	5,276	1,076	1,093	1,098	15,499	5,249	5,250	5,256	971	14,717	5,269	5,267	5,244	8,636
FG	Total SNPs after LD-pruning and removing MAGIC hits	17,980	3,040	2,455	723	738	700	2,470	3,229	2,224	2,089	715	603	432	3,471	2,077	2,126	3,089	378	3,465	2,326	2,278	2,080	2,977
	Total # P<0.05 in follow-up	1,166	206	173	50	48	42	172	202	136	139	32	38	30	228	144	128	205	31	219	154	131	133	190
	q-value at P=0.05	0.77	0.74	0.71	0.72	0.71	0.79	0.71	0.80	0.81	0.74	0.89	0.76	0.70	0.76	0.72	0.82	0.74	0.59	0.78	0.74	0.82	0.77	0.77
	Total # P<0.05 in follow-up and consistent direction	701	135	112	27	31	29	102	115	82	88	17	27	25	128	86	68	131	22	126	103	81	72	96
	Binomial test P-value	5.01E-12	9.73E-06	1.30E-04	6.72E-01	5.95E-02	1.95E-02	1.78E-02	5.72E-02	2.03E-02	2.15E-03	8.60E-01	1.39E-02	3.25E-04	7.35E-02	2.41E-02	5.36E-01	8.33E-05	2.95E-02	3.04E-02	3.38E-05	8.51E-03	3.86E-01	9.42E-01
FI	Total SNPs after LD-pruning and removing MAGIC hits	17,783	3,026	2,449	723	733	697	2,454	3,173	2,190	2,061	711	600	431	3,403	2,048	2,093	3,055	377	3,407	2,303	2,260	2,044	2,977
	Total # P<0.05 in follow-up	1,167	207	156	53	46	40	156	247	173	175	57	43	47	261	160	145	250	38	230	167	142	137	173
	q-value at P=0.05	0.76	0.72	0.78	0.65	0.74	0.81	0.78	0.63	0.63	0.58	0.58	0.67	0.43	0.64	0.62	0.67	0.61	0.49	0.74	0.67	0.79	0.72	0.83
	Total # P<0.05 in follow-up and consistent direction	706	122	88	40	24	17	83	168	114	118	36	31	34	170	102	84	179	23	143	106	87	79	95
	Binomial test P-value	7.58E-13	1.22E-02	1.28E-01	2.69E-04	8.83E-01	4.30E-01	4.71E-01	1.52E-08	3.50E-05	4.66E-06	6.27E-02	5.40E-03	3.09E-03	1.15E-06	6.29E-04	6.73E-02	6.17E-12	2.56E-01	2.70E-04	6.19E-04	9.04E-03	8.71E-02	2.24E-01
FI (adjusted for BMI)	Total SNPs after LD-pruning	16,501	2,771	2,276	674	664	647	2,222	3,008	2,141	1,998	655	597	409	3,225	1,985	2,060	3,029	357	3,111	2,085	2,047	2,078	2,735
	Total # P<0.05 in follow-up	1,103	188	154	57	40	49	149	250	169	160	43	53	54	230	133	136	237	28	224	151	137	129	172
	q-value at P=0.05	0.75	0.70	0.72	0.56	0.73	0.65	0.74	0.60	0.63	0.62	0.73	0.56	0.37	0.70	0.74	0.72	0.63	0.52	0.68	0.69	0.73	0.77	0.79
	Total # P<0.05 in follow-up and consistent direction	647	114	78	44	24	29	86	157	103	85	32	42	41	150	80	83	158	20	146	95	82	81	99
	Binomial test P-value	9.76E-09	4.33E-03	9.36E-01	4.71E-05	2.68E-01	2.53E-01	7.11E-02	6.21E-05	5.46E-03	4.77E-01	1.91E-03	2.25E-05	1.75E-04	4.58E-06	2.38E-02	1.26E-02	3.18E-07	3.57E-02	6.50E-06	1.89E-03	2.60E-02	4.65E-03	5.63E-02
2hGlu	Total SNPs after LD-pruning and removing MAGIC hits	17,015	2,979	2,414	715	722	684	2,360	2,987	2,044	1,938	702	570	420	3,179	1,922	1,970	2,881	359	3,244	2,242	2,183	1,907	2,878
	Total # P<0.05 in follow-up	974	176	138	36	61	27	119	179	116	111	37	32	21	169	114	103	144	21	195	117	119	95	171
	q-value at P=0.05	0.87	0.59	0.85	0.87	0.59	0.92	0.92	0.83	0.87	0.84	0.89	0.74	0.81	0.88	0.84	0.89	0.87	0.80	0.83	0.89	0.91	0.82	0.95
	Total # P<0.05 in follow-up and consistent direction	561	106	77	20	48	15	61	92	59	62	20	15	12	83	66	46	79	8	116	73	68	57	94
	Binomial test P-value	2.37E-06	8.15E-03	2.02E-01	6.18E-01	7.67E-06	7.01E-01	8.55E-01	7.65E-01	9.26E-01	2.55E-01	7.43E-01	8.60E-01	6.64E-01	8.78E-01	1.11E-01	3.25E-01	2.79E-01	3.83E-01	9.76E-03	9.34E-03	1.42E-01	6.42E-02	2.21E-01

Table 3.2: *Directional consistency of associations between discovery and follow-up studies for glycaemic traits. The number of SNPs nominated to the Metabochip for follow-up of particular phenotypes by each consortium is shown, alongside the number of SNPs where p-value < 0.05 in follow-up studies. The number of those SNPs showing consistent direction is also shown, as well as the p-value for the binomial test comparing this number to the null expectation (50%). In addition, the q-value from FDR analyses at p-value = 0.05 is also shown. From Scott et al. 2012*[18]*.*

3.1.2 The Cross-Consortia Pleiotropy Group

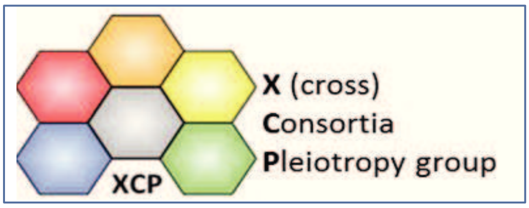


Figure 3.4: *Symbol of the XC-Pleiotropy group.*

The Cross-Consortia Pleiotropy Group (XC-Pleiotropy group, figure 3.4 for the symbol) is a "consortium of consortia" that was initiated in 2011 with the aim of investigating patterns of established multi-phenotype associations across the human genome for cardiometabolic traits and disease outcomes.

The Consortium serves as a platform for multiple GWAS consortia of cardiometabolic phenotypes (table 3.3) to share their meta-analyses results. All data are regulated by appropriate ethics oversight from their respective institutional review boards.

GWAS Consortium Name	Abbreviation	Phenotypes	Main Reference	Reference number
Diabetes Genetics Replication and Meta-Analysis	DIAGRAM	Type 2 diabetes	Morris et al. 2012; Voight et al. 2010	19, 108
Genetics of Body Fat Percentage	-	Body fat percentage	Kilpelainen et al. 2011	132
Genetic Investigation of Anthropometric Traits	GIANT	Height, BMI, waist circumference, waist-to-hip ratio	Speliotes et al. 2010, Heid et al. 2009, Lango Allen et al. 2010	16, 126, 137
Genetics of Blood Pressure	Global BPgen	Systolic and diastolic blood pressure, hypertension	Newton-Cheh et al. 2009; Ehret et al. 2011	152, 152
Global Lipids Genetics Consortium	GLGC	Total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides	Teslovich et al. 2010	145
Meta-Analyses of Glucose and Insulin-Related Traits	MAGIC	Fasting glucose and insulin with and without adjustment for BMI, two-hour glucose, fasting proinsulin, glycated Hemoglobin	Dupuis et al. 2010; Manning et al. 2012; Soranzo et al. 2009; Strowbridge et al. 2011; Saxena et al. 2010	117, 118, 119, 120, 121

Table 3.3: *GWAS Consortium partners of the XC-pleiotropy group.*

The XC-Pleiotropy group's objectives are to explore results of these meta-analyses to clarify several questions about pleiotropy; to this aim its participants are divided in different, but interacting, working groups.

One of the main objectives is to understand whether pleiotropic loci can be discovered by testing existing GWAS data using multiple-phenotype mapping methods, establishing (1) what is the potential of existing univariate analyses, (2) what is the best methodology to detect new pleiotropic associations from them, and (3) which approaches could be applied to verify the hypotheses of pleiotropy at already known cardiometabolic-phenotype loci.
Defining the fraction of established loci for metabolic traits and diseases with discernible pleiotropic effects is thus a key point, as well as evaluating the effects of pleiotropic loci in the context of established epidemiology, in particular verifying if individual pleiotropic effects are consistent with epidemiological expectation, and if pleiotropic loci form clusters of phenotype correlations that

match the epidemiological expectation. The group aims to aid the interpretation of variation at established genomic regions with evidence of pleiotropic effects, and to the evaluation of the structure of pathways around similar pleiotropic loci.
The consortium aims, in addition, to dissect the underlying architecture for those genomic regions showing adjacent multiple signals of univariate associations with cardiometabolic traits and disorders through, for example, the development of methods for distinguishing allelic heterogeneity from potential pleiotropy.

In the context of the XC-Pleiotropy group, and on its behalf, I started my work on the study of pleiotropy and I developed my PhD project in an attempt to achieve some of the above listed objectives of the consortium.

3.1.3 Aims of my PhD project

Since I started my University studies, I developed a deep interest in human genetics.
During the first year of my PhD I became involved in work investigating the genetic background of complex human diseases: I worked in the project investigating the association between genetic variants and Aggressive Periodontitis, a complex human disease involving the Immune system, with a particular focus on detecting the genotype-genotype interactions underlying disease predisposition.
While doing this work, I realised the complexity of studying multifactorial complex phenotypes and the necessity of developing appropriate methodological and statistical approaches and to explore new areas of research for the analysis of the genetic data.
During the period from November 2011 to December 2012, I undertook research training at the Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford. It was there that I started studying Type 2 Diabetes (T2D) and glycaemic traits in non-diabetic individuals, as well as the framework of multiple T2D-related cardiometabolic and inflammatory phenotypes; in this context we also developed the project on the study of pleiotropy.

Deepening the study of present GWAS for cardiometabolic phenotypes, it was clear that there is considerable overlap between associated loci, as reported, for example, in our work on glycaemic loci described above[18]; but the patterns of multi-phenotype associations resulted very complex and this is evident in chapters “2.3.2_Evidences of CP effects in cardiometabolic phenotypes” and “2.3.3_Relationships within cardiometabolic phenotypes”. This complexity of the observed metabolic trait associations within univariate analyses might be due to several underlying factors, as explained in chapter “2.1.2_Cross-Phenotype association and definition of pleiotropy”, including pleiotropy.
The phenomenon of pleiotropy refers to genetic variants exerting their effects on multiple phenotypes (in our case cardiometabolic); combinations of such effects might, or might not, follow epidemiological expectations and therefore add complexity to the aetiology of complex human

traits and disease outcomes. Our idea is that the dissection of pleiotropy will help uncover the mechanistic basis of the pathogenetic processes leading to T2D and cardiac diseases; moreover, the definition of specific sets of effects on combinations of cardiometabolic and inflammatory phenotypes might highlight novel biological pathways, targets for translational research, for therapeutic intervention, and for the understanding of the pathophysiology of human metabolism.

Based on this hypothesis, and in collaboration with the XC-pleiotropy group, my PhD project mainly focused on exploration of the pleiotropic effects at common variants across the genome on cardiometabolic phenotypes, with the objective of understanding how DNA sequence variation influences risk of metabolic diseases, with a particular focus on the impact of variants that influence multiple phenotypes and the mechanisms underlying those multiple effects.
The research has been divided into three specific aims, and thus three sub-projects:
(1) first of all, we wanted to explore established multi-phenotype effects at cardiometabolic loci from published results of univariate meta-analyses, defining clusters of loci with similar multiple-phenotype effects, comparing them to known epidemiological expectations, and identifying enriched biological networks within the most interesting clusters;
(2) secondly, we applied a strategy for dissecting the architecture of established cardiometabolic loci showing multiple associations for a better definition of the underlying mechanisms of these multi-phenotype effects, and for the discernment of potential pleiotropy from allelic heterogeneity;
(3) the third sub-project aimed to develop and apply a statistical strategy for multivariate analyses of CP phenomena in cohorts from the ENGAGE consortium to verify *a priori* hypotheses of pleiotropy, and to discover new uncovered multiple associations.

3.2 Project 1: Clustering and pathway analysis of univariate GWAS results for the detection of pleiotropic effects

3.2.1 Introduction and Aim

As reported in the "2_Literature Review" section of this thesis, cardiometabolic continuous traits are related to phenomena of dysmetabolism, which are considered as epidemics in the world.
Since the first studies on cardiometabolic disorders, it was noticed that several of them commonly clustered together and in 2004, metabolic syndrome (MetS) was defined[158].
Metabolic disorders and related traits have been studied in genome-wide association studies (GWAS) during the past seven years, resulting successful in the identification of common genetic variants associated with these phenotypes: several hundreds of loci have been identified (187 variants/108 loci for lipids, 99 variants/67 loci for glycaemic traits, 59 variants/53 loci for obesity, 65 variants/46 loci for blood pressure and hypertension, 85 variants/64 loci for T2D). A subset of these variants has shown to be associated with more than one of these phenotypes, thus corresponding to potentially pleiotropic loci. However, the patterns of phenotype associations observed in GWAS at individual cardiometabolic risk-loci are highly variable and, in addition, the overlap of genetic associations is not always consistent with epidemiological correlations (for a more complete description of all these aspects, see chapter "2.3_Overview of genetics of cardiometabolic phenotypes").

In this study, on behalf of the XC-Pleiotropy Group, we aimed to extend the analysis applied in Scott et al. 2012[18] to investigate patterns of multiple cardiometabolic phenotype associations across the genome using existing univariate analysis results.
First, we wanted to test the capability to detect groups of loci with shared cross-phenotype effects by analysing simultaneously individual effects on multiple traits and diseases extracted from existing data and using unexplored simple statistical and graphical instruments.
Our objective was also to evaluate pleiotropic loci in the context of established epidemiology, verifying when potential pleiotropic loci form clusters of phenotype correlations that match epidemiological expectations and when not, considering the difference in magnitudes of observed effects between related phenotypes and how this can influence the power to detect pleiotropic associations.
Using univariate GWAS meta-analysis data for established loci, we wanted to achieve a systems-level understanding of the role of potentially pleiotropic loci by exploring functional interactions between codified proteins through pathway analysis. These connections form networks that enable viewing of a given set of genes as something more than just a static collection of distinct genetic functions. Protein association network information can aid in the interpretation of functional genomics data and, furthermore, has also proven surprisingly useful for the detection and characterisation of disease genes, both for Mendelian and for complex diseases[162,163].

We aimed to test different methods to identify specific mechanisms evaluating the structure of pathways and networks around potential pleiotropic loci. To this purpose, we used several software packages that reconstruct networks enriched for connectivity across clusters of loci using information from literature, protein-protein interaction databases, expression and annotation databases. This analysis can help in answering some important questions. For example, are there clusters of traits and respective pleiotropic loci that impact the same pathways? Which pathways are more enriched within potential pleiotropic loci? Can pathway connectivity in multi-phenotype networks suggest gene candidates for causality or tissues of action underlying the association signals?

To summarise, in the present project, we undertook (1) the examination of associations at established cardiometabolic loci with epidemiologically correlated cardiometabolic phenotypes, by grouping shared patterns of individual trait or disease effects; subsequently we (2) compared the observed combinations of effects at identified groups of loci with our expectations based on epidemiological knowledge of cardiometabolic phenotypes; finally we (3) defined pathways and gene networks involved in the phenotypic variability within the identified association pattern groups.

3.2.2 Materials and Methods

3.2.2.1 Starting data: cardiometabolic univariate meta-analyses results

Through the XC-Pleiotropy Group we have priority access to association summary statistics from published GWAS discovery meta-analysis on cardiometabolic phenotypes.

These data were shared by six cardiometabolic trait and disease consortia as reported in table 3.3. Each study was approved by their local ethics board and each participant provided written, informed consent.

We used already published genome-wide meta-analysis association studies results for 22 cardiometabolic phenotypes, 20 quantitative traits and 2 diseases, in European samples from the six international consortia as reported in table 3.4: 5 traits from GIANT, 1 from the Body Fat Percentage consortium , 3 phenotypes from the Global BPgen consortium, 4 from the GLGC, 8 from MAGIC and 1 disease from DIAGRAM.

For 3 traits from GIANT (HIP, WC, WHR), and for 4 traits form MAGIC (FG, FI, HOMAB, HOMAIR) consortium, we also considered phenotypic refinements through adjustment for BMI (HIPadjBMI, WCadjBMI, WHRadjBMI, FGadjBMI, FIadjBMI, HOMABadjBMI, HOMAIRadjBMI), raising the number of evaluated phenotypes to 29.

Sample sizes for phenotypes varied from 10,382 individuals for fasting proinsulin to 183,727 for height. We employed the GWAS meta-analysis association results for these phenotypes to extract effects and p-values of established associated SNPs.

Consortium (abbreviation)	Complete phenotype name	Abbreviation	Paper of publication	Paper reference number	Sample size (average)
GIANT	Body Max Index	BMI	Speliotes et al. 2010	16	108,156
	Waist Circumference	WC	-	-	74,825
	Hip Circumference	HIP	-	-	66,712
	Waist-Hip Ratio	WHR	-	-	66,326
	Waist Circumference adjusted for BMI	WCadjBMI	-	-	75,084
	Hip Circumference adjusted for BMI	HIPadjBMI	-	-	-
	Waist-Hip Ratio adjusted for BMI	WHRadjBMI	Heid et al. 2009	126	113,636
	Height	HEIGHT	Lango Allen et al. 2010	137	183,727
-	Body fat percentage	PCBFAT	Kilpelainen et al. 2011	132	31,159
Global BPgen	Diastolic Blood Pressure	DBP	Newton-Cheh et al. 2009	152	28,466
	Systolic Blood Pressure	SBP	Newton-Cheh et al. 2009	152	28,424
	Hypertension	HTN	Newton-Cheh et al. 2009	152	16,820
GLGC	High Density Lipoprotein	HDL	Teslovich et al. 2010	145	88,754
	Low Density Lipoprotein	LDL	Teslovich et al. 2010	145	84,685
	Total Cholesterol	TC	Teslovich et al. 2010	145	88,754
	TryGlicerides	TG	Teslovich et al. 2010	145	85,691
MAGIC	2 hour Glucose adjusted for BMI	HGLUadjBMI	Saxena et al. 2010	118	42,854
	2 hour Insulin adjusted for BMI	HINSadjBMI	-	-	-
	Fasting Glucose	FG	Manning et al. 2012	119	50,510
	Homeostasis Model Assessment for Beta cell function	HOMAB	Manning et al. 2012	119	-
	Fasting Insulin	FI	Manning et al. 2012	119	44,972
	Homeostasis Model Assessment for Insulin Resistance	HOMAIR	Manning et al. 2012	119	-
	Fasting Glucose adjusted for BMI	FGadjBMI	Manning et al. 2012	119	51,785
	Homeostasis Model Assessment for Beta cell function adjusted for BMI	HOMABadjBMI	Manning et al. 2012	119	-
	Fasting Insulin adjusted for BMI	FIadjBMI	Manning et al. 2012	119	46,271
	Homeostasis Model Assessment for Insulin Resistance adjusted for BMI	HOMAIRadjBMI	Manning et al. 2012	119	-
	Fasting Pro-insulin	PROINS	Strowbridge et al. 2011	121	10,382
	Glycated Haemoglobine	HBA1C	Soranzo et al. 2009	120	30,587
DIAGRAM	Type 2 Diabetes	T2D	Voight et al. 2010	19	26,288

Table 3.4: *GWAS discovery meta-analyses for cardiometabolic phenotypes used in the present study.*

3.2.2.2 Selection of variants at cardiometabolic loci

As first step of this study, after a systematic literature search using PubMed and NHGRI catalogue[7], we listed all genome-wide significant (p-value < $5x10^{-8}$) SNPs reported from published GWAS for cardiometabolic phenotypes (before October 2012); secondary signals, that are additional peak signals of association detected after conditioning the genome-wide association analysis on previously detected main signals, were included. For a complete list of these SNPs see Appendix tables 1, 2, 3, 4, 5 and 6.

Among 687 identified association signals, there were 623 distinct polymorphisms.

Using SNAP internet tool[164], we calculated the pair-wise linkage disequilibrium (LD) between adjacent polymorphisms using 1000 Genomes CEU data (pilot phase)[165] as reference panel. Redundant SNPs were then removed using an LD cut-off of $r^2 \geq 0.8$. The resulting set of 547 SNP variants was used for subsequent analyses.

3.2.2.3 Alignment of multi-phenotype effects and meta-analysis of multiple association

Omnibus p-value calculation through Fisher's omnibus test as a simple multi-phenotype meta-analysis

A meta-analysis combines association summary statistics from different studies to provide a summary result and it can be applied to different phenotype analyses, in our case for CP effect detection.

We decided to apply one of the simplest meta-analytical approaches based on aggregation of p-values across phenotypes in different studies: the Fisher's omnibus test[49].

For each cardiometabolic variant a cumulative association statistic S_{cum} was calculated with the following formula:

$$S_{cum} = -2 \times \sum_{i=1}^{N} ln p_i.$$

The statistic was calculated from univariate p-values of all 29 cardiometabolic phenotypes from GWAS meta-analysis results. When a variant was not reported for a particular phenotype, its value was considered as missing and S_{cum} was calculated only for the remaining phenotypes.

S_{cum} follows the χ^2 distribution with $2N$ df[50] and tests the null hypothesis that the genetic variant is not associated with any phenotype versus the alternative hypothesis that it is associated with at least one phenotype. As we already knew that each of the selected variants was associated with at least one cardiometabolic phenotype, we used S_{cum} to verify the presence of multiple significant or suggestive associations at same variants where cumulative p-value resulted more significant than the single univariate ones.

Z-score calculation

From GWAS meta-analyses results of the 29 available phenotypes we extracted the summary statistics for the 547 listed cardiometabolic SNPs and we aligned the effects based on the HDL rising allele. HDL was chosen as reference arbitrarily.

For each listed SNP we obtained the z-score value for each cardiometabolic phenotype as calculated from beta (β) and standard error (SE) summary statistics from GWAS meta-analyses results, with the following formula:

$$z_{score} = \frac{\beta}{SE};$$

z-score was used to take into account the size of the effect (represented by β parameter) and its significance (represented by the division for the SE). An absolute value of z-score more than or equal to 5.45 corresponds to a p-value less or equal to the GW significance threshold (p-value = 5×10^{-8}). A positive value of z-score means increasing effect, while negative values are indexes of decreasing effects.

We decided to do not apply a multi-phenotype meta-analysis of the effect statistics (as for example z-score) for two main reasons:

(1) on one hand, fixed-effects meta-analysis assumes that the tested genetic variant has the same underlying effect on each phenotype, and that the observed differences are due to chance alone; this assumption is not applicable to multiple phenotypes considered in this study, since we know from epidemiological observations that some of them may have opposite effects (for example HDL and other lipids); therefore, the application of a fixed-effects meta-analysis on our data would have represented an excessive approximation that is far from the reality;

(2) on the other hand, random-effects meta-analysis or subset-based meta-analysis, which allow the genetic effect to differ across phenotypes (the first method) or to be opposite (the second method) would lead to an excessive loss of power, due to the substantial number of phenotypes included in the study.

We therefore analysed multi-phenotype z-score statistics using a different approach, as explained below.

Used software

R software[166] (R version 3.0.1 (2013-05-16)) was used to run mentioned statistical analyses. It is available at http://cran.r-project.org/.

3.2.2.4 Clustering of cardiometabolic loci effects on multiple phenotypes

Clustering method

As described above, we obtained a cardiometabolic multi- phenotype combined effect matrix of z-score values for the list of cardiometabolic SNPs.

Using this matrix of data, we applied a hierarchical agglomerative clustering method, using Euclidean distance between effects, to group together variants with more similar behaviour on cardiometabolic phenotypes and, thus, to identify clusters of cardiometabolic variants with shared multiple effects. We opted for this method because we did not know how many groups of similar loci we could observe within our data. In fact, in contrast with other clustering algorithms such as k-means or k-medoids clustering, hierarchical clustering approach does not require any *a priori* specification of the number of groups to be searched.

Agglomerative clustering algorithms begin with every of the N observations representing a singleton cluster. At each of the $N - 1$ steps, the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next highest level. This union can be graphically represented by two branches joining the two clusters into a unique node: the graph obtained in this manner from the hierarchical clustering analysis is called "dendrogram". Through hierarchical clustering, the entire hierarchy represents an ordered sequence of groupings.

Single linkage (SL) agglomerative clustering takes the intergroup dissimilarity to be that of the closest (least dissimilar) pair; this is also often called the nearest-neighbour technique. Complete linkage (CL) agglomerative clustering (furthest-neighbour technique) takes the intergroup dissimilarity to be that of the furthest (most dissimilar) pair. Group average (GA) clustering uses the average dissimilarity between groups[167].

Having different sample sizes for different phenotypes, distances between groups of loci with similar behaviour could be underestimated due to weaker effects in some phenotypes; considering this, we decided to perform the complete linkage method for hierarchical agglomerative clustering as it is based on the maximum differences, partially skipping the bias caused by different sample sizes.

The obtained hierarchical cluster was subsequently evaluated via multi-scale bootstrap resampling: 10,000 bootstrap replicates were generated to compute a probability (%, bootstrap value) as index of the strength of each dendrogram node.

Sub-cluster sets definition

As described above, hierarchical clustering methods give an ordered sequence of groupings without defining how many groups are best representative of the data. It is up to the user to decide which

level (if any) actually represents a "natural" clustering in the sense that observations within each group are sufficiently more similar than observations belonging to different groups at that level[167].
Several methods exist to calculate the best number of sub-clusters in a hierarchical cluster dendrogram, such as that proposed in the R package "dynamicTreeCut" by Peter Langfelder and colleagues[168]. However these methods are extremely dependent on input parameters in our dataset, leading to very different results as a consequence of minimal changes in their setting.
Therefore, we chose to apply a constant height cut-off at three different levels of the Euclidean distance and to then compare the three different sets of groups obtained. The chosen cut-off levels were at the 25% of Euclidean distance (cut-off A), at the 20% of Euclidean distance (cut-off B), and at the 15% of Euclidean distance (cut-off C).

Used software
R software[166] packages hclust and pvclust (R version 3.0.1 (2013-05-16)) were used to run clustering analyses and groups definition. R is available at http://cran.r-project.org/. For a description of the hclust package see http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html; for a description of the pvclust package see http://www.is.titech.ac.jp/~shimo/prog/pvclust/.

In parallel, we performed hierarchical complete clustering using the Genesis software[169], a package originally developed as a platform independent Java package of tools to simultaneously visualise and analyse a whole set of gene expression experiments. Results from this parallel analysis were compared with those obtained with R packages. The Genesis software is available at http://genome.tugraz.at/genesisclient/genesisclient_description.shtml.

3.2.2.5 Pathway analysis

We considered the obtained groups of SNPs with similar cardiometabolic multi- phenotype effects from each of the three sets based on different height cut-offs of the Euclidean distance applied to the results of the cluster analysis. For each of these groups of clustered variants we wanted to verify the presence of enriched common biological networks and test the significance of those enrichments. We thus conducted a pathway analysis using different web tools.

DAPPLE
DAPPLE (Disease Association Protein-Protein Link Evaluator) is a programme that looks for significant physical connectivity among proteins encoded by genes according to protein-protein interactions reported in the literature. It is based on the InWeb database[95], which combines reported protein interactions from the Molecular INTeraction database (MINT), the Biomolecular Interaction Network Database (BIND), IntAct, Kyoto Encyclopedia of Genes and Genomes (KEGG) annotated protein-protein interactions (PPrel), KEGG Enzymes involved in neighbouring steps (ECrel), Reactome and others. It is particularly developed to study genes in loci associated with diseases, as its hypothesis is that causal genetic variation affects a limited set of underlying mechanisms that are detectable by protein-protein interactions.
Contrary to the majority of tools for pathway analysis, DAPPLE takes as input a list of seed SNPs or

genomic regions, and applies an algorithm to define the nearest genes in a flanking region defined by the user; therefore it was particularly applicable for our study where an input was represented by the list of all SNPs in a defined sub-cluster with common multi- phenotype effects.

After defining nearest input genes, DAPPLE uses the information included in the databases mentioned above[170] for proteins encoded by these genes to build direct and indirect interaction networks. In direct interactions, any two associated proteins can be connected by exactly one edge, while in indirect interactions, associated proteins can be connected via common interactor proteins not present in the input data, but shared among associated proteins.

DAPPLE represents the constructed network in a graphical image as reported in the example in figure 3.5: input genes are represented as coloured circles, while additional connectors are in grey.

Furthermore, DAPPLE calculates several metrics to evaluate network properties and assesses the statistical significance of these network connectivity parameters using a within-degree node-label permutation method. The calculated metrics can be divided into two categories: edge metric and node metrics. The edge metric is the direct network connectivity parameter, defined as the number of edges in the direct network. We interpreted direct network connectivity as the frequency with which different loci harbour proteins that directly bind each other. Node metrics include associated protein direct connectivity and associated protein indirect connectivity, which refer to the number of distinct loci an associated protein can be connected to directly and indirectly, respectively, and common interactor connectivity, which refers to the average number of proteins in distinct loci bound by common interactors in indirect networks. A more tightly clustered network might be enriched for both edge and node metrics[170].

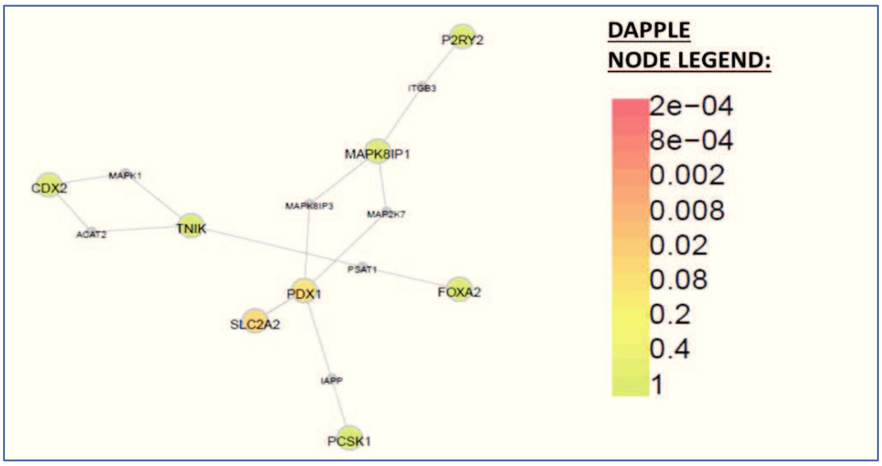


Figure 3.5: *Example of graphical output from DAPPLE pathway analysis. On the right there is the reconstructed network: coloured circles represent input genes, their colour is proportional to their p-value significance of inclusion in the network, as represented in the legend on the right. Grey circles are interactors added by the programme as connectors for indirect interactions between input genes.*

Individual scores for each protein are calculated and reported in the graphical output using a colour code (see legend in figure 3.5) for input genes. The individual protein scores for interactor factors, similarly calculated, can be used to propose candidate related genes.

Several parameters can be set; we run DAPPLE pathway analysis using the default parameters and considering genes in +/- 50kb regions flanking input SNPs.

DAPPLE is an internet tool available at http://www.broadinstitute.org/mpg/dapple/dapple.php.

STRING

The Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org/newstring_cgi/show_input_page.pl?UserId=d0QyhUDToyxf&sessionId=x9_KU35utwtG) provides uniquely comprehensive coverage and ease of access to both experimental and predicted interaction information, derived from a large number of databases: Clusters of Orthologous Groups (COG), Ensembl, IntAct, RefSeq, PubMed, Reactome, Database of Interacting Proteins (DIP), Biological General Repository for Interaction Datasets (BioGRID), MINT, KEGG, Saccharomyces Genome Database (SGD), FlyBase, SwissProt/UniProt, SwissModel, HUGO, Online Mendelian Inheritance in Man (OMIM), NCI/Nature Pathway Interaction Database (PID), RCSB Protein Data Bank (PDB), The Interactive Fly, BioCyc, Gene Ontology, Similarity Matrix of Proteins (SIMAP). The main strengths of STRING lie in its unique comprehensiveness, as well as in its confidence scoring calculation, and its interactive and intuitive user interface. Interactions in STRING are not limited to direct, physical interactions between two proteins, but they also account for possible genetic interactions, transcriptional or post-transcriptional regulation, contribution to larger structural assemblies, or involvement in subsequent steps in a metabolic pathway (functional interactions). The complete sets of associations are assembled into a large network, which captures the current knowledge on the functional modularity and interconnectivity[160,161]. An example is reported in figure 3.6: circles are input proteins and all variegate information about connections is represented by edges of different colours, as explained in the legend.

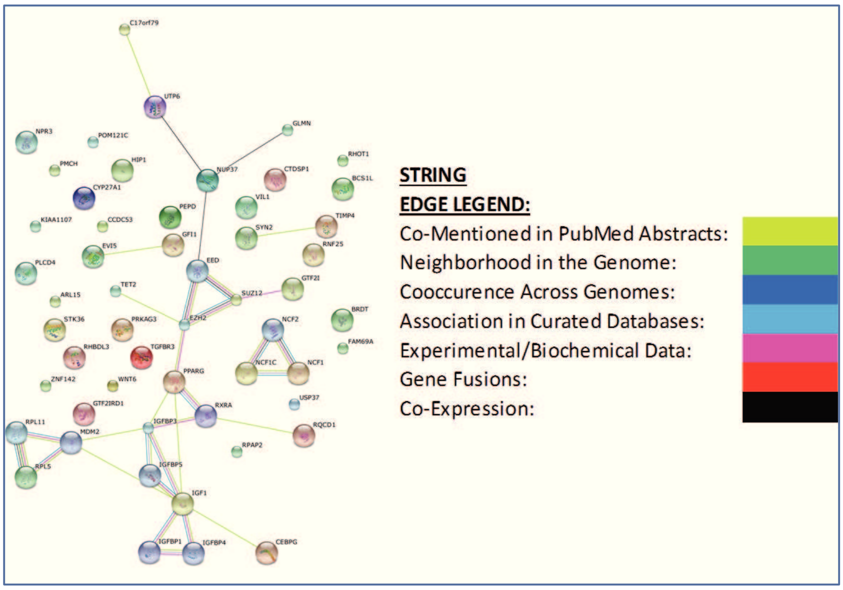


Figure 3.6: *Example of graphical output from STRING pathway analysis. On the right there is the reconstructed network: coloured circles represent input genes, for bigger circles the encoded protein structure is available, edges are coloured according to the legend on the right.*

The main limitation is that SNP IDs cannot be used as input: STRING in fact accepts gene names or protein sequences only; therefore for our analysis with this software we used genes defined from input SNPs by the DAPPLE algorithm.

STRING allows the analysis to be run using the input genes only, or in combination with a number of common interactors, as defined by the user. For our study, we primarily run STRING with no added interactors; if no significant enrichment was detected, an additional analysis with 10 interactors was performed.

STRING also calculates a series of confidence scores for identified connections, as well as a statistical enrichment analysis of any known biological function or pathway based on GeneOntology (GO) data, applying FDR or Bonferroni's correction. The most recent version of STRING (v9.1), the one used for our analyses, extends the automated mining of scientific texts for interaction information to also include full-text articles[163].
Given all STRING characteristics, we decided to adopt it as a pathway analysis tool used for this study.

Other approaches to evaluate pathways
As we did for cluster analysis, we compared the results obtained with DAPPLE and STRING using two additional tools: GeneMANIA and GOrilla.

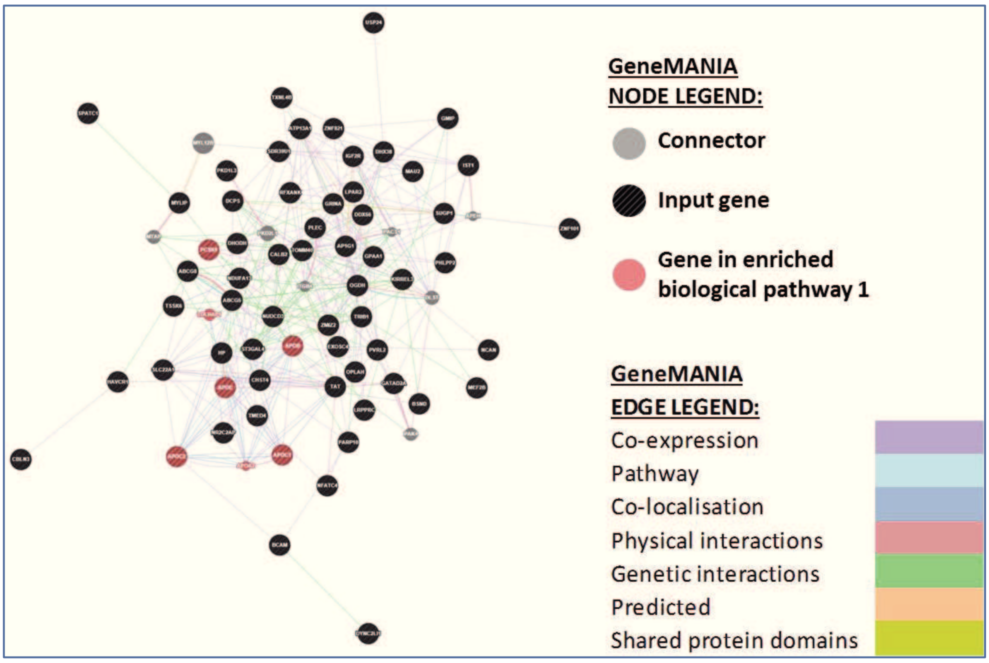


Figure 3.7: *Example of graphical output from GeneMANIA pathway analysis. On the right there is the reconstructed network: darker circles represent input genes while lighter circles are added connectors as explained in the legend on the right, above; coloured circles highlight genes involved in an enriched biological process; for bigger circles the protein information is available. Edges are coloured according to the legend below, on the right.*

GeneMANIA (http://genemania.org/) searches many large, publicly available biological datasets to find related genes. These include protein-protein, protein-DNA and genetic interactions, pathways, reactions, gene and protein expression data, protein domains and phenotypic screening profiles: Gene Expression Omnibus (GEO), BioGRID, PathwayCommons, InterPro, Simple Modular Architecture Research Tool (SMART), Protein Family (Pfam), Reactome, BioCyc, Ensembl and OMIM. GeneMANIA assigns weights to the network with the aim to maximize connectivity between all input genes using linear regression. It also provides a function that calculates GeneOntology terms enriched among the genes in the network[171]. Given its features, GeneMANIA revealed itself as a tool highly similar to STRING, we thus decided to use it to compare GO enrichment results, the structure of the network and the types of direct and indirect connection found. As represented in figure 3.7, the software builds a network of input genes and (eventually) common interactors; edges connecting nodes are coloured on the basis of the type of connection as described in the legend; the user can highlight genes that form part of specific enriched biological processes with different colours of nodes. The user can decide to run the analysis on input genes only or after adding a certain number of interactors. Similarly to procedures in STRING, we used flanking gene entries defined by the DAPPLE software from SNP rsIDs as input in this analysis and we used the same analysis settings.

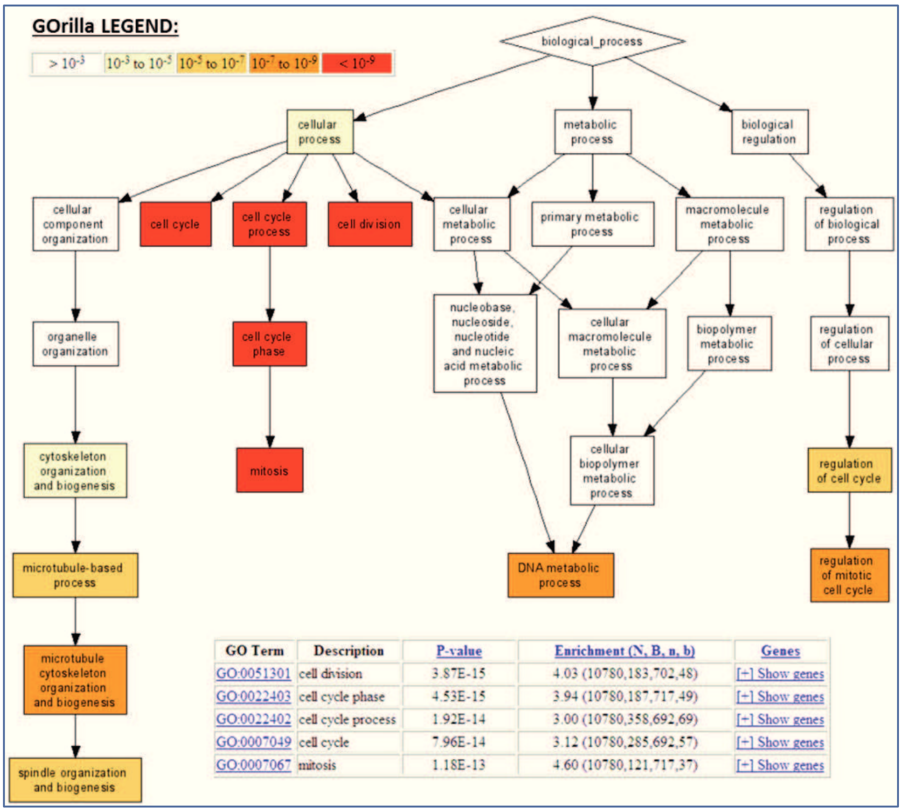


Figure 3.8: *Example of graphical output from GOrilla GO enrichment analysis. A hierarchical representation of enriched biological processes is provided with a table of significance for each of them (below); the significance is also represented in the graph through a colour code as reported in the legend above.*

GOrilla is a web-based application that identifies enriched GO terms in ranked lists of genes: it employs a flexible threshold statistical approach to identify enriched GO terms and to compute an exact p-value for the observed enrichment, taking the threshold for multiple testing into account without the need for simulations. It also produces a hierarchical structure of enriched processes, thus providing a clear view of the relations between enriched GO terms[172]. An example of GOrilla output is in figure 3.8. GOrilla was used to compare the enrichments identified through STRING, where input genes were the same as used in STRING and GeneMANIA. GOrilla is publicly available at: http://cbl-gorilla.cs.technion.ac.il.

3.2.3 Results

3.2.3.1 Alignment of meta-analysis results for cardiometabolic SNPs and Fisher's Omnibus p-value calculation

After the alignment of GWAS meta-analysis results for cardiometabolic phenotypes for the list of 547 published cardiometabolic variants, we observed that for 3 variants (rs5945326 on chromosome 23, rs3918226 on chromosome 7 and rs11066280 on chromosome 12), association summary statistics for more than half of considered phenotypes were not available in consortia meta-analyses. We decided to exclude these SNPs from subsequent analyses, reducing our SNP list to 544 variants.

From Fisher's Omnibus p-value calculation, 324 SNPs (about 60%) showed a genome-wide significant (p-value < $5x10^{-8}$) Omnibus test p-value.
For 40 of them, the significance was attributable to a very strong association with a single phenotype: 6 with a single trait within lipids (TC, HDL, LDL or TG), 5 with a single trait within glycaemic phenotypes (FG with or without adjustment for BMI, FI with or without adjustment for BMI, HOMAB with or without adjustment for BMI, HOMAIR with or without adjustment for BMI, HGLUadjBMI, HINSadjBMI, PROINS or HBA1C) and 29 with one trait within the obesity group (BMI, WC with or without adjustment for BMI, WHR with or without adjustment for BMI, HIP with or without adjustment for BMI, HEIGHT or PCBFAT).

175 variants resulted in significant Fisher's Omnibus p-value test because of multiple high associations to more than one phenotype within the same subgroup of phenotypes: 52 SNPs for lipids, 42 SNPs for glycaemic traits, 80 SNPs for obesity traits and one SNP for blood pressure (DBP, SBP and HTN).
For example, rs964184 near the *APOA1* gene was associated with all lipid traits (p-value for HDL = $5.47x10^{-47}$, p-value for LDL = $1.46x10^{-26}$, p-value for TC = $6.21x10^{-57}$, p-value for TG = $6.71x10^{-240}$) and this resulted in an Omnibus test p-value << $1x10^{-300}$, thus absolutely significant.

Importantly, there were 109 SNPs which showed significant Omnibus test p-values occurring from the combination of multiple significant univariate associations with phenotypes belonging to *different* phenotype groups (as reported in table 3.5).
The first two most significant signals in this group were rs4420638 at the *APOEC1* locus (Omnibus test p-value = $1.2x10^{-279}$) and rs1260326, near the *GCKR* gene (Omnibus test p-value = $7.1x10^{-236}$). Both showed very significant univariate association with lipid traits; in addition rs4420638 was less strongly associated with T2D (univariate p-value = $3x10^{-7}$) and with WHRadjBMI (univariate p-value = $5x10^{-4}$), while rs1260326 was also associated with many glycaemic traits, for example with FG (univariate p-value = $5x10^{-24}$), and with height (univariate p-value = $9x10^{-5}$). The combination of these multiple significant p-values resulted in a very significant omnibus multi-phenotype association.

SNP	HDL	LDL	TC	TG	PCBFAT	BMI	WC	HIP	WHR	WC ADJBMI	HIP ADJBMI	WHR ADJBMI	HEIGHT	DBP	SBP	HTN	FG	HOMAB	FI	HOMAIR	FG adjBMI	HOMAB adjBMI	FIadjBMI	HOMAIR adjBMI	HGLU adjBMI	HINS adjBMI	PROINS	HBA1C	T2D	OMNIB
rs4420638	4E-21	9E-147	5E-111	5E-22	0.293	0.245	0.03	0.644	0.002	0.005	0.324	5E-04	0.258	0.075	0.526	0.527	0.023	0.231	0.014	0.018	0.224	0.567	0.165	0.161	0.764	0.007	0.167	0.576	3E-07	**1.2365E-279**
rs1260326	0.077	2E-04	7E-27	6E-133	0.799	0.129	0.998	0.074	0.017	0.024	0.085	2E-04	9E-05	0.656	0.847	0.59	5E-24	0.07	6E-09	1E-11	1E-24	0.016	6E-13	7E-17	2E-06	0.862	0.064	0.307	0.061	**7.0958E-236**
rs1421085	3E-05	0.982	0.354	0.058	3E-14	3E-62	5E-50	5E-19	5E-19	0.02	0.154	0.043	0.058	0.115	0.283	0.376	0.373	5E-04	1E-05	3E-05	0.03	0.837	0.227	0.211	0.981	0.421	0.43	0.028	2E-09	**6.2949E-157**
rs4506565	0.78	0.092	0.027	0.028	0.018	5E-04	0.001	3E-05	0.716	0.387	0.045	0.379	0.737	0.738	0.96	0.747	7E-09	1E-09	8E-07	6E-04	5E-11	2E-08	7E-05	0.012	9E-08	0.24	1E-17	1E-05	5E-68	**1.3839E-124**
rs12243326	0.467	0.169	0.169	0.112	0.041	6E-04	0.006	1E-04	0.992	0.761	0.334	0.176	0.436	0.54	0.436	0.444	2E-08	5E-11	1E-06	2E-04	6E-11	5E-10	9E-05	0.006	1E-09	0.569	4E-15	4E-05	4E-61	**1.6148E-116**
rs174546	3E-22	2E-21	3E-22	5E-24	0.908	0.774	0.289	0.84	0.585	0.555	0.325	0.756	0.033	0.016	0.405	0.588	5E-10	3E-07	0.043	0.336	5E-09	1E-08	0.011	0.136	0.491	0.88	0.747	0.047	0.003	**1.90566E-94**
rs9987289	6E-25	2E-14	7E-23	0.021	0.14	0.388	0.131	0.161	0.008	0.633	0.002	0.015	0.881	0.734	0.381	0.633	3E-09	0.01	2E-09	3E-09	2E-07	0.064	2E-08	3E-08	0.019	0.058	0.117	0.141	0.015	**1.0072E-90**
rs12916	0.135	5E-45	9E-47	0.304	0.176	1E-04	5E-04	5E-04	0.139	0.99	0.687	0.968	0.763	0.029	0.24	0.243	0.048	0.528	0.095	0.1	0.396	0.654	0.903	0.979	0.116	0.365	0.277	0.974	0.341	**4.93225E-76**
rs10401969	0.579	7E-22	3E-38	2E-29	0.308	0.351	0.46	0.182	0.009	0.023	0.04	0.002	0.076	0.196	0.666	0.151	0.008	0.966	0.514	0.213	0.004	0.873	0.295	0.08	0.858	0.07	0.215	0.355	5E-04	**9.00743E-76**
rs10195252	9E-08	2E-06	2E-05	2E-10	0.001	0.009	0.772	1E-04	6E-07	3E-05	0.002	5E-11	0.895	0.453	0.094	0.027	0.016	0.013	3E-05	1E-05	0.001	1E-04	1E-10	5E-11	0.005	0.014	0.259	7E-04	0.012	**1.6618E-75**
rs983309	3E-19	2E-13	6E-21	0.128	0.075	0.329	0.061	0.287	0.023	0.69	0.026	0.059	0.934	0.799	0.475	0.5	8E-10	0.064	8E-08	8E-08	3E-08	0.225	3E-07	3E-07	0.068	0.276	0.127	0.034	0.039	**6.02465E-75**
rs571312	3E-08	0.997	0.611	1E-05	1E-05	2E-22	9E-19	5E-14	2E-07	0.698	0.315	0.76	3E-06	0.268	0.231	0.327	0.01	0.029	0.006	0.004	0.471	0.831	0.546	0.647	0.191	0.815	0.268	0.93	6E-04	**4.73094E-63**
rs3923113	1E-06	8E-05	5E-04	7E-08	0.002	0.004	0.895	5E-04	6E-05	2E-04	0.008	5E-08	0.867	0.766	0.127	0.028	0.022	0.035	1E-04	4E-05	0.001	7E-04	1E-09	3E-10	0.007	0.086	0.276	7E-04	0.031	**2.51823E-58**
rs2943641	2E-08	0.06	0.452	1E-07	2E-08	0.006	0.003	0.002	0.589	0.128	0.128	0.602	0.551	0.374	0.4	0.047	0.74	2E-05	2E-06	3E-05	0.176	6E-10	5E-14	3E-12	0.19	0.172	0.174	0.199	5E-05	**2.99651E-58**
rs2785980	9E-04	0.047	0.221	0.002	0.012	0.192	0.496	8E-10	2E-06	0.293	1E-09	4E-10	0.003	0.038	0.876	0.254	0.492	1E-04	7E-06	1E-04	0.218	5E-06	6E-08	1E-06	0.655	0.912	0.467	0.632	0.001	**7.93861E-51**
rs389883	0.577	2E-06	9E-13	4E-15	0.856	4E-05	0.004	1E-06	0.092	0.525	1E-05	1E-04	7E-10	0.101	2E-04	0.038	0.334	0.274	0.094	0.174	0.208	0.06	0.003	0.016	0.246	0.476	0.864	0.275	8E-04	**3.14493E-50**
rs12328675	3E-10	0.057	0.091	3E-08	0.131	0.034	0.653	0.008	0.004	0.003	0.127	1E-05	0.803	0.337	0.152	0.221	0.609	0.003	3E-05	8E-05	0.171	3E-06	1E-12	2E-11	0.166	0.023	0.166	0.012	0.036	**8.57428E-50**
rs489693	1E-06	0.099	0.186	5E-07	4E-04	5E-17	2E-15	3E-09	8E-09	0.393	0.93	0.108	5E-04	0.684	0.464	0.307	0.36	0.007	0.003	0.011	0.545	0.659	0.909	0.62	0.582	0.507	0.774	0.708	0.001	**1.1009E-48**
rs12970134	1E-05	0.532	0.451	6E-06	3E-05	3E-18	4E-15	3E-10	2E-07	0.814	0.781	0.363	6E-04	0.477	0.153	0.249	0.165	0.053	0.019	0.03	0.884	0.726	0.391	0.328	0.51	0.589	0.693	0.792	1E-04	**7.22093E-47**
rs2820436	0.006	0.072	0.278	0.006	3E-04	0.083	0.573	1E-08	6E-07	0.114	1E-06	2E-11	0.02	0.202	0.83	0.754	0.052	0.002	2E-05	2E-05	0.058	4E-04	8E-07	5E-07	0.831	0.754	0.15	0.26	0.003	**1.09387E-46**
rs11782386	6E-14	2E-09	8E-17	0.67	0.082	0.646	0.277	0.143	0.028	0.609	0.001	0.057	0.469	0.35	0.161	0.573	1E-04	0.088	4E-05	2E-04	9E-04	0.498	8E-04	0.003	4E-05	0.084	0.055	0.036	0.281	**3.50539E-46**
rs9491696	2E-05	0.629	0.769	4E-05	0.216	0.416	7E-05	0.49	4E-15	2E-06	6E-05	1E-15	0.675	0.263	0.167	0.824	0.252	3E-05	4E-04	0.001	0.116	2E-04	8E-04	0.003	0.586	0.925	0.789	0.596	0.113	**5.83865E-46**
rs143384	0.121	0.015	4E-04	0.431	0.667	0.723	0.193	6E-05	6E-04	0.01	1E-11	5E-05	5E-39	0.135	0.311	0.311	0.605	0.325	0.839	0.661	0.365	0.357	0.625	0.505	0.199	0.07	0.862	0.076	0.23	**2.4398E-45**
rs7578326	2E-07	0.359	0.81	3E-06	8E-07	0.007	0.009	0.002	0.833	0.457	0.34	0.203	0.402	0.37	0.239	0.057	0.453	3E-04	3E-05	1E-04	0.188	3E-07	3E-11	7E-10	0.088	0.109	0.317	0.2	2E-06	**2.44203E-45**
rs2247056	0.039	9E-06	4E-14	2E-15	0.178	0.004	0.066	3E-06	0.223	0.928	2E-05	0.008	5E-13	0.09	0.047	0.591	0.628	0.499	0.313	0.239	0.469	0.384	0.158	0.189	0.828	0.242	0.177	0.545	9E-04	**6.25758E-43**
rs2112347	3E-04	8E-20	7E-23	0.745	0.689	5E-08	1E-05	2E-06	0.042	0.495	0.926	0.529	0.562	0.078	0.747	0.209	0.304	0.584	0.333	0.318	0.474	0.589	0.409	0.541	0.081	0.159	0.938	0.958	0.025	**2.25793E-41**
rs6882076	0.895	2E-22	7E-28	1E-10	0.671	0.919	0.5	0.38	0.882	0.446	0.068	0.78	0.967	0.788	0.66	0.291	0.113	0.095	0.242	0.311	0.505	0.16	0.206	0.3	0.371	0.429	0.807	0.017	0.019	**4.5815E-40**
rs459193	4E-04	0.217	0.177	5E-05	0.077	0.225	0.002	0.519	4E-06	0.002	0.219	9E-06	0.073	6E-04	7E-04	0.178	4E-04	0.01	7E-05	2E-05	9E-04	0.023	1E-05	1E-06	0.086	0.055	0.86	0.632	0.021	**1.86442E-38**
rs2000999	0.534	2E-22	3E-24	6E-06	0.404	0.005	0.015	0.047	0.149	0.927	0.267	0.851	0.74	0.041	0.218	0.764	0.5	0.073	0.043	0.051	0.266	0.391	0.607	0.634	0.402	0.193	0.301	0.149	0.87	**4.91137E-37**
rs4731702	1E-15	0.017	0.213	1E-06	0.098	0.173	0.071	0.087	0.468	0.309	0.014	0.888	0.113	0.373	0.232	0.778	0.078	0.006	5E-04	1E-04	0.042	0.001	2E-05	3E-06	0.069	0.048	0.67	0.192	2E-07	**3.41775E-36**
rs17036328	0.003	0.287	0.449	6E-04	3E-04	0.019	0.004	0.223	0.034	0.6	0.484	0.347	0.048	0.88	0.468	0.847	0.007	0.01	8E-04	2E-04	2E-04	1E-04	3E-09	1E-09	0.002	0.016	0.763	0.46	4E-07	**1.76569E-35**
rs2814944	4E-09	7E-05	9E-08	0.354	0.778	0.01	2E-04	8E-06	0.054	2E-05	2E-06	0.199	6E-13	0.546	0.102	0.076	0.963	0.046	0.014	0.041	0.354	0.254	0.228	0.406	0.425	0.497	0.848	0.235	0.811	**3.36479E-35**
rs4297946	0.886	5E-19	3E-17	0.006	0.088	0.01	0.465	5E-04	0.07	0.118	0.007	0.009	0.912	0.149	0.154	0.934	0.37	0.527	0.102	0.137	0.077	0.279	0.01	0.025	0.303	0.201	0.754	0.096	0.183	**2.09129E-32**
rs6457620	0.003	1E-05	7E-10	0.004	0.202	0.132	0.741	0.013	0.063	0.979	3E-04	0.007	4E-08	0.005	0.048	0.026	0.622	0.009	0.002	0.002	0.706	0.009	0.002	0.003	0.103	0.368	0.964	0.238	0.02	**3.82104E-31**
rs3817334	1E-09	0.382	0.002	0.108	0.003	5E-11	2E-06	9E-04	0.023	0.145	0.36	0.338	5E-05	0.223	0.075	0.62	0.168	0.996	0.334	0.198	0.758	0.167	0.233	0.348	0.9	0.381	1E-10	0.113	0.049	**6.85997E-31**
rs3184504	5E-06	2E-09	3E-11	0.185	0.066	1E-04	4E-04	0.029	0.128	0.224	0.945	0.895	0.966	4E-07	8E-04	0.038	0.545	0.925	0.983	0.766	0.937	0.23	0.057	0.092	0.112	0.505	0.112	1E-04	0.627	**2.21376E-28**
rs11065987	6E-04	2E-09	7E-12	0.867	0.018	1E-04	2E-04	0.047	0.111	0.069	0.77	0.675	0.706	1E-06	4E-04	0.004	0.776	0.564	0.568	0.431	0.783	0.545	0.185	0.202	0.169	0.786	0.258	8E-05	0.263	**5.81847E-28**

Table 3.5: *SNPs with significant Omnibus test p-values possibly determined by the combination of multiple significant univariate associations with traits belonging to different trait groups.*

SNP	HDL	LDL	TC	TG	PCBFAT	BMI	WC	HIP	WHR	WC ADJBMI	HIP ADJBMI	WHR ADJBMI	HEIGHT	DBP	SBP	HTN	FG	HOMAB	FI	HOMAIR	FG adjBMI	HOMAB adjBMI	FIadjBMI	HOMAIR adjBMI	HGLU adjBMI	HINS adjBMI	PROINS	HBA1C	T2D	OMNIB
rs4765127	3E-10	9E-04	0.005	2E-08	0.004	0.002	0.616	9E-04	0.004	6E-04	0.01	8E-06	0.346	0.727	0.108	0.084	0.384	0.536	0.278	0.329	0.192	0.214	0.01	0.019	0.213	0.177	0.48	0.962	0.018	**1.25991E-27**
rs3822072	2E-06	0.023	0.548	0.007	0.004	0.121	0.402	0.034	0.032	0.208	3E-05	2E-04	0.07	0.034	0.131	0.686	0.774	0.002	0.001	0.002	0.269	9E-05	3E-06	1E-05	0.754	0.439	0.318	0.105	0.017	**4.17245E-27**
rs731839	1E-06	0.622	0.13	2E-04	0.027	0.005	0.022	0.05	0.213	0.234	0.9	0.536	0.009	0.156	0.238	0.639	0.131	0.005	5E-05	3E-04	0.028	0.001	5E-07	1E-06	0.566	0.378	0.2	0.288	0.415	**1.46958E-24**
rs7027110	0.196	0.052	0.467	0.107	0.991	0.761	0.041	0.002	0.981	1E-04	1E-07	0.868	1E-10	0.218	0.906	0.304	0.684	6E-04	4E-04	3E-04	0.618	0.001	5E-04	4E-04	0.035	0.968	0.329	0.647	0.135	**1.97004E-24**
rs11920090	0.726	0.073	0.031	4E-04	0.29	0.327	0.374	0.116	0.18	0.67	0.569	0.331	0.1	0.788	0.895	0.411	2E-09	1E-07	0.035	0.348	2E-11	7E-07	0.157	0.966	0.584	0.186	0.597	6E-04	0.051	**2.07012E-24**
rs9686661	8E-07	0.007	0.006	1E-10	0.083	0.51	0.114	0.506	4E-04	6E-04	0.663	4E-05	0.462	0.487	0.24	0.005	0.808	0.624	0.225	0.337	0.989	0.073	0.003	0.012	0.271	0.027	0.298	0.083	9E-05	**2.68908E-24**
rs1378942	0.29	6E-05	6E-05	0.745	0.083	0.016	0.009	0.121	0.022	0.119	0.785	0.111	0.068	6E-08	3E-06	4E-05	2E-04	0.008	0.96	0.664	2E-04	0.011	0.504	0.252	0.781	0.022	0.841	0.942	0.081	**3.83203E-24**
rs6450176	5E-08	0.091	0.056	2E-05	0.391	8E-05	0.002	0.354	6E-04	0.82	0.021	0.014	0.35	0.732	0.434	0.244	0.704	0.057	0.074	0.148	0.625	8E-04	3E-04	0.002	0.32	0.065	0.78	0.114	4E-05	**6.39055E-23**
rs2814982	1E-06	5E-07	5E-11	0.73	0.477	0.031	0.02	0.012	0.596	0.064	0.006	0.981	3E-08	0.671	0.714	0.974	0.634	0.063	0.005	0.02	0.925	0.109	0.016	0.042	0.513	0.616	0.606	0.494	0.494	**7.2458E-23**
rs1173771	0.513	0.415	0.376	0.045	0.158	0.525	0.004	0.052	0.19	2E-06	9E-05	0.126	8E-15	0.003	1E-04	5E-05	0.277	0.823	0.466	0.442	0.202	0.815	0.213	0.269	0.657	0.124	0.129	0.67	0.399	**6.23003E-22**
rs6569648	0.061	0.36	0.187	0.005	0.606	0.016	0.343	0.004	0.027	0.995	9E-05	0.001	9E-12	0.09	0.134	0.076	0.368	0.008	0.174	0.154	0.524	0.001	0.026	0.027	0.004	0.547	0.074	0.999	0.001	**7.63025E-22**
rs1800562	0.184	6E-10	2E-08	0.748	0.452	0.531	0.361	0.157	0.195	0.072	0.108	0.145	0.009	0.109	0.469	0.185	0.87	0.697	0.877	0.624	0.639	0.9	0.792	0.757	0.508	0.461	0.683	3E-20	0.594	**1.79237E-21**
rs2287019	1E-03	5E-04	1E-05	0.413	0.005	3E-07	1E-05	4E-05	0.034	0.182	0.452	0.968	0.93	0.412	0.806	0.517	0.001	0.718	0.321	0.241	0.045	0.059	0.432	0.714	5E-05	0.086	0.008	0.29	0.328	**2.02713E-21**
rs6912327	0.229	0.062	0.146	0.385	0.553	0.066	0.651	0.951	0.704	0.037	0.001	0.384	1E-04	0.259	0.241	0.18	0.325	1E-03	8E-05	3E-04	0.159	4E-04	1E-06	9E-06	0.295	4E-04	0.077	0.798	0.029	**2.49416E-21**
rs1055144	0.385	0.89	0.331	1E-03	0.138	0.232	0.026	0.379	5E-06	3E-07	0.512	3E-09	0.007	0.66	0.581	0.577	0.36	0.016	0.017	0.009	0.372	0.005	0.001	1E-03	0.864	0.386	0.342	0.305	0.152	**9.46117E-21**
rs13107325	7E-11	0.121	0.003	0.015	0.118	1E-07	7E-04	0.073	0.171	0.028	0.298	0.139	0.003	7E-05	1E-04	0.006	0.208	0.835	0.987	0.988	0.801	0.223	0.294	0.253	0.537	0.841	0.933	0.856	0.382	**1.28005E-20**
rs2256183	2E-04	0.003	3E-07	0.002	0.186	0.014	0.011	0.002	0.708	0.424	0.024	0.514	3E-14	0.092	0.097	0.269	0.581	0.927	0.974	0.751	0.593	0.814	0.767	0.847	0.431	0.389	0.274	0.663	2E-04	**1.29204E-20**
rs1167800	0.073	0.862	0.476	0.086	0.246	4E-05	0.006	0.02	0.326	0.535	0.589	0.294	0.634	0.649	0.47	0.599	0.655	4E-05	2E-08	2E-07	0.933	0.004	3E-05	9E-05	0.191	0.158	0.534	0.932	0.081	**5.19226E-20**
rs4865796	0.019	0.414	0.193	0.012	0.361	5E-05	0.071	0.923	0.021	0.06	7E-06	0.376	0.034	0.568	0.412	0.139	0.727	0.031	0.047	0.148	0.534	4E-04	2E-04	0.003	0.103	0.061	0.166	0.003	2E-05	**2.03673E-19**
rs10423928	6E-04	0.002	7E-04	0.187	0.011	2E-06	1E-04	2E-04	0.047	0.259	0.402	0.866	0.867	0.26	0.787	0.392	1E-03	0.732	0.223	0.12	0.064	0.06	0.4	0.71	3E-06	0.177	0.004	0.284	0.421	**2.24762E-19**
rs849134	0.583	0.603	0.659	0.106	0.902	0.057	0.899	0.436	0.721	7E-05	3E-04	0.104	3E-13	0.071	0.036	0.031	0.153	0.334	0.457	0.519	0.019	0.572	0.973	0.937	0.056	0.627	0.609	0.021	3E-10	**6.01894E-19**
rs11605924	4E-04	0.83	0.271	0.119	0.681	0.832	0.776	0.771	0.47	0.222	0.914	0.41	0.48	0.173	0.957	0.868	3E-13	1E-04	0.532	0.565	2E-13	5E-04	0.75	0.208	0.201	0.622	0.537	0.721	0.007	**6.11186E-19**
rs2277862	0.028	6E-06	4E-10	0.002	0.513	0.487	0.743	0.047	0.01	0.345	1E-04	0.007	6E-10	0.037	0.301	0.048	0.789	0.744	0.546	0.833	0.969	0.745	0.999	0.77	0.653	0.479	0.221	0.448	0.083	**7.85865E-19**
rs12444979	6E-04	0.406	0.753	0.228	0.204	4E-11	2E-07	2E-06	0.001	0.254	0.608	0.919	0.067	0.909	0.761	0.728	0.767	0.01	0.014	0.023	0.577	0.106	0.234	0.355	0.384	0.042	0.497	0.201	0.065	**1.09261E-18**
rs10761731	3E-07	4E-04	0.001	3E-12	0.446	1E-03	0.014	0.827	0.009	0.642	0.036	0.112	0.027	0.063	0.284	0.197	0.33	0.537	0.307	0.322	0.418	0.961	0.822	0.822	0.057	0.506	0.003	0.656	0.916	**1.32042E-18**
rs6759321	0.005	1E-06	1E-08	0.627	0.595	0.008	7E-04	0.002	0.064	0.317	0.595	0.673	0.01	0.87	0.711	0.107	0.64	1E-03	2E-04	6E-04	0.609	0.097	0.097	0.134	0.258	0.991	0.686	0.858	0.286	**3.0433E-18**
rs605066	3E-08	0.021	0.165	3E-06	0.088	0.603	0.278	0.93	0.001	7E-04	0.835	6E-05	0.508	0.069	0.06	0.351	0.177	0.134	0.01	0.019	0.097	0.31	0.03	0.055	0.027	0.505	0.779	0.014	0.943	**3.76734E-18**
rs9804646	0.648	1E-08	2E-16	4E-09	0.646	0.919	0.77	0.665	0.494	0.323	0.119	0.361	0.555	0.357	0.55	0.305	0.76	0.08	0.49	0.436	0.911	0.132	0.644	0.558	0.031	0.117	0.668	0.154	0.13	**7.65128E-18**
rs7941030	3E-08	3E-06	2E-10	0.985	0.794	7E-04	0.002	0.439	0.004	0.953	0.006	0.158	0.209	0.381	0.577	0.2	0.645	0.645	0.86	0.892	0.931	0.073	0.05	0.104	0.764	0.315	0.468	0.077	0.078	**8.1178E-18**
rs10838687	2E-14	0.687	0.013	0.152	0.018	0.02	0.508	0.957	0.716	0.01	0.312	0.055	0.068	0.879	0.796	0.422	3E-05	0.003	0.09	0.212	1E-04	0.007	0.282	0.537	0.444	0.695	0.933	0.876	0.38	**6.21287E-17**
rs442177	3E-07	1E-03	4E-04	9E-12	0.686	0.114	0.137	0.249	0.157	0.559	0.086	0.463	6E-04	0.387	0.489	0.961	0.358	0.149	0.056	0.025	0.813	0.363	0.277	0.151	0.121	0.414	0.108	0.799	0.212	**1.57923E-16**
rs1530559	0.012	2E-04	6E-05	0.601	0.204	0.012	0.008	0.055	0.203	0.02	0.744	0.447	0.008	0.359	0.792	0.221	0.793	7E-05	7E-05	1E-04	0.847	0.006	0.016	0.015	0.502	0.73	0.442	0.255	0.775	**1.80071E-16**
rs1495743	0.849	5E-04	9E-09	4E-14	0.053	0.231	0.066	0.618	0.025	0.345	0.589	0.035	0.502	0.484	0.328	0.98	0.029	0.968	0.472	0.235	0.031	0.743	0.558	0.342	0.667	0.362	0.753	0.023	0.035	**3.14264E-16**
rs1325598	0.86	0.256	0.367	0.291	0.988	0.049	0.185	0.061	0.266	0.759	0.058	0.035	2E-08	0.096	0.661	0.77	0.971	0.024	0.008	0.018	0.512	0.002	2E-04	6E-04	5E-04	5E-04	0.977	0.187	0.111	**6.93668E-16**
rs6784615	2E-05	0.068	0.408	0.011	0.853	0.19	0.967	0.017	2E-04	0.023	0.016	9E-08	0.095	0.227	0.447	0.922	0.838	0.092	0.061	0.048	0.603	0.039	0.004	0.006	0.966	0.133	0.641	0.531	0.003	**1.38246E-15**
rs3792752	0.41	0.187	0.154	0.603	0.51	0.123	0.71	0.556	0.159	6E-04	0.242	0.024	3E-09	0.383	0.154	0.106	0.219	0.002	0.007	0.02	0.164	1E-04	3E-04	0.002	0.329	0.835	0.594	0.323	0.82	**2.8146E-15**
rs879882	0.553	0.002	6E-05	1E-04	0.878	0.565	0.109	0.015	0.305	0.001	8E-05	0.318	8E-07	0.594	0.915	0.05	0.4	0.569	0.232	0.211	0.304	0.361	0.092	0.16	0.933	0.158	0.773	0.005	0.006	**3.08792E-15**

Table 3.5: *Continuation.*

SNP	HDL	LDL	TC	TG	PCBFAT	BMI	WC	HIP	WHR	WC ADJBMI	HIP ADJBMI	WHR ADJBMI	HEIGHT	DBP	SBP	HTN	FG	HOMAB	FI	HOMAIR	FG adjBMI	HOMAB adjBMI	FIadjBMI	HOMAIR adjBMI	HGLU adjBMI	HINS adjBMI	PROINS	HBA1C	T2D	OMNIB
rs1708299	0.174	0.339	0.213	0.306	0.961	0.27	0.71	0.49	0.314	2E-04	9E-06	0.936	1E-17	0.376	0.108	0.23	0.857	0.421	0.201	0.18	0.513	0.781	0.752	0.595	0.049	0.897	0.899	0.044	9E-04	**4.35413E-15**
rs3123629	0.345	2E-06	6E-09	1E-06	0.175	0.013	0.061	0.399	0.75	0.765	0.652	0.435	0.184	0.363	0.764	0.15	0.234	8E-04	0.13	0.114	0.181	0.005	0.404	0.401	0.425	0.519	0.088	0.403	0.273	**1.12458E-14**
rs2898290	0.468	0.915	0.279	6E-05	8E-04	0.039	0.483	0.971	0.661	0.007	0.006	0.567	0.001	0.009	0.418	0.476	0.378	0.034	0.081	0.027	0.64	8E-04	4E-05	2E-05	0.981	0.807	0.287	0.505	0.093	**1.29584E-14**
rs2737229	0.295	8E-07	2E-08	0.012	0.533	0.017	0.012	7E-04	0.653	0.496	0.016	0.051	0.413	0.136	0.635	0.306	0.952	0.184	0.218	0.343	0.503	0.998	0.555	0.346	0.574	0.379	1E-05	0.055	0.921	**4.10837E-14**
rs1961456	0.825	2E-04	2E-09	3E-11	0.726	0.479	0.196	0.68	0.082	0.928	0.707	0.167	0.188	0.465	0.385	0.671	0.136	0.262	0.21	0.085	0.125	0.545	0.366	0.224	0.554	0.779	0.327	0.001	0.052	**5.41223E-14**
rs4607103	0.058	0.902	0.943	0.187	0.017	0.006	0.047	0.003	0.048	0.04	0.008	2E-05	0.127	0.449	0.994	0.958	0.004	0.965	0.232	0.236	6E-04	0.714	0.026	0.02	0.039	0.086	0.866	0.189	1E-04	**6.80518E-14**
rs11613352	4E-08	5E-04	0.003	4E-10	0.398	0.019	0.065	0.012	0.352	0.337	0.338	0.506	0.381	0.785	0.988	0.908	0.404	0.337	0.666	0.667	0.233	0.691	0.849	0.886	0.197	0.003	0.232	0.047	0.191	**1.07773E-13**
rs492602	0.732	8E-08	2E-10	3E-04	0.71	0.049	0.009	2E-04	0.908	0.218	0.001	0.464	0.857	0.309	0.933	0.875	0.546	0.78	0.819	0.923	0.343	0.869	0.196	0.35	0.471	0.179	0.148	0.067	0.617	**2.06728E-13**
rs11153594	0.389	3E-09	2E-10	0.005	0.969	0.625	0.136	0.612	0.738	0.669	0.058	0.262	6E-04	0.496	0.681	0.964	0.946	0.134	0.03	0.055	0.59	0.134	0.037	0.086	0.28	0.752	0.552	0.225	0.442	**3.1625E-13**
rs12946454	0.015	0.833	0.839	0.047	0.158	0.612	0.58	0.596	0.486	0.012	0.131	0.861	3E-07	8E-04	6E-06	4E-04	0.08	0.455	0.136	0.146	0.093	0.775	0.238	0.264	0.003	0.057	0.05	0.589	0.452	**3.21035E-13**
rs1173766	0.603	0.894	0.598	0.055	0.245	0.189	0.004	0.039	0.229	1E-04	7E-04	0.205	1E-08	0.016	7E-04	2E-04	0.34	0.979	0.659	0.565	0.191	0.959	0.226	0.252	0.985	0.06	0.256	0.375	0.517	**3.38179E-13**
rs9488822	0.251	3E-08	2E-10	2E-04	0.437	0.856	0.558	0.451	0.27	0.662	0.064	0.045	0.006	0.327	0.731	0.672	0.593	0.141	0.067	0.097	0.395	0.135	0.077	0.153	0.692	0.68	0.971	0.229	0.473	**9.07395E-13**
rs6495122	0.437	2E-04	4E-04	0.986	0.032	0.048	0.08	0.54	0.197	0.3	0.77	0.327	0.225	4E-05	2E-04	5E-04	0.009	0.022	0.951	0.69	0.014	0.024	0.617	0.38	0.994	0.021	0.184	0.846	0.354	**1.56034E-12**
rs6457821	0.006	0.011	8E-04	0.443	0.59	0.256	0.359	0.354	0.33	0.003	0.075	0.2	2E-11	0.467	0.833	0.876	0.22	0.113	0.179	0.429	0.448	0.042	0.029	0.136	0.069	0.03	0.522	0.478	0.243	**1.87149E-12**
rs7134594	7E-15	0.562	1E-04	0.61	0.566	0.25	0.067	3E-04	0.751	0.591	0.059	0.082	0.33	0.853	0.105	0.578	0.261	0.117	0.086	0.036	0.481	0.324	0.169	0.108	0.689	0.632	0.432	0.6	0.624	**1.94457E-12**
rs974801	0.007	4E-04	0.064	0.857	0.457	0.887	0.493	0.919	0.686	0.551	0.406	0.744	5E-04	0.176	0.396	0.677	0.511	0.015	1E-04	0.003	0.519	0.003	1E-05	5E-04	0.056	0.021	0.942	0.496	0.515	**2.08595E-12**
rs11847697	4E-04	0.59	0.589	0.016	0.005	1E-08	3E-06	0.007	0.004	0.344	0.362	0.621	0.425	0.053	0.191	0.355	0.796	0.161	0.186	0.086	0.51	0.958	0.682	0.981	0.273	0.628	0.173	0.194	0.187	**2.62646E-12**
rs2925979	2E-11	0.946	0.494	9E-05	0.03	0.118	0.385	0.07	0.14	0.054	0.969	0.011	0.011	0.546	0.926	0.811	0.146	0.649	0.294	0.265	0.198	0.724	0.155	0.102	0.246	0.023	0.352	0.853	0.002	**4.02169E-12**
rs3786897	5E-05	0.215	0.034	0.244	0.892	0.02	0.007	0.934	5E-05	8E-05	0.154	2E-05	0.282	0.029	0.123	0.462	0.401	0.436	0.144	0.197	0.104	0.36	0.054	0.055	0.526	0.998	0.456	0.882	0.306	**4.44012E-12**
rs17271305	0.554	0.443	0.626	0.297	0.06	0.255	0.242	0.527	0.109	0.986	0.037	0.181	7E-06	0.08	0.865	0.491	0.003	0.228	0.911	0.696	2E-04	0.966	0.09	0.038	1E-06	0.599	0.004	0.221	0.006	**1.26588E-11**
rs2336725	1E-04	0.092	0.709	0.129	0.947	0.779	0.968	0.034	0.026	0.928	4E-04	0.005	4E-08	0.604	0.851	0.714	0.004	0.397	0.784	0.53	0.013	0.197	0.913	0.706	0.444	0.864	0.227	0.039	2E-04	**1.46144E-11**
rs7178424	0.191	0.294	0.444	0.263	0.158	0.585	0.693	0.15	0.127	0.665	0.005	0.093	2E-07	0.286	0.851	0.363	0.018	0.292	0.681	0.918	0.001	0.741	0.409	0.3	1E-05	0.539	9E-04	0.436	8E-04	**2.3541E-11**
rs10037512	0.717	0.935	0.903	0.56	0.937	0.43	0.754	0.325	0.597	0.531	0.034	0.474	4E-09	0.711	0.558	0.784	0.001	0.345	5E-04	2E-04	0.002	0.559	4E-04	3E-04	0.612	0.915	0.461	0.632	0.496	**6.03674E-11**
rs1799945	0.618	0.424	0.52	0.116	0.914	0.138	0.005	0.065	0.016	0.003	0.032	0.026	0.456	3E-05	3E-04	4E-05	0.813	0.766	0.766	0.798	0.907	0.475	0.535	0.689	0.891	0.871	0.415	1E-04	0.027	**1.30429E-10**
rs6795735	0.083	0.821	0.922	0.155	0.138	0.033	0.269	0.021	0.004	0.011	0.025	7E-08	0.29	0.114	0.367	0.796	0.064	0.62	0.56	0.498	0.017	0.574	0.346	0.261	0.082	0.281	0.102	0.431	2E-04	**1.30999E-10**
rs10010325	0.013	0.021	0.485	0.525	0.754	0.556	0.411	0.774	0.413	0.261	0.021	0.418	2E-06	0.124	0.185	0.957	0.57	0.014	5E-04	0.014	0.671	0.006	3E-04	0.021	0.044	0.097	0.61	0.77	0.429	**1.66356E-10**
rs5017948	9E-07	0.858	0.084	0.597	0.325	0.361	0.801	0.093	0.138	0.228	0.002	0.26	5E-06	0.143	0.103	8E-04	0.074	0.523	0.792	0.722	0.069	0.389	0.578	0.954	0.329	0.043	0.16	0.519	0.031	**5.13077E-10**
rs11063069	0.719	0.024	0.007	0.009	0.874	0.127	0.002	0.1	0.137	0.068	0.103	0.472	6E-04	0.517	0.798	0.091	0.026	0.5	0.236	0.205	0.027	0.688	0.11	0.06	0.908	0.204	0.031	0.59	2E-04	**9.53056E-10**
rs2154319	0.212	0.595	0.351	0.765	0.869	0.131	0.071	0.022	0.901	0.589	0.019	0.173	4E-10	0.905	0.529	0.501	0.878	0.57	0.335	0.563	0.776	0.219	0.022	0.067	0.007	9E-04	0.418	0.592	6E-04	**1.803E-09**
rs2929282	3E-06	0.767	0.641	2E-11	0.703	0.011	0.003	0.493	0.021	0.119	0.455	0.21	0.085	0.662	0.506	0.056	0.904	0.629	0.74	0.932	0.897	0.966	0.904	0.941	0.212	0.646	0.571	0.437	0.322	**8.10885E-09**
rs12940887	0.007	0.119	0.136	3E-04	0.027	0.121	0.64	0.555	0.531	0.033	0.254	0.646	0.001	5E-06	0.002	0.023	0.515	0.254	0.57	0.438	0.561	0.469	0.7	0.622	0.443	0.706	0.038	0.762	0.631	**1.51992E-08**
rs3829109	0.613	0.128	0.664	0.065	0.491	0.631	0.34	0.026	0.431	0.119	0.011	0.578	0.001	0.018	0.023	0.189	0.012	0.073	0.48	0.592	0.008	0.024	0.337	0.457	0.019	0.117	0.24	0.213	0.004	**1.66288E-08**
rs11597086	7E-04	0.043	6E-05	0.027	0.026	0.556	0.011	0.047	0.513	0.081	0.02	0.862	2E-04	0.932	0.786	0.471	0.485	0.417	0.936	0.872	0.56	0.243	0.528	0.589	0.936	0.433	0.682	0.198	3E-04	**2.54612E-08**
rs645040	3E-06	0.223	0.352	3E-08	0.652	0.056	0.251	0.107	0.785	0.998	0.402	0.628	0.077	0.607	0.745	0.365	0.945	0.026	0.019	0.061	0.776	0.045	0.072	0.137	0.871	0.51	0.279	0.85	0.146	**2.99114E-08**
rs7225700	0.119	4E-09	1E-06	0.172	0.558	0.993	0.048	0.21	0.946	0.012	0.006	0.979	0.003	0.608	0.406	0.771	0.888	0.509	0.466	0.637	0.788	0.458	0.612	0.626	0.828	0.366	0.56	0.147	0.656	**4.75231E-08**

Table 3.5: *Continuation.*

At the *FTO* locus, rs1421085 is the third most significant SNP (Omnibus test p-value = 6.3×10^{-157}) with its primary association with BMI (p-value = 3×10^{-62}) and with other obesity-related traits, followed by significant association with T2D (p-value = 2×10^{-9}) and other suggestive associations with glycaemic traits and with HDL. As we know from the literature, it was demonstrated that these multiple associations at *FTO* variants are attributable to the association with BMI, which mediates all the others. This is confirmed by our results since the significant associations disappeared when the traits are adjusted for BMI (for WC, HIP, WHR, HOMAB, FI and HOMAIR).

Of particular interest are 86 variants which showed almost equivalent multiple associations (when difference of order of magnitude at univariate associations was no more than 10) with different phenotypes, for example rs10195252 at *GRB14* locus: it is comparably associated with TG (p-value = 2×10^{-10}), WHRadjBMI (p-value = 5×10^{-11}), FiadjBMI (p-value = 1×10^{-10}) and HOMAIRadjBMI (p-value = 5×10^{-11}) at a GW significance level; moreover this variant presented additional suggestive associations with HDL, LDL, TC, HIP, WCadjBMI and HBA1C. All together these associations led to an omnibus p-value that increased in significance: 2×10^{-75}.

3.2.3.2 <u>Evaluation of multi- phenotype effects and association significance at cardiometabolic loci through complete hierarchical clustering</u>

To identify groups of loci with similar patterns of multi- phenotype effects, to clarify the degree of connection between them, and to shed light on the types of multiple associations in comparison with epidemiological expectations, we decided to consider z-score values from cardiometabolic GWAS meta-analysis results and to apply a complete hierarchical clustering algorithm.

Complete hierarchical cluster obtained from the matrix of z-scores is represented in figure 3.9 as a dendrogram of 544 included variants. In figure 3.9, below the dendrogram, the heatmap of multiple cardiometabolic trait effects is also reported as it visually represents the combination of multiple effects and their hierarchical organisation. The heatmap is built based on a colour code from bright yellow (very significant p-value < 5×10^{-8}, positive effect) to bright blue (very significant p-value < 5×10^{-8}, negative effect) with intermediate black colour for non-significant associations.
The Approximately Unbiased (AU) estimate of bootstrap value (%), obtained by multiscale 10,000 bootstrap resampling, is reported for each node of the dendrogram in figure 3.9.
We initially observed that nodes at the highest levels of the dendrogram, which represent separations between bigger groups of loci, are poorly supported, while nodes that separate smaller groups are in general well supported (bootstrap value > 65%). This result can be interpreted as the fact that cardiometabolic phenotype loci share same multi- phenotype effects within small groups, each of which probably contributes to the same pathway that influences the phenotypes.

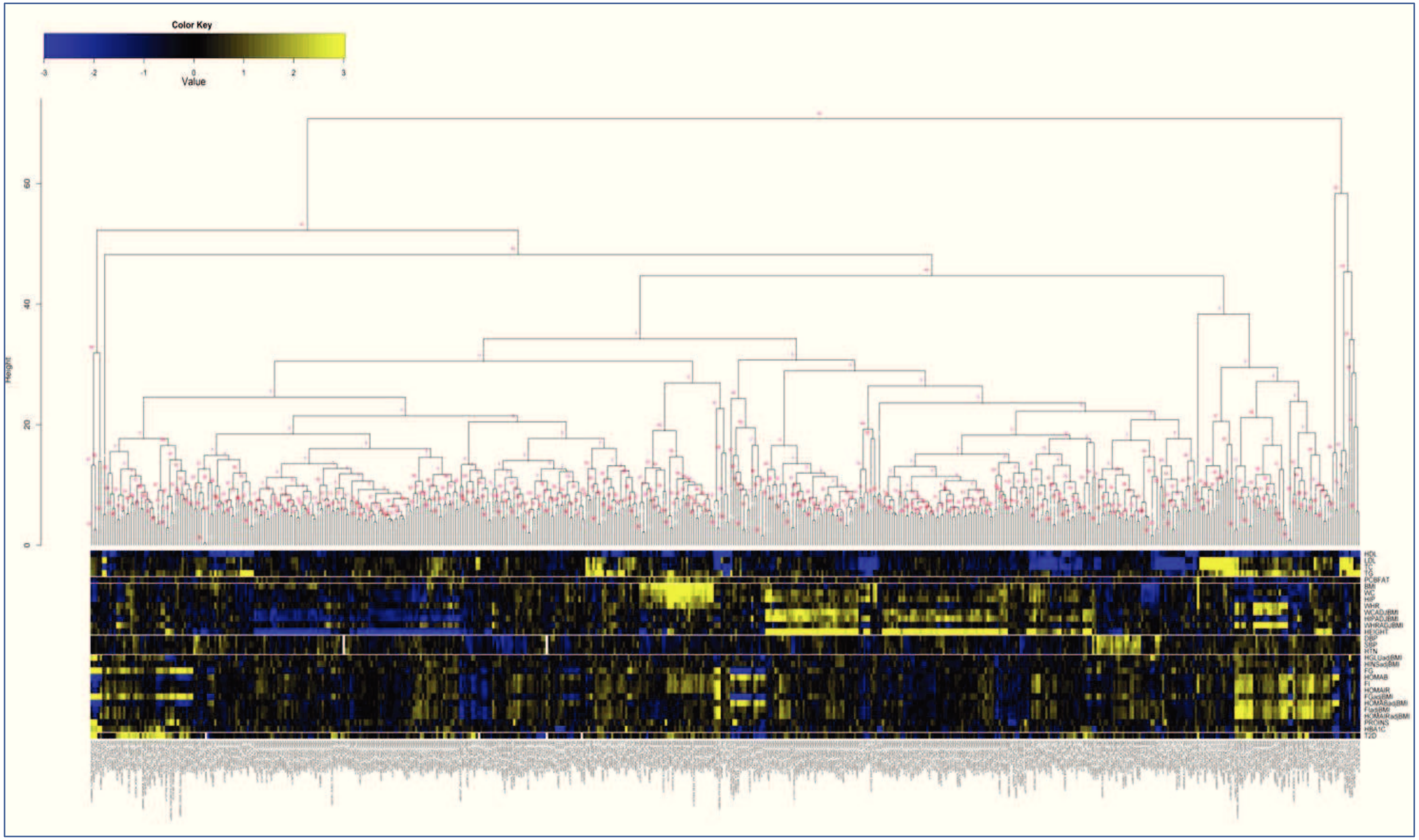


Figure 3.9: *Complete hierarchical cluster dendrogram obtained from cardiometabolic z-scores data of 544 included variants with bootstrap value (%) for each node. The heatmap of multiple cardiometabolic phenotype effects is reported below the dendrogram: each row is a cardiometabolic phenotype (labelled on the right) and each column is a variant (labelled below); a colour code is used in each cell to represent the direction and the intensity of effects: yellow represents positive direction, while blue represents negative direction of effects; the brightness of the colour is directly proportional to the intensity and the significance of the effect.*

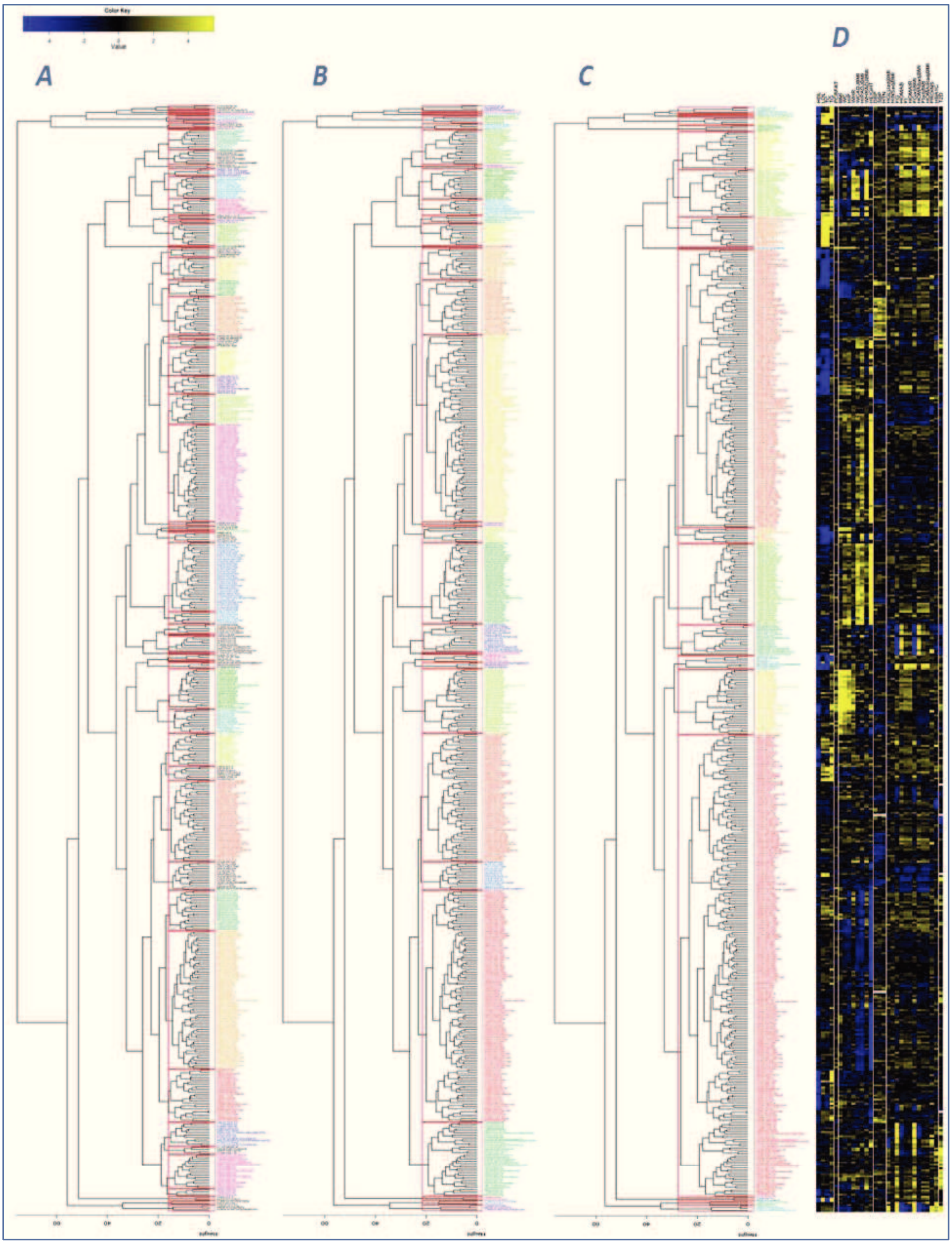


Figure 3.10: *The three sub-sets of the cluster are represented as obtained using the three different Euclidean distance threshold:* ***A****. threshold C at 15% of Euclidean distance,* ***B****. threshold B at 20% of Euclidean distance and* ***C****. threshold A at 25% of Euclidean distance.* ***D****. Heat map of the clustered multi-phenotype effects.*

3.2.3.3 Definition of sub-clusters of loci with shared effects and Pathway analyses

Using three different thresholds of the Euclidean distance (25%, 20% and 15%) we defined three sets of sub-clusters derived from the total cluster that we had obtained from the hierarchical clustering approach. The three sets are represented in figure 3.10 A, B and C respectively.

Set A (25% of Euclidean distance) contained 19 sub-clusters with a mean number of 28.63 SNPs each; set B (20% of Euclidean distance) had 30 groups and an average of 18.13 SNPs for each; set C (15% of Euclidean distance) includes 57 sub-sets containing a mean of 9.54 SNPs each. Each of defined groups of SNPs with similar cardiometabolic multi- phenotype effects obtained through one of the thresholds above was interrogated through pathway analysis using the four different internet software tools described above (figures from 3.11 to 3.25 represent the most interesting one).
By examining the structure and the multiple effect architecture of identified sub-clusters, we could recognise some trends in the patterns of multi- phenotype effects.
A summary of trends of multiple effects and of the results of pathway analyses for the four software tools are reported in Appendix table 7. In general, GOrilla was less useful for discovering enriched pathways, probably because it is not well suited for small lists of genes used as input, such as in most of the groups of our sets. STRING and GeneMANIA were the most useful tools and resulted in agreement for the majority of our analyses.

In the sections below, the trends are categorised and described with special focus on those sub-clusters showing significant result in pathway analyses.

Sub-clusters of cardiometabolic loci without a uniform trend of multi- phenotype effects
First of all, in some groups, especially the biggest ones obtained preferentially through the wider threshold of Euclidean distance (cut-off A), it was not possible to identify a uniform trend of multiple effects, but rather a unique effect on a single phenotype or on a few phenotypes, while other phenotypes showed different effects .

An example is represented in figure 3.11A (we called this group "H25_6"): in this sub-cluster we recognised a common trend of increased glycaemic traits, in particular HOMAB, FI and HOMAIR, which is maintained also after adjustment for BMI, accompanied by a common trend of lipids, in particular with a decrease of HDL and an increase of TG. Within this group of loci we could recognise two separate effects of obesity/anthropometric traits: the first half of the group showed a strong increase of height, while the second half did not present this characteristic, but instead showed low BMI, WC and WHR (these last two effects were also maintained after adjustment for BMI). The second half is thus concordant with the description of healthy obesity/unhealthy leanness (HOUL): in fact, decreased adiposity (both BMI and central obesity) is present together with low levels of HDL cholesterol and high levels of TG, glycaemic traits and high T2D risk.
The total group includes variants originally identified as associated with height or with glycaemic traits; pathway analysis revealed a significant excess of direct and indirect connections (p-value from DAPPLE software = 0.02 and 0.002 respectively, figure 3. 11B for network representation) between

putative genes in the proximity of these variants with an enrichment of the viral reproduction pathway (from STRING: p-value = $7x10^{-7}$, q-value after FDR correction = 0.008; see red rectangles in figure 3. 11C).

This enrichment was not confirmed by other internet tools, such as GOrilla and GeneMANIA. Moreover the significance was lost when the group was further sub-divided in two groups using a smaller cut-off of Euclidean distance that divided anthropometric/obesity effects.

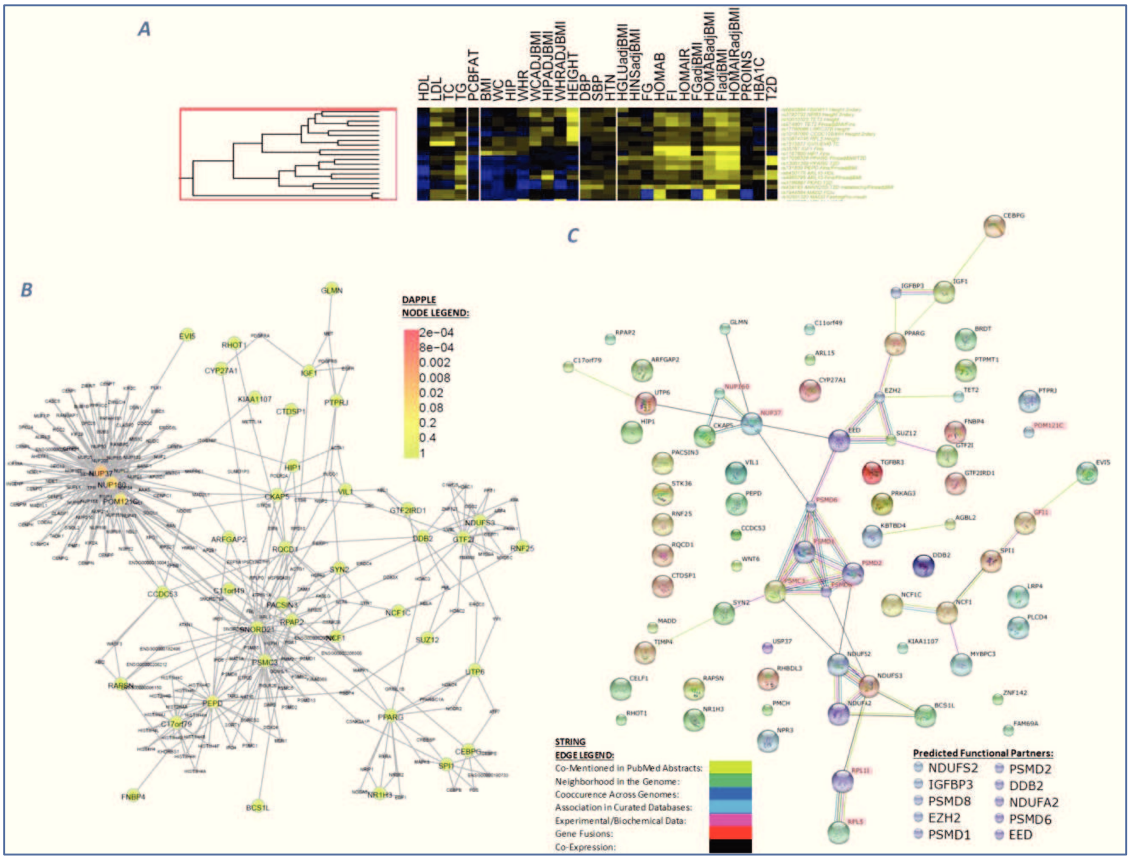


Figure 3.11: *Example of a sub-cluster of cardiometabolic loci obtained with cut-off A of the whole cluster and without a uniform trend of multiple effects.* ***A****. Zoom on the heat-map of this sub-cluster;* ***B****. Network obtained through pathway analysis with DAPPLE software;* ***C****. Network obtained through pathway analysis with STRING software: 10 connectors are added by the programme (see legend below), red rectangles highlight the name of genes involved in the most significant enriched biological process, edge colour are explained in the legend below. Comparable results were obtained with GeneMANIA software.*

Sub-clusters of cardiometabolic loci characterised by an effect on a single phenotype or on a specific subgroup of phenotypes

Some sub-clusters were characterised by an effect on a single phenotype or on a specific subgroup of phenotypes (lipids, glycaemic, blood pressure, obesity) with a uniform trend of multiple effects along all the included loci.

An interesting example is represented in figure 3.12 (H25_4). This highly supported group (bootstrap value = 93%) contains four SNPs (rs4420638, rs629301, rs6511720, rs1367117) that map near 15 genes (within 100 kb), as indicated by DAPPLE software.

The effect of the four variants is limited to lipids only, with low HDL cholesterol and high TC, LDL and TG: a combination of effects that is consistent with epidemiological expectation (figure 3.12A).

The group of loci reported a significant excess of direct connections (DAPPLE p-value = 0.001) and of simultaneous interactions of input genes to common connectors (DAPPLE p-value = 0.003, figure 3.12B).

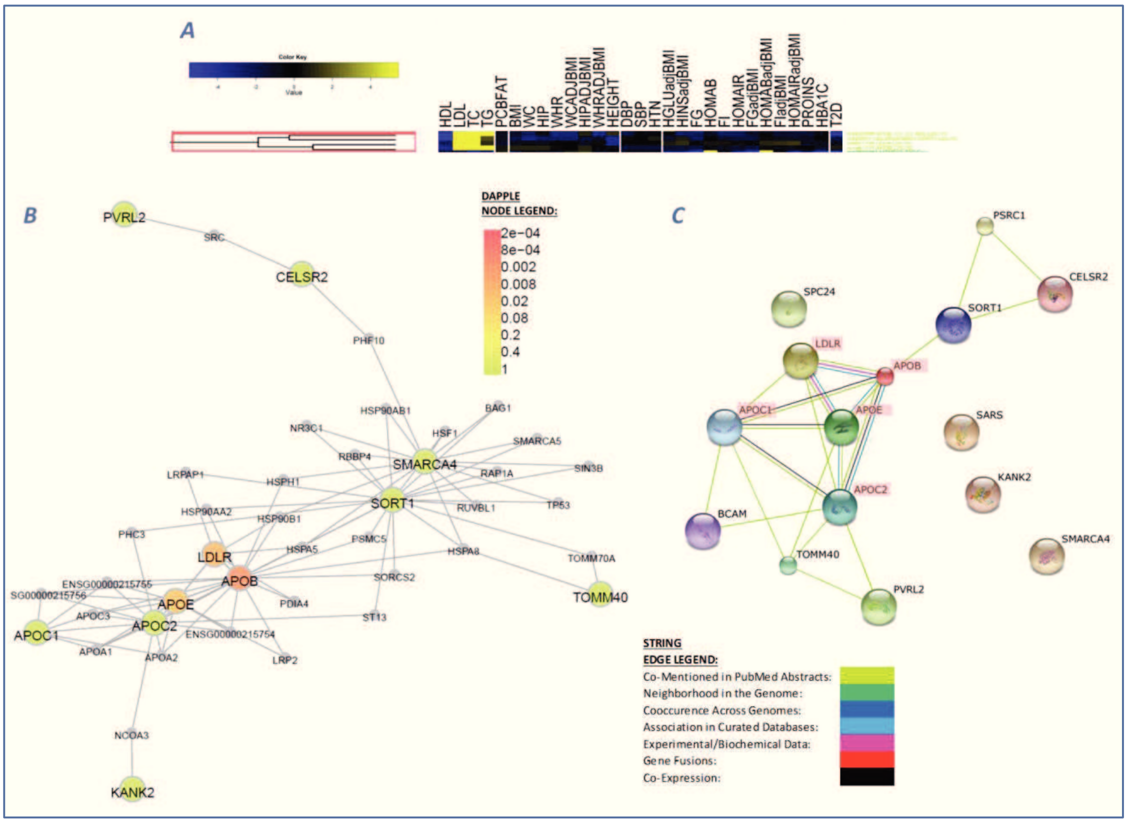


Figure 3.12: *Example of a sub-cluster of loci with effects on lipids only.* ***A****. Zoom on the heat-map of this sub-cluster;* ***B****. Network obtained through pathway analysis with DAPPLE software;* ***C****. Network obtained through pathway analysis with STRING software: no connector is added, red rectangles highlight the name of genes involved in the most significant enriched biological process. Same result was obtained with GeneMANIA software.*

This high significance was confirmed by STRING and GeneMANIA, which both indicated enrichment for plasma lipoprotein particle clearance and remodelling (STRING q-value after FDR correction = 4.25×10^{-8}, GeneMANIA FDR q-value = 2.45×10^{-7}) and regulation of phospholipid catabolic process (STRING FDR q-value = 3.69×10^{-5}), without the addition of further interactors (figure 3.12C). This enrichment was attributable to five genes, as reported by both STRING and GeneMANIA: *APOC1*, *APOC2*, *APOE*, *APOB* and *LDLR*.

The APOC1 protein modulates the interaction of APOE with beta-migrating VLDL (very-low density lipoproteins), while APOC2 is a component of VLDL that activates the enzyme lipoprotein lipase: VLDL becomes thus LDL. LDLR is a low density lipoprotein receptor placed at the cell membrane: it binds LDL, the major cholesterol-carrying lipoprotein of plasma, thanks to the mediation of ligands APOE and APOB, and it transports it into cells by endocytosis.

Another example is reported in figure 3.13. As in the previous group of loci, in this one (called H25_8) it is also possible to notice a strong effect on lipids, but in this second case on LDL, TG and TC only (figure 3.13A). This sub-cluster had high support after multiscale bootstrap of clustering results (bootstrap value = 87%).

The sub-cluster is composed of 15 variants originally associated with lipids, at genome-wide significance, and significantly enriched for indirect connection between input genes (DAPPLE p-value = 0.05, figure 3.13B).

Also in this case STRING and GeneMANIA agreed reporting significant enrichment of some biological processes, in particular of negative regulation of cholesterol and sterol transport (STRING FDR q-value = 2.53×10^{-9}, GeneMANIA FDR q-value = 4.15×10^{-5}) with the involvement of genes such as *APOE*, *APOB*, *APOC1*, *APOC2*, *ABCG5*, *ABCG8*, *PCSK9* and *MYLIP* (figure 3.13C). The highlighted connections are significant without adding any other connector to input genes.

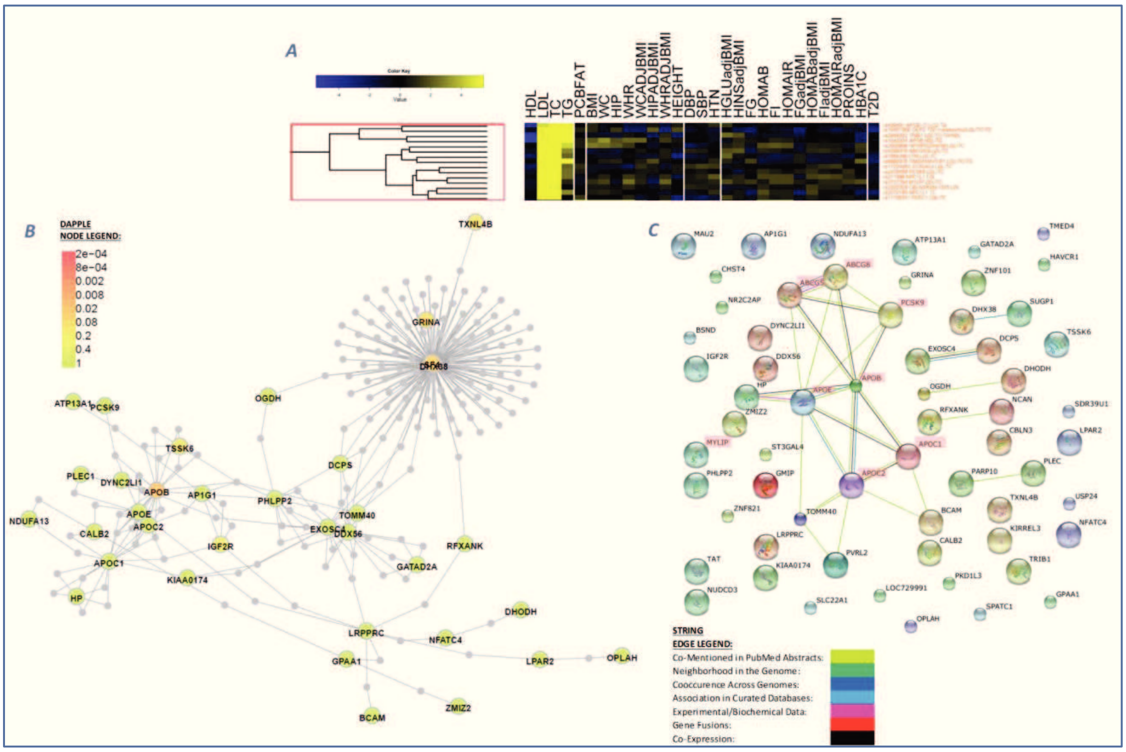


Figure 3.13: A second example of a sub-cluster of loci with effects on lipids only. ***A****. Zoom on the heat-map of this sub-cluster;* ***B****. Network obtained through pathway analysis with DAPPLE software;* ***C****. Network obtained through pathway analysis with STRING software: no connector is added, red rectangles highlight the name of genes involved in the most significant enriched biological process. Same result was obtained using GeneMANIA software.*

A third example is represented by sub-cluster H15_22 in figure 3.14. Here, 19 SNPs have a strong effect on blood pressure and hypertension (figure 3.14A), although bootstrap analysis did not support it. This lack of support could be attributable to heterogeneous minor effects on other phenotypes.

The DAPPLE software did not reveal any significant interaction for this group, but instead STRING highlighted an enrichment of 11 input genes for heterocycle metabolic processes (FDR q-value = 0.006, figure 3.14B).

Interestingly, in this case, GeneMANIA suggested a different enriched pathway for some of the same genes (*ADM*, *NPPB*, *GUCY1A3*, *GUCY1B3*) when 10 interactors were included in the analysis: circulatory system process pathway (FDR q-value = 0.04, figure 3.14C). No significant enrichment was observed using the GOrilla software.

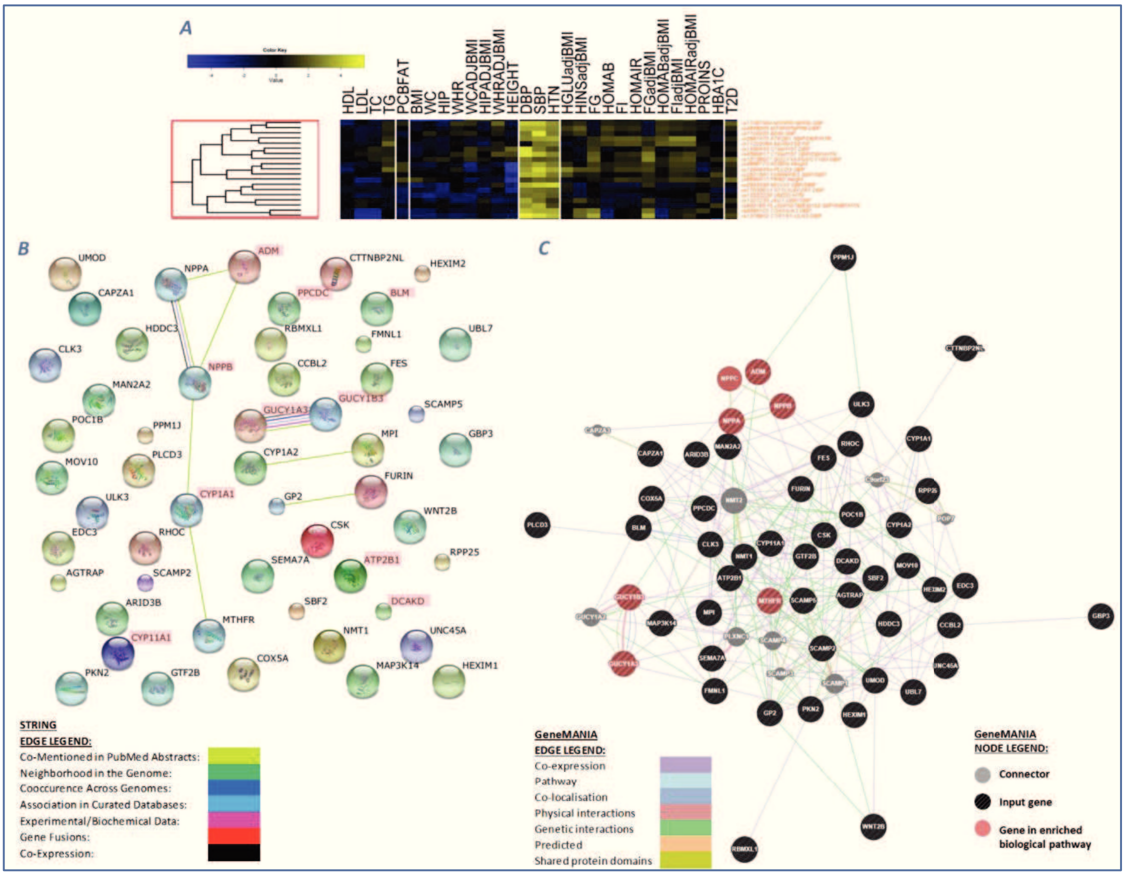


Figure 3.14: *Sub-cluster of BP associated loci.* ***A****. Heat-map of cardiometabolic effects;* ***B****. Network obtained using STRING software, red rectangles highlight the name of genes involved in heterocycle metabolic biological process;* ***C****. Network obtained using GeneMANIA software with 10 connectors added by the programme, red circles highlight the name of genes involved in circulatory system process.*

Sub-clusters with unexpected effects on a specific subgroup of phenotypes

Some of the observed sub-clusters were characterised by multiple effects on phenotypes belonging to the same subgroups of phenotypes, but with unexpected directions or combinations of these effects. We were interested to discover if these complex combinations of outcomes could be indexes of unpredicted biological processes.

For example, figure 3.15 represents a group of 14 SNPs (H15_25) with strong effects on lipids, but with a strange pattern: in fact we could observe a strong effect leading to lower LDL and TC, but also unexpectedly to lower HDL, and less strong effect on TG (figure 3.15A).

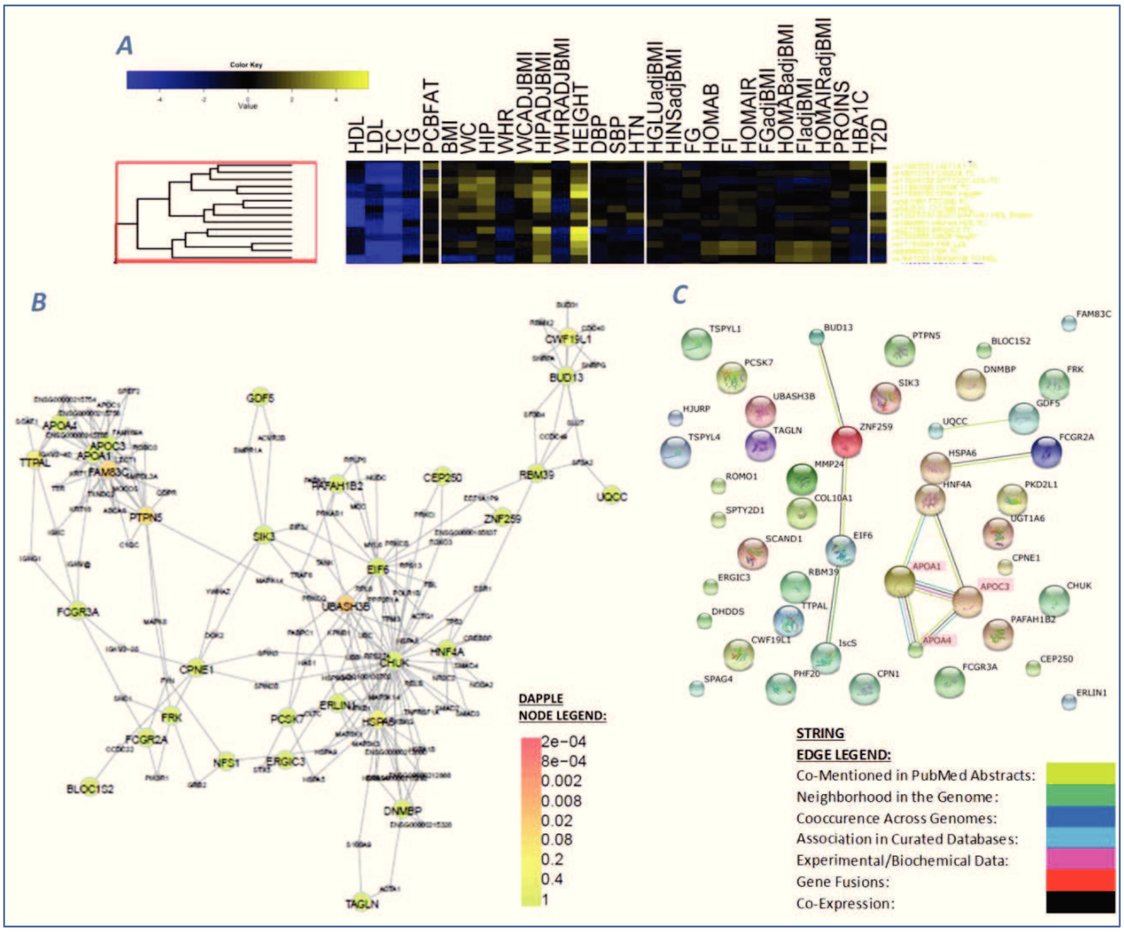


Figure 3.15: *Sub-cluster of lipids-associated loci with strange pattern of effects.* ***A****. Heat-map of the effects;* ***B****. Network obtained through pathway analysis with DAPPLE software with excess of direct edges and common interactors;* ***C****. Network obtained using STRING software, red rectangles highlight the name of genes involved in the most significant enriched biological process; comparable results were obtained using GeneMANIA programme.*

The sub-cluster was not supported by bootstrap analysis, but this could be due to heterogeneity of minor effects on other phenotypes, as for example on height and glycaemic traits.
This group of variants in DAPPLE led to a list of 45 genes with high degree of direct edges between input nodes (DAPPLE p-value = 0.04) and of common interactors (DAPPLE p-value = 0.007, figure 3.15B). In the network, the presence of factors such as *UBASH3B* (ubiquitin associated and SH3 domain containing B), *HSPA6* (heat shock protein 6), *PTPN5*, *FAM83* (family with sequence similarity

83 member E) and *TTPAL* (tocopherol transfer protein-like) was significant; some proteins encoded by these genes are involved in proteins transport and folding. Pathway analysis was significant using STRING software with an enrichment of phospholipid efflux and protein-lipid complex assembly (p-value = 2.37×10^{-6}, FDR q-value = 0.02, figure 3.15C), a result that was confirmed also using GeneMANIA (FDR q-value = 0.01), but not using GOrilla.

Another example of epidemiologically unexpected effects on related phenotypes is a sub-cluster which includes variants mapping near known T2D and glycaemic-associated loci such as *ADCY5*, *CDKN2A/B*, *PCSK1*, *ARAP1* and others (figure 3.16A, cluster H25_13). This sub-cluster presented a singular pattern of effects on glycaemic traits: in fact we could observe a strong decrease of FG flanked by a complete opposite increase of β-cell function (HOMAB) and, with less intensity, of FI; this outcome was not mediated by an association with BMI as it was maintained after BMI adjustment.

This picture can be explained by a defect on the functionality of β-cells (rather than on insulin resistance) which causes an impaired production of insulin, even if high levels of glucose in the blood are present, leading to an inadequate response, and thus to a persistent hyperglycaemia and risk of developing T2D (a suggestive effect of this increased risk can be observed in figure 3.16A). In fact, if we considered the effects attributable to the alternative alleles of the reported loci, we would observe high FG, but low FI and HOMAB, and thus suggestive high T2D risk.

This group of variants was supported by a bootstrap value of 46%, a quite low value that could be attributable to the effects of rs174546 near the *FADS1* locus: this variant in fact is differentiated by the rest of the group as it shows additional strong effects on lipids. In the dendrogram, rs174546 firstly separated from the rest of the group and, when we excluded it, the bootstrap value raised to 68%.

When we analysed the group in a pathway analysis, DAPPLE did not find enriched connections, but just three networks, as represented in figure 3.16B. The STRING software, instead, revealed enrichment for response to carbohydrate stimulus pathway (p-value = 1.09×10^{-6}, FDR q-value = 0.01) and pancreas development (p-value = 3.05×10^{-6}, FDR q-value = 0.02, figure 3.12C). In addition, GeneMANIA was significant for peptide transport (FDR q-value = 0.014) and insulin secretion (FDR q-value = 0.014) processes when 10 interactors were added to input genes (figure 3.16D).

Using a smaller threshold (cut-off C) of the Euclidean distance, this sub-cluster was further subdivided in two parts, H15_35 and H15_37. H15_37 was remarkably significant in pathway analysis, and it could be responsible also for the total significance of the bigger sub-cluster H25_15 to which it belongs (figure 3.17A). In fact, this group had a strong bootstrap value (90%) and borderline degree of direct connections (DAPPLE p-value = 0.06). In STRING the group of loci was particular enriched for pancreas development biological process (p-value = 9×10^{-7}, FDR q-value = 0.008, figure 3.17B). The same result was confirmed by GeneMANIA, where peptide hormone secretion and pancreas development were proposed as enriched pathways with comparable significance (FDR q-value = 0.015). Highlighted biological processes involved in four particular factors: *PDX1* (pancreatic and duodenal homeobox 1), *FOXA2* (forkhead box A2), *SLC2A2* and *PCSK1*.

These are all factors involved in insulin/proinsulin secretion and β-cell/pancreatic islets development.

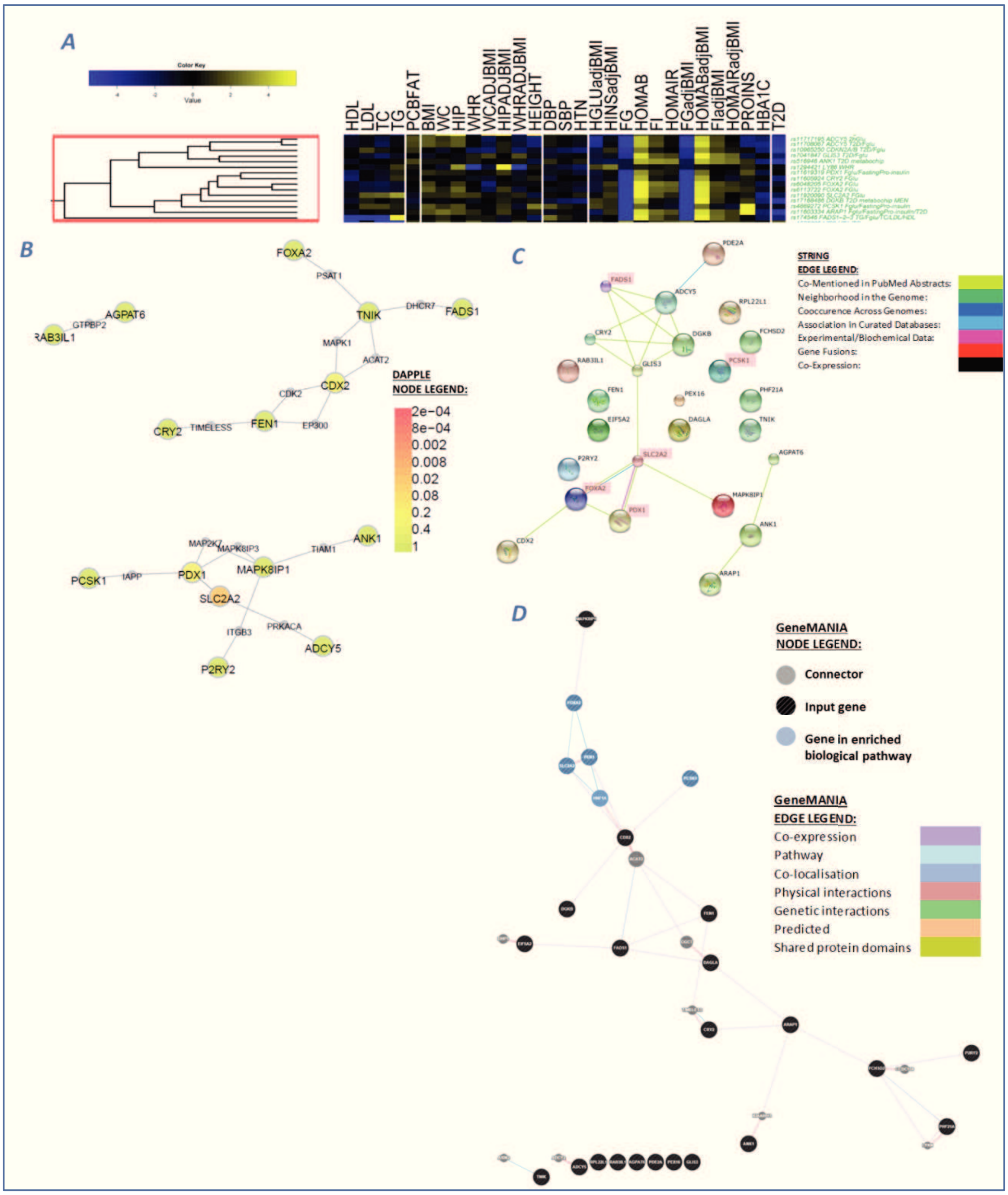


Figure 3.16: *Sub-cluster of loci with a strange pattern of effects on glycaemic traits.* ***A****. Heat-map of the effects;* ***B****. Three networks obtained through pathway analysis with DAPPLE software;* ***C****. Network obtained using STRING software, red rectangles highlight the name of genes involved in carbohydrate stimulus pathway; no connectors are added.* ***D****. Network obtained using GeneMANIA software, blue circles highlight the name of genes involved in peptide transport process.*

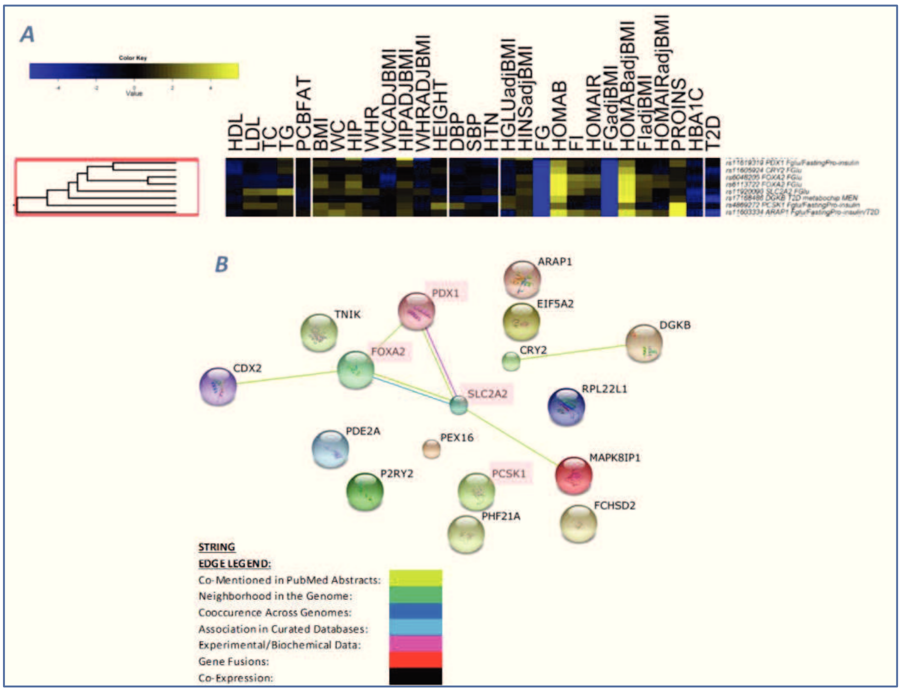


***Figure 3.17:** Sub-cluster derived from a further subdivision of group in figure 3.16. **A.** Heat-map of the effects; **B**. Network obtained using STRING software, red rectangles highlight the name of genes involved in pancreas development and insulin signalling processes; similar results were obtained using GeneMANIA programme.*

Sub-clusters with multiple effects consistent with the definition of MetS

In the whole cluster of multiple cardiometabolic effects, we distinguished sub-clusters of loci with multiple effects on multiple phenotypes belonging to different groups of related phenotypes; by analysing the patterns of those effects, we identified groups of loci which behaved in a way that is consistent with metabolic syndrome definition (MetS).

As defined in chapter "2.3.3.1_Proposed models: Metabolic Syndrome", MetS is characterised by the concurrent presence of some cardiometabolic phenotypes that cluster together: increased risk of T2D, increased obesity, high blood pressure high triglycerides, low HDL-cholesterol levels and presence of insulin resistance[157]. Some of the identified sub-clusters in our data presented several of these aspects together.

The group in figure 3.18 is one example: this sub-cluster of variants (H25_12) is characterised by a strong positive effect on height, accompanied by a general increase of obesity-related traits (WC, HIP and WHR with or without adjustment for BMI), but not of BMI. Even if of minor intensity, the main effects are combined with a MetS-compatible trend of all other cardiometabolic phenotypes, especially glycaemic traits (figure 3.18A).

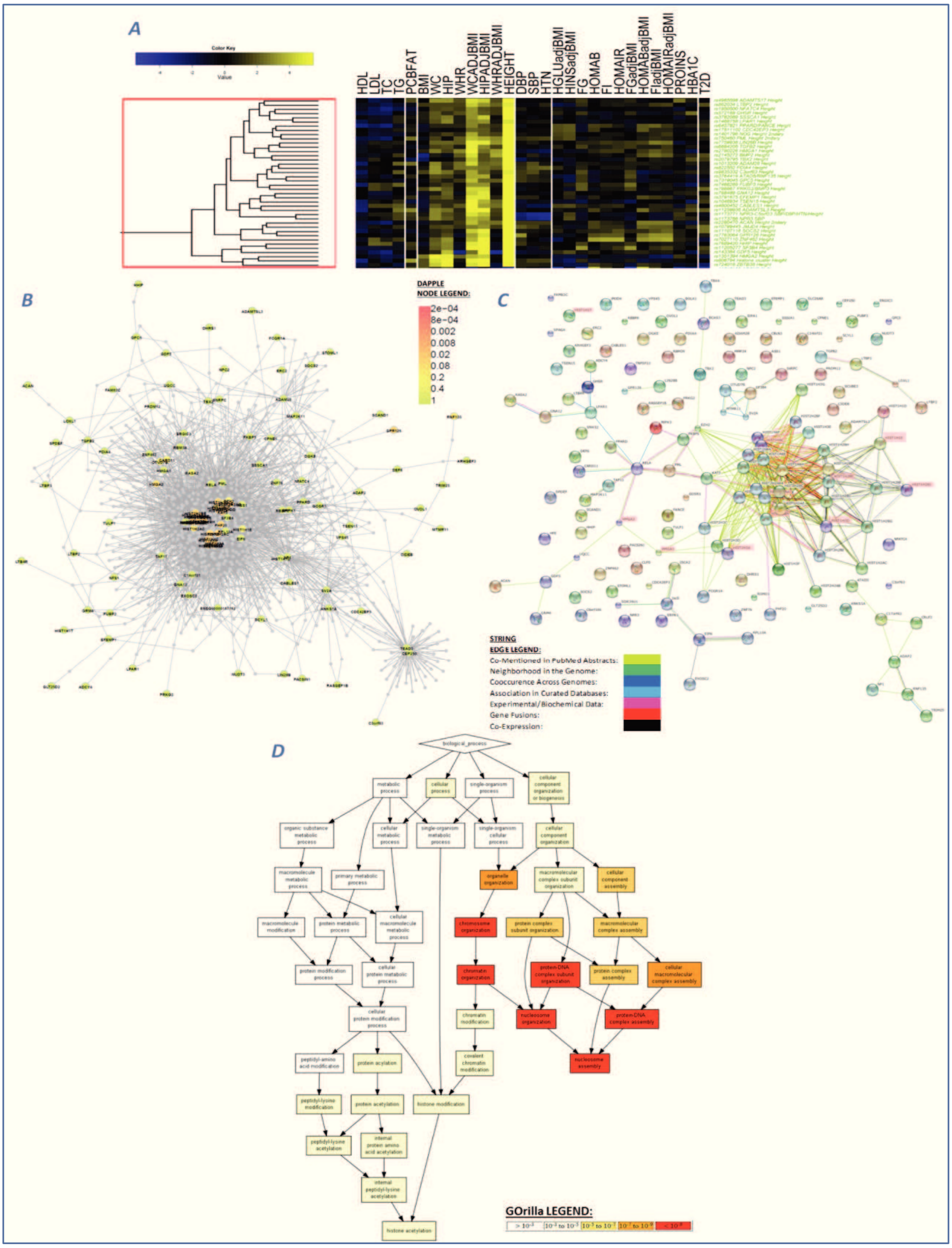


Figure 3.18: *Sub-cluster of loci with a pattern of multiple effects that is suggestively compatible with MetS definition.* ***A****. Heat-map of the effects;* ***B****. Network obtained through pathway analysis with DAPPLE software;* ***C****. Network obtained using STRING software, red rectangles highlight genes involved in chromatin assembly process; GeneMANIA gave comparable results.* ***D****. Network of biological processes reconstructed using GOrilla software.*

This sub-cluster was enriched for direct connections between genes near input variants (DAPPLE p-values = 0.001), but also for indirect connections (DAPPLE p-value = 0.01) and for common interactors (DAPPLE p-value = 0.001), as reported in figure 3.18B. The strong significance is attributable to the numerous identified genes within a histone cluster element in the DNA.
Pathway analysis with other software tools revealed concordance of results which supported this hypothesis: STRING showed a strong enrichment for chromatin assembly process (p-value = 4.24×10^{-11}, FDR q-VALUE = 4.72×10^{-7}, figure 3.18C) and GeneMANIA agreed with this result (FDR q-value = 6.83×10^{-7}, data not shown), proposing also the nucleosome related pathway as a more general enriched biological process (FDR q-value = 4.94×10^{-7}).

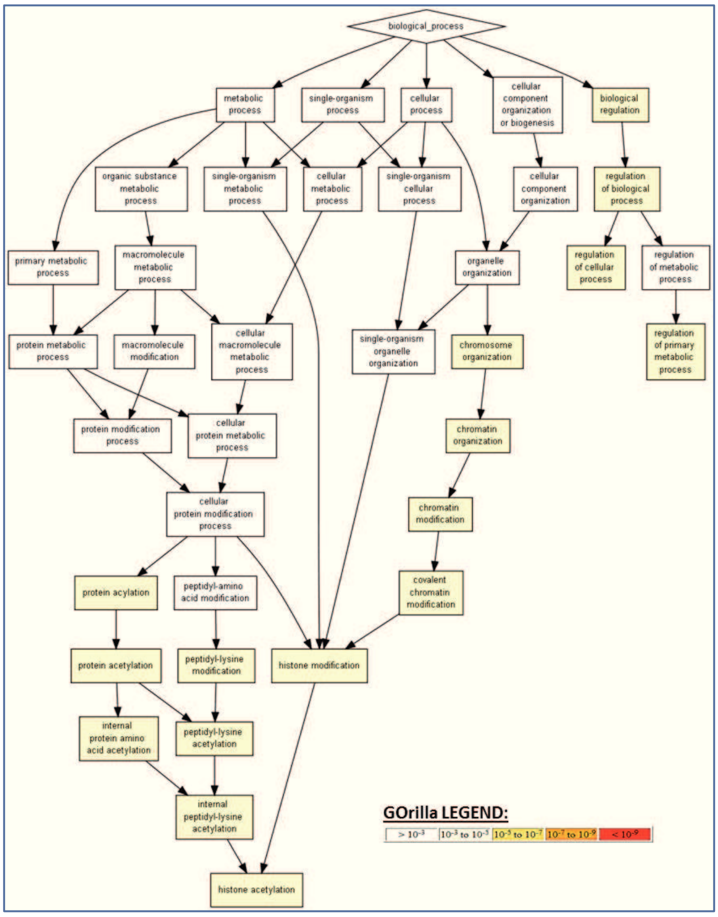


Figure 3.19: *Network of biological processes reconstructed using GOrilla software starting from data of sub-cluster described in figure 3.18, after removal of histone cluster genes. Two parallel processes are proposed: histone modification and regulation of cellular process.*

The GOrilla software produced comparable results, with the reconstruction of a highly significant (p-value = 1.55×10^{-13}, FDR q-value = 2.99×10^{-10}) enriched pathway, as reported in figure 3.18D, which involves histone cluster genes and other genes (*EZH2*, *ZNF462*, *KAT5*, *HMGA2*, *PHF20*, see below for a description) in chromatin organisation, and nucleosome assembly pathways.

Histone cluster genes were attributable to only one associated variant in this group, rs80674; therefore when we removed it from variants included in the group, the results of the pathway analysis changed, shifting to an enrichment of aging process in STRING (STRING p-value = 5.84×10^{-6}, FDR q-value = 0.07) and GeneMANIA (even if for GeneMANIA, the FDR q-value was not significant = 0.5). After this removal, GOrilla maintained histone modification as a common biological process, but with less significance (p-value = 2×10^{-5}, FDR q-value = 0.04), and involving only *EZH2* (enhancer of zeste homolog 2), *KAT5* (K(lysine) acetyltransferase 5), *HMGA2* (high mobility group AT-hook 2) and *PHF20* (PHD finger protein 20); it also proposed regulation of cellular process pathway (p-value

= $6x10^{-5}$, FDR q-value = 0.023) as a common pathway for 35 input genes (see figure 3.19).

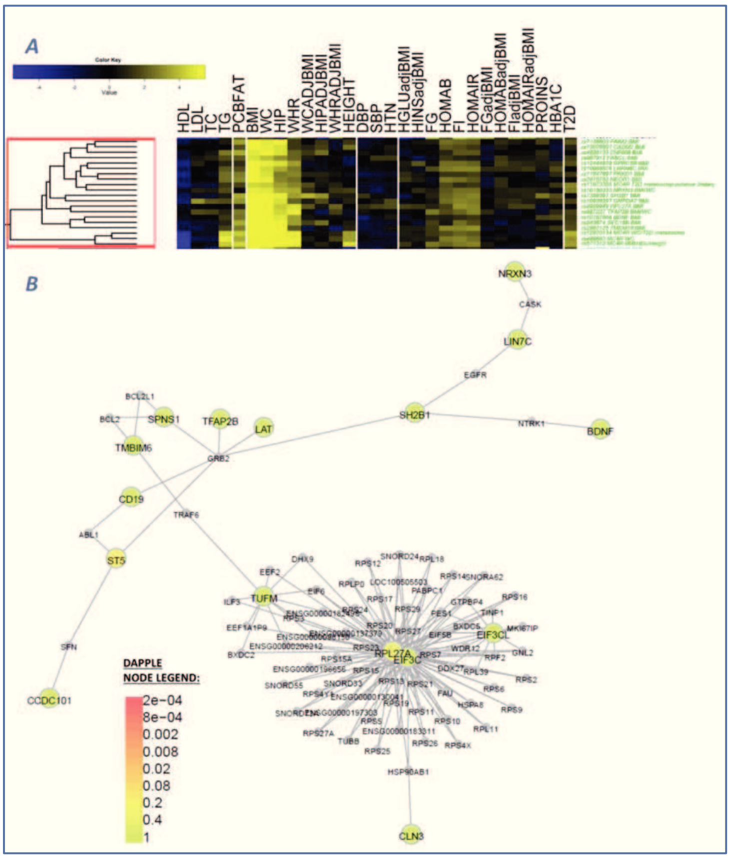


Figure 3.20: *Sub-cluster of loci with a pattern of multiple effects consistent with MetS definition.* ***A.*** *Heat-map of the effects;* ***B.*** *Network obtained through pathway analysis with DAPPLE software.*

A second example of a MetS compatible sub-cluster is H15_42 in figure 3.20: it is a highly supported (bootstrap value = 95%) group of 20 BMI-associated variants with a pronounced increasing effect on BMI and also on WC, HIP and WHR, even if this effect resulted mediated by BMI association. In fact no remarkable association was found for WCadjBMI, HIPadjBMI and WHRadjBMI. Suggestive increase was also reported for some glycaemic traits (HOMAB, FI and HOMAIR) and for T2D risk, as well as for TG and PBFAT; a decrease instead was observed for HDL (figure 3.20A). The 20 included SNPs were near 27 genes (within flanking 100kb regions) and their analysis in DAPPLE revealed a borderline significance for an excess of common interactors (p-value = 0.07, figure 3.20B). No significant enrichment was observed using other tools for pathway analysis.

Another interesting sub-cluster is represented in figure 3.21 (group H15_53): this is a group of 17 SNPs with a strong effect on T2D. An increasing effect, even if of minor significance, was observed also for the other traits, but not for HDL. The trend described in this picture can be interpreted as epidemiologically expected (figure 3.21A). Pathway analysis using the DAPPLE software did not reveal significant excess of connections; while using STRING (figure 3.21B) and GeneMANIA (figure 3.21C), with the addition of 10 interactors between input genes, we obtained significant enrichment for regulation of cell cycle process including interphase, G1 phase, and mitosis (STRING p-value = $1.49x10^{-15}$, STRING FDR q-value = $1.66x10^{-11}$, GeneMANIA FDR q-value = 0.03).

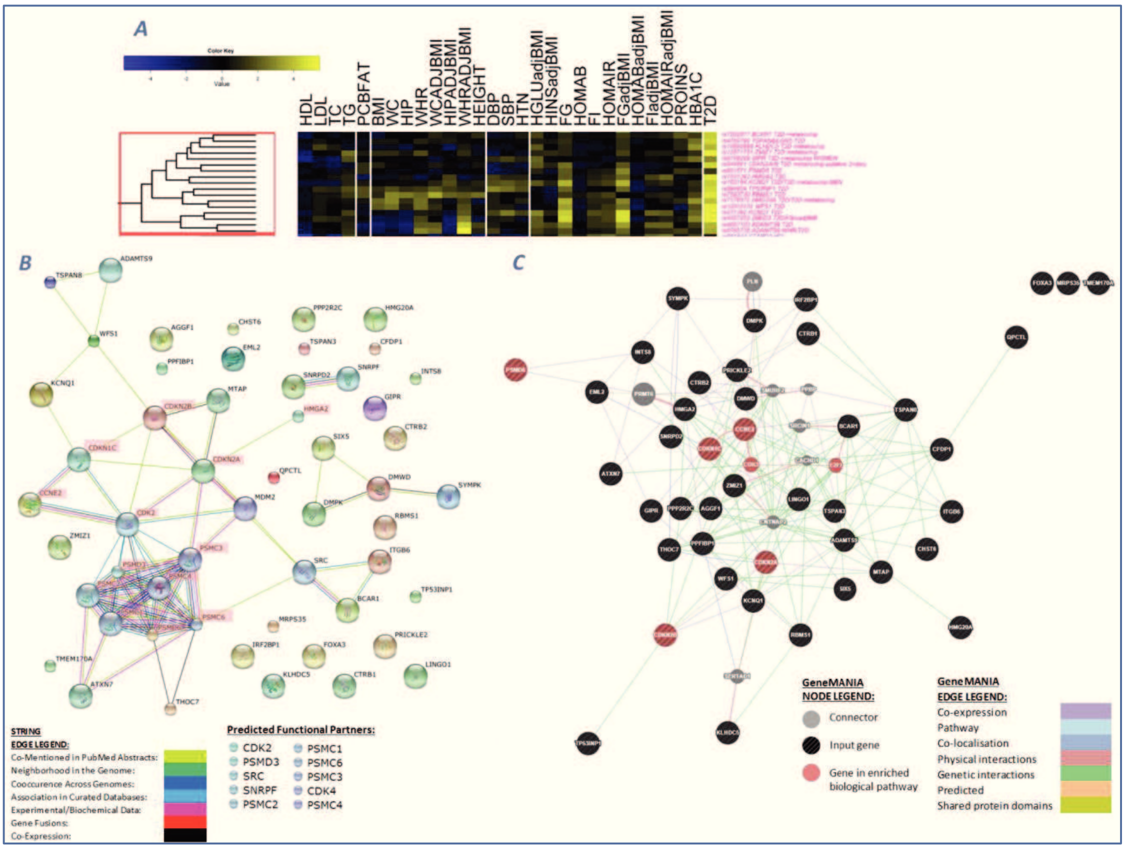


Figure 3.21: *Sub-cluster of loci with a pattern of multiple effects that is compatible with MetS definition.* ***A****. Heat-map of the effects;* ***B****. Network obtained through pathway analysis using STRING software, 10 interactors are added (blue and light-blue circles in the legend), red rectangles highlight the name of genes involved in the most significant enriched biological process;* ***C****. Network obtained using GeneMANIA software with comparable results.*

Sub-clusters with multiple unexpected effects

Other sub-clusters of loci showed multiple effects on phenotypes belonging to different groups of related phenotypes, but not following our expectations according to the epidemiological definition of MetS or to epidemiological expectations. Several of these groups were, in fact, were characterised by particular combinations of multiple effects, which can reveal novel involved pathways that should be considered in the knowledge of cardiometabolic phenotypes.

A first example is the group of 23 SNPs that we called H25_7: it presented with a varied pattern of strong effects, significantly supported when we applied a bootstrap test on the cluster (bootstrap value = 68%). Describing the observed pattern of effects in an ordered manner (see figure 3.22A), we can firstly see an expected effect on lipids, with very low HDL, high LDL (that brings to a general high level of TC), and very high TG. This is accompanied by increased glycaemic trait levels and T2D risk and a less strong increase of blood pressure and hypertension risk. These described effects are expected, but obesity-related traits revealed an unpredictable behaviour: body fat percentage and

BMI, in fact, were decreased.

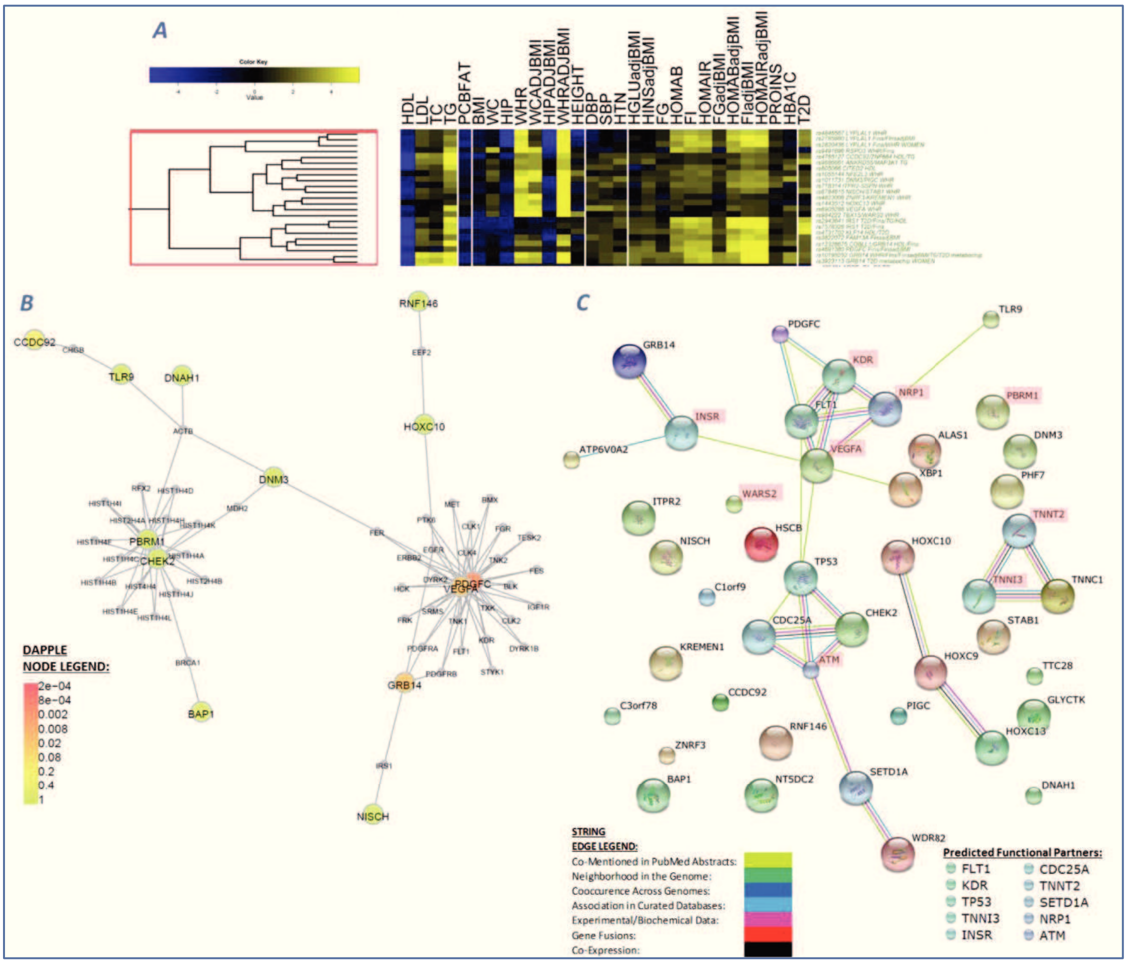


Figure 3.22: *Sub-cluster of loci with a HOUL pattern of multiple effects.* ***A****. Heat-map of the effects;* ***B****. Network obtained through pathway analysis with DAPPLE software.* ***C****. Network obtained using STRING software, 10 interactors are added (blue and light-blue circles in the legend), red rectangles highlight the name of genes involved in cardiovascular development process.*

This picture can be interpreted as lean individuals, but with high lipids levels and a compromised state of metabolic health (T2D, HTN); we called this state healthy obesity/unhealthy leanness (HOUL). By exploring the patterns of effects on obesity traits in more detail, it was possible to notice that low BMI is followed by low hip circumference, maintained also after BMI adjustment, but high WHR and WCadjBMI: this observation revealed that the HOUL condition described by the variants in this group is attributable to an overall normal or low BMI, but with high levels of central adiposity. This characteristic is typical of an "apple shaped" body type, where fat is predominantly deposited on the visceral region of the waist, as explained in figure 3.23, a status usually associated with "dysmetabolism" and cardiovascular diseases.

When analysing the included SNPs in pathway analysis, DAPPLE recognised 33 flanking genes with strong indirect connections (p-value = 0.01, figure 3.22B). This discovery was also supported by the STRING software, which revealed enrichment for vascular endothelial growth factor signalling pathway (FDR q-value = $5.75x10^{-5}$) and cardiovascular development (FDR q-value = 0.018) when 10 connectors were added to input genes (as described in figure 3.22C). Other tools did not replicate the results.

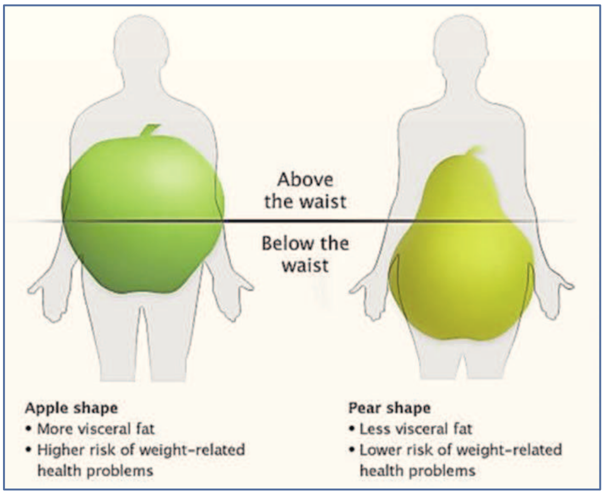


Figure 3.23: Graphical representation of "apple shaped" (on the left) and "pear shaped" (on the right) body type. In "apple shape" fat is more visceral, while in "pear shape" it is predominantly deposited on the hips and buttocks.

A second example is in figure 3.24A (group H25_11): here an unexpected pattern on lipids with low levels of TG, TC and LDL, accompanied also by normal or low levels of HDL, manifested together with high levels of BMI and height. The described combination of effects can be explained by a medical case of obesity without any effects on lipidemia (compatible with HOUL definition). The group was highly supported in bootstrap analysis (bootstrap value = 84%) and revealed a high level of common interactors in DAPPLE pathway analysis (p-value = 0.001). GeneMANIA suggested two possible enriched pathways for the group: ER to Golgi transport vesicle membrane pathway (FDR q-value = $3.49x10^{-8}$) and immune response-activating cell surface receptor signalling pathway (FDR q-value = $3.53x10^{-6}$, figure 3.24B), but this was not confirmed by other pathway analyses.

Finally, the sub-cluster in figure 3.25A (H15_43) is another example of unusual multi-phenotype effects with high BMI and obesity traits, low HDL, but also low blood pressure, and low glycaemic traits (FI, HOMAB and HOMAIR) after adjustment for BMI. This is a particular case of HOUL where hypotension manifests in obese individuals. The group was particularly enriched for indirect connections between input genes (DAPPLE p-value = 0.02, figure 3.25B). STRING confirmed this result with a significant enrichment of DNA damage response as signal transduction by p53 class mediator when 10 interactors are added to the pathway (p-value = $6.49x10^{-10}$, FDR q-value = $6.98x10^{-6}$, figure 3.25C). Another suggestive enriched pathway was regulation of protein metabolic process (p-value = $3.81x10^{-8}$, FDR q-value = $4.62x10^{-5}$, figure 3.25C).

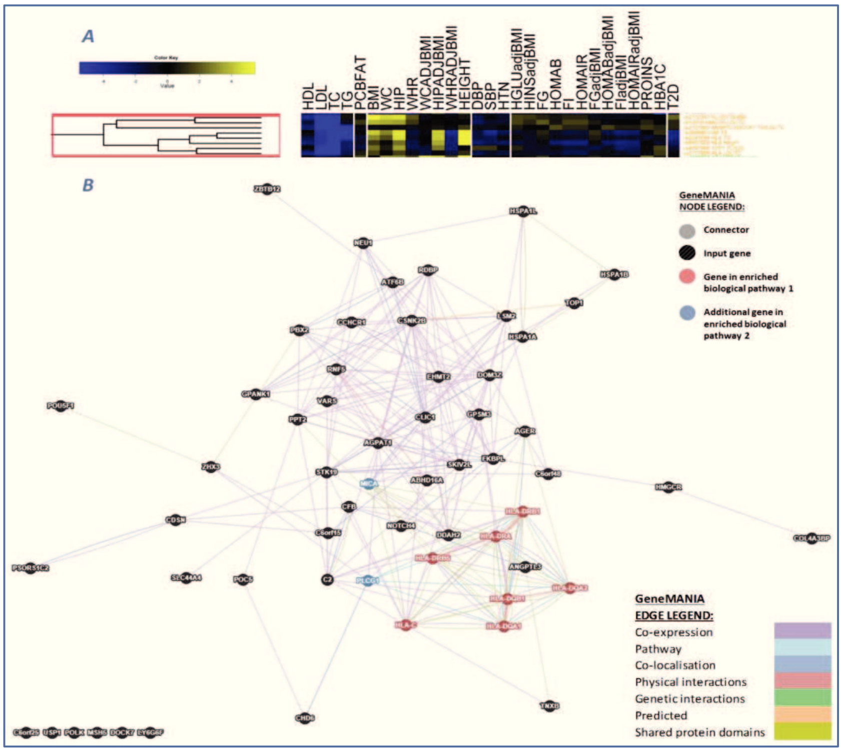


Figure 3.24: *Sub-cluster of loci with unexpected pattern of multiple effects.* ***A.*** *Heat-map of the effects;* ***B.*** *Network obtained through pathway analysis with GeneMANIA software, red circles highlight the name of genes involved in ER to Golgi transport vesicle membrane pathway, blue circles highlight the name of genes that, together with the red ones, are involved in immune response-activating cell surface receptor signalling.*

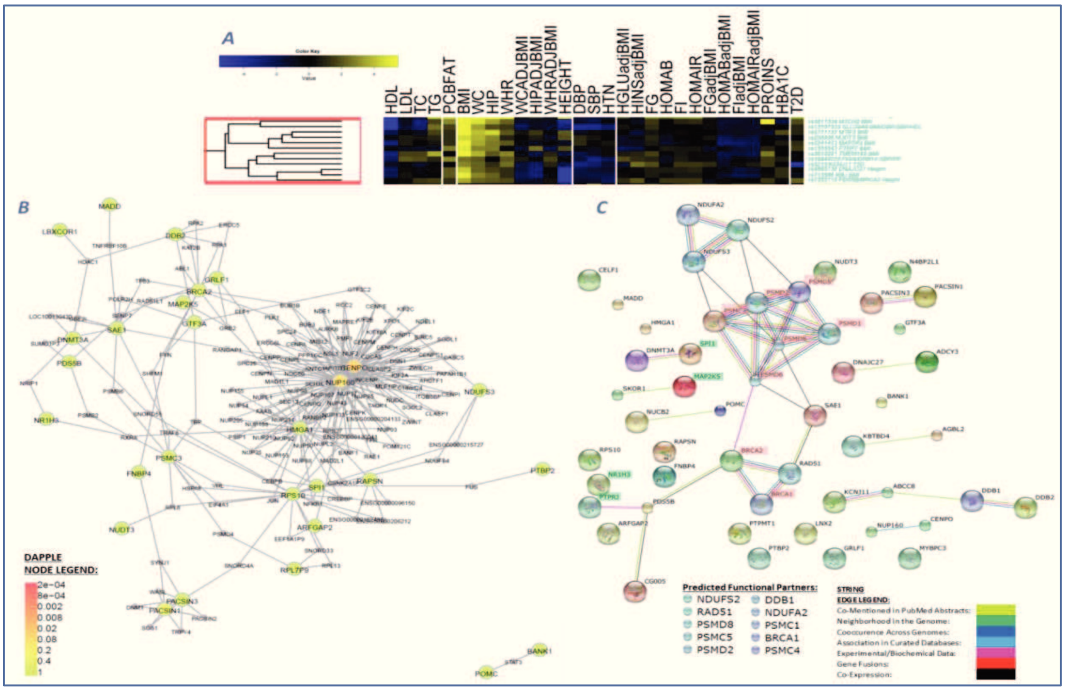


Figure 3.25: *Sub-cluster of loci with a HOUL pattern of multiple effects.* ***A.*** *Heat-map of the effects;* ***B.*** *Network obtained through pathway analysis with DAPPLE software.* ***C.*** *Network obtained through pathway analysis with STRING software, 10 interactors are added (blue and light-blue circles in the legend); red rectangles highlight the name of genes involved in DNA damage response, green rectangles highlight additional genes that, together with red ones, are involved in regulation of protein metabolic process.*

3.2.4 Discussion

In this study we aimed to explore patterns of multiple cross-phenotype effects for cardiometabolic traits and diseases across the genome by analysing known cardiometabolic loci in existing univariate analysis data and to discover possible biological processes which have a role in the regulation of metabolism and, thus, an influence on risk of cardiometabolic diseases.

To have a general view of the extent of noteworthy multiple effects in our data, we applied a simple meta-analysis approach on p-values of association from univariate analyses. 109 variants of the analysed 544 (20%) showed significant Omnibus test p-values as result of the combination of multiple significant univariate associations in unrelated phenotypes. 86 (15.8%) of these SNPs showed multiple signals of almost equivalent significance in different phenotypes. Some of them are already known signals for multiple traits or diseases reported by the literature, such as those near *GRB14, KLF14, IRS1* and *C6orf106*[19,81,119,99,139]; others were novel, especially because most of them did not reach genome-wide significant levels of univariate association for secondary traits, but gave highly significant Omnibus p-values when single trait results were combined: variants near *PPP1R3B*, *PPARG*, *MTCH2* (mitochondrial carrier 2), *PEPD* (peptidase D), *ZNF462* are just a few examples.
Fisher's omnibus p-value test was useful to highlight variants at established cardiometabolic loci with multiple associations, revealing that around the 15% of known cardiometabolic-associated SNPs could be potentially pleiotropic on phenotypes characterising different aspects of metabolism (for example obesity and blood pressure, or lipids and glycaemic levels).
Nonetheless, this method has some limits. First of all this approach does not take into account the effects, but just the p-values, therefore it did not shed light on the modalities of multiple association in comparison with epidemiological expectations. Additionally, it did not allow us to easily identify groups of loci with similar patterns of multiple effects and to clarify the degree of connection between them, useful information for studying biological pathway enrichment.

To remedy for these limitations, and to achieve the aims of this research, we therefore considered z-score values from 29 cardiometabolic GWAS meta-analysis results and we applied a clustering analysis of multiple effects. We identified several groups of loci with similar patterns of multi-phenotype effects (see figures 3.9 and 3.10). Our results suggested that cardiometabolic loci predominantly share same multiple effects within little groups (average size: 7, from groups with bootstrap value ≥ 65%), most of which are probably representing a distinct mechanism that participates to the characterisation of involved phenotypes.
Among identified groups of effects, we were able to distinguish and categorise five different behavioural trends: (1) sub-clusters of cardiometabolic loci without a uniform trend of multi-phenotype effects, (2) sub-clusters of cardiometabolic loci characterised by effects on a single phenotype or on a specific group of related phenotypes, (3) sub-clusters with unexpected effects on a specific group of related phenotypes, (4) sub-clusters with multiple effects consistent with the definition of MetS, (5) sub-clusters with multiple epidemiologically unexpected effects.
Within these categories, several sub-clusters were particularly interesting because their

combination of effects or the functional connections between genes near their included variants, suggested unintuitive or peculiar biological processes involved.

An example is the group in figure 3.15: pathway analysis of variants included in this group suggested that a perturbed process of protein folding and transport related to the creation of protein-lipids complexes and involving factors such as UBASH3B, HSPA6, PTPN5, FAM83, TTPAL, APOE and APOC, may have strong effects on all lipids, bypassing the normal difference between HDL and the other lipid traits.

Another group of genes (figures 3.16, 3.17), among which *PDX1*, *FOXA2*, *SLC2A2* and *PCSK*, implicated in insulin/proinsulin secretion and β-cell/pancreatic islets development, confirmed the hypothesis that defects in the functionality of β-cells (rather than on insulin resistance), which cause an impaired production of insulin even if high levels of glucose are present in the blood, may lead to an hyperglycaemic status with consequent increased risk of developing T2D. This result thus supports the idea, already reported in literature, that β-cell dysfunction may be an important factor in T2D pathogenesis[19,99].

MetS is the clinical definition of a certain combination of cardiometabolic and inflammatory phenotypes, characterised by increased risk of T2D, increased obesity, high BP, high triglycerides, low HDL-cholesterol levels and presence of insulin resistance[157]. It is the most common, and thus epidemiologically expected, clinical manifestation for cardiometabolic phenotypes.

Our results highlighted that MetS is just one possible relationship, and that biological processes involved in cell cycle and cell processes may be of key importance in the determination of its status (figures 3.19, 3.21).

Alternatives exist, for example metabolically healthy obesity or unhealthy leanness, as we observed in our data (groups of loci in figures 3.22, 3.24 and 3.25).

HOUL individuals are normal weight patients who present dysmetabolic characteristics such as high lipids, LDL and/or glucose levels in the blood and high blood pressure, until the development of out-and-out metabolic diseases such as T2D, HTN or CAD; alternatively, HOUL status may describe obese individuals without any other cardiometabolic disorder and with normal levels of lipids, cholesterol, glycaemic traits and blood pressure, therefore in an excellent status of metabolic health.

From our results, a consistent number of cardiometabolic loci (at least 43, and up to 70, if we consider also other loci with suggestive effects) showed a pattern of effects which is compatible with HOUL. The majority of these loci (64 of 70, 91.43%) were genome-wide significant in Fisher's omnibus test (p-value $\leq 5\times10^{-8}$). From the analysis of factors encoded by these loci, cardiovascular development, DNA damage response and regulation of protein metabolism and transport, seem to be key biological processes involved in the determination of such cases.

In conclusion, this study enabled analysis of the extent of cross-phenotype effects of cardiometabolic variants and allowed identification of groups of loci with shared patterns of exerted effects. Pathway analysis revealed that some of these groups are enriched for loci that impact the same biological processes. These pathways may be expected, for example regulation of lipids metabolism or cholesterol transport for groups of loci with strong effects on lipids (figures 3.12, 3.13

and 3.15), or circulatory system processes for genes near blood pressure-association signals (figure 3.14); but sometimes the highlighted processes are counterintuitive, for example regulation of cellular process for a group of loci with effects on obesity and anthropometric traits (figures 3.18, 3.19).
In some cases, connectivity in multi-phenotype networks was useful in suggesting genes that are more likely for causality or tissues of action underlying the association signals (see the example described in figure 3.12). In some other cases, enriched networks were significant only in the presence of additional interactors that could be further investigated as candidate factors for association with implicated phenotypes.

The approach used in this first project revealed highly useful in recognising cross-phenotype effects using univariate GWAS results and in characterising these associations in terms of causal genes and biological mechanisms involved, contributing to shedding light on the processes that regulate physiological aspects of metabolism or that contribute to the risk of developing cardiometabolic diseases.
Nevertheless, this method has some limitations and cannot uncover all the aspects concerning the study of pleiotropy. First of all, it does not provide a measure of statistical significance of the best model that represents the pattern of multiple effects for each variant or groups of variants. Secondly, this approach does not allow discovery of novel variants across the genome, besides those already associated with at least one cardiometabolic phenotype in univariate studies, since it used already established SNPs from single-phenotype GWASs; this limit leaves out polymorphisms which could have a strong overall multiple effect without standing out in univariate GWAS analyses for single phenotypes and which, therefore, may contribute to part of the missing heritability of complex phenotypes. Finally, the undertaken study revealed cross-phenotype effects, but was not able to discern the real genetic mechanisms behind them. In other words, it did not discern real pleiotropy from mediation, even it did not deal with the interpretation of multi- phenotype signals which lie in common genomic regions, distinguishing multi-phenotype allelic heterogeneity from real overlapping signals.
To solve the additional issues described, different approaches must be developed and other methods must be tested on different types of data, as we applied in the following described sub-projects.

3.3 Project 2: Validating pleiotropy, and analysis of locus architecture in potential pleiotropic regions

3.3.1 Introduction and Aim

As we have already described in previous chapters (see “2.3_Overview of genetics of cardiometabolic phenotypes”), cardiometabolic phenotypes have complex aetiology and are epidemiologically correlated.
In the past years, GWAS have identified hundreds of novel susceptibility loci for cardiometabolic diseases and, interestingly, their findings have highlighted multiple loci that are associated with more than one cardiometabolic phenotype, suggesting shared molecular pathways[20]. In some cases, the same variant has shown an association with more than one phenotype; in other cases, distinct nearby markers have indicated a multi-phenotype association pattern for a genomic region.

The specific genetic mechanisms underlying the shared physiology of metabolic phenotypes remain poorly understood, rendering the comprehensive analysis of multiple phenotypes an important area of investigation. Additionally, the mechanisms of genetic multi-phenotype effects are specific at each locus and require individual investigation. As different mechanisms have different implications for disease risk and pathogenesis, it is crucial to design approaches for studying them and verifying the hypotheses of pleiotropy at already known loci; in particular, the development of analytical and statistical tools to distinguish and study CP effects of cardiometabolic risk loci will permit clarification of the common genetic basis of these phenotypes.

In the study described in the precedent section, we analysed similar patterns of multi-phenotype effects within single DNA variants, but we did not consider the possibility that adjacent variants with effects on different phenotypes could be representative of the same pleiotropic region, and thus that they were part of the same association signal. However, we observed that multiple variants near the same gene, even if not in high LD ($r^2 < 0.8$), usually show similar effects on cardiometabolic phenotypes.
When two or more SNPs in the same region show a multi-phenotype association signal, the pattern of association may occur, either due to overlapping signals, where the variants tag the same functional region, or because of multi- phenotype allelic heterogeneity, where the identified variants co-localise in the same genomic region but represent independent signals.

To address the challenge of distinguishing overlapping signals from multi-phenotype allelic heterogeneity in established cardiometabolic loci, contributing to the dissection and characterisation of the genetic architecture of the corresponding genomic regions, we systematically investigated shared genetic associations across multiple phenotypes by utilizing GWAS results for 21 cardiometabolic traits and diseases, available in the XC-Pleiotropy group, and applying approximate conditional analysis on the regions showing multiple signals of association for different phenotypes.

3.3.2 Materials and methods

For an overview of the workflow and main results, see figure 3.26.

3.3.2.1 Identification of variants with multi-phenotype cardiometabolic associations

To identify known autosomal SNPs genome-wide significantly (p-value < $5x10^{-8}$) associated with two or more cardiometabolic phenotypes, we performed a systematic literature search using PubMed and the NHGRI catalogue[7]. We selected published associations (before October 2012) from GWAS meta-analyses in Europeans and non-Europeans for 19 quantitative traits and two disease phenotypes. More specifically, for glycaemic traits: fasting glucose (FG) with and without adjustment for BMI, fasting insulin (FI) with and without adjustment for BMI, two-hour glucose (2hGlu), fasting proinsulin (PROINS) and glycated haemoglobin (HbA1c)[18,117-121]; for anthropometric/obesity traits: height, body mass index (BMI), waist circumference (WC) and waist to hip ratio (WHR) with and without adjustment for BMI[16,126,129-131,133,134,137]; body fat percentage

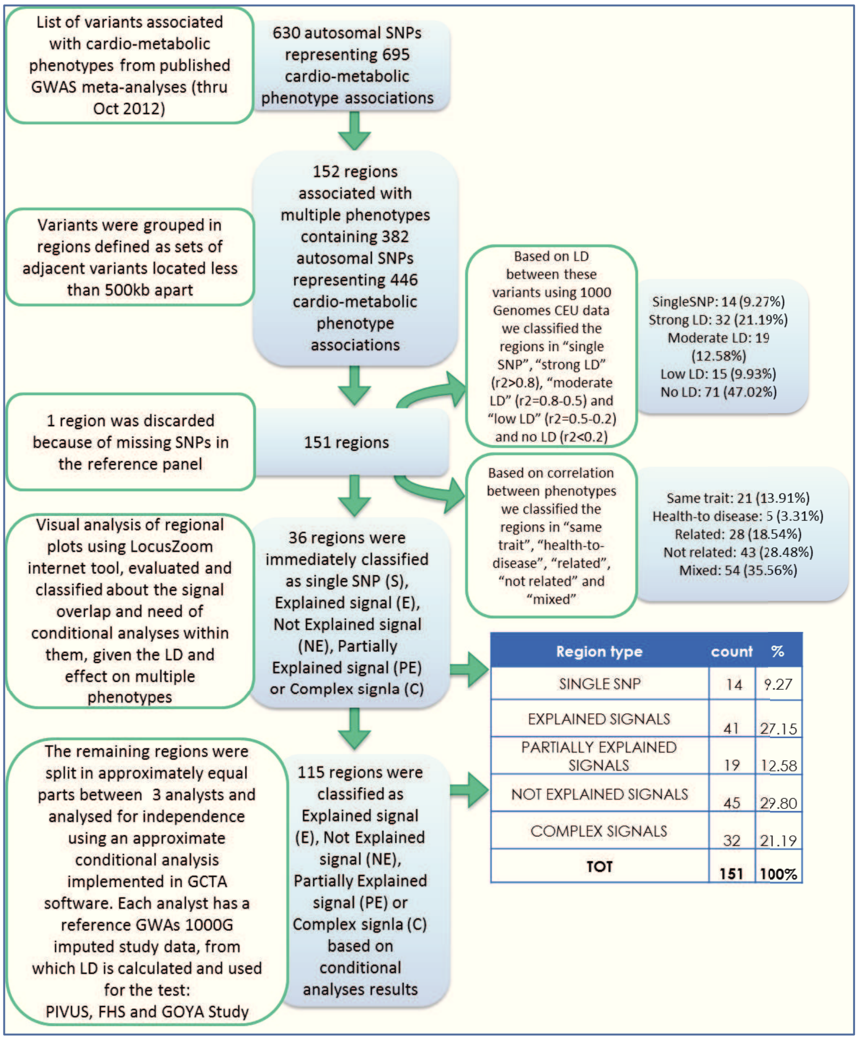


Figure 3.26: *Workflow followed for the analysis of genetic architecture of cardiometabolic associated loci.*

(PCBFAT)[132]; for lipids: high density lipoprotein (HDL), low density lipoprotein (LDL), triglycerides (TG) and total cholesterol (TC)[139,145]; for blood pressure phenotypes: systolic (SBP) and diastolic (DBP) blood pressure, pulse pressure (PS) mean arterial pressure (MAP) and hypertension (HTN, disease phenotype)[147-154,173,174]; Type 2 Diabetes (T2D)[19,108,109,112,117,175,176,110,177,178]. In total, 695 cardio-metabolic SNP-phenotype associations for 630 genome-wide autosomal SNPs were identified. For a complete list of these SNPs, see Appendix tables 1, 2, 3, 4, 5 and 6.

3.3.2.2 Definition and characterization of genomic regions with multi-phenotype association signals

Genomic region definition

To facilitate the dissection of the genetic architecture of multi-phenotype association signals, we assigned the variants into genomic regions. We defined two variants as belonging to the same region if they were located less than 500 kb apart from each other. We labelled each defined region with the name of nearest gene/s.

Region categorisation based on Linkage Disequilibrium

We estimated LD between each two variants based on the 1000 Genomes CEU reference panel (pilot phase)[165] and then we used the lowest pairwise LD value observed within each region to roughly classify the regions into five categories: 1) "Single SNP region" – a single SNP associated with multiple phenotypes; 2) "Strong LD region"' - Distinct SNPs associated with cardiometabolic phenotypes, but in strong LD ($r^2 > 0.8$); 3) "Moderate LD region"- Distinct SNPs in moderate LD (r^2 = 0.5-0.8) associated with cardiometabolic phenotypes; 4) "Low LD region" - Distinct SNPs associated with cardiometabolic phenotypes, in low LD (r^2 = 0.2-0.5); 5) "No LD region" - Distinct SNPs not in LD ($r^2 < 0.2$) associated with cardiometabolic phenotypes. Pairwise LD was evaluated using SNAP internet tool[164].

Region categorisation based on correlation between associated traits

We used 3204 individual-level data from the Framingham cohort study to calculate a Pearson's correlation matrix between available cardiometabolic traits. The traits included were: BMI, WC and WCadjBMI, HIP and HIPadjBMI, WHR and WHRadjBMI, height, FG and FGadjBMI, HOMAB and HOMABadjBMI, FI and FIadjBMI, HOMAIR and HOMAIRadjBMI, 2hGlu (GLUC2H) and 2hGlu with BMI adjustment (GLUC2HRadjBMI), 2 hour insulin with (INS2HRadjBMI) and without (INS2HR) BMI adjustment, HbA1c, HDL, LDL, TG, TC, SBP and SBP adjusted for BMI (SBPadjBMI), DBP and DBP adjusted for BMI (DBPadjBMI).

Traits were adjusted for sex, age (and squared age in some cases); some phenotypes were log-transformed. A detailed description of phenotype definition, transformation and applied exclusions are given in table 3.6.

We defined groups of highly-correlated traits using a threshold of the absolute value of correlation (indicated as "r") >= 0.5. We applied this definition to the analysed genomic data, distinguishing regions associated with the same trait (ST) or with highly-correlated traits (R) from those associated with non-highly correlated traits (NR), or from mixed ones (regions associated with both highly and

Trait	Transformation	Adjustment for BMI	Covariates
FI	log-transformed	with & without	sex, age
HOMAIR	log-transformed	with & without	sex, age
HOMAB	log-transformed	with & without	sex, age
FG	untransformed	with & without	sex, age
2hglu	untransformed	with & without	sex, age
2 hour insulin (2hIns)	log-transformed	with & without	sex, age
HbA1c	untransformed	without	sex, age
DBP	untransformed	with & without	sex, age, age^2
SBP	untransformed	with & without	sex, age, age^2
HDL	normalized	without	sex, age, age^2
LDL	normalized	without	sex, age, age^2
TC	normalized	without	sex, age, age^2
TG	normalized	without	sex, age, age^2
WHR	untransformed	with & without	sex, age, age^2
WC	untransformed	with & without	sex, age, age^2
HIP	untransformed	with & without	sex, age, age^2
BMI	untransformed	without	sex, age, age^2
HEIGHT	untransformed	without	sex, age

Table 3.6: *Detailed information about traits used for calculation of correlation matrix in FHS cohort.*

non-highly correlated traits).

We reserved particular consideration to type 2 diabetes (T2D) and hypertension (HTN), which are disease outcomes, based on their pathological relationships with physiological traits: as T2D arises from high levels of FG, these two phenotypes were classified as pathophysiological related in a healthy variation-to-disease manner (HD). The same definition was applied for DBP/SBP and HTN.

3.3.2.3 Regional plots examination for genome-wide associations

We used published genome-wide meta-analysis association results for 19 quantitative traits and two disease phenotypes, in European samples from the six international consortia which shared their result data within the XC-Pleiotropy group (see table 3.3):

- Height, BMI, WC, WHR and WHRadjBMI from GIANT (Genetic Investigation of ANthropometric Traits),
- PCBFAT,
- DBP, SBP and HTN from the Global BPgen consortium (Global Blood Pressure genetics Consortium),
- HDL, LDL, TC and TG from the GLGC (Global Lipids Genetics Consortium),
- FG, FGadjBMI, FI, FIadjBMI, 2hGlu adjusted for BMI (HGLUadjBMI), PROINS and HbA1c from MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium) and
- T2D from DIAGRAM (Diabetes Genetics Replication And Meta-analysis Consortium).

See table 3.4 for a description of these traits.

We employed GWAS meta-analysis association results for these phenotypes to visualize the multi-phenotype association signals at the defined genomic regions: the $-\log_{10}$(p-value) of the associations of each genomic variant within each region with the corresponding cardiometabolic phenotypes were plotted using the LocusZoom software[179]. This visualisation allowed us to select regions which needed to be further evaluated through approximate conditional analysis.

3.3.2.4 Approximate Conditional Analysis

To assess whether each pair of variants within a genomic region represented independent associations or shared signals, we performed approximate conditional analyses for the corresponding phenotypes by using the Genome-Wide Complex Trait Analysis (GCTA) tool[180]. GCTA implements a conditional analysis of phenotype associations using GWAS meta-analysis summary statistics while incorporating LD information from a reference sample as explained in Yang et al. 2012[181]. In this way, it allows the calculation of a new p-value of association for a SNP of interest with a particular phenotype, corrected for the effect of another adjacent SNP or group of SNPs on the same phenotype that could influence the association of the primary SNP, based on the extent of the LD between them.

The analytical work was split in three parts, each estimating LD between the SNPs from a local population sample of European ancestry from which the individual level genotype data were available: the Framingham Heart Study (FHS, N = 2,459); Genetics of Overweight Young Adults (GOYA, N = 801)[182] and Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS, N = 949)[183]. Additional details about studied cohorts are reported in table 3.7.

Short study name	FHS	GOYA	PIVUS
Long study name	Framingham Heart Study	Genetics of Overweight Young Adults	Prospective Investigation of the Vasculature in Uppsala Seniors
Total sample size	2,459	801	949
Ethnicity	European descent	European descent	Northern European
Country	USA	Danemark	Sweden
Genotyping array	Affymetrix 500K and MIPS 50K	Illumina 610K Quad array	Illumina BeadStation 500GX, Metabochip (custom Illumina iSelect genotyping array)
Imputation software	MACH	MACH	IMPUTE2
Imputation panel	HapMap release 22 (CEU individuals)	HapMap release 22 (CEU individuals)	HapMap release 22 (CEU individuals)
Reference	see web site	182	183
Web Site	http://www.framinghamheartstudy.org/	-	http://www.medsci.uu.se/pivus/

Table 3.7: *Detailed information about cohorts used in approximate conditional analysis.*

We firstly evaluated the attainment of comparable results with the use of the three different cohorts.

Then, for each of those regions that needed to be further evaluated, we proceeded with the approximate conditional analysis for each variant and for each phenotype that the region had been shown to be associated with, conditioning on other variants lying in the same region.

Based on visualisation of association signals of some regions and on results from approximate conditional analyses of the remaining regions, we classified them into five categories: 1) “Single SNP region” (S), as described for LD analysis; 2) “Explained signals region” (E) – Distinct SNPs associated with cardiometabolic phenotypes but underlying the same signal, as when conditioning the association of one variant on the others, the association signal considerably decreased; 3) “Not Explained signals region” (NE) - Distinct signals of association with cardiometabolic phenotypes, such that when conditioning the association of one variant on the others, the association signal did not decrease; 4) “Partially Explained signals region” (PE) - When conditioning the association of one variant on the others, the association signal decreased but it remained significant or it did not decrease consistently compared to original significance; 5) “Complex signals region” (C) – More than two distinct SNPs showing mixed behaviours when conditioning the association of one variant on the others.

We follow this general scheme to define NE, PE and E signals:

If the original p-value was	Conditional p-value for NE signal	Conditional p-value for PE signal	Conditional p-value for E signal
$p \le 10^{-8}$	$10^{-8} \le p \le 10^{-6}$	$10^{-6} \le p \le 10^{-5}$	$p \ge 10^{-5}$
$10^{-8} \le p \le 10^{-6}$	$10^{-6} \le p \le 10^{-4}$	$10^{-4} \le p \le 10^{-3}$	$p \ge 10^{-3}$
$10^{-6} \le p \le 10^{-4}$	$10^{-4} \le p \le 10^{-2}$	$10^{-2} \le p \le 5x10^{-2}$	$p \ge 5x10^{-2}$
$10^{-4} \le p \le 10^{-2}$	$10^{-2} \le p \le 5x10^{-2}$	$5x10^{-2} \le p \le 8x10^{-2}$	$p \ge 5x10^{-2}$

Anyway, every region was then individually evaluated and its classification was refined taking into account the combination of behaviours of all signals after conditional analysis.

3.3.3 Results

Our study was subdivided in different sequential steps. For a better comprehension, figure 3.26 represents the workflow with the main results.

3.3.3.1 <u>Genomic regions with multi- phenotype cardiometabolic associations and their descriptive characterisation</u>

We gathered information about 630 established SNPs associated with at least one cardiometabolic phenotype with p-value less than $5x10^{-8}$ (see Appendix tables 1, 2, 3, 4, 5 and 6) and we grouped them into regions where each variant was distant from the adjacent ones by less than 500kb. We identified 152 regions associated with multiple phenotypes including 382 autosomal SNPs representing 446 cardiometabolic associations.
The region near *NOS3/TMEM176A* locus contained SNPs not present in the HapMap CEU reference panel and neither in the 1000 Genomes CEU reference panel (pilot phase), therefore this region was discarded for the following analyses. In total we obtained 151 regions to further investigate (see table 3.8).

We explored LD patterns within each region using 1000 Genomes CEU reference panel data[165] to calculate the LD between every couple of adjacent SNPs.
Based on the lowest pairwise LD value observed within each region, we identified 14 (9.27%) regions containing the same SNP associated with more than one phenotype, and 32 (21.19%) containing variants in strong LD ($r^2 > 0.8$). Of the remaining regions, 19 (12.58%) contained variants in moderate LD, while 15 regions (9.93%) contained variants in low LD. Finally, in 71 (47.02%) regions we observed variants with no LD ($r^2 < 0.2$). LD-based category assignment for each region is reported in table 3.8. This preliminary description helped us in initially outlining the possible mechanisms behind multi-phenotype associations in defined genomic regions: for example it is easier to exclude allelic heterogeneity in regions containing a single SNP showing multi- phenotype association signals, or multiple SNPs in strong LD.

Locus	Chr	Associated phenotypes						N° of SNPs	LD type	Relationship between phenotypes
SINGLE SNP REGIONS										
TMEM57/LDLRAP1*	1	LDL	TC					1	SINGLE SNP	R
PCSK9*	1	LDL	TC					1	SINGLE SNP	R
CELSR2/PSRC1/SORT1*	1	LDL	TC					1	SINGLE SNP	R
PROX1*	1	T2D	FG					1	SINGLE SNP	HD
MOSC1*	1	LDL	TC					1	SINGLE SNP	R
IRF2BP2/TOMM20*	1	LDL	TC					1	SINGLE SNP	R
SLC39A8*	4	BMI	DBP	SBP	HDL			1	SINGLE SNP	MIXED
MYLIP*	6	LDL	TC					1	SINGLE SNP	R
TFAP2B*	6	BMI	WC					1	SINGLE SNP	R
CYP7A1*	8	LDL	TC					1	SINGLE SNP	R
PLEC1*	8	LDL	TC					1	SINGLE SNP	R
ABCA1*	9	HDL	TC					1	SINGLE SNP	NR
HP/HPR/DHX38*	16	LDL	TC					1	SINGLE SNP	R
CSPG3/CILP2/PBX4*	19	T2D	LDL	TC	TG			1	SINGLE SNP	MIXED
EXPLAINED SIGNALS REGIONS										
Locus	**Chr**	**Associated phenotypes**						**N° of SNPs**	**LD type**	**Relationship between phenotypes**
ANGPTL3/DOCK7*	1	TG	LDL	TC				2	STRONG LD	MIXED
GALNT2	1	HDL	TG					2	STRONG LD	NR
RBJ/DNAJC27	2	BMI	HEIGHT					2	MODERATE LD	NR
GCKR	2	2hGlu	TC	TG	T2D	FG	FI	2	STRONG LD	MIXED
BCL11A*	2	T2D						2	STRONG LD	ST
IRS1	2	T2D	FI	TG	HDL	PCBFAT		7	MODERATE LD	MIXED
ULK4	3	DBP						2	STRONG LD	ST
ADCY5	3	T2D	FG	2hGlu				2	MODERATE LD	R
WFS1*	4	T2D						2	STRONG LD	ST
FGF5	4	DBP	SBP	HTN				2	STRONG LD	R
PDGFC	4	FI	FIadjBMI					2	STRONG LD	ST
ZBED3*	5	T2D	FGadjBMI					2	STRONG LD	HD
TIMD4/HAVCR1	5	LDL	TC	TG				2	STRONG LD	MIXED
CDKAL1	6	T2D	BMI	FG				4	MODERATE LD	MIXED
FRK	6	TC	LDL					2	MODERATE LD	R
RSPO3	6	WHRadjBMI	FI					2	STRONG LD	NR
DNAH11	7	TC	LDL					2	MODERATE LD	R
KLF14*	7	HDL	T2D					2	STRONG LD	NR
LPL*	8	HDL	TG					2	STRONG LD	MIXED
SLC30A8*	8	T2D	FG	PROINS				3	STRONG LD	MIXED
TRIB1*	8	LDL	TC	TG	HDL			3	STRONG LD	MIXED
GLIS3	9	T2D	FG					2	STRONG LD	HD
ABO*	9	TC	LDL					2	STRONG LD	R
C10orf107	10	SBP	DBP	HTN				2	MODERATE LD	R
CYP17A1	10	SBP	DBP	HTN				3	STRONG LD	R
GPAM*	10	LDL	TC					2	STRONG LD	R
SPTY2D1*	11	LDL	TC					2	STRONG LD	R
ARAP1/CENTD2	11	FG	PROINS	T2D				2	STRONG LD	MIXED
UBASH3B	11	TC	HDL					2	STRONG LD	NR
LRP1*	12	TG	HDL					2	STRONG LD	NR
ATP2B1*	12	SBP	DBP	HTN				4	STRONG LD	R
SH2B3/BRAP	12	SBP	DBP	LDL	TC			3	MODERATE LD	MIXED
CCDC92/ZNF664	12	HDL	TG					2	STRONG LD	NR
NRXN3	14	BMI	WC					2	STRONG LD	R
CYP1A1/CSK/ULK3	15	DBP						2	MODERATE LD	ST
HMG20A*	15	T2D						2	STRONG LD	ST
FTO*	16	FI	HDL	BMI	PCBFAT	T2D		4	STRONG LD	MIXED
PLCD3/ACBD4	17	SBP	HEIGHT					2	MODERATE LD	NR
FOXA2	20	FG						2	MODERATE LD	ST
MAFB*	20	TC	LDL					2	STRONG LD	R
PLTP	20	TG	HDL					2	MODERATE LD	NR
PARTIALLY EXPLAINED SIGNALS REGIONS										
Locus	**Chr**	**Associated phenotypes**						**N° of SNPs**	**LD type**	**Relationship between phenotypes**
ST7L/CAPZA1/MOV10	1	DBP	SBP					2	NO LD	R
G6PC2	2	FG	HbA1C					2	MODERATE LD	R
ADAMTS9	3	WHRadjBMI	T2D					2	LOW LD	MIXED
MSL2L1/PCCB	3	TG	HEIGHT					2	NO LD	NR
LCORL	4	HEIGHT						2	NO LD	ST
HHIP	4	HEIGHT						2	NO LD	ST
NPR3	5	SBP	DBP	HTN	HEIGHT			3	NO LD	MIXED
HMGCR/FLJ35779	5	LDL	TC	BMI				2	LOW LD	MIXED
JAZF1	7	HEIGHT	T2D					2	LOW LD	NR
MLXIPL	7	TG	HDL					2	STRONG LD	NR
GATA4	8	SBP	TG					2	LOW LD	NR
NAT2	8	TC	TG					2	MODERATE LD	NR
TTC39B	9	HDL	TC					2	MODERATE LD	NR
CPN1/CHUK	10	HEIGHT	TC					2	LOW LD	NR
MTNR1B	11	T2D	HbA1C	FG				2	LOW LD	MIXED
ST3GAL4	11	LDL	TC					2	STRONG LD	R
SRR	17	T2D	PROINS					2	LOW LD	NR
ZNF652	17	HEIGHT	SBP	DBP				3	LOW LD	MIXED
CSH1/GH1	17	HEIGHT						2	NO LD	ST

Table 3.8: *Genome-wide regions grouped based on the classification of multiple signals, with associated phenotypes and number of SNPs included. LD classification and phenotype relationship classification are shown. Continuation in the following page. * = this region was not analysed through approximate conditional analysis.*

Locus	Chr	Associated phenotypes								N° of SNPs	LD type	Relationship between phenotypes
NOT EXPLAINED SIGNALS REGIONS												
GVI1/EVI5/RPL5	1	TC	HEIGHT							2	LOW LD	NR
DNM3/PIGC	1	HEIGHT	WHRadjBMI							3	NO LD	MIXED
APOB	2	HDL	TG	LDL	TC					2	NO LD	MIXED
EFEMP1	2	HEIGHT								2	LOW LD	ST
CCDC108/IHH	2	HEIGHT								2	NO LD	ST
GHSR/FNDC3B	3	HEIGHT								2	NO LD	ST
ARL15	5	FI	HDL	FladjBMI						2	NO LD	MIXED
ANKRD55/MAP3K1	5	T2D	FladjBMI	TG						2	NO LD	NR
FGF18/FBXW11	5	HEIGHT								2	NO LD	ST
BOD1/CPEB4	5	HEIGHT	WHRadjBMI							2	NO LD	NR
LY86/RREB1	6	WHRadjBMI	FG							2	NO LD	NR
HFE/HIST1H4C	6	SBP	DBP	HTN	HbA1C	LDL	TC	HEIGHT		3	NO LD	MIXED
VTA1/GPR126	6	HEIGHT								2	NO LD	ST
SLC22A1/LPA	6	LDL	TC	TG	HDL					3	NO LD	MIXED
ANK1	8	T2D	HbA1C							3	NO LD	MIXED
PLAG1/SDR16C5	8	HEIGHT								2	NO LD	ST
TRPS1	8	HDL	TC							2	NO LD	NR
QSOX2/DNLZ	9	HEIGHT	FG							2	NO LD	NR
VPS26A/HK1	10	T2D	HbA1C							2	NO LD	NR
PPIF	10	HEIGHT (secondary signal)	T2D	HEIGHT						3	NO LD	MIXED
DUSP8/LSP1/TNNT3	11	T2D	BP							2	NO LD	NR
ADM/AMPD3	11	SBP	HDL							2	NO LD	NR
PLEKHA7/NUCB2/KCNJ11	11	SBP	HEIGHT	T2D						3	NO LD	NR
CRY2/LRP4/NR1H3	11	FG	HDL							2	NO LD	NR
*SERPINH1/DGAT2**	11	HEIGHT	HDL							2	NO LD	NR
APOA1/C3/A4/A5/BUD13	11	HDL	LDL	TC	TG					3	NO LD	MIXED
*PDE3A/SLCO1C1**	12	HDL	HEIGHT							2	NO LD	NR
*STAT2/GLS2**	12	HEIGHT	FGadjBMI							2	NO LD	NR
HMGA2	12	T2D	HEIGHT							2	NO LD	NR
SOCS2/CRADD	12	HEIGHT								2	NO LD	ST
HNF1A/TCF1	12	LDL	TC	T2D						2	NO LD	MIXED
SBNO1	12	HDL	HEIGHT							2	NO LD	NR
PDS5B/BRCA2/KL	13	LDL	HEIGHT	FG						3	NO LD	NR
SPRY2	13	T2D	PCBFAT							2	NO LD	NR
NFATC4/CBLN3/KIAA1305	14	HEIGHT	LDL							2	NO LD	NR
LOXL1/PML	15	HEIGHT								2	NO LD	ST
ACAN	15	HEIGHT								2	NO LD	ST
FURIN/FES/PRC1	15	SBP	DBP	T2D						2	NO LD	MIXED
GPRC5B/GP2/UMOD	16	BMI	HTN							2	NO LD	NR
NOG	17	HEIGHT								2	NO LD	ST
ANGPTL4/ADAMTS10	19	HDL	HEIGHT							2	LOW LD	NR
PEPD/KCTD15	19	T2D	FI	FladjBMI	BMI					3	NO LD	MIXED
APOEC1/C2	19	TG	HDL	LDL	TC					2	NO LD	MIXED
GDF5/ERGIC3	20	HEIGHT	TC							2	NO LD	NR
FITM2/R3HDML/HNF4A	20	T2D	HDL	TC						3	NO LD	MIXED
COMPLEX SIGNALS REGIONS												
Locus	**Chr**	**Associated phenotypes**								**N° of SNPs**	**LD type**	**Relationship between phenotypes**
MTHFR/NPPB/CLCN6	1	DBP	SBP							3	LOW LD	R
LYPLAL1	1	FI	WHR (only in women)	FladjBMI	HEIGHT	WHRadjBMI				6	LOW LD	MIXED
THADA/ABCG5/8	2	T2D	LDL	TC						3	NO LD	MIXED
FIGN	2	SBP	BP	DBP						3	NO LD	R
COBLL1/GRB14	2	T2D (only in women)	WHRadjBMI	FI	FladjBMI	TG	HDL			5	NO LD	MIXED
PPARG/RAF1	3	T2D	FladjBMI	TC						4	NO LD	MIXED
IGF2BP2/ETV5	3	FG	2hGlu	T2D	HEIGHT	BMI				4	NO LD	MIXED
TET2	4	FladjBMI	FI	HEIGHT						3	MODERATE LD	MIXED
PCSK1/ERAP2	5	FG	PROINS	BMI	2hGlu					5	NO LD	MIXED
MICA/HLA	6	HEIGHT (secondary signal)	TG	HEIGHT	SBP	DBP	HTN	LDL	TC	7	NO LD	MIXED
HMGA1/C6orf107/UHRF1BP1	6	HEIGHT	BMI	TC	HDL	FladjBMI	FI			7	NO LD	MIXED
DGKB/TMEM195	7	T2D (only in men)	FG							3	NO LD	HD
GCK/NPC1L1	7	HbA1C	2hGlu	T2D	FG	TC	LDL			5	NO LD	MIXED
PPP1R3B	8	FG	FI	HDL	FladjBMI	LDL	TC	2hGlu		5	MODERATE LD	MIXED
CDKN2A/B	9	T2D (secondary signal)	FG							3	NO LD	HD
CACNB2	10	SBP	DBP	HTN						3	NO LD	R
HHEX/IDE/CYP26A1	10	T2D	TG							3	NO LD	MIXED
TCF7L2	10	FG	T2D	FI	2hGlu					3	MODERATE LD	MIXED
KCNQ1	11	T2D	HEIGHT							4	NO LD	MIXED
MADD/MTCH2/SLC39A13/OR4S1	11	PROINS	FG	BMI	HEIGHT					6	NO LD	MIXED
FADS1/2/3	11	TG	FG	TC	LDL	HDL				4	STRONG LD	MIXED
TBX5/TBX3	12	DBP								3	NO LD	ST
MTIF3/PDX1	13	BMI	FG	PROINS						3	NO LD	NR
LIPC	15	HDL (secondary signal)	TC	TG						3	NO LD	MIXED
FAM148B/VPS13C/C2CD4A/B	15	2hGlu	HEIGHT	PROINS	T2D	FG				5	LOW LD	MIXED
CETP	16	LDL	HDL	TC	TG					4	MODERATE LD	MIXED
GOSR2/OSBPL7	17	SBP	LDL	TC						3	NO LD	MIXED
DYM/LIPG	18	HEIGHT	HDL	TC						3	NO LD	NR
MC4R	18	BMI	HDL	HEIGHT	WC	T2D				6	NO LD	MIXED
*LDLR/DOCK6/LOC55908**	19	LDL	TC	HDL						2	NO LD	MIXED
GIPR/QPCTL	19	T2D (only in women)	2hGlu	FG	BMI					4	NO LD	MIXED
TOP1	20	FG	TC	LDL						3	LOW LD	MIXED

Table 3.8: *Continuation.*

Based on the correlation matrix between cardiometabolic traits, calculated from Framingham cohort study and represented in figure 3.27, we identified groups of highly correlated traits.

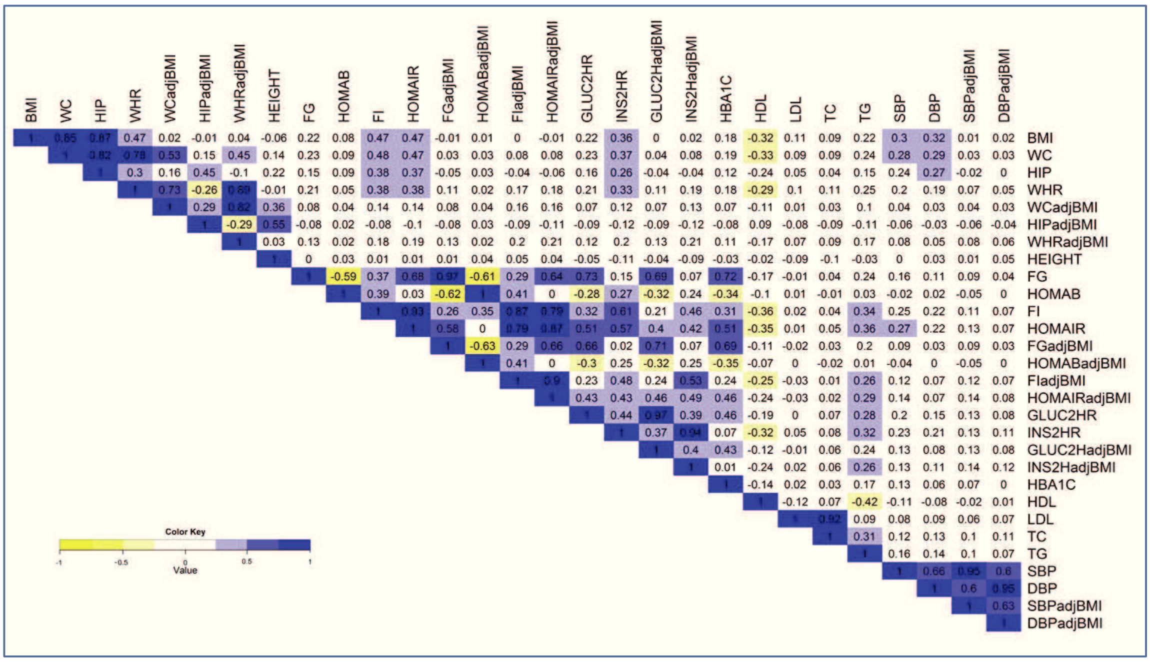


Figure 3.27: *Correlation matrix between cardiometabolic traits calculated from data of 3204 individuals from the Framingham cohort study. Colours are proportional to the level of pairwise correlation, as explained in the legend below, on the left.*

BMI was highly correlated with other anthropometric and obesity traits, such as WC (correlation value r = 0.85), HIP (r = 0.87) and borderline with WHR (r = 0.47). However this correlation disappeared when WC, HIP and WHR were adjusted for BMI. WC and WHR were highly correlated with and without BMI adjustment (r = 0.78 and r = 0.82, respectively), while the relationship between WC and HIP was evident only without the adjustment for BMI. HIP demonstrated a relationship with HEIGHT only when adjusted for BMI (r = 0.55). A borderline correlation was also seen for BMI and WC with FI and HOMAIR, but it disappeared after BMI adjustment.

We highlighted strong correlations between glycaemic traits: negative correlation is observable between FG and HOMAB (r = -0.59) and it is maintained also when the traits were adjusted for BMI; FG was positively correlated with HOMAIR, 2hGlu and HBA1C (r = 0.68, 0.73 and 0.72, respectively), while FI with HOMAIR and 2hIns (r = 0.93 and 0.61, respectively): these results did not significantly change after BMI adjustment. HOMAIR showed interesting correlations with HBA1C, 2hGlu and 2hIns (r = 0.51, 0.57 and 0.51, respectively) that became borderline (just below the defined threshold of r = 0.5) after BMI adjustment.

Interestingly, among lipids we observed a uniform negative trend of HDL in its correlation with all the other metabolic traits (even if not more than 0.50 as an absolute value) and this is consistent with the definition of MetS[157]. A high positive correlation was instead evident between TC and LDL (r = 0.92). Finally, we could observe that DBP and SBP were highly correlated (r = 0.66), even when BMI-adjusted.

Combining the information on correlation between phenotypes, we better characterised observed

multi- phenotype associated loci (table 3.8).
In general, as we could expect from epidemiological data, we observed an excess of highly related traits (27 R, HD and ST regions out of 46, 58.70%) in regions containing single SNPs or SNPs in very strong LD ($r^2 > 0.8$); while regions with variants in low LD or no LD ($r^2 < 0.5$) had an excess of non-related associated phenotypes or mixed phenotypes (66 NR and MIXED regions out of 86, 76.74%).
In total, we identified 97 regions (64.24%) that contained variants associated with not highly correlated phenotypes (NR and MIXED, see table 3.8). The remaining 54 regions (35.76%) were characterised by multiple associations with the same phenotype, or highly related phenotypes (R regions, r > 0.5), or by "health-to-disease" phenotypes.

3.3.3.2 Visualisation of the association signals

To complete the first part of our study for descriptive analyses of the 151 regions, we undertook a visual inspection of their patterns of multi- phenotype association signals, using regional plots of the association p-values from GWAS meta-analysis results shared within the XC-Pleiotropy group.
Combining the observation of association signals with the information about LD and about correlation between associated phenotypes, we were able to immediately interpret the multi-phenotype association of 36 regions, distinguishing multiple signals that were clearly shared between phenotypes from those that were distinctly separate signals located within the same genomic region.
As already described in the previous sub-chapter, 14 (9.27%) regions contained single SNP (S) showing associations with multiple phenotypes.

18 regions, from the 151 analysed, showed a clear pattern of explained signals (E): as we can observe in the example represented in figure 3.28, the two association signals at *GPAM* (glycerol-3-phosphate acyltransferase) region are led by two different SNPs (rs1129555 for LDL, p-value = 2.14×10^{-9} in our data, and rs2255141 for TC, p-value = 2.03×10^{-10}), but both represent the same association pattern of a group of SNPs in high LD (coloured points in the figures) that is identical for the two traits, LDL and TC. Additionally, the two variants (rs1129555 and rs2255141) are in high LD with each other ($r^2 = 0.96$). In this case approximate conditional analysis was not necessary, and we interpreted the region as containing a shared signal of association. All of the 18 regions classified here as E were also classified as "strong LD" regions; most of them (11) were associated with related phenotypes (R, or HD, or ST) as in the described example.

For three regions the presence of two separate signals of association was highly visible. An example is reported in figure 3.29: the *SERPINH1/DGAT2* region contains two associations at two different SNPs, rs11236530 for HDL (p-value = 0.005 in our data) and rs634552 for height (p-value = 1.35×10^{-9}); from regional plot visualisation, we observed that the two SNPs highlight two separate signals with completely different surrounding LD patterns and divided by a recombination hotspot (recombination rate ≈ 30 cM/Mb). As confirmation of our observation, all the three regions contain couples of variants not in LD ($r^2 < 0.2$). We decided to categorise them as Not Explained signals (NE) without running approximate conditional analysis.

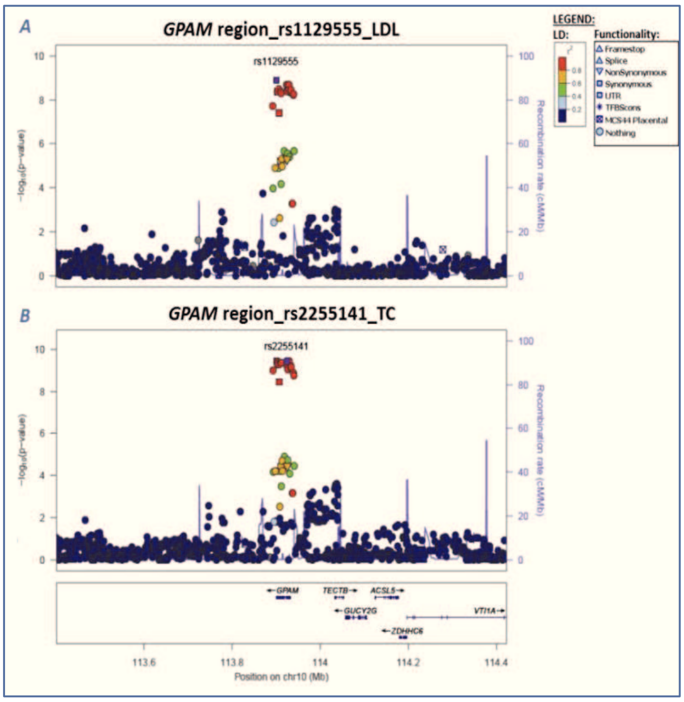


Figure 3.28: *Regional plots of associations at the GPAM locus, as example of region with a clear pattern of Explained signals (E) observable just from the graphical visualisation. -log_{10} of p-values of association are plotted for all variants included in the region; each point is a variant, violet point is the main one, the others are represented following a colour code proportional to the LD with the main variant, and a shape code consistent with functionality, as described in the legend.* ***A****: GPAM locus was associated with LDL (violet square points rs1129555, p-value = $2.14x10^{-9}$ in our data),* ***B****: but also with TC (violet circle points rs2255141, p-value = $2.03x10^{-10}$ in our data). The two signals are similar. Below the plots: genes lying in the region.*

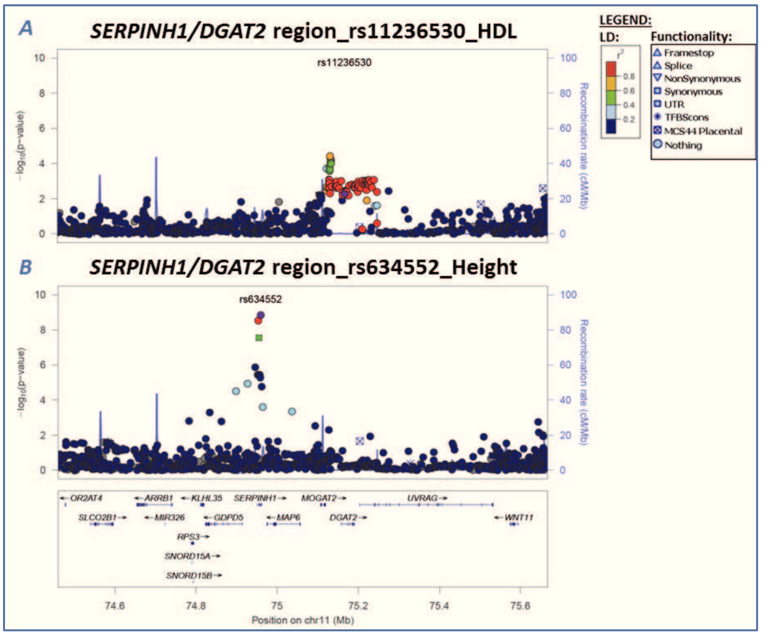


Figure 3.29: *Regional plots for the SERPINH1/DGAT2 region as example of clear pattern of Not Explained signals (NE) observable just from the graphical visualisation.* ***A****: SERPINH1/DGAT2 locus was associated with HDL (violet circle points rs11236530, p-value = 0.005 in our data),* ***B****: but also with height (violet circle points rs634552, p-value = $1.35x10^{-9}$). The two signals are clearly different and separated by a recombination hotspot (recombination rate ≈ 30 cM/Mb). Below the plots: genes lying in the region.*

Through the visualisation of association patterns only, we classified the *LDLR/DOCK6/LOC55908* (low density lipoprotein receptor/dedicator of cytokinesis 6/locus 55908) region as a Complex signal (C), since it presented the same variant (rs6511720) associated with both, LDL and TC, and another variant (rs737337) associated with HDL, but with a completely different signal of association, separated from the previous one by a recombination hotspot with recombination rate of ≈ 50 cM/Mb. The two variants are not in LD (r^2 = 0.008, data not shown).

3.3.3.3 Approximate Conditional Analysis

For the remaining 115 regions, comprising 257 markers, we were not able to interpret the patterns of multi-phenotype associations just using descriptive analyses; we thus decided to apply an analytical approach using approximate conditional analysis: it is a statistical method that allows calculating a signal conditioned on other signals in the same locus, using the summary data results from association analyses and the LD information as input. Conditioning was done on every marker in a cross- phenotype manner.

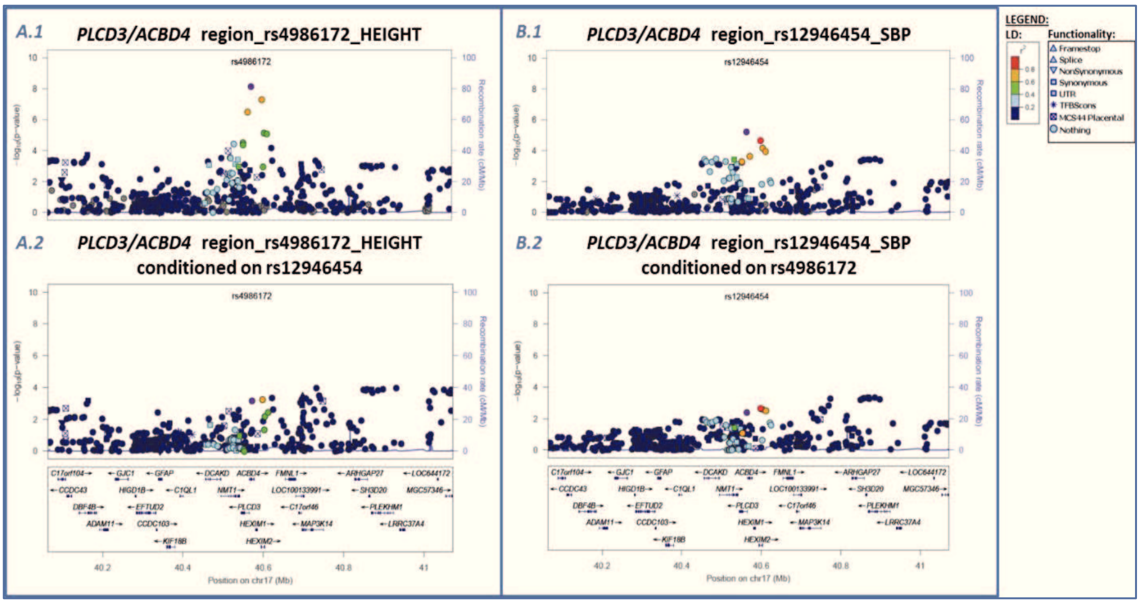


Figure 3.30: *Regional plots of the PLCD3/ACBD4 locus before (**A.1** and **B.1**) and after (**A.2** and **B.2**) approximate conditional analysis. **A.1**: This locus was associated with height (violet circle points rs4986172, p-value = $7.12x10^{-9}$) **B.1**: and with SBP (violet circle points rs12946454, p-value = $6.05x10^{-6}$ in our data). **A.2**: Conditioning the height (rs4986172) signal for the SBP one (rs12946454), regional association for height decreased below genome-wide significance level (conditional p-value = $7x10^{-4}$). **B.2**: Conditioning the SBP (rs12946454) signal for the height one (rs4986172), regional association for SBP significantly decreased (conditional p-value = 0.004). This region was thus classified as Explained signals (E).*

After approximate conditional analysis, we classified 23 regions as Explained signals (E): in these regions the analysis decreased significantly the association signals below genome-wide level. An example is reported in figure 3.30, where *PLCD3/ACBD4* (phospholipase C delta 3/acyl-CoA binding domain containing 4) locus was associated with height (rs4986172, p-value = $7.12x10^{-9}$) and with SBP (rs12946454, p-value = $6.05x10^{-6}$ in our data) and it was not clear if the two variants represented the same signal of association for the two traits.

After conditioning the height signal for the SBP one, the regional association for height decreased below genome-wide significance level (rs4986172 conditional p-value = $7x10^{-4}$); and conditioning the SBP signal for the height one, regional association for SBP significantly decreased (rs12946454 conditional p-value = 0.004). From this result, in this region the association signal attributable to one variant for one trait is explainable by the other variant, originally associated with the other trait. This feature is exclusively statistical: its biological explanation may not be immediate and may require further analysis for the identification of the causal gene and its functional characterisation in relationship with height and SBP.

42 regions showed Not Explained signals (NE) as in *LY86/RREB1* (lymphocyte antigen 86/ras responsive element binding protein 1) region (figure 3.31): it contains two different SNPs, one associated with WHRadjBMI (rs1294421, p-value = 5×10^{-8}) and one with FG (rs17762454, p-value = 1×10^{-5} in our data); after approximate conditional analysis on the FG variant, WHRadjBMI association signal did not change (rs1294421, conditional p-value = 6.33×10^{-8}) and this was true also for FG-association signal after conditioning on the WHR variant (rs17762454, conditional p-value = 8×10^{-6}).

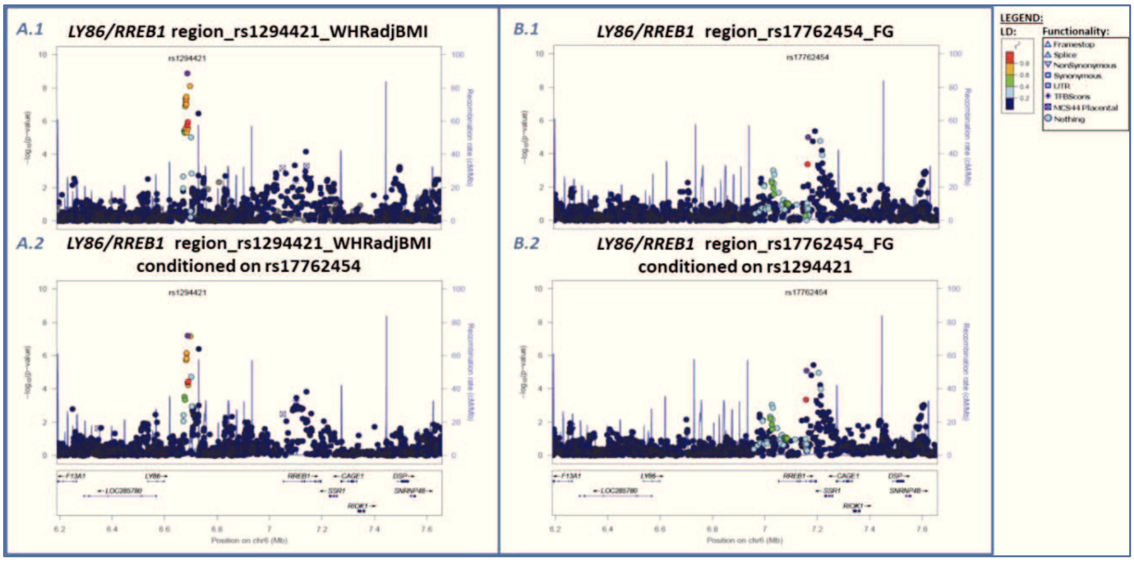


Figure 3.31: *Regional plots of the LY86/RREB1 locus before (**A.1** and **B.1**) and after (**A.2** and **B.2**) approximate conditional analysis. **A.1**: This locus was associated with WHRadjBMI (violet circle points rs1294421, p-value =* 5×10^{-8} *in our data) **B.1**: and with FG (violet circle points rs17762454, p-value =* 1×10^{-5}*). **A.2**: Conditioning the WHRadjBMI (rs1294421) signal for the FG one (rs17762454), regional association for WHRadjBMI did not decrease (conditional p-value =* 6.33×10^{-8}*). **B.2**: Either conditioning the FG (rs17762454) signal for the WHRadjBMI one (rs1294421), regional association for FG did not significantly decrease (conditional p-value =* 8×10^{-6}*). This region was thus classified as Not Explained signals (NE).*

For 19 regions the association signal at one phenotype was not completely explained by the signals observed for other phenotypes within the same region; based on this unclear profile, and on our inability to interpret this kind of multiple association signal, we decided to classify these regions as "Partially Explained" (PE). An example is the *JAZF1* locus in figure 3.32: here two variants in low LD ($r^2 = 0.48$) are associated, one with height (rs1708299, p-value = 1.8×10^{-17}) and another with T2D (rs849134, p-value = 3.22×10^{-10}); approximate conditional analysis on the T2D variant (rs849134) considerably decreased the height association signal, but it remained near the genome-wide significance level (rs1708299, conditional p-value = 3.32×10^{-7}); the same behaviour was observed for the T2D association signal (rs849134) after conditioning on the height variant (conditional p-value = $6x10^{-8}$).

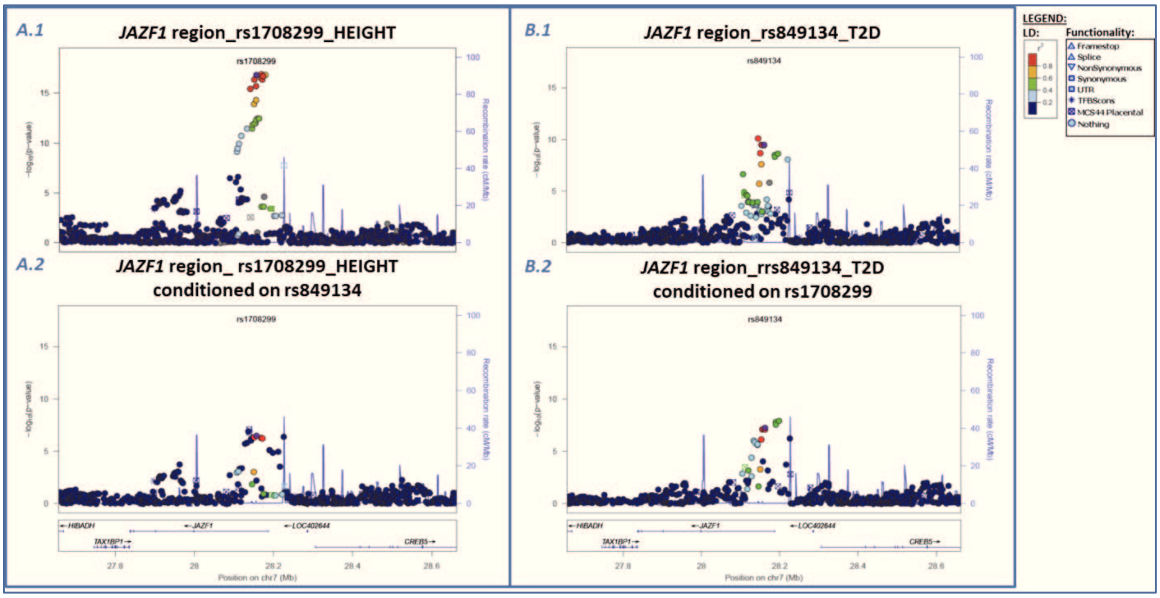
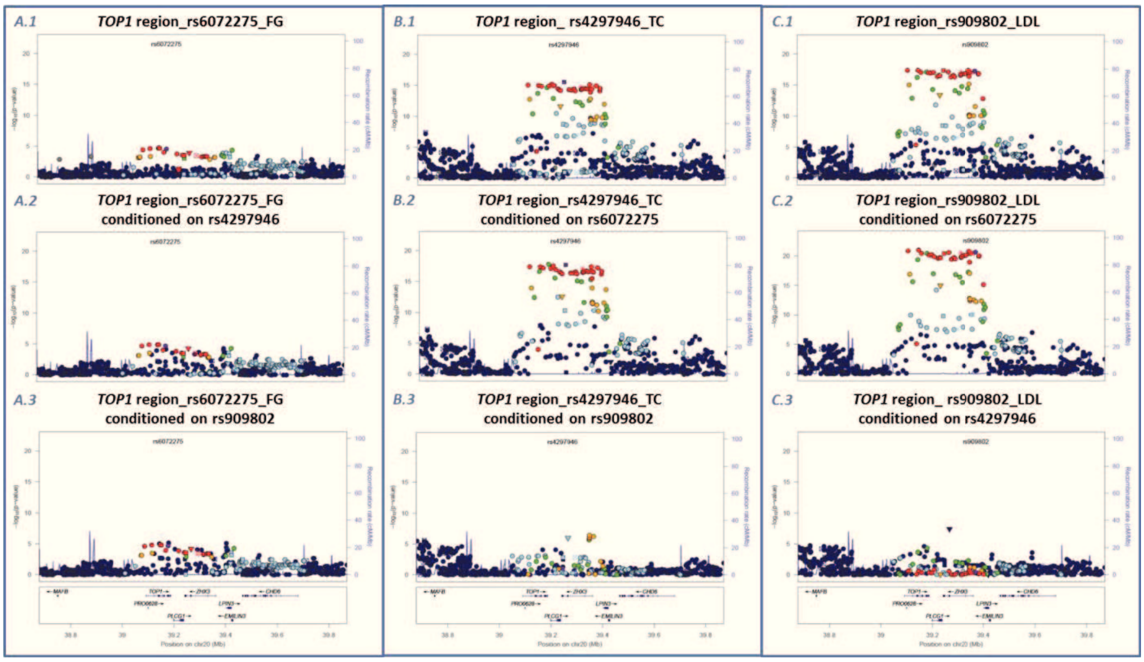


Figure 3.32: *Regional plots of the JAZF1 locus before (**A.1** and **B.1**) and after (**A.2** and **B.2**) approximate conditional analysis. **A.1**: This locus was associated with HEIGHT (violet circle points rs1708299, p-value = $1.8x10^{-17}$ in our data) **B.1**: and with T2D (violet circle points rs849134, p-value = $3.22x10^{-10}$). **A.2**: Conditioning the HEIGHT (rs1708299) signal for the T2D one (rs849134), regional association for HEIGHT decreased, but still near G-W significance level (conditional p-value = $3.32x10^{-7}$). **B.2**: Conditioning the T2D (rs849134) signal for the HEIGHT one (rs1708299), regional association for T2D decreased, but remained at a borderline G-W significance level (conditional p-value = $6x10^{-8}$). This region was thus classified as Partially Explained signals (PE).*

The remaining 31 regions were mixed, that is they included both explained and unexplained signals (complex, C). An example is represented in figure 3.33: the region of the *TOP1* locus contains three variants, rs6072275 associated with FG (p-value = $3x10^{-5}$), rs4297946 associated with TC (p-value = $2.76x10^{-17}$) and rs909802 associated with LDL (p-value = $3x10^{-19}$); we observed that the FG signal (rs6072275) conditioned on the TC one (rs4297946) and on the LDL one (rs909802) did not change its pattern of association (conditional p-value = $2x10^{-5}$ for both analyses); we obtained the same result even when we conditioned the TC signal (rs4297946) and the LDL signal (rs909802) on the FG variant (rs6072275, conditional p-value became $2x10^{-18}$ for TC and $2.1x10^{-21}$ for LDL). We noted, instead, a decrease of the association signal when we condition the TC association (rs4297946) on the LDL variant (rs909802), resulting in a conditional p-value = 0.09, and also when conditioning the LDL signal (rs909802) on the TC variant (rs4297946) (conditional p-value = 0.4). This pattern can be explained as an independent signal of association between both rs6072275 - rs4297946 and rs6072275 - rs909802 and, instead, a shared association between rs4297946 and rs909802. This is confirmed by pairwise LD values between these variants from 1000Genomes reference panel ($r^2_{rs6072275\text{-}rs4297946}$ = 0.246, $r^2_{rs6072275\text{-}rs909802}$ = 0.230 and $r^2_{rs4297946\text{-}rs909802}$ = 0.935).



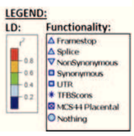


Figure 3.33: *Regional plots of TOP1 locus before (**A.1, B.1** and **C1**) and after (**A.2, B.2, C.2, A.3, B.3** and **C.3**) approximate conditional analysis.* ***A.1:*** *this locus was associated with FG (rs6072275, p-value = $3x10^{-5}$ in our data),* ***B.1****: TC (rs4297946, p-value = $3x10^{-17}$).* ***C.1****: and LDL (rs909802, p-value = $3x10^{-19}$). We observed that the FG signal conditioned on the TC (**A.2**) and LDL (**A.3**) variants did not change the pattern of association; we obtained the same result even when we conditioned TC (**B.2**) and LDL (**B.2**) on the FG variant. We noted, instead, a decrease of the association signal when we condition TC on the LDL variant (**B.3**) (conditional p-value = 0.09) and LDL on the TC variant (**C.3**) (conditional p-value = 0.4). We thus classified this region as Complex signals (C).*

3.3.3.4 Final interpretation of cardiometabolic loci architecture

Table 3.8 reports the complete classification of studied genomic regions.

Starting from 151 regions, we identified 14 (9.27% of the total) as single SNP (S) ones; the majority of them showed associations with related phenotypes (11 R, or HD; 78.60%) and only three were reported in the literature as associated with multiple not highly correlated phenotypes: rs13107325 at *SLC39A8* locus associated with SBP and DBP, BMI and HDL; rs1883025 at *ABCA1* locus associated with HDL and TC; and rs10401969 at *CSPG3/CILP2/PBX4* (chondroitin sulfate proteoglycan 3/cartilage intermediate layer protein 2/pre-B-cell leukemia homeobox 4) locus associated with LDL and TC, but also with TG and T2D.

We defined 41 (27.15% on the total) regions with Explained signals (E), that is with potential shared patterns of multi-phenotype associations; all of them contained variants in strong (29 with $r^2 > 0.8$) or moderate (12 with $r^2 > 0.5$) LD. 21 (51%) showed multiple associations with highly related phenotypes (R and HD) or with the same phenotype, and 20 (49%) with at least with two non-related phenotypes (NR or MIXED). The latter are the most interesting regions as they are potentially pleiotropic; they include: *ANGPTL3/DOCK7, GALNT2, RBJ/DNAJC27* (DnaJ homolog subfamily C member 27)*, GCKR, IRS1, TIMD4/HAVCR1* (T-cell immunoglobulin and mucin domain containing 4/ hepatitis A virus cellular receptor 1)*, CDKAL1, RSPO3* (R-spondin 3)*, KLF14, LPL, SLC30A8, TRIB1* (tribbles pseudokinase 1)*, ARAP1/CENTD2, UBASH3B, LRP1, SH2B3/BRAP* (SH2B adaptor protein 3/BRCA1 associated protein)*, CCDC92/ZNF664* (coiled-coil domain containing 92/zinc finger protein 664)*, FTO, PLCD3/ACBD4* and *PLTP*.

32 (21.19% on the total) regions contained Complex signals (C), in other words, some explained and some unexplained. 6 (19%) were associated with related phenotypes, while 26 (81%) showed mixed associations or associations with non-highly related phenotypes. Between them, the most interesting overlapping, and thus potential pleiotropic, signals were: *LYPLAL1* (lysophospholipase-like 1) for its association with FI and WHRadjBMI; *COBLL1/GRB14* associated with FI, T2D, WHR, TG and HDL; *PPARG/RAF1* (*RAF1* is v-raf-1 murine leukemia viral oncogene homolog 1) in T2D and FI; *TET2* (tet methylcytosine dioxygenase 2) associated with FI and height; *MICA/HLA* (MHC class I polypeptide-related sequence A/major histocompatibility complex) for its associations with height and TG; *HMGA1/C6orf107/UHRF1BP1* (UHRF1 binding protein 1) in TC, HDL and height; *PPP1R3B* associated with FG, FI HDL, LDL and TC; *FADS1/2/3* for TG, TC, LDL and FG; *MC4R* for its effects on BMI, WC, height, HDL and T2D; and finally *GIPR/QPCTL* (glutaminyl-peptide cyclotransferase-like) and its association with 2hGlu and BMI.

45 (29.8% on the total) regions contained Not Explained signals (NE), suggestive of independence between included variants and thus of multi-phenotype allelic heterogeneity. Our inspection of regional plots and approximate conditional analyses was supported by the fact that all NE regions contained variants with no LD (r^2 < 0.2), or just low LD (r^2 < 0.5). In addition, NE regions were predominantly associated with lowly correlated phenotypes (35.78%).

Also, after approximate conditional analysis, for 19 (12.58% on the total) regions we were not able to understand if the multiple signals of association overlapped or not, as the conditional analysis led to a decrease of the original association signal, but not to a complete loss of the significance of association. We defined these regions as Partially Explained (PE) signals.

3.3.4 Discussion

Discerning the real genetic mechanisms behind cross-phenotype effects is an important phase for evaluation and quantification of the shared genetic basis and physiology of phenotypes, including pathogenesis and disease risk.
In the precedent project we analysed combinations of CP effects at single DNA variants, but we realised that it is also important to consider the architecture of whole loci showing multiple phenotype associations. In particular we were interested in verifying the possibility of pleiotropy for

cardiometabolic phenotypes, and distinguishing it from allelic heterogeneity.
When two or more SNPs in the same region show a multi-phenotype association signal, the pattern of associations may occur either due to overlapping signals of association, where the variants tag the same functional region, or to multi-phenotype allelic heterogeneity, where the identified variants co-localise in the same genomic region but represent independent signals.
In the present project, we systematically applied descriptive and statistical analyses, using GWAS results for 21 cardiometabolic phenotypes (available within the XC-Pleiotropy group) and LD information estimated from 1000Genome CEU reference panel and from three European ancestry cohorts. Our aim was to discern multi-phenotype allelic heterogeneity from real overlapping signals at each genomic locus containing multiple cardiometabolic phenotype associations, and thus dissect and characterise the genetic architecture of the corresponding regions.

Our results highlighted that a substantial proportion (29.8%) of metabolic phenotype loci incorporate complex patterns of potential multi-phenotype allelic heterogeneity: in fact, we observed that the presence of multiple cardiometabolic phenotype effects could be explained by suggestive independent signals of associations in 45 genomic regions out of the 151 analysed. They could underlie different causal genes that are involved in the determination of distinct phenotypes through separate functional mechanisms. An example is the *LY86/RREB1* region, described in figure 3.31: two non-overlapping signals are present at this region in association with WHRadjBMI and FG; the WHRadjBMI-associated variant (rs1294421) maps nearer *LY86* gene, which encodes for a protein that participates in the innate immune response; the FG variant (rs17762454) instead maps within *RREB1* sequence, a gene encoding a transcription factor that binds to the RAS-responsive elements of gene promoters. These two genes could separately influence WHRadjBMI and FG.

Moreover, approximate conditional analysis of multi-phenotype effects allowed the definition of at least 87 (57.62%) genomic regions with the same associated genetic variant, or variants attributable to the same causal mutation, affecting multiple cardiometabolic phenotypes. For them, in fact, our analyses confirmed the overlap between multi-phenotype effects, thus, suggestive for pleiotropy, and we can therefore exclude allelic heterogeneity as genetic mechanism leading to multiple associations.
Of these regions, those where shared associations are with non-highly correlated phenotypes (42, 27.8% of the total) are particularly relevant because it is less probable that their genetic association for one phenotype might reflect associations for the other phenotype, in the sense that the effect is partially or totally explained through the association with the other phenotype.
Within this group, some noteworthy examples are: a single missense variant (rs13107325) at *SLC39A8* locus, well-known variants at *GCKR*, *IRS1*, *CDKAL1, RSPO3, KLF14, SH2B3/BRAP, FTO, PLCD3/ACBD4*, *LYPLAL1*, *COBLL1/GRB14*, *PPARG/RAF1*, *TET2*, *HMGA1/C6orf107/UHRF1BP1*, *PPP1R3B*, *FADS1/2/3*, *MC4R* and *GIPR/QPCTL*.

Even if we can exclude allelic heterogeneity at these loci, the real mechanisms of multiple effects cannot be inferred from our analysis, as it might be pleiotropy, but possibly something else. For example, our approach is not able to verify the presence of mediation among associated phenotypes at one locus.

In addition, our analyses were not able to clarify the pattern of multi-phenotype associations at 19 regions.

Another limitation of our approach is the use of an "approximate" conditional analysis: in fact, the method implemented in the GCTA software is, of course, highly useful since it works directly on genome-wide meta-analysis results instead than on cohort-level data (that are less publicly available), but it incurs in an approximation due to the use of an external reference panel, instead of data from the original cohorts, for LD estimation and calculation of conditional p-values. In this analysis, we tried to limit, as much as possible, the errors of this approximation by using, as reference, cohorts having the same ancestry as samples analysed in GWAS meta-analyses. Previous studies, in fact, demonstrated that this method leads to results that are consistent with those obtained with exact conditional analysis directly on cohort-level data, when the reference sample is from the same general population as the discovery sample, even if independent[181].

In addition, we used three different cohorts for our analyses and, even if two of the three used cohorts had a sample size below the recommended value of 2,000 individuals[181], we evaluated the attainment of comparable results using them separately: our results demonstrated robust with respect to the choice of reference samples.

The approaches developed in this project and in the previous one have the limit of not allowing the discovery of novel variants across the genome, besides those already established from single-phenotype GWASs. We thus did not considered polymorphisms which could have a strong overall multiple effect on more than one phenotype, without standing out in univariate GWAS analyses, in the dissection of loci architecture. In the following section, I will present a third project where we applied a multivariate GWAS meta-analysis, with the aim of identifying novel variants associated with multiple cardiometabolic phenotypes.

3.4 Project 3: A multivariate approach for the study of pleiotropy within cardiometabolic phenotypes

3.4.1 Introduction and Aim

In the previous sections we have described approaches for studying multi-phenotype effects and for deepening the knowledge about their mechanisms using available results from univariate data.
Nevertheless, analyses of individual phenotypes are typically limited by (1) the reduced power, arising from the known differences in the magnitude of the observed effects and in sample sizes of phenotype-specific meta-analyses; (2) the increased heterogeneity between larger numbers of genetic studies included, especially in the low and rare allele frequency range; (3) the explanation of a reduced proportion of phenotypic variability, due also to a limited power in detecting low frequency variants and rare variants; (4) a limited capacity in defining multi-phenotype models of association and in interpreting biological functional roles of genetic loci in associated phenotypes.
Another of our aims was thus to apply other powerful multivariate methods directly on cohort-level data because this strategy can lead us to the discovery of novel unknown variants with evidences of cross-phenotype effects at a genome-wide level, and provides the possibility of evaluating the hypothesis of pleiotropy through calculation and comparison of test statistics.
Since the XC-Pleiotropy group has only GWAS meta-analysis result data for cardiometabolic phenotypes at its disposal, we collaborated within the ENGAGE consortium to perform our analyses on its cohort-level data.

The statistical evaluation of multi-phenotype effects through comprehensive modelling and systematic analysis across the genome is challenging. Multivariate association methods have emerged as computationally feasible in large-scale studies, and powerful for dissecting the genetic mechanism at loci associated with several phenotypes[62].
In this third project, we thus undertook multi-phenotype analysis by extending the MultiPhen[62] methodology from O'Reilly and colleagues (see chapter "2.2.2.3_Multivariate approaches" for a description of the original method) and implementing it in a new software: PLEIOTROPY, which models allelic effects on multiple correlated phenotypes. Simulations demonstrated that this method increases power to detect novel associations over single-phenotype analysis by allowing for correlation between phenotypes[62].

We undertook this project following a two-stage study design, which allowed implementation of two complementary approaches: firstly, (1) in stage one we applied a multivariate approach for a genome-wide multi-phenotype analysis and meta-analysis of imputed data up to the 1000 Genomes Project reference panel[165] for four plasma lipids (TG, TC HDL and LDL) and BMI to evaluate comprehensively genetic effects on multiple correlated metabolic phenotypes; secondly, (2) in stage two, detailed follow-up analyses at two known loci were conducted in a comprehensive set of cardiometabolic phenotypes for systematic investigation of the mechanism that underlies the multi-phenotype effects observed at these loci in the genome-wide analysis. This was achieved by employing a two-step multi-phenotype analysis approach that allowed model selection of the best combination of phenotypes that fits the data. In step-one of the analysis, we included cohorts with a wide range of phenotypes available and we investigated the effects of variants at these two loci on this wide range of traits and diseases simultaneously at a study level and across all cohorts through meta-analysis. Based on the best models prioritised in the step-one meta-analysis, we selected the traits that could be tested in step-two of the analysis, including an additional set of cohorts with a smaller number of phenotypes available.

For this second analysis we chose to evaluate the *FTO* and *FADS1* loci. In fact, variants at the *FADS1* gene have been significantly associated in the literature with lipid phenotypes[141,145], fasting glucose[117], resting heart rate[184], inflammatory bowel disease[185] and Crohn's disease[186] in single-phenotype GWAS, making it highly feasible as a pleiotropic candidate. We have already illustrated the numerous associations attributed to the BMI-locus *FTO*[16]: T2D[102], lipids[139], FI[18] and, as secondary effect, risk of coronary artery disease; Mendelian randomisation approaches have demonstrated that variants at *FTO* influence metabolic phenotypes through their effect on adiposity measured by BMI[89,90]. *FTO* was thus a good candidate for the study of pleiotropic effects, and it allowed us to verify if our approach gave results comparable with those from Mendelian randomisation and, thus, if it was appropriate also to distinguishing mediation from potential pleiotropy.

3.4.1.1 The ENGAGE consortium

Figure 3.34: *Logo of the ENGAGE consortium.*

ENGAGE (European Network for Genetic and Genomic Epidemiology) is a research project started in January 2008, funded by the European Commission under the 7th Framework Programme-Health Theme and with duration of five years (http://www.euengage.org/, see figure 3.34 for the logo of the consortium).

The ENGAGE Consortium is composed by 24 leading research organizations and two biotechnology and pharmaceutical companies across Europe, and in Canada and Australia, and it integrates and analyses one of the largest ever human genetics dataset (more than 80,000 genome-wide association scans and DNAs and serum/plasma samples from over 600,000 individuals).

ENGAGE aims to translate the wealth of data emerging from large-scale research in genetic and genomic epidemiology from European (and other) population cohorts into information relevant to future clinical applications in medicine. The concept of ENGAGE is to enable European researchers to identify large numbers of novel susceptibility genes that influence metabolic, behavioural and cardiovascular traits, and to study the interactions between genes and life style factors. The final goal is to investigate the origins and causes of diseases, and to demonstrate that findings from these studies can be used as diagnostic indicators for common diseases, and will help to understand better risk factors, disease progression and why people differ in responses to therapeutic treatment.

In collaboration with the ENGAGE consortium, we had the possibility to work using large cohort's data for our study of multivariate association analysis for cardiometabolic phenotypes.

3.4.2 Stage one: Genome-wide multi-phenotype meta-analysis of lipids five-trait and BMI

3.4.2.1 Materials and Methods

Studies

The genome-wide analysis included 19 GWAS in different cohorts with up to 51,725 individuals. The studied cohorts included 58BC[187], deCODE[116,128,145], DGI[188], DIL, EGCUT_370[189,190], EGCUT_omni[189,190], FINRISK[191], Finnish Twin Cohort[192], Health 2000 GENMETS sub-study[193], Helsinki Birth Cohort Study[194,195], KORAF4[196], Leiden Longevity Study[197], NFBC66[198-200], NTR[201], PIVUS[183], TWINGENE, ULSAM[202] and Cardiovascular Risk in Young Finns Study[203] (table 3.9). All participants were adults and of European ancestry. All subjects provided informed consent and all studies were approved by local ethics committees.

Genotyping and quality control

Contributing GWAS included in undertaken analyses were genotyped with a range of genome-wide arrays (table 3.9).

The quality criteria for filtering of poorly genotyped individuals prior to imputation in each study included: (1) call rate < 93%; (2) sex-discrepancies; (3) ethnic outliers; (4) excess of heterozygosity; (5) known relatedness and (6) MDS (multidimensional scaling) outliers.

The quality criteria for filtering of low quality SNPs were: 1) minor allele frequency (MAF) < 1%; 2) call rate < 95%, or < 99% if SNP has MAF < 5%; 3) failure of Hardy-Weinberg Equilibrium (HWE) exact test (precise threshold depending on studies); 4) sex chromosome SNPs.

Imputation was performed using the 1000 Genomes Project Phase 1 interim, including 2,188 haplotypes from "ALL" populations released in June 2011 or later release of the reference panel[165]. A total of 38 million autosomal SNPs were imputed using IMPUTE2 (with the exception of the deCODE cohort, for which deCODEs own software was used). Study-specific information regarding genotyping platforms and imputation methods are listed in table 3.9.

Short study name	Long study name	Ethnicity	Country	References	Sample size	Genotyping array	Imputation software	Reference panel used for	Website
58BC	1958 British Birth Cohort	White European	UK	187	2,556	Affy6.0 & Illumina 1M	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.ucl.ac.uk/ich/research-ich/mrc-cech/cohort-studies/1958
deCODE	deCODE study	White European	Iceland	116,128,145	14,558	Illumina Human Hap and Omni chips	deCODEs own software	1000 Genomes Phase I (interim)	http://www.decode.com/
DGI	Diabetes Genetics Initiative	White European	Sweden and Finland	188	2,539	Affymetrix GeneChip® Human Mapping 500K Array Set	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.broadinstitute.org/diabetes
DIL	-	White European	UK	-	2,334	Illumina HumanHap550	IMPUTE 2	1000 Genomes Phase I (interim)	-
EGCUT_370	Estonian Genome Center, University of Tartu	White European	Estonia	189,190	833	Illumina HumanHap 300	IMPUTE2	1000 Genomes Phase I (interim)	www.biobank.ee
EGCUT_omniX	Estonian Genome Center, University of Tartu	White European	Estonia	189,190	613	Illumina OmniExpress	IMPUTE2	1000 Genomes Phase I (interim)	www.biobank.ee
FINRISK	FINRISK	White European	Finland	191	1,371	Illumina Human610-Quad	IMPUTE 2	1000 Genomes Phase I (interim)	www.ktl.fi/finriski
FTC	Finnish Twin Cohort	White European	Finland	192	408	Illumina Human670-QuadCustom	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.nationalbiobanks.fi/index.php/studies2/30-finnish-twin-cohort
GenMets	Health2000 GenMets Study	White European	Finland	193, Health and functional capacity in finland, baseline results of the health 2000 health examination survey. 2004. National Public Health Institute.	767/809 cases/controls	Illumina Human610-Quad	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.nationalbiobanks.fi/index.php/studies2/8-health2000
HBCS	Helsinki Birth Cohort Study	White European	Finland	194,195	1,277	Illumina Human670-QuadCustom	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.thl.fi/en_US/web/en/project?id=23572
KORA F4	Cooperative Health Research in the Region of Augsburg	white european	Germany	196	1,633	Affymetrix 6.0	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.helmholtz-muenchen.de/en/kora-en/kora-homepage/index.html
LLS	Leiden Longevity Study	White European	The Netherlands	197	1,769	Illumina Human660W-Quad and Illumina OmniExpress	IMPUTE 2	1000 Genomes Phase I (interim)	https://www.lumc.nl/con/2095/83047/86636/86648/
NFBC1966 (anthro+fasting)	Northern Finland Birth Cohort 1966	White European	Finland	198	5,202	Illumina HumanCNV-370DUO Analysis BeadChip	IMPUTE 2	1000 Genomes Phase I (interim)	http:kelo.oulu.fi/NFBC/
NFBC66 (lipids)	Northern Finland Birth Cohort Study 1966	White European	Finland	198-200	5,202	Illumina Infinium 370cnvDuo	IMPUTE 2	1000 Genomes Phase I (interim)	http:kelo.oulu.fi/NFBC/
NTR	Netherlands Twin Register	White European	Netherlands	201	5,810	-	IMPUTE 2	1000 Genomes Phase I (interim)	www.tweelingenregister.org
PIVUS	Prospective Investigation of the Vasculature in Uppsala Seniors	White European	Sweden	183	793	Merged Metabochip and Omni Express	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.medsci.uu.se/pivus/pivus.htm
Twingene	Twingene	White European	Sweden	-	5,562	Illumina Human OmniExpress	IMPUTE 2	1000 Genomes Phase I (interim)	http://ki.se/ki/jsp/polopoly.jsp?l=en&d=9610
ULSAM	Uppsala Longitudinal Study of Adult Men	White European	Sweden	202	1,003	Merged Metabochip and Omni 2.5M	IMPUTE 2	1000 Genomes Phase I (interim)	http://www.pubcare.uu.se/ULSAM";"http://www.pubcare.uu.se/ULSAM
YFS	The Cardiovascular Risk in Young Finns Study	White European	Finland	203	1,888	Illumina Human670-QuadCustom	IMPUTE 2	1000 Genomes Phase I (interim)	http://youngfinnsstudy.utu.fi/

Table 3.9: *Characteristics of cohorts used for stage one genome-wide multi-phenotype analysis.*

Traits

To investigate the multi-phenotype effects across the genome, information about five traits was used: BMI, HDL, LDL, TC and TG. Measurement of BMI followed standard procedures in all studies; BMI was then inverse normal transformed in men and women separately and sex-specific residuals after adjustment for age, squared age and other study specific covariates, including principal components and centre effects in multi-centric studies, were calculated. Lipid traits (HDL, LDL, TC and TG) were measured from serum or plasma extracted from whole blood, typically using standard enzymatic methods. If LDL was not directly measured, it was calculated using Friedewald's Equation (LDL = TC–HDL–TG/5) for only those with TG below 400 mg/dl, otherwise set to missing. Lipid measurements deviating more than 5 standard deviations from the mean were set to missing. Individuals were excluded if they were receiving lipid-lowering medication at the time of sampling. After applying all these criteria, the lipid phenotypes were defined in men and women separately as the inverse normal transformed residuals resulting from the regression of the lipid measurement on age, squared age and other study specific covariates.

Statistical analysis

To investigate the effect of directly genotyped and imputed variants on the five traits simultaneously, we extended and implemented in a new software (called PLEIOTROPY) the recently published MultiPhen multivariate method[62]. This method was particularly appropriate for our study for three main reasons: (1) it utilises a robust multiple logistic regression to identify the linear combination of the traits most associated with the genotype at a SNP, (2) it allows the combination of both dichotomous and continuous phenotypes, as it makes no assumptions of their distribution and, finally, (3) it allows the analysis of correlated phenotypes as the beta coefficient for each phenotype is adjusted for the other phenotypes in the model, taking into account their correlation.
In a standard genetic association study, a linear regression of the quantitative trait on SNP genotypes is usually performed. However, in the current method, for joint analysis of *K* phenotypes, we modelled the genotype, G_{ij}, of the *i*th individual, at the *j*th variant, coded as 0, 1 or 2, according to the number of minor alleles it carries, as a linear function of phenotype values, *yi*, in a logistic regression framework. Specifically,

$$g^{-1}(G_{ij}) = \alpha_j + \beta_j y_i,$$

where g^{-1} was the logit link function, α_j was the intercept, and β_j was a vector of phenotype regression coefficients for the *j*th variant. For imputed variants, we assigned the genotype with maximal posterior probability. Under this model, we obtained maximum-likelihood estimates (and standard errors) of the phenotype regression coefficients and the corresponding deviance D_j defined as:

$$D_j = 2 \times (l_j - l_0)$$

with an approximate chi-squared distribution with *n* degrees of freedom, where l_j is the log-likelihood of the *j*th logistic regression model and l_0 is the log-likelihood for the null model.

We carried out GWAS in *N* cohorts for *K* (five in this case) phenotypes jointly; then we applied a multi-phenotype fixed-effects inverse-variance weighted meta-analysis of the *N* cohorts by obtaining a combined deviance, $\sum_n D_{jn}$, having an approximate chi-squared distribution with *NK* degrees of freedom and we selected variants with significant multi-phenotype effects based on a p-value$_{LRT}$ < $5x10^{-8}$ (p-value of the Likelihood Ratio Test for joint association).

To parse the meta-analysis results, to allow us to evaluate the ability of the five-trait model to dissect multi-phenotype effects at the genome-wide significant loci, and to identify the main drivers of the observed multi-phenotype associations at each locus, two further analyses were conducted: 1) fixed-effects inverse-variance weighted meta-analysis of each trait regression coefficients from the multi-trait model using GWAMA software[204]; 2) conditional analysis for each trait conditioning on the remaining four traits in each study followed by fixed-effects inverse-variance weighted meta-analysis using GWAMA in the same set of studies as those used in the multi-phenotype meta-analysis.

3.4.2.2 <u>Results</u>

We undertook a genome-wide association analysis of multi-phenotype effects through joint-modelling of BMI, HDL, LDL, TC and TG in up to 51,527 individuals within each of 19 European ancestry GWAS with 1000 Genomes-imputed data, followed by meta-analysis across all studies.
Through fixed-effects inverse-variance weighted meta-analysis of estimates of phenotype regression coefficients from the multi-phenotype model, we detected 26 multi-phenotype association signals achieving genome-wide significance (p-value$_{LRT}$ < $5x10^{-8}$, table 3.10 and figure 3.35).
All signals are localized within or near loci previously associated with lipid traits in large-scale meta-analyses of single-trait GWAS[145,205] (figure 3.36).

Locus	SNP	Chr	Position (b37)	Effect Allele	Other Allele	EAF	N	N Cohorts	pLTR	pBMI	pHDL	pLDL	pTC	pTG
CELSR2	rs12740374	1	109817590	G	T	0.22	33,200	16	2.6x10-47	0.002362	3.6x10-08	0.108898	0.001165	8.6x10-06
PCSK9	rs11591147	1	55505647	G	T	0.01	46,549	18	1.6x10-37	0.025089	5.0x10-06	0.090381	8.4x10-04	3.4x10-04
PCSK9	rs191448950	1	55584844	G	A	0.01	42,477	17	1.5x10-52	0.035975	3.9x10-04	0.006649	3.5x10-04	2.3x10-04
DOCK7	rs61775910	1	62993403	G	A	0.31	34,095	17	4.4x10-18	0.085002	0.573222	0.006796	1.6x10-04	3.9x10-04
APOB	rs563290	2	21288226	G	A	0.17	35,121	18	1.1x10-25	0.002962	0.009373	0.10829	0.061752	0.212319
GCKR	rs1260326	2	27730940	T	C	0.36	35,448	18	1.1x10-43	3.2x10-07	0.016247	0.847902	0.619513	3.0x10-28
ABCG8	rs4953023	2	44074000	G	A	0.06	49,325	19	1.7x10-17	0.503171	0.224011	0.189971	0.020232	0.128563
MTHFD2L	4-75180409	4	75180409	T	C	0.01	36,208	13	5.6x10-09	0.509875	0.252448	0.99921	0.052683	0.673972
HMGCR	rs10474433	5	74616843	T	C	0.34	50,539	19	2.7x10-22	0.02872	0.086567	0.060543	0.030685	0.047951
MLXIPL	rs2240466	7	72856269	G	A	0.12	36,017	17	1.4x10-14	1.8x10-10	0.104985	0.700401	0.359004	2.3x10-17
TRIB1	rs2954021	8	126482077	A	G	0.47	36,127	17	3.2x10-19	2.5x10-10	0.115768	0.965079	0.13703	7.9x10-06
LPL	rs139315015	8	19893297	A	G	0.09	50,233	18	1.3x10-42	1.7x10-11	2.7x10-07	0.893757	0.749625	1.5x10-12
PPP1R3B	rs4841132	8	9183596	A	G	0.11	34,309	15	2.6x10-16	0.333951	4.1x10-05	0.28034	0.198573	0.218652
ABCA1	rs2575876	9	107665739	G	A	0.22	34,728	16	7.6x10-22	0.662752	2.6x10-7	0.072887	0.017059	0.251427
APOA1	rs964184	11	116648917	G	C	0.13	36,783	18	3.2x10-98	3.0x10-17	7.4x10-05	0.00441	6.1x10-04	3.0x10-23
MADD	rs7109147	11	47338384	C	T	0.36	36,932	18	2.6x10-11	0.227393	3.2x10-05	0.858387	0.684621	0.556788
FADS1	rs174550	11	61571478	T	C	0.37	36,922	18	2.7x10-33	0.358195	0.007006	0.023035	4.1x10-07	2.3x10-18
LIPC	rs1532085	15	58683366	A	G	0.40	36,511	18	2.9x10-72	0.170515	2.0x10-39	0.040511	0.028331	6.7x10-18
CETP	rs3764261	16	56993324	C	A	0.31	34,047	17	4.0x10-212	4.4x10-23	2.6x10-63	0.004808	0.221944	3.0x10-05
NUTF2	rs111315946	16	67889793	G	C	0.14	37,027	18	4.2x10-11	0.041327	1.4x10-06	0.367479	0.345876	0.977064
HPR	rs12445401	16	72148419	A	G	0.19	34,978	17	2.7x10-09	0.269564	0.479789	0.053008	0.611345	0.82832
LIPG	rs4939883	18	47167214	T	C	0.18	36,934	18	1.2x10-10	0.318471	1.6x10-05	0.08205	0.110444	0.217432
LDLR	rs8106503	19	11196886	T	C	0.10	26,697	15	6.4x10-56	0.023991	3.6x10-04	0.00946	0.017508	0.002633
CILP2	rs3794991	19	19610596	C	T	0.09	49,399	17	1.1x10-14	0.331952	0.486681	0.214443	0.002063	0.029126
APOE	rs1065853	19	45413233	G	T	0.05	35,338	13	2.8x10-298	0.023336	6.0x10-21	9.3x10-08	1.5x10-07	1.7x10-45
HNF4A	rs1800961	20	43042364	C	T	0.04	28,816	16	2.2x10-08	0.634416	3.3x10-7	0.656425	0.89849	0.319165
PLTP	rs6065906	20	44554015	T	C	0.17	33,399	16	1.5x10-08	1.2x10-7	0.072091	0.201962	0.117541	2.5x10-08

Table 3.10: *26 multi-phenotype association signals achieving genome-wide significance obtained through joint-modelling of BMI, HDL, LDL, TC and TG in 19 European ancestry cohorts. Position is based on build 37 of NCBI database; EAF: Effect allele Frequency; p: p-value.*

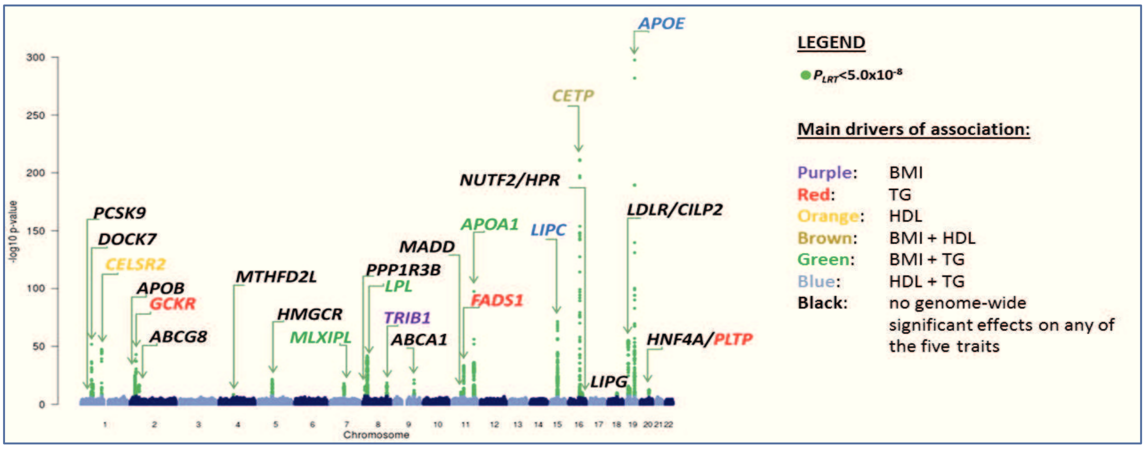


Figure 3.35: *Loci with genome-wide significant joint effects on BMI and lipid traits in the multi-phenotype meta-analysis. Loci are colour-assigned to groups defined based on the main drivers of the observed associations identified through fixed-effects inverse-variance weighted meta-analysis of estimates of trait regression coefficients from the multi-phenotype model.*

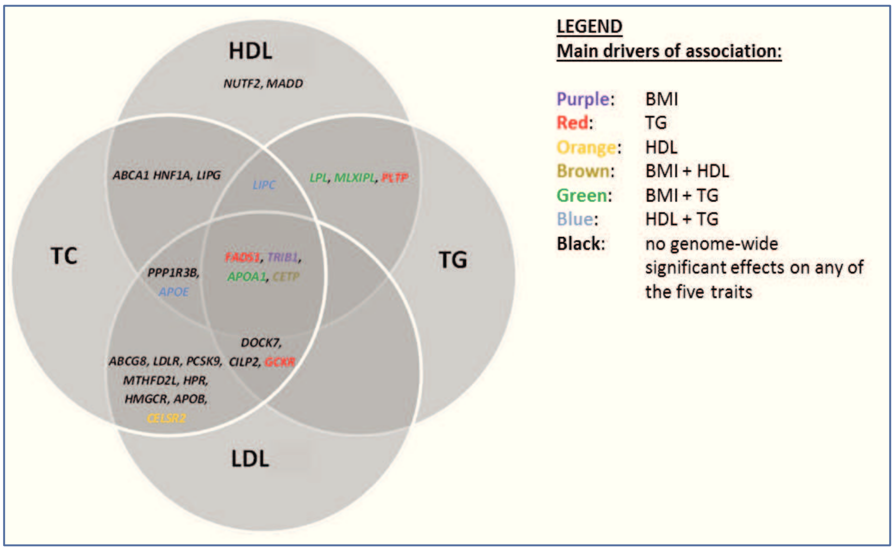


Figure 3.36: *Genome-wide significant associations of the 26 loci with one or multiple lipid traits in previous single-trait meta-analyses. Venn diagram illustrates that all 26 loci have been previously associated with one or more lipid traits in single-trait GWAS meta-analyses*[144,201]*. Loci are located according to their GW significant effects in single-trait analyses and are coloured according to the main drivers of the observed associations within the present fixed-effects inverse-variance weighted meta-analysis of estimates of trait regression coefficients from the multi-phenotype model.*

We observed that at 11 of these loci, associations were driven by an individual trait or two-trait effects; in other words, the individual effects for one or two traits were genome-wide significant in the multi-phenotype meta-analysis. In particular, we observed genome-wide significant effects of variants at:

1) *TRIB1* on BMI,
2) *CETP* on BMI/HDL,
3) *MLXIPL* (MLX interacting protein-like), *LPL*, *APOA1* on BMI/TG (see figure 3.35 and 3.37),
4) *GCKR*, *FADS1*, *PLTP* on TG,
5) *CELSR2* (cadherin EGF LAG seven-pass G-type receptor 2) on HDL and
6) *LIPC* (lipase member C)*,* *APOE* on HDL /TG (table 3.10 and figures 3.35 and 3.38).

The genome-wide significant effects of *TRIB1*, *CETP*, *MLXIPL*, *LPL* and *APOA1* on BMI through meta-analysis of individual trait estimates from the multi-phenotype model were observed for the first time and were missed by previously published single-trait meta-analyses[16,127]. These effects were revealed in the model, where the BMI effect estimates were adjusted for the four plasma lipids, thus taking into account trait correlation.

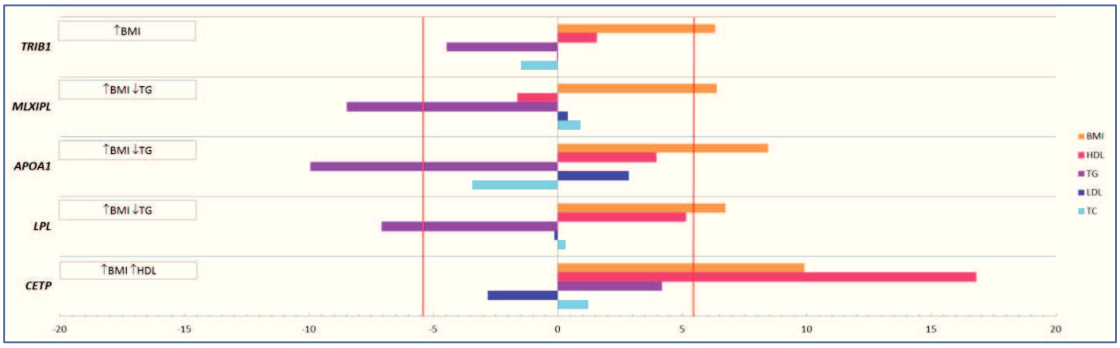


Figure 3.37: *Five loci with genome-wide significant effects on BMI in multi-trait association analysis, coloured bars represent z-score values of associations from multi-phenotype meta-analysis.*

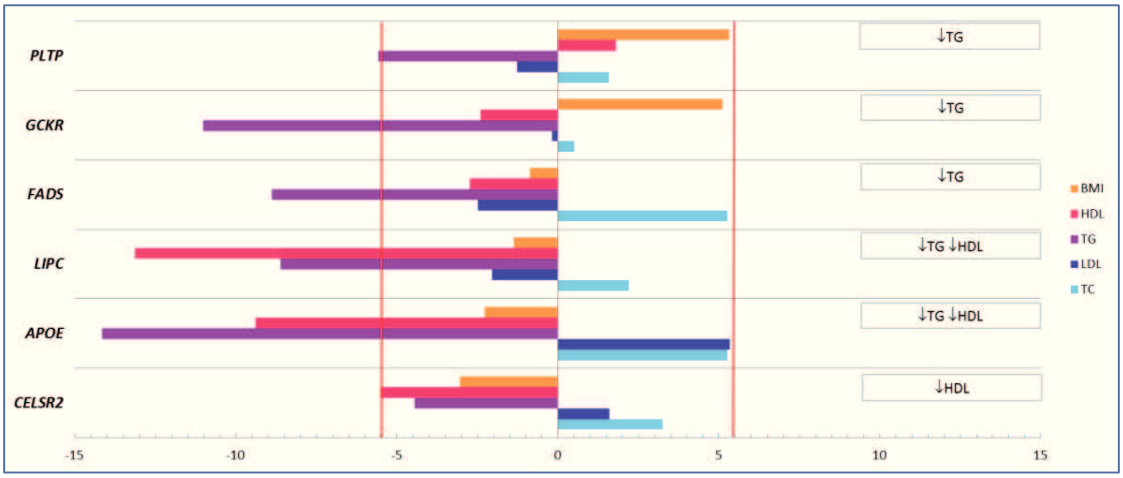


Figure 3.38: *Six loci with identified within the multi-trait association analysis with associations driven by effects on lipids, coloured bars represent z-score values of associations from multi-phenotype meta-analysis.*

At the remaining 15 loci, multiple traits contributed to the signal and, as such, the main drivers of the observed associations could not be determined, suggesting potential pleiotropic effects (table 3.10, figure 3.39).

To evaluate the ability of the five-trait model to dissect multi-phenotype effects at the genome-wide significant loci, we also performed conditional analyses within the same set of studies for each trait, with adjustment for the remaining four traits in each study, and we combined the study-specific results in fixed effects inverse-variance weighted meta-analyses.

The conditional analysis confirmed the presence of associations with individual traits highlighted by the multi-phenotype meta-analysis at most of the above mentioned loci: *TRIB1*, *GCKR*, *PLTP*, *CELSR2*, *CETP*, *MLXIPL*, *APOA1*, *LIPC*, *APOE*.

On the other hand, most of the effects on the remaining lipid traits highlighted by the single-trait GWAS were attenuated or disappeared after conditioning (table 3.11).

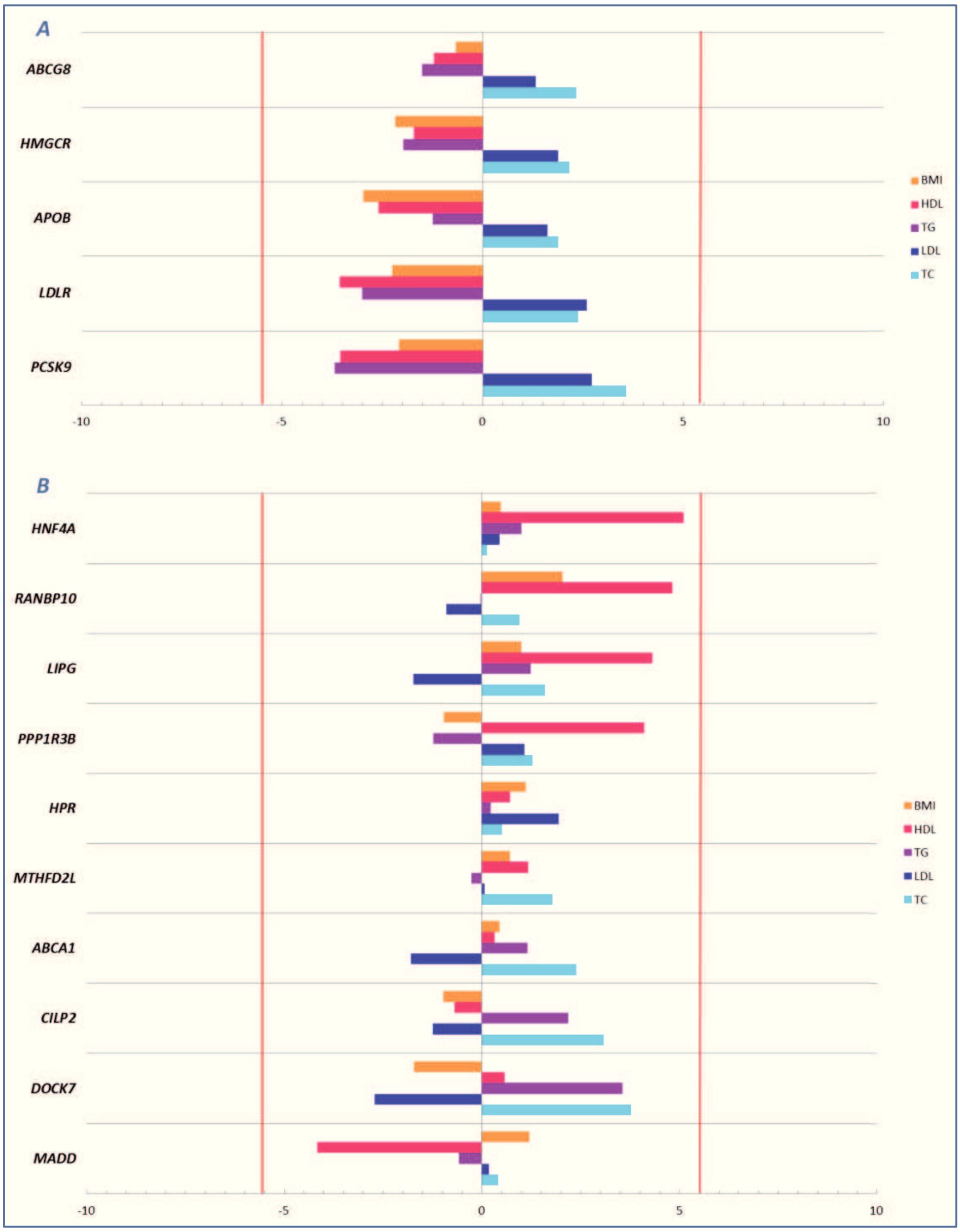


Figure 3.39: *15 pleiotropic loci where multiple traits contributed to the signal of association.* ***A.*** *Group of loci with low BMI, HDL and TG and high LDL and TC, none of these effects reached GW significance by itself.* ***B.*** *Loci with other trend of multiple effects that do not reach GW significance by itself. Coloured bars represent z-score values of associations from multi-phenotype meta-analysis.*

We evaluated multi-phenotype effects of 49 established BMI- and WHR-associated loci within our five-trait meta-analysis. Among these loci, we observed suggestive significance of joint effects only for *FTO* (p-value$_{LRT_rs1558902}$ = 8.9x10^{-7}, r^2=0.965 with rs9939609 used in follow-up analysis): only BMI reached genome-wide significance in the multi-phenotype meta-analysis of this variant (p-value$_{rs1558902}$ = 6.4x10^{-15}) and no effect was observed for the four lipid traits in accordance with the evidence for mediation through adiposity at this locus[89,90].

							Multi-trait meta-analysis																									Single-trait meta-analysis				
										BMI				HDL				LDL				TC				TG				BMI	HDL	LDL	TC	TG
Locus	SNP	Chr	Position (b37)	Effect Allele	Other Allele	EAF	N	N Cohorts	pLTR	Effect	SE	Pvalue	Pcond	Effect	SE	Pvalue	Pcond	Effect	SE	Pvalue	Pcond	Effect	SE	Pvalue	Pcond	Effect	SE	Pvalue	Pcond	Pvalue	Pvalue	Pvalue	Pvalue	Pvalue
CELSR2	rs12740374	1	109817590	G	T	0.22	33,200	16	2.57E-47	-0.030807	0.010128	0.002362	0.164934	-0.106165	0.019261	3.63E-08	0.0000613	0.05948	0.037102	0.108898	0.000166	0.136116	0.041899	0.001165	0.000689	-0.077625	0.017433	0.0000086	0.309599	0.419271	2.21E-06	2.21E-55	9.02E-37	0.529667
PCSK9	rs11591147	1	55505647	G	T	0.01	46,549	18	1.62E-37	-0.075507	0.033703	0.025089	0.832104	-0.299267	0.065535	0.00000503	0.086578	0.198142	0.117009	0.090381	6.92E-08	0.447266	0.133873	0.00084	1.08E-09	-0.206027	0.057433	0.000337	0.688377	0.967115	0.022193	2.01E-88	3.24E-68	0.157483
PCSK9	rs191448950	1	55584844	G	A	0.01	42,477	17	1.55E-52	-0.064484	0.030744	0.035975	0.766613	-0.198317	0.055899	0.000391	0.418931	0.265571	0.097826	0.006649	0.0000583	0.399984	0.111828	0.00035	0.000483	-0.17939	0.048673	0.00023	0.505234	0.808992	0.037213	4.74E-93	1.16E-74	0.161699
DOCK7	rs61775910	1	62993403	G	A	0.31	34,095	17	4.42E-18	-0.015515	0.009008	0.085002	0.133816	0.009954	0.017669	0.573222	0.133173	-0.09363	0.034582	0.006796	0.506351	0.146343	0.038805	0.000164	0.000609	0.055973	0.01577	0.000389	0.000297	0.682208	0.021712	0.0000111	2.57E-17	9.16E-21
APOB	rs563290	2	21288226	G	A	0.17	35,121	18	1.08E-25	-0.031991	0.010762	0.002962	0.024057	-0.054257	0.020877	0.009373	0.155102	0.065265	0.040641	0.10829	0.013191	0.085858	0.045959	0.061752	0.000578	-0.023341	0.018716	0.212319	0.846425	0.860189	0.40749	7.36E-31	9.1E-27	0.007804
GCKR	rs1260326	2	27730940	T	C	0.36	35,448	18	1.09E-43	0.043466	0.008501	0.000000324	0.03456	-0.039238	0.016323	0.016247	0.575389	-0.006145	0.032042	0.847902	0.028741	0.017945	0.036138	0.619513	0.000069	-0.164442	0.014909	2.96E-28	2.88E-23	0.174331	0.018219	0.222493	0.000000174	3.44E-57
ABCG8	rs4953023	2	44074000	G	A	0.06	49,325	19	1.72E-17	-0.009468	0.014141	0.503171	0.065765	-0.031866	0.026208	0.224011	0.804877	0.06358	0.048512	0.189971	0.201343	0.126346	0.054403	0.020232	0.007566	-0.035861	0.023597	0.128563	0.323714	0.754151	0.411545	1.05E-32	1.9E-32	0.515406
MTHFD2L	4-75180409	4	75180409	T	C	0.01	36,208	13	5.63E-09	-0.033643	0.051047	0.509875	0.985533	-0.10599	0.092618	0.252448	0.022345	-0.000181	0.182987	0.99921	0.001018	-0.401603	0.207264	0.052683	0.000321	0.035158	0.083563	0.673972	0.51561	0.472185	0.000991	1.9E-12	2.92E-17	0.01288
HMGCR	rs10474433	5	74616843	T	C	0.34	50,539	19	2.66E-22	0.015528	0.007098	0.02872	0.566793	0.022306	0.013016	0.086567	0.623903	-0.045092	0.024025	0.060543	0.054896	-0.058056	0.02686	0.030685	0.857002	0.023207	0.011733	0.047951	0.078422	0.21928	0.820104	2.19E-41	1.36E-36	0.345131
MLXIPL	rs2240466	7	72856269	G	A	0.12	36,017	17	1.34E-14	-0.069574	0.010896	1.77E-10	0.000039	0.032837	0.020256	0.104985	0.901751	-0.014612	0.037974	0.700401	0.780558	-0.038931	0.042443	0.359004	0.080017	0.153015	0.018037	2.31E-17	3.38E-09	0.022845	0.001354	0.020654	0.709361	1.51E-21
TRIB1	rs2954021	8	126482077	A	G	0.47	36,127	17	3.21E-19	0.051199	0.008086	2.51E-10	0.000000824	0.02468	0.015692	0.115768	0.00451	-0.001349	0.030814	0.965079	0.617066	-0.051727	0.034789	0.13703	0.683161	-0.063834	0.014278	0.0000079	0.00000224	0.140721	1.67E-05	2.16E-14	1.13E-18	4.4E-27
LPL	rs139315015	8	19893297	A	G	0.09	50,233	18	1.26E-42	-0.080908	0.012012	1.69E-11	0.001499	-0.118016	0.022943	0.000000275	0.00000185	0.005583	0.041809	0.893757	0.021896	-0.014943	0.046822	0.749625	0.0511	0.140142	0.019798	1.52E-12	0.0000175	0.479022	2.45E-53	0.031594	0.073029	5.83E-69
PPP1R3B	rs4841132	8	9183596	A	G	0.11	34,309	15	2.6E-16	-0.013005	0.01346	0.333951	0.290446	0.101294	0.024673	0.0000408	0.00000412	0.051861	0.048041	0.28034	0.011853	0.069532	0.054087	0.198573	0.000209	-0.028371	0.023065	0.218652	0.644854	0.072013	1E-18	2.36E-11	7.81E-17	0.006492
ABCA1	rs2575876	9	107665739	G	A	0.22	34,728	16	7.6E-22	0.004317	0.009898	0.662752	0.297798	0.019215	0.061322	0.000000264	0.00072	-0.068318	0.03809	0.072887	0.365332	0.102263	0.042861	0.017059	0.000511	0.020204	0.017617	0.251427	0.572749	0.50645	7.38E-22	0.000424	6.48E-14	0.207101
APOA1	rs964184	11	116648917	G	C	0.13	36,783	18	3.21E-98	0.100467	0.011886	3E-17	0.0000123	0.088177	0.022239	0.0000741	0.00000444	0.125192	0.043955	0.00441	0.000134	-0.171162	0.049909	0.000608	0.000216	-0.212291	0.021356	2.96E-23	2.44E-27	0.936689	3.4E-23	3.56E-11	5.96E-22	7.82E-104
MADD	rs7109147	11	47338384	C	T	0.36	36,932	18	2.55E-11	0.010049	0.008325	0.227393	0.352476	-0.067329	0.01617	0.0000317	0.00000664	0.005654	0.03169	0.858387	0.275889	0.014527	0.035763	0.684621	0.252964	-0.008545	0.014542	0.556788	0.107813	0.019245	3.86E-15	0.000665	0.201481	0.0000593
FADS1	rs174550	11	61571478	T	C	0.37	36,922	18	2.7E-33	-0.007637	0.008312	0.358195		-0.043407	0.016093	0.007006		-0.072011	0.031678	0.023035		0.181229	0.03575	0.000000407		-0.128076	0.01464	2.29E-18		0.02298	1.35E-11	6.48E-18	3.04E-15	1.79E-17
LIPC	rs1532085	15	58683366	A	G	0.40	36,511	18	2.88E-72	-0.01129	0.008238	0.170515	0.513884	-0.211537	0.016093	2.01E-39	2.55E-18	-0.064262	0.031368	0.040511	0.083257	0.077678	0.035422	0.028331	0.00085	-0.125067	0.014499	6.72E-18	0.0000956	0.158757	2.77E-65	0.124208	5.33E-15	0.0012
CETP	rs3764261	16	56993324	C	A	0.31	34,047	17	4E-212	-0.090686	0.009159	4.4E-23	0.00000708	-0.299023	0.017795	2.59E-63	1.57E-41	0.095374	0.033814	0.004808	0.000024	-0.046823	0.038338	0.221944	0.181002	-0.065359	0.015655	0.0000301	0.015858	0.918585	5.7E-167	2.34E-10	0.0000634	0.0000136
NUTF2	rs111315946	16	67889793	G	C	0.14	37,027	18	4.24E-11	-0.023569	0.011551	0.041327	0.257638	-0.109205	0.022652	0.00000145	0.00000232	0.039664	0.044013	0.367479	0.741103	-0.046963	0.049822	0.345876	0.803047	0.000575	0.020002	0.977064	0.05515	0.160387	5.27E-21	0.789561	0.007959	0.013562
HPR	rs12445401	16	72148419	A	G	0.19	34,978	17	2.75E-09	-0.01154	0.010452	0.269564	0.271611	-0.014317	0.020259	0.479789	0.737575	-0.076723	0.039651	0.053008	0.000192	-0.022668	0.044605	0.611345	0.000255	-0.003969	0.018304	0.82832	0.1901	0.772514	0.407844	1.28E-17	7.8E-18	0.001709
LIPG	rs4939883	18	47167214	T	C	0.18	36,934	18	1.24E-10	0.010464	0.010489	0.318471	0.645592	0.088197	0.020445	0.0000162	0.000014	-0.070893	0.040768	0.08205	0.399292	0.073205	0.045863	0.110444	0.08265	0.022921	0.018585	0.217432	0.637784	0.274208	1.96E-25	0.521689	0.0000387	0.139993
LDLR	rs8106503	19	11196886	T	C	0.10	26,697	15	6.37E-56	-0.034915	0.015465	0.023991	0.418991	-0.108576	0.030429	0.000362	0.000164	0.150579	0.058012	0.00946	0.000109	0.157835	0.06642	0.017508	0.0000067	-0.081371	0.027045	0.002633	0.642007	0.731207	0.360399	7.86E-89	1.3E-62	0.1664
CILP2	rs3794991	19	19610596	C	T	0.09	49,399	17	1.06E-14	-0.011619	0.011976	0.331952	0.928845	-0.015229	0.021893	0.486681	0.001197	-0.04888	0.039375	0.214443	0.000635	0.135373	0.04392	0.002063	0.005793	0.042503	0.019478	0.029126	0.111474	0.980303	0.680915	3.42E-18	3.65E-23	1.18E-17
APOE	rs1065853	19	45413233	G	T	0.05	35,338	13	2.8E-298	-0.042938	0.01893	0.023336	0.133632	-0.35361	0.03763	5.97E-21	0.760563	0.344251	0.064424	9.32E-08	0.002952	0.388545	0.073875	0.000000148	0.0000312	-0.471843	0.033318	1.74E-45	0.000000432	0.43453	3.22E-10	0	4.45E-203	1.22E-24
HNF4A	rs1800961	20	43042364	C	T	0.04	28,816	16	2.2E-08	0.010048	0.021129	0.634416	0.143275	0.204167	0.039982	0.000000335	0.000161	0.033951	0.076315	0.656425	0.532339	0.010928	0.08568	0.89849	0.333895	0.036102	0.036242	0.319165	0.806944	0.557646	1.34E-16	0.025404	0.00000189	0.977517
PLTP	rs6065906	20	44554015	T	C	0.17	33,399	16	1.52E-08	0.046362	0.00873	0.000000112	0.033653	0.028018	0.015578	0.072091	0.330938	-0.035847	0.028095	0.201962	0.311103	0.049114	0.03138	0.117541	0.127754	-0.078857	0.014131	2.46E-08	0.0000125	0.58057	1.5E-11	0.912008	0.424536	0.000000011

Table 3.11: *26 multi-phenotype association signals achieving genome-wide significance within the five-trait meta-analysis. Conditional analyses and single-trait meta-analysis within the same set of studies for each trait are also shown. Position is based on build 37 of NCBI; EAF: Effect allele Frequency; p: p-value.*

Short study name	Long study name	References	Website	Number of subjects with FTO and BMI data	Longitudinal data	Age at BMI	Mean BMI at baseline	Proportion women (%)	N with ever type 2 diabetes	N with ever any acute stroke or transient ischemic attack	N with ever ischemic stroke	N with ever hypertension	N with ever coronary heart disease	N with total cholesterol	N with systolic blood pressure	N with triglycerites	N with c-reactive protein	N with LDL cholesterol	N with HDL cholesterol	N with fasting glucose	N with diastolic blood pressure	N with 2h post OGTT glucose	Glucose: mean (SD) / mmol/L	2hG: mean (SD) / mmol/L	HDL-C: mean (SD) / mmol/L	LDL-C: mean (SD) / mmol/L	Triglycerides: mean (SD) / mmol/L	Total cholesterol: mean (SD) / mmol/L	Systolic blood pressure: mean (SD) / mmHg	Diastolic blood pressure: mean (SD) / mmHg	CRP: mean (SD) / mg/L
EGCUT	Estonian Genome Centre of the University of Tartu	189,190	www.biobank.ee	11282	No	45.74 (18.34)	26.48 (5.52)	56.0%	1132	323	128	4362	537	2362	11277	1868	-	1872	1921	1757	11277	301	5.38 (0.79)	5.92 (1.20)	1.36 (0.70)	3.60 (1.63)	0.42 (0.50)	5.50 (1.18)	130.81 (21.75)	80.52 (12.92)	-
FR02	Finnish Risk factor survey 2002	191	www.ktl.fi/finriski	8142	Yes	47.957 (13.12)	26.91 (4.68)	53.3%	746	228	157	3569	383	7549	8142	7549	8119	7454	7549	-	8142	-	-	-	1.51 (0.42)	3.45 (0.95)	1.40 (0.94)	5.60 (1.07)	137.13 (22.02)	80.46 (12.53)	2.49 (5.23)
FR07	Finnish Risk factor survey 2007	191	www.ktl.fi/finriski	5900	Yes	50.45 (13.93)	27.13 (4.88)	53.3%	567	121	83	2874	196	5066	5877	3991	5900	3990	3991	3872	5874	3827	5.72 (0.46)	6.16 (1.65)	1.46 (0.35)	3.28 (0.82)	1.16 (0.74)	5.32 (0.99)	138.92 (22.69)	81.01 (12.55)	2.43 (5.03)
FR92	Finnish Risk factor survey 1992	191	www.ktl.fi/finriski	5536	Yes	44.39 (11,32)	26.13 (4.46)	53.9%	629	253	175	2415	410	5451	5537	5450	932	5330	5451	-	5536	-	-	-	1.40 (0.35)	3.55 (0.986)	1.50 (1.07)	5.62 (1.11)	136.84 (20.98)	82.18 (12.93)	4.05 (7.63)
FR97	Finnish Risk factor survey 1997	191	www.ktl.fi/finriski	6747	Yes	47.79 (13.22)	26.63 (4.61)	53.3%	818	303	235	3191	516	6594	6807	6594	6457	6480	6594	-	6808	-	-	-	1.40 (0.36)	3.48 (0.92)	1.48 (1.03)	5.54 (1.05)	137.63 (21.78)	83.52 (12.47)	2.38 (5.91)
KORAF3	Cooperative Health Research in the Region of Augsburg, KOoperative Gesundheitsforschung in der Region Augsburg	196	http://www.helmholtz-muenchen.de/en/kora-en/kora-homepage/index.html	2976	No	56.92 (12.76)	27.61 (4.62)	52.3%	238	-	-	1476	-	231	2985	231	243	231	231	231	2985	-	6.41 (1.18)	-	1.51 (0.46)	3.48 (0.91)	1.52 (1.27)	5.81 (1.02)	135.21 (22.48)	84.95 (11.61)	0.45 (0.85)
KORAF4	Cooperative Health Research in the Region of Augsburg, KOoperative Gesundheitsforschung in der Region Augsburg	196	http://www.helmholtz-muenchen.de/en/kora-en/kora-homepage/index.html	3009	No	56.08 (13.26)	27.62 (4.81)	51.5%	214	-	-	1158	-	3008	3018	3007	3018	3007	3007	2990	3018	2724	6.17 (1.20)	7.06 (2.47)	1.44 (0.37)	3.51 (0.90)	1.42 (1.02)	5.58 (1.02)	126.94 (21.19)	78.25 (10.95)	0.25 (0.53)
MPP	Malmö Prevention Project	206	http://www.ludc.med.lu.se/research-units/diabetes-and-endocrinology/sample-collections/malmoe-prevention-project-mpp/	13616	No	45.2 (7.01)	24.28 (3.30)	33.3%	-	-	-	4700	-	10880	9853	10870	-	-	243	13615	9858	7370	5.46 (0.554)	5.64 (1.47)	1.55 (0.37)	-	1.27 (0.78)	5.61 (1.04)	127.1 (14.2)	83.9 (8.8)	-
NFBC1966	Northern Finland Birth Cohort 1966	198-200	http:kelo.oulu.fi/NFBC/	4775	Yes	31.17 (0.35)	24.70 (4.28)	51.8%	123	33	13	419	17	4566	4769	4565	4755	4551	4566	4322	4762	-	5.70 (0.63)	-	1.55 (0.38)	3.00 (0.89)	1.18 (0.73)	5.06 (0.99)	125.21 (13.88)	77.69 (11.60)	2.01 (3.66)
NFBC1986	Northern Finland Birth Cohort 1986	198-200	http:kelo.oulu.fi/NFBC/	5285	Yes	16.00 (0.37)	21.22 (3.48)	51.0%	-	-	-	17	-	5110	5281	5110	5247	5110	5110	4789	5281	-	5.18 (0.73)	-	1.40 (0.29)	2.25 (0.57)	0.84 (0.42)	4.26 (0.79)	115.48 (12.73)	67.69 (7.58)	0.99 (2.85)
PIVUS	Prospective Investigation of the Vasculature in Uppsala Seniors	183	http://www.medsci.uu.se/pivus/pivus.htm	979	No	70.19 (0.17)	27.07 (4.3)	49.8%	34	35	-	144	27	784	975	784	972	782	784	855	975	-	5.57 (0.56)	-	1.52 (0.43)	3.40 (0.84)	1.23 (0.56)	5.4 (0.98)	149.7 (22.7)	78.8 (10.2)	3.2 (4.8)
ULSAM	Uppsala longitudinal study of adult men	202	http://www.pubcare.uu.se/ULSAM	1175	Yes	49.6 (0.6)	24.8 (3.0)	0.0%	48	274	167	438	271	1128	1175	1128	1082	917	917	1123	1175	908	5.50 (0.50)	6.90 (1.80)	1.40 (0.40)	5.20 (1.20)	1.8 (0.83)	6.8 (1.3)	131.4 (16.8)	82.6 (10.5)	3.3 (4.7)
WTCCCCont	Wellcome Trust Case Control Consortium 1958 Birth Cohohrt	207	www.wtccc.org.uk	5443	No	46(0)	27.37 (4.8)	45.7%	113	-	-	1427	-	5352	5430	5341	2687	5041	5345	-	5430	-	-	-	1.11 (0.40)	2.95 (0.90)	2.10 (1.2)	5.00 (1.1)	139.46 (24.0)	81.14 (13.4)	2.20 (4.32)
QIMR-AUSTRALIA	Twin studies at the Queensland Instutite of Medical Research	208-210	http://genepi.qimr.edu.au/	11827	No	35.61 (17.41)	24.12 (5.12)	57.2%	-	-	-	-	-	8315	-	8311	-	7962	8278	-	-	-	-	-	1.52 (0.42)	3.31 (0.93)	1.89 (1.24)	5.67 (1.05)	-	-	-
deCODE	deCODE genetics sample set	116,128,145	http://www.decode.com/	36896	No	59.1 (18.0)	27.2 (5.3)	63.8%	2126	-	2366	8248	3568	18393	16726	17099	24128	16297	17009	12017	-	-	5.25 (0.60)	-	1.45 (0.42)	3.69 (1.08)	1.46 (0.94)	5.82 (1.17)	135.4 (20.4)	-	38.7 (65.8)
DGIcases	Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research	188	http://www.broadinstitute.org/scientific-community/science/projects/diabetes-genetics-initiative/diabetes-genetics-initiative	1602	No	64.42 (10.32)	28.50 (4.40)	49.8%	-	201	-	1101	447	1455	1584	1455	-	426	1414	-	1583	-	-	-	1.20 (0.31)	3.96 (1.04)	1.99 (1.43)	5.81 (1.18)	149.2 (20.8)	84.1 (10.2)	-
DGIcontrols	Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research	188	http://www.broadinstitute.org/scientific-community/science/projects/diabetes-genetics-initiative/diabetes-genetics-initiative	1508	No	58.61 (10.16)	26.70 (3.78)	52.2%	-	26	-	666	36	1416	1503	1416	-	710	1406	1387	1502	1017	5.32 (0.52)	5.64 (1.32)	1.39 (0.34)	4.03 (0.92)	1.32 (0.69)	5.93 (1.09)	135.9 (18.8)	81.5 (9.9)	-
NESDA	Netherlands Study of Depression and Anxiety	201	http://www.nesda.nl/en/	1927	No	41.90 (12.52)	25.65 (5.04)	67.6%	95	-	-	-	-	1813	-	1810	1901	1795	1806	1722	-	-	5.03 (0.58)	-	1.63 (0.44)	3.13 (1.00)	1.29 (0.85)	5.11 (1.06)	-	-	2.84 (5.13)
NTR	Netherlands Twin Register	211	www.tweelingenregister.org	5416	No	42.55 (14.76)	25.25 (4.30)	61.2%	240	-	-	-	-	5032	-	5032	4958	5021	5030	4821	-	-	5.35 (0.53)	-	1.42 (0.38)	3.10 (0.95)	1.32 (0.82)	5.12 (1.06)	-	-	3.30 (6.54)
HELENA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MONALISA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DIL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MONICA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PPP	Prevalence, Prediction and Prevention of diabetes	212	-	4355	No	47.90 (15.63)	26.31 (4.44)	53.8%	160	49	-	1778	210	3923	4355	3924	-	3881	3923	4173	4355	4144	5.28 (0.57)	5.24 (1.59)	1.42 (0.39)	3.30 (0.944)	1.26 (0.748)	5.30 (1.06)	129.3 (17.2)	79.1 (9.9)	-
RS	Rotterdam Study	213	http://www.epib.nl/research/ergo.htm	5745	Yes	69.0 (8.80)	26.3 (3.69)	58.7%	1178	149	-	3273	1557	3382	5791	3230	5567	3140	3331	3295	5790	-	5.88 (1.32)	-	1.34 (0.37)	3.75 (0.88)	1.51 (0.73)	6.59 (1.22)	139.2 (22.3)	73.7 (11.5)	3.38 (6.8)
Twingene	Cardiovascular risk factor study of Swedish twin pairs	-	http://ki.se/ki/jsp/polopoly.jsp?l=en&d=9610	6386	Yes	65.4(8.3)	26.2(4.2)	45.0%	640	461	254	3771	25	5401	6101	5398	6489	5322	5401	5657	6044	-	5.40 (0.60)	-	1.4(0.41)	3.90 (0.93)	1.4(0.9)	5.94(1.1)	139.3(19.8)	81.6(10.5)	3.42(7.4)
TwinsUK	TwinsUK	214	http://www.twinsuk.ac.uk/	4829	No	52.79394(14.42)	26.0598 (5.06)	0.0%	80	-	-	572	-	4245	2646	4194	4035	4183	4247	4517	2646	966	4.66 (0.58)	6.78 (0.40)	1.47 (0.40)	3.41 (1.09)	1.05 (0.70)	5.44 (1.23)	121.41 (15.91)	76.60 (10.56)	2.65 (4.70)

Table 3.12: *Phenotypic details of cohorts studied for follow-up multi-phenotype analysis.*

3.4.3 Stage two: Multi-phenotype follow-up analysis of two selected loci, FTO and FADS1, to dissect the mechanism of multi-phenotype effects

3.4.3.1 Materials and Methods

Studies

Two loci, *FTO* and *FADS1*, were selected for detailed follow-up analyses of the genetic mechanisms of effects. The step-one analyses consisted of 12 studies with up to 72,247 individuals: EGCUT[189,190], FINRISK92/FINRISK97/FINRISK02/FINRISK07[191], KORAF3/KORAF4[196]; MPP[206], NFBC1966 and NFBC86[198-200], PIVUS[183] and ULSAM[202] (table 3.12).

The step-two analysis included step-one plus 14 additional studies, which were: 58BC-WTCCC[207], AUSTWINS[208-210], deCODE[116,128,145], DGI[188], DIL, HELENA, MONALISA, MONICA, NTR/NESDA[201,211], PPP[212], Rotterdam Study[213], TWINSUK[214] (table 3.12). The maximum number of individuals in stage-two analysis was 167,984. All participants were adults or adolescents (HELENA cohort) and of European ancestry. All subjects provided informed consent and all studies were approved by local ethics committees.

SNPs and proxies at *FTO* and *FADS1*

We considered two lead SNPs (rs9939609 for *FTO* and rs174550 for *FADS1*). If those were not genotyped or imputed within a specific cohort, proxies were analysed instead (table 3.13). To define proxies, we used linkage disequilibrium estimates from 1000 Genomes Pilot 1 European samples[165] with r^2 range between 0.5 and 1.

Nearest gene	Lead SNP	Chr	Position (b36)	Effect allele	Other allele	Proxy used	r2 with lead SNP
FTO	rs9939609	16	52378028	T	A	rs17817712	1
						rs3751812	1
						rs8050136	1
FADS1	rs174550	11	61328054	T	C	rs174547	0.93
						rs102275	0.932
						rs174546	0.965

Table 3.13: *FTO and FADS1 variants and their proxies tested in the follow-up multi-phenotypes analysis.*

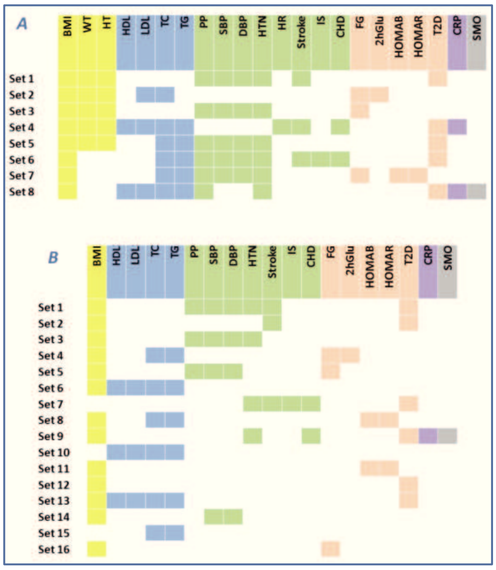


Figure 3.40: *Sets of phenotypes tested using multi-phenotype analysis at FTO and FADS1 loci **A**. in step-one **B**. and in step-two.*

Sets of analysed phenotypes

The following groups of phenotypes were investigated in step-one (figure 3.40A) and step-two (figure 3.40B) analyses:

- Anthropometric: BMI, weight, height;
- Lipid traits: HDL, LDL, TC, TG;
- Cardiovascular: pulse pressure (PP), systolic/diastolic blood pressure (SBP, DBP), hypertension (HTN), heart rate (HR), stroke, ischemic stroke (IS), coronary heart disease (CHD);
- Glycaemic: fasting glucose (FG),

2hGlu, homeostasis model assessments of beta-cell function and insulin resistance(HOMAB, HOMAIR), T2D;

- Inflammation: C-reactive protein (CRP) and
- Addiction: smoking behaviour (SMO).

Phenotype	Definition	Exclusions
Body mass index (BMI)	Defined as weight (kg)/height2 (m2). Trait is inverse normal transformed separately in men and women	None
Weight (WT)	Trait (kg) is inverse normal transformed separately in men and women	None
Height (HT)	Gender-specific Z-scores from residual (standardized residuals are calculated from raw height adjusted for age and study-specific covariates)	Outliers +/-4SD
HDL cholesterol (HDL)	Trait is untransformed fasting in mmol/l	Patients on lipid-lowering medication
LDL cholesterol (LDL)	Trait is untransformed fasting in mmol/l. For individuals with lipid-lowering medication adjust observed values by dividing LDL values by 0.7.	None
Total cholesterol (TC)	Trait is untransformed, fasting in mmol/l	Patients on lipid-lowering medication
Triglycerides (TG)	Natural log-transformed, fasting in mmol/l	Patients on lipid-lowering medication
Pulse pressure (PP)	Defined as the difference between systolic blood pressure and diastolic blood pressure measured in mmHg. For individuals on anti-hypertensive treatment observed values were first adjusted by +15 mmHg for SBP and +10 mmHg for DBP	None
Systolic blood pressure (SBP)	Trait is untransformed in mmHg.	None
	Observed values were adjusted by +15 mmHg for SBP for individuals on anti-hypertensive treatment	
Diastolic blood pressure (DBP)	Trait is untransformed in mmHg.	None
	Observed values were adjusted by +10 mmHg for DBP for individuals on anti-hypertensive treatment	
Hypertension (HTN)	SBP ≥140mmHg, DBP ≥90mmHg, or on anti-hypertensive treatment	None
Heart rate (HR)	Measured by ECG or peripheral pulse (usually at wrist) in beats per minute (bpm)	None
Ever any acute stroke (Stroke)	Stroke or transient ischemic attack at any time point either defined from hospital discharge registry or cause of death registry (main diagnosis); or from adjudicated events	None
	ICD-8 codes: 430-436	
	ICD-9 codes: 430-436	
	ICD-10 codes: I60-I64+G45	
	Note: Self-reported events were not considered useful	
Ever ischemic stroke (IS)	IS at any time point either defined from hospital discharge registry or cause of death registry (main diagnosis); or from adjudicated events	None
	ICD-8 codes: 432-434	
	ICD-9 codes: 433-434	
	ICD-10 codes: I63	
	Note: Self-reported events were not considered useful	
Ever coronary heart disease (CHD) (acute myocardial infarction or unstable angina)	CHD at any time point either defined from hospital discharge registry or cause of death registry (main diagnosis); or from validated events	None
	ICD-8 codes: 410, 411	
	ICD-9 codes: 410, 411B	
	ICD-10 codes: I20.0, I21, I22	
	Note: Self-reported events were not considered useful	
Fasting glucose (FG)	Trait is inverse normal transformed in mmol/L. Glucose measurements made in blood were adjusted by multiplying FG values by 1.13, as glucose concentration in blood is lower than in plasma	Diabetics (T2D, T1D) ['diagnosed', on diabetes treatment (oral and insulin), and/or FPG >=7 mmol/L)], non-fasting, pregnant
2h post OGTT Glucose (2hGlu)	Trait is inverse normal transformed in mmol/L	
Homeostasis model assessment of percent beta cell function/ homeostasis model assessment of insulin resistance (HOMAIR/HOMAB)	Traits are inverse normal transformed (unitless).HOMAIR=FG (mmol/L) x FI (mU/L)/ 22.5	
	HOMAB= 20 x FI (mU/L)/ [FG (mmol/L) -3.5]	
Type 2 diabetes (T2D)	Fasting blood glucose ≥7 mmol/L or anti-diabetic treatment. Self-reported diabetes	Type I diabetics, pregnant at blood sampling
C-reactive protein (CRP)	Trait is natural log-transformed, measured using high-sensitivity assays, in mg/l	Known inflammatory disease or acute infection (at time of blood sampling)
Smoking behavior (SMO)	Defined as ever/never smoker	None

Table 3.14: *Traits/outcomes (their definition and exclusions applied at a study level) tested in the follow-up multi-phenotype analysis.*

Based on prior knowledge of the genetic effects at *FTO* and *FADS1* on specific phenotypes, eight phenotype sets of continuous traits and dichotomous disorder (maximum 12 phenotypes within each set) were formed and tested to allow maximisation of sample sizes and meaningful combinations of phenotypes in the step-one analysis (figure 3.40A). Based on the best models prioritised within each set through meta-analysis across step-one studies, 16 new phenotype sets (maximum seven phenotypes within each set) were formed and tested in step-two analysis (figure 3.40B). To avoid confounding by age, sex and study-specific variables (e.g. study site and geographical covariates), the residuals from a linear regression model on these variables were used for quantitative traits. A detailed description of phenotype definition, normalization and exclusions applied are given in table 3.14.

Statistical analysis

In analyses of selected loci with sets of phenotypes tested using the PLEIOTROPY software (as explained above in chapter "3.4.2_Stage one: Genome-wide multi-phenotype meta-analysis of lipids five-trait and BMI"), we aimed to select the optimal model from the set of all alternative models for each genetic variant of interest, based on an appropriate fit measure. With a set of *K* phenotypes, there are 2^k possible models. As the best model(s) were selected at the meta-analysis step, each cohort was asked to fit all 2^k logistic regression models for each variant and phenotype set considered. The Bayesian Information Criterion (BIC) score was selected as the optimal model fit statistic as it adds a penalty to the likelihood ratio to optimise the trade-off between added complexity and explained variance by adding more phenotypes to the model. BIC is defined as:

$$BIC_j = 2l_j + (s_j + 1) \times \log(n)$$

Where l_j is the log-likelihood of the *j*th logistic regression model, s_j is the number of phenotypes in the model and *n* is the sample size (note that for a null model with intercept only, $BIC_0 = -2l_0 + log\ (n)$, where l_0 is the log-likelihood for the null model).

We calculated meta-analysis BIC scores and null BICs using two different approaches: (1) just summing BIC (sumBIC) and null BIC (sumBICnull) estimates from all cohorts contributing in the meta-analysis; (2) based on the sum of log-Likelihood and sum of null log-Likelihood using data from all contributing cohorts:

$$BIC_j = -2 \times \sum_i l_{ij} + K + log \sum_i size_i;$$
$$BICnull = -2 \times \sum_i l_{i0} + log \sum_i size_i,$$

where:

$\sum_i l_{ij}$ = sum of log-Likelihoods from all *i* contributing cohorts,

K = count of phenotypes in given model,

$\sum_i size_i$ = sum of sample sizes from all *i* contributing cohorts,

$\sum_j logl_{j0}$ = sum of null LogLikelihoods.

As this second calculation seemed to be biased towards more complicated models, we used the first approach. Step-one and step-two meta-analyses were thus performed by summing the BIC scores across all participating studies separately for each genetic marker and each model. While comparing alternative models for the same dependent variable (here: genetic marker), the model with smallest BIC within each set and locus was selected. More in details, the best models within the specific sets were selected based on (summed BIC_j < summed BIC_0) across all studies.

FTO rs9939609

STEP ONE

Phenotype s	Top models	N	BIC	BICnull	sumBIC	sumBICnull	LTR	P_{LTR}
Set 1	BMI	71065	191526.84	191807.4	191706.22	191897.2	291.7	**2.11907E-65**
	WT	71065	191582.6	191807.4	191761.93	191897.2	235.94	3.01994E-53
	BMI+WT	71065	191512.21	191807.4	191781.55	191897.2	317.5	1.13698E-69
	HT+WT	71065	191513.09	191807.4	191782.19	191897.2	316.62	1.76539E-69
	BMI+DBP	71065	191514.33	191807.4	191783.41	191897.2	315.38	3.28173E-69
Set 2	BMI	15249	41005.2	41038.09	41049.19	41060.08	42.52	**6.99632E-11**
	WT	15249	41020.02	41038.09	41064.13	41060.08	27.7	-
	BMI+2hGlu	15249	41007.3	41038.09	41073.29	41060.08	50.06	-
	BMI+TG	15249	41007.66	41038.09	41073.65	41060.08	49.7	-
	BMI+FG	15249	41010.18	41038.09	41076.07	41060.08	47.18	-
Set 3	BMI	33682	90518.59	90634.12	90633.467	90691.47	125.964	**3.13112E-29**
	WT	33682	90549.62	90634.12	90664.594	90691.47	94.93	1.97238E-22
	BMI+WT	33682	90514.78	90634.12	90687.024	90691.47	140.194	3.60794E-31
	BMI+FG	33682	90516.18	90634.12	90688.362	90691.47	138.794	7.26551E-31
	BMI+DBP	33682	90516.58	90634.12	90688.841	90691.47	138.392	8.88299E-31
Set 4	BMI	11864	31680.7	31724.14	31747.17	31757.46	52.82	**3.6556E-13**
	WT	11864	31690.76	31724.14	31757.21	31757.46	42.76	6.18847E-11
	BMI+WT	11864	31675.91	31724.14	31775.63	31757.46	67	-
	BMI+Stroke	11864	31678.24	31724.14	31778.03	31757.46	64.66	-
	HT+WT	11864	31679.98	31724.14	31779.65	31757.46	62.92	-
Set 5	BMI	54120	145401.49	145600.5	145568.438	145684	209.928	**1.42421E-47**
	WT	54120	145443.51	145600.5	145610.475	145684	167.906	2.12096E-38
	BMI+DBP	54120	145386.77	145600.5	145637.192	145684	235.542	7.12368E-52
	BMI+SBP	54120	145393.37	145600.5	145643.803	145684	228.942	1.93142E-50
	BMI+T2D	54120	145393.25	145600.5	145643.847	145684	229.068	1.81349E-50
Set 6	BMI	50789	136438.47	136630.4	136565.9	136694.1	202.744	**5.26077E-46**
	BMI+DBP	50789	136428.7	136630.4	136619.76	136694.1	223.344	3.17297E-49
	BMI+IS	50789	136430	136630.4	136621.21	136694.1	222.044	6.07796E-49
	BMI+Stroke	50789	136432.32	136630.4	136623.41	136694.1	219.726	1.93689E-48
	BMI+SBP	50789	136433.55	136630.4	136624.76	136694.1	218.496	3.58258E-48
Set 7	BMI	19086	50986.44	51045.56	51074.58	51089.6	68.978	**9.9568E-17**
	BMI+FG	19086	50983.61	51045.56	51115.89	51089.6	81.664	-
	BMI+SBP	19086	50983.88	51045.56	51116.21	51089.6	81.388	-
	BMI+DBP	19086	50984.24	51045.56	51116.53	51089.6	81.026	-
	BMI+TG	19086	50988.75	51045.56	51121.13	51089.6	76.516	-
Set 8	BMI	31976	85646.49	85772.15	85784.1284	85841.04	136.03738	**1.95809E-31**
	BMI+TC	31976	85645.67	85772.15	85852.0781	85841.04	147.22362	-
	BMI+SMO	31976	85646.1	85772.15	85852.6044	85841.04	146.79334	-
	BMI+CRP	31976	85648.74	85772.15	85855.1893	85841.04	144.15644	-
	BMI+TG	31976	85649.98	85772.15	85856.4436	85841.04	142.9201	-

STEP TWO

Phenotype s	Top models	N	BIC	BICnull	sumBIC	sumBICnull	LTR	P_{LTR}
Set 1	BMI	95,649	257866.4	258247.7	258082.9	258355.9	392.8	**2.03385E-87**
	BMI+DBP	95,649	257854.9	258247.7	258179.2	258355.9	415.6	5.67031E-91
	BMI+T2D	95,649	257856	258247.7	258180.1	258355.9	414.6	9.34876E-91
	BMI+SBP	95,649	257857.6	258247.7	258181.9	258355.9	413	2.0806E-90
	BMI+Stroke	95,649	257859.9	258247.7	258184.2	258355.9	410.8	6.25048E-90
Set 2	BMI	140,892	379892.8	380502.8	380110.9	380611.8	621.8	**3.0357E-137**
	BMI+T2D	140,892	379884.2	380502.8	380211.9	380611.8	642.2	3.5322E-140
	BMI+Stroke	140,892	379886.6	380502.8	380214.1	380611.8	639.8	1.1727E-139
	BMI+Stroke+T2D	140,892	379878.3	380502.8	380314.9	380611.8	659.8	1.0928E-142
	T2D	140,892	380446.1	380502.8	380664.5	380611.8	68.4	-
Set 3	BMI	116,131	312569.5	313025.5	312922.6	313201.9	467.6	**1.0667E-103**
	BMI+SBP	116,131	312550.6	313025.5	313079.5	313201.9	498.2	6.5651E-109
	BMI+DBP	116,131	312551.5	313025.5	313080.6	313201.9	497.2	1.0824E-108
	BMI+PP	116,131	312553.2	313025.5	313082.4	313201.9	495.6	2.4089E-108
	BMI+HTN	116,131	312569.3	313025.5	313098.4	313201.9	479.6	7.181E-105
Set 4	BMI	18,681	50050.58	50111.84	50132.13	50152.73	71.1	**3.3957E-17**
	BMI+TC	18,681	50047.28	50111.84	50169.88	50152.73	84.24	-
	BMI+TG	18,681	50048.24	50111.84	50170.8	50152.73	83.26	-
	BMI+FG	18,681	50049.88	50111.84	50172.4	50152.73	81.62	-
	TC	18,681	50108.48	50111.84	50190.17	50152.73	13.2	-
Set 5	BMI	63,921	171649.8	171885.3	171905.8	172013.2	246.52	**1.48981E-55**
	BMI+FG	63,921	171636.8	171885.3	172020.4	172013.2	270.7	-
	BMI+SBP	63,921	171644.6	171885.3	172028.5	172013.2	262.82	-
	BMI+PP	63,921	171645.5	171885.3	172029.2	172013.2	261.98	-
	BMI+DBP	63,921	171647.4	171885.3	172031	172013.2	260.06	-
Set 6	BMI	76,521	204958.3	205231.1	205296.6	205400.2	284	**1.00913E-63**
	BMI+TG	76,521	204938.8	205231.1	205446.4	205400.2	314.6	-
	BMI+TC	76,521	204942.7	205231.1	205450.3	205400.2	310.8	-
	TG	76,521	205200	205231.1	205538.4	205400.2	42.2	-
	TC	76,521	205211	205231.1	205549.3	205400.2	31.2	-
Set 7	T2D	166,650	449959.6	450017.9	450148.9	450113.1	70.2	-
	HTN	166,650	449994	450017.9	450183.2	450113.1	35.8	-
	Stroke	166,650	450013.4	450017.9	450203.6	450113.1	16.4	-
	CHD	166,650	450020.3	450017.9	450209.4	450113.1	9.6	-
	IS	166,650	450022	450017.9	450212.1	450113.1	7.8	-
Set 8	BMI	29,508	78832.3	78947.23	79019.28	79040.73	125.22	**4.55537E-29**
	BMI+TG	29,508	78820.9	78947.23	79101.42	79040.73	146.92	-
	BMI+TC	29,508	78827.69	78947.23	79108.23	79040.73	140.14	-
	TG	29,508	78929.49	78947.23	79116.56	79040.73	28.04	-
	TC	29,508	78941.54	78947.23	79128.53	79040.73	15.98	-
Set 9	BMI	68,606	184504.4	184775.8	184676.9	184862.1	282.54	**2.0994E-63**
	BMI+SMO	68,606	184498.5	184775.8	184757	184862.1	299.58	8.85176E-66
	BMI+CRP	68,606	184500.2	184775.8	184758.8	184862.1	297.92	2.02999E-65
	BMI+T2D	68,606	184504.4	184775.8	184762.9	184862.1	293.72	1.65772E-64
	BMI+CHD	68,606	184510.1	184775.8	184768.7	184862.1	287.96	2.95312E-63
Set 10	TG	78,565	210719	210749.4	211057.9	210919.1	41.6	-
	TC	78,565	210728.7	210749.4	211067.6	210919.1	32	-
	TC+TG	78,565	210706.3	210749.4	211214.9	210919.1	65.6	-
Set 11	BMI	32,489	86840.4	86970.17	87030.46	87065.27	140.16	**2.45598E-32**
Set 12	BMI	161,417	434639.3	435318.8	434966.9	435482.6	691.4	**2.2174E-152**
	BMI+T2D	161,417	434622.8	435318.8	435114.7	435482.6	720	4.508E-157
	T2D	161,417	435248.4	435318.8	435576.3	435482.6	82.4	-
Set 13	BMI	71,071	190454.9	190701.9	190749.5	190849.1	258.12	**4.40876E-58**
	BMI+TG	71,071	190440.2	190701.9	190881.6	190849.1	284.04	-
	BMI+T2D	71,071	190440.6	190701.9	190882.1	190849.1	283.58	-
	BMI+TC	71,071	190442.1	190701.9	190883.8	190849.1	282.08	-
	T2D	71,071	190668.2	190701.9	190962.6	190849.1	44.78	-
Set 14	BMI	114,297	307609.6	308086	307960.8	308261.4	488	**3.8814E-108**
	BMI+SBP	114,297	307590.3	308086	308116.4	308261.4	519	1.9979E-113
	BMI+DBP	114,297	307590.8	308086	308117	308261.4	518.6	2.4403E-113
	BMI+DBP+SBP	114,297	307576.4	308086	308278.2	308261.4	544.6	-
	SBP	114,297	308049.6	308086	308400.2	308261.4	48.2	-
Set 15	TG	122,512	329464.9	329511.6	329825.9	329692.1	58.4	-
	TC	122,512	329488.9	329511.6	329849.8	329692.1	34.6	-
	TC+TG	122,512	329449.5	329511.6	329991	329692.1	85.6	-
Set 16	BMI	22,515	60361.52	60432.95	60460.56	60482.47	81.44	**1.80664E-19**

FADS1 rs174550

STEP ONE

Phenotype s	Top models	N	BIC	BICnull	sumBIC	sumBICnull	LTR	P_{LTR}
Set 1	HT	72247	190652.36	190672.3	190832.02	190762.1	31.12	-
	DBP	72247	190664.34	190672.3	190844.16	190762.1	19.14	-
	WT	72247	190666.28	190672.3	190845.56	190762.1	17.2	-
	BMI	72247	190666.66	190672.3	190846.04	190762.1	16.82	-
	Stroke	72247	190666.86	190672.3	190846.39	190762.1	16.62	-
Set 2	TC+TG	15586	39969.06	40014.31	40035.01	40036.4	64.56	**9.57137E-15**
	TC	15586	39997.85	40014.31	40041.81	40036.4	26.12	-
	TC+WT+TG	15586	39958.22	40014.31	40046.33	40036.4	85.06	-
	TG	15586	40004.89	40014.31	40048.85	40036.4	19.08	-
	WT	15586	40009.07	40014.31	40053.12	40036.4	14.9	-
Set 3	FG	34804	90744.77	90757.78	90859.66	90815.27	23.46	-
	PP	34804	90751.94	90757.78	90866.9	90815.27	16.3	-
	SBP	34804	90753.55	90757.78	90868.45	90815.27	14.68	-
	WT	34804	90753.48	90757.78	90868.48	90815.27	14.76	-
	BMI	34804	90757.96	90757.78	90872.89	90815.27	10.28	-
Set 4	LDL	11892	32191.62	32224.51	32258.16	32257.77	42.26	-
	TC	11892	32197.43	32224.51	32263.97	32257.77	36.46	-
	TC+TG	11892	32170.59	32224.51	32270.34	32257.77	72.68	-
	LDL+TG	11892	32172.2	32224.51	32271.99	32257.77	71.06	-
	HDL+LDL	11892	32172.48	32224.51	32272.25	32257.77	70.8	-
Set 5	TC+TG	55296	146217.63	146408.6	146468.252	146492.1	212.796	**6.19349E-47**
	TC	55296	146335.65	146408.6	146502.755	146492.1	83.854	-
	TG	55296	146359.86	146408.6	146527.041	146492.1	59.638	-
	HT+TC+TG	55296	146203.55	146408.6	146537.762	146492.1	237.794	-
	TC+WT+TG	55296	146205.72	146408.6	146540.032	146492.1	235.624	-
Set 6	TC+TG	51958	137949.7	138115.5	138140.8	138179.3	187.54	**1.88889E-41**
	TC	51958	138053.4	138115.5	138181	138179.3	72.94	-
	PP+TC+TG	51958	137945.1	138115.5	138200	138179.3	203	-
	T2D+TC+TG	51958	137945.4	138115.5	138200.3	138179.3	202.64	-
	SBP+TC+TG	51958	137947.6	138115.5	138202.4	138179.3	200.5	-
Set 7	TC	19291	51212.47	51251.12	51300.84	51295.27	48.52	-
	TC+TG	19291	51172.15	51251.12	51304.62	51295.27	98.72	-
	FG	19291	51238.15	51251.12	51326.49	51295.27	22.84	-
	TG	19291	51241.85	51251.12	51330.25	51295.27	19.14	-
	FG+TC+TG	19291	51153.78	51251.12	51330.48	51295.27	126.96	-
Set 8	TC+TG	32026	86237.32	86389.83	86443.8749	86458.71	173.26086	**2.38167E-38**
	TC	32026	86323.29	86389.83	86460.8235	86458.71	76.91432	-
	TG	32026	86358.34	86389.83	86496.2413	86458.71	41.86648	-
	PP+TC+TG	32026	86227.6	86389.83	86502.9781	86458.71	193.35754	-
	HTN+TC+TG	32026	86229.77	86389.83	86505.2163	86458.71	191.18938	-

STEP TWO

Phenotype s	Top models	N	BIC	BICnull	sumBIC	sumBICnull	LTR	P_{LTR}
Set 1	BMI	94,105	247707.6	247720.6	247922.8	247828	24.4	-
	DBP	94,105	247713.4	247720.6	247928.4	247828	18.6	-
	Stroke	94,105	247715	247720.6	247930.2	247828	17	-
	PP	94,105	247717.9	247720.6	247933	247828	14.2	-
	SBP	94,105	247720.1	247720.6	247935.4	247828	12	-
Set 2	BMI	145,213	384184.8	384197.3	384419.5	384314.3	24.4	-
	Stroke	145,213	384192.2	384197.3	384426.3	384314.3	17	-
	T2D	145,213	384197.6	384197.3	384431.8	384314.3	11.6	-
	BMI+Stroke	145,213	384179.7	384197.3	384531.4	384314.3	41.4	-
	BMI+T2D	145,213	384181.9	384197.3	384532.9	384314.3	39.2	-
Set 3	DBP	117,671	309921.1	309951.7	310260.9	310121.6	42.2	-
	BMI	117,671	309934.5	309951.7	310274.6	310121.6	28.8	-
	HTN	117,671	309944.5	309951.7	310284.6	310121.6	18.8	-
	PP	117,671	309945.8	309951.7	310285.8	310121.6	17.6	-
	SBP	117,671	309946.2	309951.7	310286	310121.6	17.2	-
Set 4	TC+TG	21,328	55476.37	55570.76	55621.86	55619.36	114.34	-
	TC	21,328	55540.13	55570.76	55637.18	55619.36	40.6	-
	TG	21,328	55543.91	55570.76	55640.94	55619.36	36.84	-
	FG+TC+TG	21,328	55467.47	55570.76	55661.59	55619.36	133.2	-
	FG	21,328	55567.08	55570.76	55664.21	55619.36	13.66	-
Set 5	FG	65,899	172370.3	172409.9	172624.7	172537.1	50.72	-
	DBP	65,899	172396.5	172409.9	172650.7	172537.1	24.54	-
	BMI	65,899	172400	172409.9	172654.4	172537.1	21	-
	SBP	65,899	172400.1	172409.9	172654.6	172537.1	20.88	-
	PP	65,899	172400.8	172409.9	172655.2	172537.1	20.18	-
Set 6	TC+TG	73,133	193762.2	194116.9	194263.3	194283.7	377.14	**1.27377E-82**
	TC	73,133	193978.8	194116.9	194312.8	194283.7	149.32	-
	TG	73,133	194010.5	194116.9	194344.7	194283.7	117.58	-
	BMI+TC+TG	73,133	193755.4	194116.9	194423.4	194283.7	395.14	-
	BMI	73,133	194104	194116.9	194438	194283.7	24.16	-
Set 7	HTN	167,984	446668.8	446669.2	446856.9	446763.1	12.6	-
	IS	167,984	446668.8	446669.2	446857.8	446763.1	12.4	-
	T2D	167,984	446670.7	446669.2	446859.9	446763.1	10.6	-
	Stroke	167,984	446671.8	446669.2	446860.8	446763.1	9.6	-
	CHD	167,984	446674.9	446669.2	446864	446763.1	6.4	-
Set 8	TC	31,170	82589.69	82680.77	82778.69	82775.34	101.42	-
	TC+TG	31,170	82496.25	82680.77	82779.83	82775.34	205.22	-
	TG	31,170	82646.44	82680.77	82835.64	82775.34	44.68	-
	BMI	31,170	82673.11	82680.77	82862.11	82775.34	18	-
	BMI+TC	31,170	82580.58	82680.77	82864.15	82775.34	120.88	-
Set 9	BMI	71,738	191147.8	191154.5	191336.3	191248.8	17.9	-
	HTN	71,738	191149.3	191154.5	191337.7	191248.8	16.44	-
	T2D	71,738	191154.2	191154.5	191342.4	191248.8	11.56	-
	CRP	71,738	191155.1	191154.5	191343.4	191248.8	10.64	-
	SMO	71,738	191156.5	191154.5	191344.8	191248.8	9.22	-
Set 10	TC+TG	75,111	199038.2	199410.4	199540.2	199577.6	394.72	**1.93928E-86**
	TC	75,111	199262.7	199410.4	199597.3	199577.6	158.94	-
	TG	75,111	199302.5	199410.4	199637.2	199577.6	119.22	-
Set 11	BMI	34,300	90795.03	90802.48	90987.11	90898.55	17.9	-
Set 12	BMI	161,726	427302.1	427318.4	427631.1	427482.4	28.2	-
	T2D	161,726	427313	427318.4	427641.3	427482.4	17.4	-
	BMI+T2D	161,726	427293.8	427318.4	427785.9	427482.4	48.6	-
Set 13	TC+TG	67,706	179752.8	180084.7	180188.1	180229.8	354.2	**1.22025E-77**
	TC	67,706	179954.5	180084.7	180244.8	180229.8	141.28	-
	TG	67,706	179984.9	180084.7	180275.1	180229.8	110.96	-
	T2D+TC+TG	67,706	179732.7	180084.7	180313	180229.8	385.36	-
	BMI+TC+TG	67,706	179747.8	180084.7	180328.2	180229.8	370.3	-
Set 14	DBP	118,797	312687.5	312722.1	313041.4	312899	46.2	-
	BMI	118,797	312703.3	312722.1	313057.4	312899	30.4	-
	SBP	118,797	312715.1	312722.1	313069	312899	18.8	-
	BMI+DBP	118,797	312668	312722.1	313199.2	312899	77.4	-
	DBP+SBP	118,797	312677.3	312722.1	313208.4	312899	68.2	-
Set 15	TC+TG	120,072	316696.9	317142.2	317231.9	317320.7	468.8	**1.5899E-102**
	TC	120,072	316979.2	317142.2	317336.1	317320.7	174.8	-
	TG	120,072	317004.3	317142.2	317360.8	317320.7	149.6	-
Set 16	BMI	22,928	59757.57	59760.59	59856.8	59810.27	13.08	-

Table 3.15: *Step-one (top) Step-two (bottom) multi-phenotype meta-analysis top five models within each of the tested sets for FTO (rs9939609, left) and FADS1 (rs174550, right) variants.*

3.4.3.2 Results

To further investigate the mechanisms that underlie the observed multi-phenotype effects, and to extend the applied methodological framework to model the combination and selection of a wide range of cardiometabolic phenotypes within multi-phenotype analysis, we employed a two-step multi-phenotype analysis approach at two selected loci: *FTO* (rs9939609 or its proxy r^2=1) and *FADS1* (rs174550 or its proxy, $r^2 \geq 0.93$) as reported in table 3.13.

Through multi-phenotype modelling in step-one, for *FTO* the best model was always the model with BMI alone: all the eight different sets of phenotypes confirmed this result reporting a p-value$_{LTR}$ which ranged from 6.99×10^{-11} (for set 2) to 2.11×10^{-65} (for set 1, see table 3.15). In step-one analysis of *FADS1*, only in four sets of phenotypes the best model was not the null model , that is only in four sets summed BIC_j was minor than summed BIC_0; in all these sets the best model included TG and TC together (p-value$_{LTR}$ from 9.57×10^{-15} for set 2 to 6.19×10^{-47} for set 5, see table 3.15).

Through multivariate analysis in step-two, in the analysis of *FTO* (sample size from 18,681 to 161,417), for set 7 and set 15, the best model was the null one. On the other hand, within the remaining tested phenotype sets, we confirmed that the best model included only BMI (as reported in figure 3.41A and in table 3.15) with a maximum significant p-value$_{LTR}$ = 2.21×10^{-152} for set 12 and a minimum significant p-value$_{LTR}$ = 3.39×10^{-17} for set 4. This result confirmed a previously reported mediation effect of BMI for the examined phenotypes[89,90] at the *FTO* locus, providing a proof of principle for the applied method in discerning mediation from potential pleiotropy.

At *FADS1* (N = 67,706 to 120,072), in step-two analysis the best model for eight sets was the null one. The best model that emerged, instead, within the remaining four sets, even in this second step, included the two lipids: TC and TG. The p-value$_{LTR}$ of this model varied from a maximum significance of 1.59×10^{-102} for set 15 and a minimum significance of 1.22×10^{-77} for set 13. Therefore, at *FADS1* locus pleiotropy between TG and TC was highly supported, as well as mediation through them for other phenotypes (figure 3.41B and in table 3.15).

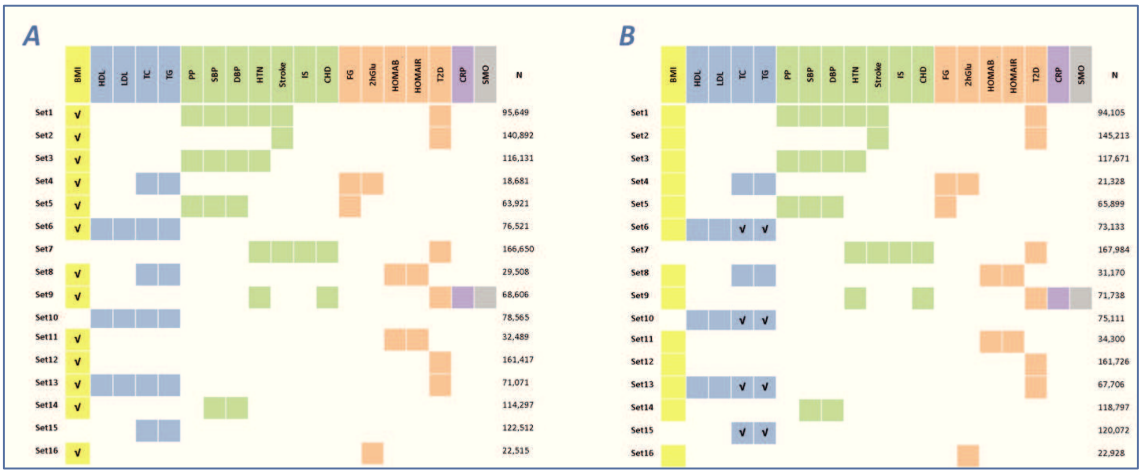


Figure 3.41: *Phenotype sets and best models from multi-phenotype modelling at **A**) FTO and **B**) FADS1 loci in step-two follow-up analysis. Every possible combination of phenotypes was tested for each set. Ticked boxes indicate the traits of the best model within each phenotype set. On the right sample size is reported for each tested set.*

3.4.4 Discussion

In the current study, we have extended methodology for simultaneous multi-phenotype analysis in individual studies to allow large-scale genome-wide multi-phenotype association testing of imputed genetic variants, meta-analysis and, moreover, phenotypic modelling and selection for individual loci, with the aim of dissecting the mechanism of multi-phenotype effects.
We applied and tested our approach with the employment of two complementary multi-phenotype approaches: 1) a genome-wide multivariate analysis for the detection of novel, potential, pleiotropic variants and 2) a follow-up analysis at two selected loci for the decomposition of the mechanism of multi-phenotype effects at each specific locus by considering a wide range of phenotypes and making model selection.

Genome-wide analysis was limited by the number of phenotypes available across all contributing studies and thus permitted testing of a single phenotypic set that included BMI, TG, TC, HDL and LDL. We only tested a model where all phenotypes were included, since testing all possible alternative models would have been too computationally demanding.
Anyway, by combining correlated phenotypes that are likely to share biological pathways, we have been able to reveal multi-phenotype effects at 26 lipid loci and highlight previously unknown effects on BMI at five of these loci.
Despite the fact that *TRIB1*, *CETP*, *MLXIPL*, *LPL* and *APOA1* have not been previously associated with BMI in single-trait GWAS, there is evidence that points to a link between several of them and obesity. In particular, a variant at *TRIB1* (rs2980879, r^2 = 0.4 with our variant, rs2954021) has been previously associated, at a genome-wide significance, with adiponectin, which is a highly abundant adipose-derived plasma protein that modulates several metabolic processes[215]. A study in middle-aged individuals has demonstrated that changes in adiposity-related measures, such as BMI, waist circumference and visceral fat area, correlated with a rise in adiponectin levels[216]. Furthermore, studies have reported an increase in CETP (cholesteryl ester transfer protein) activity and mass in obese compared to non-obese controls and weight reduction normalizing the altered CETP levels[217-219]. One of these studies has also suggested that elevated plasma PLTP (plasma lipid transfer protein) levels in obese patients might be the direct outcome of adiposity *per se*[220]. PLTP is one of the loci where we also observed suggestive effects on BMI (p-value = 1.2×10^{-7}) in multi-phenotype meta-analysis, which, interestingly, has not been observed in single-trait published GWAS meta-analysis for BMI[16,126], or in our BMI meta-analysis in the same set of individuals. ChREBP (also known as MLXIPL) is a major determinant of adipose tissue fatty acid synthesis and glycolysis and a recently discovered isoform, ChREBP-β, has been demonstrated to correlate strongly with ChREBP activity[221]. The expression of this isoform has been shown to be markedly reduced in obese and obese-diabetic compared to non-obese controls[222]. Another study provided evidence of a decrease in LPL expression primarily due to an increase in BMI indicating a different transcriptional regulation between obese and lean subjects[223]. Our study indicates that, although these loci have not been previously associated with BMI per se in single-trait GWAS, joint modelling of the five correlated traits might have been able to capture unmeasured products, mediators or traits that are not

considered here and are related to obesity. Therefore, our approach resulted useful in discovering novel variants across the genome which could have strong effect without standing out in univariate GWAS analyses for single phenotypes and may contribute to explain part of the missing heritability of complex phenotypes.

Follow-up analysis at two selected loci allowed the decomposition of the mechanism of multi-phenotype effects at each specific locus by considering a wide range of phenotypes and making model selection. Model selection in the follow-up analysis at each of the two selected loci was done at meta-analysis rather than individual-study level; such a strategy was advantageous as it allowed avoiding bias in meta-analysis results due to low power to detect effects within individual studies.
Using this second approach, we demonstrated mediation of the effects through adiposity, measured by BMI, at *FTO* and through TC and TG at *FADS1*. A recent Mendelian randomization study of the effects of *FTO*-derived adiposity on 24 cardiometabolic disease outcomes and traits suggested causal relationship between BMI and multiple cardiometabolic phenotypes, further supporting a mediation effect[90]. Nevertheless, the mediation effects observed at *FTO* and *FADS1* cannot rule out the possibility of pleiotropic effects on other untested phenotypes at these loci and should be subject to further research.
The observed independent effects of *FADS1* on TC and TG are supported by separate biochemical pathways in which the enzyme fatty acid delta-5 desaturase (FADS1) is involved. In particular, FADS1 plays a role in the synthesis of omega-6 fatty acids where it inserts to eicosatrienoyl-CoA a fourth double bound between carbons of the fatty acid chain generating arachidonyl-CoA. Such polyunsaturated fatty acids of the Acyl-CoAs are directly used in the formation of glycerolipids like TG and phosphatidylcholines. As cholesterol is hydrophilic (due to its hydroxyl), it cannot be easily transported or stored. The hydroxyl group of cholesterol and the fatty acid of a phosphatidylcholine are necessary to form a cholesterol ester in blood. The resulting apolar cholesterol ester can be stored and transported with lipoproteins. Therefore, FADS1 is used for the formation of TG (directly) and cholesterol esters (via phosphatidylcholines). This study did not provide evidence for independent effects of *FADS1* on HDL, LDL, BMI or T2D, consistently with the discussed function of FADS1. Therefore, associations of *FADS1* with lipoproteins, HDL and LDL, BMI and T2D observed in previous GWAS are likely to be mediated by its effect on TC and TG, having pleiotropic effect on these last two traits.

In this third project we have demonstrated that modelling of multiple correlated phenotypes can help in the discovery and characterisation of complex phenotype loci, otherwise missed by the standard univariate approaches. This study has also highlighted that the systematic evaluation of multi-phenotype effects through multivariate analysis can uncover some of the possible mechanisms of genetic effects at individual loci, for example mediation, and can provide novel insights into the pathophysiological processes underlying metabolic trait variability.
Regrettably, a potential limitation of the tested method is the assessment of multi-phenotype effects that is possible only in the context of those phenotypes available in the participating studies. Another limit is that this approach is time consuming and multiple model evaluation is less feasible

at a genome-wide level. Finally, as it considers variants one at a time, this study does not cover the problem of individualising potential pleiotropic genomic regions and interpreting their multiple associations, distinguishing for example phenomena of multi-phenotype allelic heterogeneity.

4 Final discussion and conclusions

4.1 Main conclusions of our study

4.1.1 Hypothesis about pleiotropic effects on metabolic phenotype

In the past years, a wealth of genetic data for cardiometabolic phenotypes highly increased together with the internationally combined effort of researchers for the identification of associated genetic loci. The discoveries of these studies highlighted the complex relationships between metabolic traits and diseases: it was, in fact, clear that numerous overlaps exist between associated loci, but the patterns of multi-phenotype associations were variable and not always consistent with epidemiological expectations. This complexity of the observed cardiometabolic phenotype associations can be due to several underlying factors, such as pleiotropy, allelic heterogeneity, phenotypic mediation, gene-gene and gene-environment interaction.
The large efforts in the past have enlightened our understanding of biology of these metabolic phenotypes, but they have also suggested the need for further analyses.

The idea that we developed in the projects presented in this thesis is that the dissection of cross-phenotype effects, and in particular of pleiotropy, will help uncovering the mechanistic basis of physiological processes governing cardiometabolic quantitative traits and of pathogenetic processes leading to metabolic diseases.
This research will increase our understanding of the extent of shared genetics among traits and diseases and our global understanding of phenotypes as a range of inter-related manifestations of biological mechanisms rather than as isolated events.
The definition of specific sets of effects on combinations of cardiometabolic phenotypes might clarify known physiological and pathophysiological mechanisms and highlight novel biological pathways, targets for translational research, for therapeutic intervention, and for the understanding of the pathophysiology of human metabolism.

Thanks to the collaboration with the XC-pleiotropy group and the ENGAGE consortium, my PhD project mainly focused on the dissection of pleiotropic effects at common variants across the genome on cardiometabolic phenotypes.

4.1.2 What we discovered in developed projects

The research presented in this thesis has been divided into three specific projects:
(1) Exploration of established multi-phenotype effects at cardiometabolic loci from published univariate meta-analyses, defining clusters of loci with similar multiple effects, comparing them to known epidemiological expectations, and identifying enriched biological networks within the most interesting groups of loci;

(2) Dissection of the architecture of established cardiometabolic loci showing multiple phenotype associations for a better definition of the underlying mechanisms of multi-phenotype effects and for the discernment of potential pleiotropy from allelic heterogeneity;
(3) Development and application of a statistical strategy for multivariate analyses of CP phenomena using cohorts data from the ENGAGE consortium to discover new uncovered multiple associations and to follow-up GWAS meta-analysis at two loci.

Specific results and conclusions for each of these sections have been already reported in precedent chapters. In general, we can group our findings in several primary points.

Both univariate and multivariate approaches can be applied for the study of pleiotropy
From a methodological perspective, in the presented projects, we developed different approaches to address the issue of pleiotropy that allowed us to undertake deeper analyses of data obtained through univariate GWAS meta-analyses. Moreover, to address limitations of single-phenotype analyses, we applied a multivariate joint analysis of multiple correlated phenotypes that brought to several advantages, including the ability to take into account correlation between phenotypes, a boost in power, an improved precision of parameter estimates, and the identification of novel candidate genes.

Cardiometabolic phenotypes share genetic background
This fact was formerly suggested by a comparison of results from univariate GWAS reported in the literature[7,20], and it was confirmed also from our study results.
Starting from the preliminary analysis that we reported in Scott et al. 2012[18], we noted a considerable number of glycaemic loci associated with other metabolic phenotypes; particularly, fasting insulin loci associated also with lipid levels (lower HDL and higher triglycerides).
Through the application of a multi-phenotype meta-analysis, and of approaches for graphical visualisation of multiple effects on association results from univariate analyses, as well as of a multivariate GWAS and meta-analysis method, many variants at cardiometabolic loci have been highlighted with interesting multiple associations characterising different aspects of metabolism (for example obesity and blood pressure, or lipids and glycaemic levels, or obesity and lipids levels).
The application of conditional analysis has also underlined that multiple associations, not necessarily at the same variants, but also at adjacent variants, may underlie shared genetic causes between different phenotypes.

Cardiometabolic phenotype loci can be grouped according to the combination of their multi-phenotype effects
Our efforts aiming at the evaluation of the effects of hundreds of established cardiometabolic genetic variants on more than 20 respective phenotypes through single-phenotype summary GWAS results suggested that loci fall into multiple groups according to the alterations of correlated metabolic phenotypes. MetS is just one possible combination of effects and several other unexpected combinations might be observed, for example healthy obesity/unhealthy leanness, lower height/skeletal growth and higher total/HDL-cholesterol, high BMI/obesity and low HDL/blood pressure/glycaemic traits.

Genetic loci with similar cardiometabolic effects are involved in shared biological pathways

Pathway analysis revealed that some groups of loci with similar cardiometabolic effects are also enriched for factors that impact the same biological processes. These pathways may be expected - for example, regulation of lipids metabolism or cholesterol transport for groups of loci with strong effects on lipids, or circulatory system processes for genes near blood pressure-association signals - but sometimes also counterintuitive, as for example regulation of cellular processes for a group of loci with effects on obesity and anthropometric traits.

This enriched connectivity was particularly true for small groups of loci (around 10-20 members) and revealed potential candidate genes or tissues of action that are more likely for causality.

Many T2D loci are related to beta-cell function

The pathophysiological abnormalities observed in T2D patients include processes reflecting both insulin resistance (IR) and beta-cell function. For example, from the comparison that we reported in Scott et al. 2012[18], the insulin-raising allele was also associated with lower HDL and higher triglyceride levels; for some loci we also observed association with high levels of glucose, as well as of insulin, of β-cell functionality and insulin-resistance homeostasis, all hallmark combination in insulin-resistant individuals. On the other hand, we observed a group of loci implicated in insulin/proinsulin secretion and β-cell/pancreatic islets development which, if altered, cause an impaired production of insulin even if high levels of glucose are present in the blood, supporting the hypothesis that defects in the functionality of β-cells (rather than on insulin resistance), may lead to an hyperglycaemic status with consequent increased risk of developing T2D[19,99].

There is a causal relationship between adiposity and cardiometabolic phenotypes

Through the comparison of univariate GWAS meta-analysis results for multiple phenotypes and through multivariate analyses within the ENGAGE consortium, we investigated the effects of the *FTO* locus on many metabolic and cardiovascular phenotypes, and demonstrated that the association between *FTO* and cardiometabolic phenotypes is mediated by adiposity and, thus, that there is a causal effect of adiposity (measured by BMI) on other phenotypes. These results confirmed previous conclusions reported in the literature and obtained by using Mendelian randomisation approaches[89,90]. There could be many other loci with similar effects and at which dissection of their effects on multiple phenotypes is required; an example is *FADS1*, at which we observed strong effects on lipids and glycaemic traits and, after multivariate analysis, we concluded that multiple effects of this locus on cardiometabolic phenotypes are due to its independent effect on total cholesterol and triglycerides.

Many cardiometabolic phenotype associated variants constitute potential multi-phenotype allelic heterogeneity

Our results highlighted that a substantial proportion of metabolic phenotype loci incorporate complex patterns of potential multi-phenotype allelic heterogeneity. This result suggests that it is important to take into account this mechanism when evaluating cross-phenotype effects at genomic loci.

4.1.3 What remains uncovered, future directions for the study of pleiotropy and its applications

4.1.3.1 <u>Additional methods and fields to explore</u>

Our GWAS approaches, undertaken in the projects explained in this thesis, presented some limitations: (1) they poorly capture low frequency and rare variants, even if imputed data were used; (2) identified common variant signals have modest estimated effects on phenotypes and explain only a limited proportion of phenotypic variability, this can be partially due to the fact that more effects on other phenotypes remain uncovered; (3) identified cross-phenotype effects and the analysis of their underlying mechanisms remain to be confirmed with further analysis and through functional characterisation.
To overcome these limitations, several approaches can be adopted in the future.

Extending observations of CP effects to a wider range of phenotypes is an emerging area, for example. One of its next challenges lies in the development of robust meta-analytical approaches for data derived from multi-phenotype univariate and multivariate analyses, with special ramification focused on detection of low frequency and rare variants, such as collapsing tests[224,225] or aggregation methods[226].
Systematic and unbiased phenome-wide association studies (PheWASs) then, where a SNP with an established association with a phenotype is tested for association with hundreds of other phenotypes, are now underway[6]. An example is PAGE: The Population Architecture using Genomics and Epidemiology network[9].

As sequencing methods are becoming faster and cheaper, the field will move towards sequencing-based association studies. Through them, we will have the opportunity to directly identify the causal alleles underlying CP effects, and thus to distinguish between their different types more accurately. Sequencing will also allow us to better interrogate lower-frequency variants[6].

Functional characterisation of identified variants showing cross-phenotype effects (as explained in chapter “2.2.3.3_Functional characterisation”) and understanding the underlying mechanism remains a major challenge in the field.
Although many resources are available for characterising protein-coding variants, experiments in animal or cellular models are generally necessary to establish causality.
Moreover, new publicly available databases, such as the Encyclopedia of DNA Elements (ENCODE) project, provide valuable resources for characterising non protein-coding variants and regulatory elements[93].
In addition, examining eQTLs in relevant tissues for each phenotype of a cross-phenotype effect can help to elucidate the functional consequence and to distinguish between mediation and pleiotropy[6].

Finally, high-throughput “omics” data are rapidly becoming available with lowered costs and improvements in technology. Overall, omics data brings a promise of novel biomarker identification based on patterns of change in tissue DNA methylation, microRNAs, transcriptome, proteasome and metabolome. Defining ways for combining omics data with genetic data in relation to multiple phenotype effects may help better uncover complex mechanisms behind phenotypic variability.

4.1.3.2 Clinical implications of cross-phenotype effects and pleiotropy

Our research represents a new way of relating genetic variability to metabolic health, considering phenotypes as an organic network of complex interactions, rather than single phenomena, and it aims to contribute to a better understanding of dysmetabolism, with the definition of target groups of patients for the application of more specific therapies, with consequent reduction of adverse reactions and remarkable impact on patients' health and on public health costs for prevention and management of such conditions.

In this context we highlight the importance of pleiotropy in human quantitative traits and diseases and, more generally, of understanding cross-phenotype effects, which can provide insight into the mechanisms of shared physiology and pathophysiology.
A better clarification of pleiotropic phenomena will have several impacts on different field.
For example, from an evolutionary point of view, it will help the reconstruction of evolutionary processes that led to pleiotropy; its application in physiology will allow to discover models of regulation for different tissues and different periods of life, to shed light on the underlying cellular processes that are behind phenotypes, and to discover novel biological processes and new interactions between factors.

The idea of stratified medicine, through translational research techniques and understanding of the physiopathology of diseases at a molecular level, has unified researches from various fields in the development of new drugs and personalised therapies, based on genetic and epigenetic profile, gene expression and exposition to influencing factors. Research of pleiotropy will highly contribute to these efforts.
First, it may have clinically relevant implications for the classification (nosology) of medical disorders, and the goal of an aetiology-based classification may become more feasible.
The growing catalogue of genetic variants with pleiotropic effects will have important implications for genetic testing and personal genomics: clinicians and medical genetics professionals will take into account that genetic tests for one disease may reveal information for risks of other diseases. Moreover, distinguishing between cross-phenotype effects caused by single versus multiple independent causal variants can improve the accuracy of genetic tests and the interpretation of results[6].

Characterising the molecular mechanisms of cross-phenotype effects will undoubtedly expand our understanding of the underlying biology of complex diseases and will have clinical implications for drug discovery[6]: on one hand, drugs developed for one disorder could be repurposed to treat another disorder, if the therapeutic target is found to be common to the biology of both disorders;

on the other hand, new information about pleiotropy and mediation can be used for the development of new medicines followed by clinical trials and also for preventive measures, as the use of diagnostic biomarkers or new targets of action.

4.2 Main conclusion of my PhD experience

The 3-years programme of the PhD in Evolutionary and Environmental Biology conducted at the Department of Life Sciences and Biotechnology of the University of Ferrara highly contributed to my formation as a researcher in the field of human genetics and bioinformatics.

Of particular importance was my training at the Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford, where I started working with international large-scale genetic analyses and meta-analyses of quantitative metabolic traits/diseases.

During the PhD period, I significantly advanced my knowledge in programming languages, and in the use of programmes for large-scale genome-wide genetic analysis, dealing also with the newest analytical approaches and statistical techniques. I applied this knowledge on high scientific impact research projects, which led and will lead us to publications in important scientific journals[18,227].

During the PhD, I successfully applied for several grants (Italian 5x1000 funds for the research, European Foundation for the Study of Diabetes travel grant, ENGAGE Exchange and mobility program, funds for Internationalisations projects) that allowed me to create a strong collaborative network between my group at the University of Ferrara and other researchers in Europe; in particular, I established an active and productive connection with Doctor Inga Prokopenko and her group at the Imperial College of London, with Professor Andrew Morris and his group at the WTCHG of Oxford, and with Doctor Reedik Magi from the Estonian Genome Center, Tartu, Estonia. I also worked in collaboration with international consortia for the study of diabetes, metabolic phenotypes and their epidemiology (MAGIC, DIAGRAM, XC-Pleiotropy group, ENGAGE).

My junior leadership in pleiotropy projects within the XC-Pleiotropy group (Projects 1 and 2 of this thesis) allowed me to improve my capacity in leading and managing research, as well as my communicating and writing skills.

During the PhD, I had the possibility to participate in numerous advanced courses and workshops, as well as to attend international congresses in Europe, where I presented some of the described results. I was also involved in several academic efforts, such as tutor activities and students training.

In conclusion, the PhD experience gave me a solid background, fundamental for the continuation of my research projects and of my scientific career and for my education as independent researcher.

5 Appendix tables

SNP	CHR	Position HG18 (UCSC Build 36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs10923931	1	120319482	T2D	t	g	0.058	*NOTCH2*	Voight et al (Nature Genetics 2010)	
rs340874	1	212225879	T2D	c	t	0.542	*PROX1*	Dupuis et al (Nature Genetics 2010)	
rs780094	2	27594741	T2D	c	t	0.6	*GCKR*	Dupuis et al (Nature Genetics 2010)	
rs11899863	2	43472323	T2D	c	t	0.917	*THADA*	Voight et al (Nature Genetics 2010)	
rs7578597	2	43586327	T2D	t	c	0.908	*THADA*	Voight et al (Nature Genetics 2010)	
rs243088	2	60422249	T2D_metabochip_MEN	t	a	0.45	*BCL11A*	Morris et al (Nature Genetics 2012)	
rs243021	2	60438323	T2D	a	g	0.45	*BCL11A*	Voight et al (Nature Genetics 2010)	
rs7593730	2	160879700	T2D	c	t	0.825	*RBMS1*	Qi et al (Human Molecular Genetics 2010)	
rs3923113	2	165210095	T2D_metabochip_WOMEN	a	c	0.642	*GRB14*	Kooner et al (Nature Genetics 2011), Morris et al (Nature Genetics 2012)	South Asian
rs13389219	2	165237122	T2D_metabochip	c	t	0.6	*GRB14*	Morris et al (Nature Genetics 2012)	
rs7578326	2	226728897	T2D	a	g	0.675	*IRS1*	Voight et al (Nature Genetics 2010)	
rs2943641	2	226801989	T2D	c	t	0.667	*IRS1*	Voight et al (Nature Genetics 2010)	
rs13081389	3	12264800	T2D	a	g	0.967	*PPARG*	Voight et al (Nature Genetics 2010)	
rs1801282	3	12368125	T2D	c	g	0.908	*PPARG*	Voight et al (Nature Genetics 2010)	
rs7612463	3	23311454	T2D	c	a	0.933	*UBE2E2*	Yamauchi et al (Nature Genetics 2010)	Japaneese
rs831571	3	64023337	T2D	c	t	0.758	*PSMD6*	Cho et al (Nature Genetics 2012)	East Asian
rs6795735	3	64680405	T2D	c	t	0.517	*ADAMTS9*	Voight et al (Nature Genetics 2010)	
rs4607103	3	64686944	T2D	c	t	0.8	*ADAMTS9*	Voight et al (Nature Genetics 2010)	
rs11708067	3	124548468	T2D	a	g	0.8	*ADCY5*	Dupuis et al (Nature Genetics 2010)	
rs1470579	3	187011774	T2D	c	a	0.275	*IGF2BP2*	Voight et al (Nature Genetics 2010)	
rs16861329	3	188149155	T2D	c	t	0.883	*ST6GAL1*	Kooner et al (Nature Genetics 2011)	South Asian
rs10010131	4	6343816	T2D	g	a	0.667	*WFS1*	Voight et al (Nature Genetics 2010)	
rs1801214	4	6353923	T2D	t	c	0.667	*WFS1*	Voight et al (Nature Genetics 2010)	
rs459193	5	55842508	T2D_metabochip	g	a	0.75	*ANKRD55*	Morris et al (Nature Genetics 2012)	
rs4457053	5	76460705	T2D	g	a	0.317	*ZBED3*	Voight et al (Nature Genetics 2010)	
rs7754840	6	20769229	T2D	c	g	0.3	*CDKAL1*	Voight et al (Nature Genetics 2010)	
rs10440833	6	20796100	T2D	a	t	0.25	*CDKAL1*	Voight et al (Nature Genetics 2010)	
rs9470794	6	38214822	T2D	c	t	0.108	*ZFAND3*	Cho et al (Nature Genetics 2012)	East Asian
rs1535500	6	39392028	T2D	t	g	0.5	*KCNK16*	Cho et al (Nature Genetics 2012)	East Asian
rs17168486	7	14864807	T2D_metabochip_MEN	t	c	0.142	*DGKB*	Morris et al (Nature Genetics 2012)	
rs6960043	7	15019385	T2D_metabochip_putative_2ndary	c	t	0.508	*DGKB*	Morris et al (Nature Genetics 2012)	
rs2191349	7	15030834	T2D	t	g	0.558	*DGKB*	Dupuis et al (Nature Genetics 2010)	
rs849134	7	28162747	T2D	a	g	0.508	*JAZF1*	Voight et al (Nature Genetics 2010)	
rs4607517	7	44202193	T2D	a	g	0.217	*GCK*	Dupuis et al (Nature Genetics 2010)	
rs6467136	7	126952194	T2D	g	a	0.508	*GCC1/PAXA4*	Cho et al (Nature Genetics 2012)	East Asian
rs972283	7	130117394	T2D	g	a	0.542	*KLF14*	Voight et al (Nature Genetics 2010)	
rs516946	8	41638405	T2D_metabochip	c	t	0.8	*ANK1*	Morris et al (Nature Genetics 2012)	
rs896854	8	96029687	T2D	t	c	0.492	*TP53INP1*	Voight et al (Nature Genetics 2010)	
rs13266634	8	118253964	T2D	c	t	0.717	*SLC30A8*	Voight et al (Nature Genetics 2010)	
rs3802177	8	118254206	T2D	g	a	0.717	*SLC30A8*	Voight et al (Nature Genetics 2010)	
rs7041847	9	4277466	T2D	a	g	0.542	*GLIS3*	Cho et al (Nature Genetics 2012)	East Asian
rs17584499	9	8869118	T2D	t	c	0.225	*PTPRD*	Tsai et al (Plos Genetics 2010)	Chinese
rs944801	9	22041670	T2D_metabochip_putative_2ndary	c	g	0.575	*CDKN2A/B*	Morris et al (Nature Genetics 2012)	
rs10965250	9	22123284	T2D	g	a	0.758	*CDKN2A/B*	Voight et al (Nature Genetics 2010)	
rs10811661	9	22124094	T2D	t	c	0.742	*CDKN2A/B*	Voight et al (Nature Genetics 2010), Morris et al (Nature Genetics 2012)	
rs13292136	9	81141948	T2D	c	t	0.942	*CHCHD9/TLE4*	Voight et al (Nature Genetics 2010)	
rs2796441	9	83498768	T2D_metabochip	g	a	0.617	*TLE1*	Morris et al (Nature Genetics 2012)	
rs12779790	10	12368016	T2D	g	a	0.225	*CDC123/CAMK1D*	Voight et al (Nature Genetics 2010)	
rs1802295	10	70601480	T2D	t	c	0.342	*VPS26A*	Kooner et al (Nature Genetics 2011)	South Asian
rs12571751	10	80612637	T2D_metabochip	a	g	0.558	*ZMIZ1*	Morris et al (Nature Genetics 2012)	
rs1111875	10	94452862	T2D	c	t	0.592	*HHEX/IDE*	Voight et al (Nature Genetics 2010)	
rs5015480	10	94455539	T2D	c	t	0.583	*HHEX/IDE*	Voight et al (Nature Genetics 2010)	
rs7903146	10	114748339	T2D	t	c	0.308	*TCF7L2*	Voight et al (Nature Genetics 2010), Grant et al (Nature Genetics 2006)	
rs2334499	11	1653425	T2D	t	c	0.417	*DUSP8*	Kong et al (Nature 2009)	
rs231362	11	2648047	T2D	g	a	0.458	*KCNQ1*	Voight et al (Nature Genetics 2010)	
rs231361	11	2648076	T2D_metabochip_putative_2ndary	a	g	0.233	*KCNQ1*	Morris et al (Nature Genetics 2012)	
rs163184	11	2803645	T2D/T2D_metabochip_MEN	g	t	0.483	*KCNQ1*	Voight et al (Nature Genetics 2010), Morris et al (Nature Genetics 2012)	
rs5215	11	17365206	T2D	c	t	0.433	*KCNJ11*	Voight et al (Nature Genetics 2010)	
rs1552224	11	72110746	T2D	a	c	0.867	*ARAP1/CENTD2*	Voight et al (Nature Genetics 2010)	
rs1387153	11	92313476	T2D	t	c	0.233	*MTNR1B*	Voight et al (Nature Genetics 2010)	
rs10830963	11	92348358	T2D	g	c	0.217	*MTNR1B*	Voight et al (Nature Genetics 2010)	
rs11063069	12	4244634	T2D_metabochip- MEN	g	a	0.25	*CCND2*	Morris et al (Nature Genetics 2012)	
rs10842994	12	27856417	T2D_metabochip	c	t	0.833	*KLHDC5*	Morris et al (Nature Genetics 2012)	
rs1531343	12	64461161	T2D	c	g	0.1	*HMGA2*	Voight et al (Nature Genetics 2010)	
rs4760790	12	69921061	T2D	a	g	0.258	*TSPAN8/LGR5*	Voight et al (Nature Genetics 2010)	
rs7957197	12	119945069	T2D	t	a	0.85	*HNF1A/TCF1*	Voight et al (Nature Genetics 2010)	
rs1359790	13	79615157	T2D	g	a	0.733	*SPRY2*	Shu et al (Plos Genetics 2010)	
rs7163757	15	60178900	T2D	c	t	0.542	*C2CD4A*	Yamauchi et al (Nature Genetics 2010)	Japaneese
rs7178572	15	75534245	T2D	g	a	0.683	*HMG20A*	Kooner et al (Nature Genetics 2011)	South Asian
rs7177055	15	75619817	T2D_metabochip	a	g	0.708	*HMG20A*	Morris et al (Nature Genetics 2012)	
rs11634397	15	78219277	T2D	g	a	0.617	*ZFAND6*	Voight et al (Nature Genetics 2010)	
rs2028299	15	88175261	T2D	c	a	0.283	*AP3S2*	Kooner et al (Nature Genetics 2011)	South Asian
rs8042680	15	89322341	T2D	a	c	0.242	*PRC1*	Voight et al (Nature Genetics 2010)	
rs11642841	16	52402988	T2D	a	c	0.458	*FTO*	Voight et al (Nature Genetics 2010)	
rs7202877	16	73804746	T2D_metabochip	t	g	0.908	*BCAR1*	Morris et al (Nature Genetics 2012)	
rs391300	17	2163008	T2D	c	t	0.692	*SRR*	Tsai et al (Plos Genetics 2010)	Chinese
rs4430796	17	33172153	T2D	g	a	0.492	*HNF1B/TCF2*	Voight et al (Nature Genetics 2010)	
rs12970134	18	56035730	T2D_metabochip	a	g	0.275	*MC4R*	Morris et al (Nature Genetics 2012)	
rs11873305	18	56200172	T2D_metabochip_putative_2ndary	a	c	0.987	*MC4R*	Morris et al (Nature Genetics 2012)	
rs10401969	19	19268718	T2D_metabochip	c	t	0.092	*CILP2*	Morris et al (Nature Genetics 2012)	
rs3786897	19	38584848	T2D	a	g	0.608	*PEPD*	Cho et al (Nature Genetics 2012)	East Asian
rs8108269	19	50850353	T2D_metabochip_WOMEN	g	t	0.242	*GIPR*	Morris et al (Nature Genetics 2012)	
rs6017317	20	42380380	T2D	g	t	0.2	*FITM2/R3HDML/HNF4A*	Cho et al (Nature Genetics 2012)	East Asian
rs4812829	20	42422681	T2D	a	g	0.2	*HNF4A*	Kooner et al (Nature Genetics 2011)	South Asian
rs5945326	23	152553116	T2D	a	g	0.767	*DUSP9*	Voight et al (Nature Genetics 2010)	

Appendix table 1: *T2D genome-wide significant (p-value < $5x10^{-8}$) SNPs reported from published GWAS (before October 2012). PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs9727115	1	98949841	Fasting Pro-insulin_adjFG	g	a	0.575	*SNX7*	Strawbridge et al (Diabetes 2011)	
rs2779116	1	156852039	HbA1C	t	c	0.283	*SPTA1*	Soranzo et al (Diabetes 2010)	
rs340874	1	212225879	FGlu	c	t	0.542	*PROX1*	Dupuis et al (Nature Genetics 2010)	
rs2820436	1	217707303	FIns	c	a	0.658	*LYPLAL1*	Scott et al (Nature Genetics 2012)	
rs2785980	1	217767142	FIns	t	c	0.65	*LYPLAL1*	Manning et al (Nature Genetics 2012)	
rs4846565	1	217788727	FInsadjBMI	g	a	0.667	*LYPLAL1*	Scott et al (Nature Genetics 2012)	
rs1371614	2	27006378	FGlu	t	c	0.275	*DPYSL5*	Manning et al (Nature Genetics 2012)	
rs1260326	2	27584444	2hGlu	t	c	0.417	*GCKR*	Saxena et al (Nature Genetics 2010)	
rs780094	2	27594741	FGlu/FIns	c	t	0.6	*GCKR*	Dupuis et al (Nature Genetics 2010)	
rs1530559	2	135472099	FIns	a	g	0.608	*YSK4*	Scott et al (Nature Genetics 2012)	
rs10195252	2	165221337	FIns/FinsadjBMI	t	c	0.583	*COBLL1/GRB14*	Scott et al (Nature Genetics 2012)	
rs7607980	2	165259447	FIns	t	c	0.867	*COBLL1/GRB14*	Manning et al (Nature Genetics 2012)	
rs560887	2	169471394	FGlu	c	t	0.692	*G6PC2*	Dupuis et al (Nature Genetics 2010)	
rs552976	2	169499684	HbA1C	g	a	0.642	*G6PC2*	Soranzo et al (Diabetes 2010)	
rs2943634	2	226776324	FIns	c	a	0.683	*IRS1*	Manning et al (Nature Genetics 2012)	
rs2943645	2	226807424	FIns	t	c	0.658	*IRS1*	Scott et al (Nature Genetics 2012)	
rs2972143	2	226824609	FIns	g	a	0.667	*IRS1*	Scott et al (Nature Genetics 2012)	
rs17036328	3	12365484	FInsadjBMI	t	c	0.908	*PPARG*	Scott et al (Nature Genetics 2012)	
rs11715915	3	49430334	FGlu	c	t	0.775	*AMT*	Scott et al (Nature Genetics 2012)	
rs11708067	3	124548468	FGlu	a	g	0.8	*ADCY5*	Dupuis et al (Nature Genetics 2010)	
rs11717195	3	124565088	2hGlu	t	c	0.8	*ADCY5*	Saxena et al (Nature Genetics 2010)	
rs11920090	3	172200215	FGlu	t	a	0.867	*SLC2A2*	Dupuis et al (Nature Genetics 2010)	
rs7651090	3	186996086	FGlu/2hGlu	g	a	0.275	*IGF2BP2*	Scott et al (Nature Genetics 2012)	
rs3822072	4	89960292	FInsadjBMI	a	g	0.458	*FAM13A*	Scott et al (Nature Genetics 2012)	
rs974801	4	106290513	FInsadjBMI	g	a	0.4	*TET2*	Scott et al (Nature Genetics 2012)	
rs9884482	4	106301085	FIns	c	t	0.4	*TET2*	Scott et al (Nature Genetics 2012)	
rs4691380	4	157939574	FIns	c	t	0.592	*PDGFC*	Manning et al (Nature Genetics 2012)	
rs6822892	4	157954125	FInsadjBMI	a	g	0.583	*PDGFC*	Scott et al (Nature Genetics 2012)	
rs4865796	5	53308421	Fins/FInsadjBMI	a	g	0.717	*ARL15*	Scott et al (Nature Genetics 2012)	
rs459193	5	55842508	FInsadjBMI	g	a	0.75	*ANKRD55/MAP3K1*	Scott et al (Nature Genetics 2012)	
rs7708285	5	76461623	FGluadjBMI	g	a	0.333	*ZBED3*	Scott et al (Nature Genetics 2012)	
rs4869272	5	95565204	FGlu	t	c	0.675	*PCSK1*	Scott et al (Nature Genetics 2012)	
rs13179048	5	95568482	FGlu	c	a	0.675	*PCSK1*	Manning et al (Nature Genetics 2012)	
rs6235	5	95754654	Fasting Pro-insulin	g	c	0.317	*PCSK1*	Strawbridge et al (Diabetes 2011)	
rs1019503	5	96280573	2hGlu	a	g	0.55	*ERAP2*	Scott et al (Nature Genetics 2012)	
rs17762454	6	7158199	FGlu	t	c	0.225	*RREB1*	Scott et al (Nature Genetics 2012)	
rs9368222	6	20794975	FGlu	a	c	0.25	*CDKAL1*	Scott et al (Nature Genetics 2012)	
rs1800562	6	26201120	HbA1C	g	a	0.967	*HFE*	Soranzo et al (Diabetes 2010)	
rs6912327	6	34872900	FInsadjBMI	t	c	0.75	*C6orf107/UHRF1BP1*	Scott et al (Nature Genetics 2012)	
rs4646949	6	34953427	FIns	t	g	0.733	*UHRF1BP1*	Manning et al (Nature Genetics 2012)	
rs2745353	6	127494628	Fins	t	c	0.542	*RSPO3*	Scott et al (Nature Genetics 2012)	
rs2191349	7	15030834	FGlu	t	g	0.558	*DGKB/TMEM195*	Dupuis et al (Nature Genetics 2010)	
rs1799884	7	44195593	HbA1C	t	c	0.217	*GCK*	Soranzo et al (Diabetes 2010)	
rs6975024	7	44198411	2hGlu	c	t	0.217	*GCK*	Scott et al (Nature Genetics 2012)	
rs4607517	7	44202193	FGlu	a	g	0.217	*GCK*	Dupuis et al (Nature Genetics 2010)	
rs6943153	7	50759073	FGlu	t	c	0.258	*GRB10*	Scott et al (Nature Genetics 2012)	
rs1167800	7	75014132	FIns	a	g	0.533	*HIP1*	Scott et al (Nature Genetics 2012)	
rs983309	8	9215142	Fglu	t	g	0.108	*PPP1R3B*	Scott et al (Nature Genetics 2012)	
rs983309	8	9215142	FIns	t	g	0.108	*PPP1R3B*	Scott et al (Nature Genetics 2012)	
rs4841132	8	9221006	FIns	a	g	0.075	*PPP1R3B*	Manning et al (Nature Genetics 2012)	
rs4841132	8	9221006	FGlu	a	g	0.075	*PPP1R3B*	Manning et al (Nature Genetics 2012)	
rs2126259	8	9222556	FInsadjBMI	t	c	0.092	*PPP1R3B*	Scott et al (Nature Genetics 2012)	
rs11782386	8	9239197	2hGlu	c	t	0.883	*PPP1R3B*	Scott et al (Nature Genetics 2012)	
rs6474359	8	41668351	HbA1C	t	c	0.975	*ANK1*	Soranzo et al (Diabetes 2010)	
rs4737009	8	41749562	HbA1C	a	g	0.267	*ANK1*	Soranzo et al (Diabetes 2010)	
rs11558471	8	118254914	FGlu	a	g	0.708	*SLC30A8*	Dupuis et al (Nature Genetics 2010)	
rs11558471	8	118254914	Fasting Pro-insulin	a	g	0.708	*SLC30A8*	Strawbridge et al (Diabetes 2011)	
rs7034200	9	4279050	FGlu	a	c	0.542	*GLIS3*	Dupuis et al (Nature Genetics 2010)	
rs10811661	9	22124094	FGlu	t	c	0.742	*CDKN2B*	Scott et al (Nature Genetics 2012)	
rs16913693	9	110720180	FGlu	t	g	0.983	*IKBKAP*	Scott et al (Nature Genetics 2012)	
rs306549	9	134459997	Fasting Pro-insulin_WOMEN	c	g	0.292	*DDX31*	Strawbridge et al (Diabetes 2011)	
rs3829109	9	138376587	FGlu	g	a	0.625	*DNLZ*	Scott et al (Nature Genetics 2012)	
rs16926246	10	70763398	HbA1C	c	t	0.9	*HK1*	Soranzo et al (Diabetes 2010)	
rs10885122	10	113032083	FGlu	g	t	0.875	*ADRA2A*	Dupuis et al (Nature Genetics 2010)	
rs4506565	10	114746031	FGlu	t	a	0.333	*TCF7L2*	Dupuis et al (Nature Genetics 2010)	
rs7903146	10	114748339	FIns	c	t	0.692	*TCF7L2*	Strawbridge et al. (Diabetes, 2011), Scott et al (Nature Genetics 2012)	
rs12243326	10	114778805	2hGlu	c	t	0.267	*TCF7L2*	Saxena et al (Nature Genetics 2010)	
rs11605924	11	45829667	FGlu	a	c	0.533	*CRY2*	Dupuis et al (Nature Genetics 2010)	
rs10501320	11	47250375	Fasting Pro-insulin	g	c	0.742	*MADD*	Strawbridge et al (Diabetes 2011)	
rs10838687	11	47269468	Fasting Pro-insulin	t	g	0.858	*MADD*	Strawbridge et al (Diabetes 2011)	
rs7944584	11	47292896	FGlu	a	t	0.725	*MADD*	Dupuis et al (Nature Genetics 2010)	
rs1483121	11	48289936	FGlu	g	a	0.858	*OR4S1*	Manning et al (Nature Genetics 2012)	
rs174550	11	61328054	FGlu	t	c	0.625	*FADS1*	Dupuis et al (Nature Genetics 2010)	
rs11603334	11	72110633	FGlu	g	a	0.867	*ARAP1*	Scott et al (Nature Genetics 2012), Manning et al (Nature Genetics 2012)	
rs11603334	11	72110633	Fasting Pro-insulin	a	g	0.133	*ARAP1*	Strawbridge et al (Diabetes 2011)	
rs1387153	11	92313476	HbA1C	t	c	0.233	*MTNR1B*	Soranzo et al (Diabetes 2010)	
rs10830963	11	92348358	FGlu	g	c	0.217	*MTNR1B*	Dupuis et al (Nature Genetics 2010)	
rs2657879	12	55151605	FgluadjBMI	g	a	0.217	*GLS2*	Scott et al (Nature Genetics 2012)	
rs35767	12	101399699	FIns	g	a	0.9	*IGF1*	Dupuis et al (Nature Genetics 2010)	
rs10747083	12	131551691	FGlu	a	g	0.708	*P2RX2*	Scott et al (Nature Genetics 2012)	
rs11619319	13	27385599	FGlu	g	a	0.242	*PDX1*	Scott et al (Nature Genetics 2012)	
rs2293941	13	27389198	Fasting Pro-insulin	a	g	0.242	*PDX1*	Manning et al (Nature Genetics 2012)	
rs576674	13	32452302	FGlu	g	a	0.1	*KL*	Scott et al (Nature Genetics 2012)	
rs7998202	13	112379869	HbA1C	g	a	0.175	*ATP11A/TUBGCP3*	Soranzo et al (Diabetes 2010)	
rs3783347	14	99909014	FGlu	g	t	0.775	*WARS*	Scott et al (Nature Genetics 2012)	
rs17271305	15	60120272	2hGlu	g	a	0.425	*FAM148B/VPS13C/C2CD4A/B*	Saxena et al (Nature Genetics 2010)	
rs4502156	15	60170447	Fasting Pro-insulin	t	c	0.542	*FAM148B/VPS13C/C2CD4A/B*	Strawbridge et al (Diabetes 2011)	
rs11071657	15	60221254	FGlu	a	g	0.608	*FAM148B/VPS13C/C2CD4A/B*	Dupuis et al (Nature Genetics 2010)	
rs1549318	15	68896201	Fasting Pro-insulin	t	c	0.575	*LARP6*	Strawbridge et al (Diabetes 2011)	
rs1421085	16	52358455	FIns	c	t	0.458	*FTO*	Scott et al (Nature Genetics 2012)	
rs4790333	17	2209453	Fasting Pro-insulin	t	c	0.433	*SGSM2*	Strawbridge et al (Diabetes 2011)	
rs1046896	17	78278822	HbA1C	t	c	0.292	*FN3K*	Soranzo et al (Diabetes 2010)	
rs731839	19	38590905	FIns/FinsadjBMI	g	a	0.3	*PEPD*	Scott et al (Nature Genetics 2012)	
rs10423928	19	50874144	2hGlu	a	t	0.175	*GIPR*	Saxena et al (Nature Genetics 2010)	
rs2302593	19	50888474	FGlu	c	g	0.525	*GIPR*	Scott et al (Nature Genetics 2012)	
rs6113722	20	22505099	FGlu	g	a	0.942	*FOXA2*	Scott et al (Nature Genetics 2012)	
rs6048205	20	22507601	FGlu	a	g	0.925	*FOXA2*	Manning et al (Nature Genetics 2012)	
rs6072275	20	39177319	FGlu	a	g	0.158	*TOP1*	Scott et al (Nature Genetics 2012)	
rs855791	22	35792882	HbA1C	a	g	0.4	*TMPRSS6*	Soranzo et al (Diabetes 2010)	

Appendix table 2: *Glycaemic G-W significant SNPs reported from published GWAS (before October 2012). PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs2815752	1	72585028	BMI	a	g	0.675	*NEGR1*	Speliotes et al (Nature Genetics 2010)	
rs1514175	1	74764232	BMI	a	g	0.408	*TNNI3K*	Speliotes et al (Nature Genetics 2010)	
rs1555543	1	96717385	BMI	c	a	0.567	*PTBP2*	Speliotes et al (Nature Genetics 2010)	
rs984222	1	119305366	WHRadjBMI	g	c	0.567	*TBX15/WARS2*	Heid et al (Nature Genetics 2010)	
rs1011731	1	170613171	WHRadjBMI	g	a	0.425	*DNM3/PIGC*	Heid et al (Nature Genetics 2010)	
rs543874	1	176156103	BMI	g	a	0.275	*SEC16B*	Speliotes et al (Nature Genetics 2010)	
rs2605100	1	217710847	WHR_WOMEN	g	a	0.675	*LYPLAL1*	Lindgren et al (PLoS Genetics 2009)	
rs4846567	1	217817340	WHRadjBMI	g	t	0.7	*LYPLAL1*	Heid et al (Nature Genetics 2010)	
rs2867125	2	612827	BMI	c	t	0.858	*TMEM18*	Speliotes et al (Nature Genetics 2010)	
rs713586	2	25011512	BMI	c	t	0.508	*RBJ*	Speliotes et al (Nature Genetics 2010)	
rs887912	2	59156381	BMI	t	c	0.325	*FANCL*	Speliotes et al (Nature Genetics 2010)	
rs2890652	2	142676401	BMI	c	t	0.158	*LRP1B*	Speliotes et al (Nature Genetics 2010)	
rs10195252	2	165221337	WHRadjBMI	t	c	0.583	*GRB14*	Heid et al (Nature Genetics 2010)	
rs2943650	2	227000000	PCBFAT	c	t	0.333	*nearIRS1*	Kilpeläinen TO et al (Nat Gen 2011)	*
rs6784615	3	52481466	WHRadjBMI	t	c	0.975	*NISCH/STAB1*	Heid et al (Nature Genetics 2010)	
rs6795735	3	64680405	WHRadjBMI	c	t	0.517	*ADAMTS9*	Heid et al (Nature Genetics 2010)	
rs13078807	3	85966840	BMI	g	a	0.225	*CADM2*	Speliotes et al (Nature Genetics 2010)	
rs9816226	3	187317193	BMI	t	a	0.842	*ETV5*	Speliotes et al (Nature Genetics 2010)	
rs10938397	4	44877284	BMI	g	a	0.45	*GNPDA2*	Speliotes et al (Nature Genetics 2010)	
rs13107325	4	103407732	BMI	t	c	0.092	*SLC39A8*	Speliotes et al (Nature Genetics 2010)	
rs2112347	5	75050998	BMI	t	g	0.675	*FLJ35779*	Speliotes et al (Nature Genetics 2010)	
rs261967	5	95876006	BMI	c	a	0.392	*PCSK1*	Wen et al (Nature Genetics 2012)	East Asian*
rs4836133	5	124360002	BMI	a	c	0.517	*ZNF608*	Speliotes et al (Nature Genetics 2010)	
rs6861681	5	173295064	WHRadjBMI	a	g	0.325	*CPEB4*	Heid et al (Nature Genetics 2010)	
rs1294421	6	6688148	WHRadjBMI	g	t	0.592	*LY86*	Heid et al (Nature Genetics 2010)	
rs9356744	6	20793465	BMI	t	c	0.717	*CDKAL1*	Wen et al (Nature Genetics 2012)	East Asian*
rs206936	6	34410847	BMI	g	a	0.2	*NUDT3*	Speliotes et al (Nature Genetics 2010)	
rs6905288	6	43866851	WHRadjBMI	a	g	0.592	*VEGFA*	Heid et al (Nature Genetics 2010)	
rs987237	6	50911009	WC	g	a	0.083	*TFAP2B*	Lindgren et al (PLoS Genetics 2009)	
rs987237	6	50911009	BMI	g	a	0.083	*TFAP2B*	Speliotes et al (Nature Genetics 2010)	
rs9491696	6	127494332	WHRadjBMI	g	c	0.533	*RSPO3*	Heid et al (Nature Genetics 2010)	
rs1055144	7	25837634	WHRadjBMI	t	c	0.158	*NFE2L3*	Heid et al (Nature Genetics 2010)	
rs545854	8	9897490	WC	g	c	0.183	*MSRA*	Lindgren et al (PLoS Genetics 2009)	
rs10968576	9	28404339	BMI	g	a	0.358	*LRRN6C*	Speliotes et al (Nature Genetics 2010)	
rs11142387	9	72188152	BMI	c	a	0.55	*KLF9*	Okada et al (Nature Genetics 2012)	East Asian*
rs4929949	11	8561169	BMI	c	t	0.592	*RPL27A*	Speliotes et al (Nature Genetics 2010)	
rs10767664	11	27682562	BMI	a	t	0.758	*BDNF*	Speliotes et al (Nature Genetics 2010)	
rs3817334	11	47607569	BMI	t	c	0.408	*MTCH2*	Speliotes et al (Nature Genetics 2010)	
rs718314	12	26344550	WHRadjBMI	g	a	0.25	*ITPR2/SSPN*	Heid et al (Nature Genetics 2010)	
rs7138803	12	48533735	BMI	a	g	0.367	*FAIM2*	Speliotes et al (Nature Genetics 2010)	
rs1443512	12	52628951	WHRadjBMI	a	c	0.183	*HOXC13*	Heid et al (Nature Genetics 2010)	
rs4771122	13	26918180	BMI	g	a	0.267	*MTIF3*	Speliotes et al (Nature Genetics 2010)	
rs534870	13	79857208	PCBFAT	g	a	0.283	*nearSPRY2*	Kilpeläinen TO et al (Nat Gen 2011)	*
rs11847697	14	29584863	BMI	t	c	0.033	*PRKD1*	Speliotes et al (Nature Genetics 2010)	
rs10150332	14	79006717	BMI	c	t	0.267	*NRXN3*	Speliotes et al (Nature Genetics 2010)	
rs10146997	14	79014915	WC	g	a	0.267	*NRXN3*	Heard-Costa et al (PLoS Genetics 2009)	
rs2241423	15	65873892	BMI	g	a	0.808	*MAP2K5*	Speliotes et al (Nature Genetics 2010)	
rs12444979	16	19841101	BMI	c	t	0.85	*GPRC5B*	Speliotes et al (Nature Genetics 2010)	
rs12597579	16	20165368	BMI	c	t	0.908	*GP2*	Wen et al (Nature Genetics 2012)	East Asian*
rs7359397	16	28793160	BMI	t	c	0.367	*SH2B1*	Speliotes et al (Nature Genetics 2010)	
rs1558902	16	52361075	BMI	a	t	0.458	*FTO*	Speliotes et al (Nature Genetics 2010)/Heard-Costa et al (PLoS Genetics 2009)	
rs8050136	16	52373776	PCBFAT	a	c	0.45	*FTO*	Kilpeläinen TO et al (Nat Gen 2011)	*
rs571312	18	55990749	BMI	a	c	0.242	*MC4R*	Speliotes et al (Nature Genetics 2010)	
rs489693	18	56033767	WC	a	c	0.625	*MC4R*	Heard-Costa et al (PLoS Genetics 2009)	
rs12970134	18	56035730	WC	a	g	0.275	*MC4R*	Chambers et al (Nature Genetics 2008)	
rs29941	19	39001372	BMI	g	a	0.675	*KCTD15*	Speliotes et al (Nature Genetics 2010)	
rs2287019	19	50894012	BMI	c	t	0.875	*QPCTL*	Speliotes et al (Nature Genetics 2010)	
rs3810291	19	52260843	BMI	a	g	0.658	*TMEM160*	Speliotes et al (Nature Genetics 2010)	
rs4823006	22	27781671	WHRadjBMI	a	g	0.525	*ZNRF3-KREMEN1*	Heid et al (Nature Genetics 2010)	

*Not included for project 1, included just for project 3

Appendix table 3: *Obesity/anthropometrics G-W significant (p-value < $5x10^{-8}$) SNPs reported from published GWAS (before October 2012). PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs425277	1	2059032	Height	t	c	0.267	*PRKCZ*	Lango Allen et al (Nature 2010)	
rs2284746	1	17179262	Height	g	c	0.583	*MFAP2*	Lango Allen et al (Nature 2010)	
rs1738475	1	23409478	Height	c	g	0.675	*HTR1D*	Lango Allen et al (Nature 2010)	
rs4601530	1	24916698	Height	c	t	0.733	*CLIC4*	Lango Allen et al (Nature 2010)	
rs7532866	1	26614131	Height	a	g	0.7	*LIN28*	Lango Allen et al (Nature 2010)	
rs2154319	1	41518357	Height	c	t	0.175	*SCMH1*	Lango Allen et al (Nature 2010)	
rs17391694	1	78396214	Height	t	c	0.125	*GIPC2*	Lango Allen et al (Nature 2010)	
rs6699417	1	88896031	Height	t	c	0.608	*PKN2*	Lango Allen et al (Nature 2010)	
rs10874746	1	93096559	Height	c	t	0.633	*RPL5*	Lango Allen et al (Nature 2010)	
rs9428104	1	118657110	Height	g	a	0.717	*SPAG17*	Lango Allen et al (Nature 2010)	
rs11205277	1	148159496	Height	g	a	0.417	*SF3B4*	Lango Allen et al (Nature 2010)	
rs17346452	1	170319910	Height	c	t	0.183	*DNM3*	Lango Allen et al (Nature 2010)	
rs2421992	1	170507874	Height_2ndary	t	c	0.733	*DNM3*	Lango Allen et al (Nature 2010)	
rs1325598	1	175058872	Height	g	a	0.55	*PAPPA2*	Lango Allen et al (Nature 2010)	
rs1046934	1	182290152	Height	c	a	0.333	*TSEN15*	Lango Allen et al (Nature 2010)	
rs10863936	1	210304421	Height	g	a	0.5	*DTL*	Lango Allen et al (Nature 2010)	
rs6684205	1	216676325	Height	g	a	0.225	*TGFB2*	Lango Allen et al (Nature 2010)	
rs11118346	1	217810342	Height	c	t	0.525	*LYPLAL1*	Lango Allen et al (Nature 2010)	
rs10799445	1	225978506	Height	a	c	0.725	*JMJD4*	Lango Allen et al (Nature 2010)	
rs4665736	2	25041103	Height	t	c	0.442	*DNAJC27*	Lango Allen et al (Nature 2010)	
rs6714546	2	33214929	Height	g	a	0.717	*LTBP1*	Lango Allen et al (Nature 2010)	
rs17511102	2	37814117	Height	t	a	0.017	*CDC42EP3*	Lango Allen et al (Nature 2010)	
rs2341459	2	44621706	Height	t	c	0.258	*C2orf34*	Lango Allen et al (Nature 2010)	
rs12474201	2	46774789	Height	a	g	0.325	*SOCS5*	Lango Allen et al (Nature 2010)	
rs1367226	2	55943044	Height_2ndary	g	a	0.608	*EFEMP1*	Lango Allen et al (Nature 2010)	
rs3791675	2	55964813	Height	c	t	0.75	*EFEMP1*	Lango Allen et al (Nature 2010)	
rs11684404	2	88705737	Height	c	t	0.308	*EIF2AK3*	Lango Allen et al (Nature 2010)	
rs7567288	2	134151294	Height	c	t	0.183	*NCKAP5*	Lango Allen et al (Nature 2010)	
rs7567851	2	178392966	Height	c	g	0.042	*PDE11A*	Lango Allen et al (Nature 2010)	
rs1351164	2	217980143	Height	t	c	0.808	*TNS1*	Lango Allen et al (Nature 2010)	
rs10187066	2	219223003	Height_2ndary	g	a	0.675	*CCDC108/IHH*	Lango Allen et al (Nature 2010)	
rs12470505	2	219616613	Height	t	g	0.875	*CCDC108/IHH*	Lango Allen et al (Nature 2010)	
rs2629046	2	224755988	Height	t	c	0.517	*SERPINE2*	Lango Allen et al (Nature 2010)	
rs2580816	2	232506210	Height	c	t	0.742	*NPPC*	Lango Allen et al (Nature 2010)	
rs12694997	2	241911659	Height	g	a	0.683	*SEPT2*	Lango Allen et al (Nature 2010)	
rs2597513	3	13530836	Height	c	t	0.133	*HDAC11*	Lango Allen et al (Nature 2010)	
rs13088462	3	51046753	Height	c	t	0.033	*DOCK3*	Lango Allen et al (Nature 2010)	
rs2336725	3	53093779	Height	c	t	0.442	*RFT1*	Lango Allen et al (Nature 2010)	
rs9835332	3	56642722	Height	g	c	0.55	*C3orf63*	Lango Allen et al (Nature 2010)	
rs17806888	3	67499012	Height	t	c	0.908	*SUCLG2*	Lango Allen et al (Nature 2010)	
rs9863706	3	72520103	Height	c	t	0.75	*RYBP*	Lango Allen et al (Nature 2010)	
rs6439167	3	130533446	Height	c	t	0.758	*C3orf37*	Lango Allen et al (Nature 2010)	
rs9844666	3	137456906	Height	g	a	0.808	*PCCB*	Lango Allen et al (Nature 2010)	
rs724016	3	142588260	Height	g	a	0.433	*ZBTB38*	Lango Allen et al (Nature 2010)	
rs7652177	3	173451771	Height_2ndary	g	c	0.517	*GHSR*	Lango Allen et al (Nature 2010)	
rs572169	3	173648421	Height	t	c	0.3	*GHSR*	Lango Allen et al (Nature 2010)	
rs720390	3	187031377	Height	a	g	0.358	*IGF2BP2*	Lango Allen et al (Nature 2010)	
rs2247341	4	1671115	Height	a	g	0.358	*SLBP/FGFR3*	Lango Allen et al (Nature 2010)	
rs2724475	4	17555530	Height_2ndary	t	c	0.358	*LCORL*	Lango Allen et al (Nature 2010)	
rs6449353	4	17642586	Height	t	c	0.883	*LCORL*	Lango Allen et al (Nature 2010)	
rs17081935	4	57518233	Height	t	c	0.2	*POLR2B*	Lango Allen et al (Nature 2010)	
rs7697556	4	73734177	Height	t	c	0.508	*ADAMTS3*	Lango Allen et al (Nature 2010)	
rs788867	4	82369030	Height	g	t	0.292	*PRKG2/BMP3*	Lango Allen et al (Nature 2010)	
rs10010325	4	106325802	Height	a	c	0.467	*TET2*	Lango Allen et al (Nature 2010)	
rs2353398	4	145742208	Height_2ndary	a	t	0.492	*HHIP*	Lango Allen et al (Nature 2010)	
rs7689420	4	145787802	Height	c	t	0.817	*HHIP*	Lango Allen et al (Nature 2010)	
rs955748	4	184452669	Height	g	a	0.733	*WWC2*	Lango Allen et al (Nature 2010)	
rs3792752	5	32804391	Height_2ndary	g	a	0.267	*NPR3*	Lango Allen et al (Nature 2010)	
rs1173727	5	32866278	Height	t	c	0.542	*NPR3*	Lango Allen et al (Nature 2010)	
rs11958779	5	55037656	Height	g	a	0.242	*SLC38A9*	Lango Allen et al (Nature 2010)	
rs10037512	5	88390431	Height	t	c	0.583	*MEF2C*	Lango Allen et al (Nature 2010)	
rs13177718	5	108141243	Height	c	t	0.908	*FER*	Lango Allen et al (Nature 2010)	
rs1582931	5	122685098	Height	g	a	0.542	*CEP120*	Lango Allen et al (Nature 2010)	
rs274546	5	131727766	Height	g	a	0.633	*SLC22A5*	Lango Allen et al (Nature 2010)	
rs526896	5	134384604	Height	t	g	0.7	*PITX1*	Lango Allen et al (Nature 2010)	
rs4282339	5	168188818	Height	g	a	0.783	*SLIT3*	Lango Allen et al (Nature 2010)	
rs6892884	5	170948228	Height_2ndary	c	t	0.7	*FBXW11*	Lango Allen et al (Nature 2010)	
rs12153391	5	171136043	Height	c	a	0.75	*FBXW11*	Lango Allen et al (Nature 2010)	
rs889014	5	172916720	Height	c	t	0.6	*BOD1*	Lango Allen et al (Nature 2010)	
rs422421	5	176449932	Height	c	t	0.783	*FGFR4/NSD1*	Lango Allen et al (Nature 2010)	
rs6879260	5	179663620	Height	c	t	0.675	*GFPT2*	Lango Allen et al (Nature 2010)	
rs3812163	6	7670759	Height	t	a	0.5	*BMP6*	Lango Allen et al (Nature 2010)	
rs1047014	6	19949472	Height	c	t	0.25	*ID4*	Lango Allen et al (Nature 2010)	
rs806794	6	26308656	Height	a	g	0.708	*Histone_cluster*	Lango Allen et al (Nature 2010)	
rs3129109	6	29192211	Height	c	t	0.556	*OR2J3*	Lango Allen et al (Nature 2010)	
rs879882	6	31247431	Height_2ndary	t	c	0.349	*MICA*	Lango Allen et al (Nature 2010)	
rs2256183	6	31488508	Height	a	g	0.558	*MICA*	Lango Allen et al (Nature 2010)	
rs6457620	6	32771977	Height	g	c	0.487	*HLA*	Lango Allen et al (Nature 2010)	
rs4711336	6	33767024	Height_2ndary	a	g	0.467	*HMGA1*	Lango Allen et al (Nature 2010)	
rs2780226	6	34307070	Height	c	t	0.083	*HMGA1*	Lango Allen et al (Nature 2010)	
rs6938239	6	34791613	Height_2ndary	g	a	0.133	*HMGA1*	Lango Allen et al (Nature 2010)	
rs6457821	6	35510783	Height	c	a	0.983	*PPARD/FANCE*	Lango Allen et al (Nature 2010)	
rs9472414	6	45054484	Height	t	a	0.833	*SUPT3H/RUNX2*	Lango Allen et al (Nature 2010)	
rs9360921	6	76322362	Height	g	t	0.125	*SENP6*	Lango Allen et al (Nature 2010)	
rs310405	6	81857081	Height	a	g	0.492	*FAM46A*	Lango Allen et al (Nature 2010)	
rs7759938	6	105485647	Height	c	t	0.375	*LIN28B*	Lango Allen et al (Nature 2010)	
rs1046943	6	109890634	Height	a	g	0.617	*ZBTB24*	Lango Allen et al (Nature 2010)	
rs961764	6	117628849	Height	g	c	0.55	*VGLL2*	Lango Allen et al (Nature 2010)	
rs1490384	6	126892853	Height	t	c	0.433	*C6orf173*	Lango Allen et al (Nature 2010)	
rs6569648	6	130390812	Height	c	t	0.225	*L3MBTL3*	Lango Allen et al (Nature 2010)	
rs225694	6	142568835	Height_2ndary	c	g	0.267	*GPR126*	Lango Allen et al (Nature 2010)	
rs7763064	6	142838982	Height	g	a	0.692	*GPR126*	Lango Allen et al (Nature 2010)	
rs543650	6	152152636	Height	g	t	0.6	*ESR1*	Lango Allen et al (Nature 2010)	
rs9456307	6	158849430	Height	t	a	0.95	*TULP4*	Lango Allen et al (Nature 2010)	
rs798489	7	2768329	Height	c	t	0.733	*GNA12*	Lango Allen et al (Nature 2010)	
rs4470914	7	19583047	Height	t	c	0.167	*TWISTNB*	Lango Allen et al (Nature 2010)	
rs12534093	7	23469499	Height	t	a	0.708	*IGF2BP3*	Lango Allen et al (Nature 2010)	
rs1708299	7	28156471	Height	a	g	0.333	*JAZF1*	Lango Allen et al (Nature 2010)	
rs6959212	7	38094851	Height	c	t	0.7	*STARD3NL*	Lango Allen et al (Nature 2010)	

Appendix table 4: *Height G-W significant SNPs reported from published GWAS (before October 2012). Continue. PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs42235	7	92086012	Height	t	c	0.333	*CDK6*	Lango Allen et al (Nature 2010)	
rs822552	7	148281567	Height	g	c	0.217	*PDIA4*	Lango Allen et al (Nature 2010)	
rs2110001	7	150147955	Height	g	c	0.308	*TMEM176A*	Lango Allen et al (Nature 2010)	
rs1013209	8	24172249	Height	c	t	0.775	*ADAM28*	Lango Allen et al (Nature 2010)	
rs10958476	8	57258362	Height_2ndary	c	t	0.133	*SDR16C5*	Lango Allen et al (Nature 2010)	
rs7460090	8	57356717	Height	t	c	0.892	*SDR16C5*	Lango Allen et al (Nature 2010)	
rs6473015	8	78341040	Height	c	a	0.317	*PEX2*	Lango Allen et al (Nature 2010)	
rs6470764	8	130794847	Height	c	t	0.767	*GSDMC*	Lango Allen et al (Nature 2010)	
rs12680655	8	135706519	Height	c	g	0.6	*ZFAT*	Lango Allen et al (Nature 2010)	
rs7864648	9	16358732	Height	t	g	0.283	*BNC2*	Lango Allen et al (Nature 2010)	
rs11144688	9	77732106	Height	g	a	0.875	*PCSK5*	Lango Allen et al (Nature 2010)	
rs7853377	9	85742025	Height	g	a	0.167	*C9orf64*	Lango Allen et al (Nature 2010)	
rs8181166	9	88306448	Height	c	g	0.542	*ZCCHC6*	Lango Allen et al (Nature 2010)	
rs2778031	9	90025546	Height	t	c	0.25	*SPIN1*	Lango Allen et al (Nature 2010)	
rs9969804	9	94468941	Height	a	c	0.492	*IPPK*	Lango Allen et al (Nature 2010)	
rs1257763	9	95933766	Height	a	g	0.033	*PTPDC1*	Lango Allen et al (Nature 2010)	
rs473902	9	97296056	Height	t	g	0.925	*PTCH1/FANCC*	Lango Allen et al (Nature 2010)	
rs7027110	9	108638867	Height	a	g	0.233	*ZNF462*	Lango Allen et al (Nature 2010)	
rs1468758	9	112846903	Height	c	t	0.717	*LPAR1*	Lango Allen et al (Nature 2010)	
rs751543	9	118162163	Height	t	c	0.7	*PAPPA*	Lango Allen et al (Nature 2010)	
rs7466269	9	132453905	Height	a	g	0.658	*FUBP3*	Lango Allen et al (Nature 2010)	
rs7849585	9	138251691	Height	t	g	0.358	*QSOX2*	Lango Allen et al (Nature 2010)	
rs7909670	10	12958770	Height	c	t	0.458	*CCDC3*	Lango Allen et al (Nature 2010)	
rs7916441	10	80595583	Height_2ndary	g	c	0.542	*PPIF*	Lango Allen et al (Nature 2010)	
rs2145998	10	80791702	Height	t	a	0.55	*PPIF*	Lango Allen et al (Nature 2010)	
rs11599750	10	101795432	Height	c	t	0.608	*CPN1*	Lango Allen et al (Nature 2010)	
rs2237886	11	2767307	Height	t	c	0.067	*KCNQ1*	Lango Allen et al (Nature 2010)	
rs7926971	11	12654616	Height	g	a	0.525	*TEAD1*	Lango Allen et al (Nature 2010)	
rs1330	11	17272605	Height	t	c	0.4	*NUCB2*	Lango Allen et al (Nature 2010)	
rs10838801	11	48054856	Height	g	a	0.325	*PTPRJ/SLC39A13*	Lango Allen et al (Nature 2010)	
rs1814175	11	49515748	Height	t	c	0.433	*FOLH1*	Lango Allen et al (Nature 2010)	
rs5017948	11	51270794	Height	a	t	0.225	*OR4A5*	Lango Allen et al (Nature 2010)	
rs3782089	11	65093395	Height	c	t	0.967	*SSSCA1*	Lango Allen et al (Nature 2010)	
rs7112925	11	66582736	Height	c	t	0.592	*RHOD*	Lango Allen et al (Nature 2010)	
rs634552	11	74959700	Height	t	g	0.158	*SERPINH1*	Lango Allen et al (Nature 2010)	
rs494459	11	118079885	Height	t	c	0.358	*TREH*	Lango Allen et al (Nature 2010)	
rs654723	11	128091365	Height	a	c	0.625	*FLI1*	Lango Allen et al (Nature 2010)	
rs2856321	12	11747040	Height	g	a	0.425	*ETV6*	Lango Allen et al (Nature 2010)	
rs10770705	12	20748734	Height	a	c	0.375	*SLCO1C1*	Lango Allen et al (Nature 2010)	
rs2638953	12	28425682	Height	c	g	0.683	*CCDC91*	Lango Allen et al (Nature 2010)	
rs2066807	12	55026949	Height	g	c	0.075	*STAT2*	Lango Allen et al (Nature 2010)	
rs1351394	12	64638093	Height	t	c	0.533	*HMGA2*	Lango Allen et al (Nature 2010)	
rs10748128	12	68113925	Height	t	g	0.358	*FRS2*	Lango Allen et al (Nature 2010)	
rs11107116	12	92502635	Height	t	g	0.192	*SOCS2*	Lango Allen et al (Nature 2010)	
rs10859563	12	92644470	Height_2ndary	c	g	0.567	*SOCS2*	Lango Allen et al (Nature 2010)	
rs7971536	12	100897919	Height	t	a	0.483	*CCDC53/GNPTAB*	Lango Allen et al (Nature 2010)	
rs11830103	12	122389499	Height	g	a	0.175	*SBNO1*	Lango Allen et al (Nature 2010)	
rs7332115	13	32045548	Height	g	t	0.375	*PDS5B/BRCA2*	Lango Allen et al (Nature 2010)	
rs3118905	13	50003335	Height	g	a	0.725	*DLEU7*	Lango Allen et al (Nature 2010)	
rs7319045	13	90822575	Height	a	g	0.383	*GPC5*	Lango Allen et al (Nature 2010)	
rs1950500	14	23900690	Height	t	c	0.267	*NFATC4*	Lango Allen et al (Nature 2010)	
rs2093210	14	60027032	Height	c	t	0.425	*SIX6*	Lango Allen et al (Nature 2010)	
rs1570106	14	67882868	Height	c	t	0.792	*RAD51L1*	Lango Allen et al (Nature 2010)	
rs862034	14	74060499	Height	g	a	0.6	*LTBP2*	Lango Allen et al (Nature 2010)	
rs7155279	14	91555634	Height	g	t	0.642	*TRIP11*	Lango Allen et al (Nature 2010)	
rs16964211	15	49317787	Height	g	a	0.958	*CYP19A1*	Lango Allen et al (Nature 2010)	
rs7178424	15	60167551	Height	c	t	0.517	*C2CD4A*	Lango Allen et al (Nature 2010)	
rs10152591	15	67835211	Height	a	c	0.892	*TLE3*	Lango Allen et al (Nature 2010)	
rs12902421	15	69948457	Height	c	t	0.017	*MYO9A*	Lango Allen et al (Nature 2010)	
rs750460	15	72028559	Height_2ndary	g	a	0.583	*PML*	Lango Allen et al (Nature 2010)	
rs5742915	15	72123686	Height	c	t	0.542	*PML*	Lango Allen et al (Nature 2010)	
rs11259936	15	82371586	Height	c	a	0.5	*ADAMTSL3*	Lango Allen et al (Nature 2010)	
rs16942341	15	87189909	Height	c	t	0.967	*ACAN*	Lango Allen et al (Nature 2010)	
rs2280470	15	87196630	Height_2ndary	a	g	0.3	*ACAN*	Lango Allen et al (Nature 2010)	
rs2871865	15	97012419	Height	c	g	0.883	*IGF1R*	Lango Allen et al (Nature 2010)	
rs4965598	15	98577137	Height	c	t	0.258	*ADAMTS17*	Lango Allen et al (Nature 2010)	
rs11648796	16	732191	Height	g	a	0.275	*NARFL*	Lango Allen et al (Nature 2010)	
rs26868	16	2189377	Height	a	t	0.458	*CASKIN1*	Lango Allen et al (Nature 2010)	
rs1659127	16	14295806	Height	a	g	0.275	*MKL2*	Lango Allen et al (Nature 2010)	
rs8052560	16	87304743	Height	a	c	0.767	*CTU2/GALNS*	Lango Allen et al (Nature 2010)	
rs4640244	17	21224816	Height	a	g	0.658	*KCNJ12*	Lango Allen et al (Nature 2010)	
rs3110496	17	24941897	Height	g	a	0.658	*ANKRD13B*	Lango Allen et al (Nature 2010)	
rs3764419	17	26188149	Height	c	a	0.558	*ATAD5/RNF135*	Lango Allen et al (Nature 2010)	
rs17780086	17	27367395	Height	a	g	0.158	*LRRC37B*	Lango Allen et al (Nature 2010)	
rs1043515	17	34175722	Height	g	a	0.517	*PIP4K2B*	Lango Allen et al (Nature 2010)	
rs4986172	17	40571807	Height	c	t	0.717	*ACBD4*	Lango Allen et al (Nature 2010)	
rs2072153	17	44745013	Height	c	g	0.3	*ZNF652*	Lango Allen et al (Nature 2010)	
rs4605213	17	46599746	Height	c	g	0.408	*NME2*	Lango Allen et al (Nature 2010)	
rs227724	17	52133816	Height	t	a	0.258	*NOG*	Lango Allen et al (Nature 2010)	
rs1401796	17	52194758	Height_2ndary	c	a	0.539	*NOG*	Lango Allen et al (Nature 2010)	
rs2079795	17	56851431	Height	t	c	0.333	*TBX2*	Lango Allen et al (Nature 2010)	
rs2665838	17	59320197	Height	g	c	0.25	*CSH1/GH1*	Lango Allen et al (Nature 2010)	
rs2070776	17	59361230	Height_2ndary	g	a	0.65	*CSH1/GH1*	Lango Allen et al (Nature 2010)	
rs11867479	17	65601802	Height	t	c	0.25	*KCNJ16/KCNJ2*	Lango Allen et al (Nature 2010)	
rs4800452	18	18981609	Height	t	c	0.75	*CABLES1*	Lango Allen et al (Nature 2010)	
rs9967417	18	45213498	Height	g	c	0.475	*DYM*	Lango Allen et al (Nature 2010)	
rs17782313	18	56002077	Height	c	t	0.233	*MC4R*	Lango Allen et al (Nature 2010)	
rs12982744	19	2128193	Height	g	c	0.417	*DOT1L*	Lango Allen et al (Nature 2010)	
rs7507204	19	3379834	Height	c	g	0.225	*NFIC*	Lango Allen et al (Nature 2010)	
rs891088	19	7135762	Height	g	a	0.3	*INSR*	Lango Allen et al (Nature 2010)	
rs4072910	19	8550031	Height	g	c	0.567	*ADAMTS10*	Lango Allen et al (Nature 2010)	
rs2279008	19	17144303	Height	t	c	0.767	*MYO9B*	Lango Allen et al (Nature 2010)	
rs17318596	19	46628935	Height	a	g	0.408	*ATP5SL*	Lango Allen et al (Nature 2010)	
rs1741344	20	4049800	Height	c	t	0.342	*SMOX*	Lango Allen et al (Nature 2010)	
rs2145272	20	6574218	Height	g	a	0.417	*BMP2*	Lango Allen et al (Nature 2010)	
rs7274811	20	31796842	Height	g	t	0.808	*ZNF341*	Lango Allen et al (Nature 2010)	
rs143384	20	33489170	Height	g	a	0.308	*GDF5*	Lango Allen et al (Nature 2010)	
rs237743	20	47336426	Height	a	g	0.342	*ZNFX1*	Lango Allen et al (Nature 2010)	
rs2834442	21	34612656	Height	a	t	0.608	*KCNE2*	Lango Allen et al (Nature 2010)	
rs4821083	22	31386341	Height	t	c	0.875	*SYN3*	Lango Allen et al (Nature 2010)	

Appendix table 4: *Continuation.*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES	
rs12027135	1	25648320	LDL	t	a	0.558	*TMEM57/LDLRAP1*	Teslovich et al (Nature 2010)		
rs12027135	1	25648320	TC	t	a	0.558	*TMEM57/LDLRAP1*	Teslovich et al (Nature 2010)		**
rs4660293	1	39800767	HDL	a	g	0.733	*MACF1/PABPC4*	Teslovich et al (Nature 2010)		
rs2479409	1	55277238	LDL	g	a	0.325	*PCSK9*	Teslovich et al (Nature 2010)		**
rs2479409	1	55277238	TC	g	a	0.325	*PCSK9*	Teslovich et al (Nature 2010)		
rs2131925	1	62798530	TG	t	g	0.617	*ANGPTL3/DOCK7*	Teslovich et al (Nature 2010)		**
rs3850634	1	62823186	LDL	t	g	0.625	*ANGPTL3–DOCK7*	Teslovich et al (Nature 2010)		
rs3850634	1	62823186	TC	t	g	0.625	*ANGPTL3–DOCK7*	Teslovich et al (Nature 2010)		**
rs7515577	1	92782026	TC	a	c	0.817	*GFI1/EVI5*	Teslovich et al (Nature 2010)		
rs629301	1	109619829	LDL	t	g	0.667	*CELSR2/PSRC1/SORT1*	Teslovich et al (Nature 2010)		**
rs629301	1	109619829	TC	t	g	0.667	*CELSR2/PSRC1/SORT1*	Teslovich et al (Nature 2010)		**
rs1801274	1	159746369	TC	a	g	0.492	*FCGR2A*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs1689800	1	180435508	HDL	a	g	0.567	*ZNF648*	Teslovich et al (Nature 2010)		
rs2807834	1	219037216	LDL	g	t	0.683	*MOSC1*	Teslovich et al (Nature 2010)		
rs2807834	1	219037216	TC	g	t	0.683	*MOSC1*	Teslovich et al (Nature 2010)		
rs4846914	1	228362314	HDL	a	g	0.608	*GALNT2*	Teslovich et al (Nature 2010)		**
rs1321257	1	228371935	TG	g	a	0.383	*GALNT2*	Teslovich et al (Nature 2010)		**
rs514230	1	232925220	LDL	t	a	0.45	*IRF2BP2/TOMM20*	Teslovich et al (Nature 2010)		
rs514230	1	232925220	TC	t	a	0.45	*IRF2BP2/TOMM20*	Teslovich et al (Nature 2010)		
rs1042034	2	21078786	HDL	c	t	0.2	*APOB*	Teslovich et al (Nature 2010)		**
rs1042034	2	21078786	TG	t	c	0.8	*APOB*	Teslovich et al (Nature 2010)		**
rs1367117	2	21117405	LDL	a	g	0.35	*APOB*	Teslovich et al (Nature 2010)		**
rs1367117	2	21117405	TC	a	g	0.35	*APOB*	Teslovich et al (Nature 2010)		**
rs1260326	2	27584444	TC	t	c	0.417	*GCKR*	Teslovich et al (Nature 2010)		
rs1260326	2	27584444	TG	t	c	0.417	*GCKR*	Teslovich et al (Nature 2010)		**
rs4299376	2	43926080	LDL	g	t	0.342	*ABCG5/8*	Teslovich et al (Nature 2010)		**
rs4299376	2	43926080	TC	g	t	0.342	*ABCG5/8*	Teslovich et al (Nature 2010)		**
rs12464355	2	118566320	TC	a	g	0.9	*INSIG2*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs6759321	2	136039146	TC	t	g	0.242	*RAB3GAP1*	Teslovich et al (Nature 2010)		
rs10195252	2	165221337	TG	t	c	0.583	*COBLL1*	Teslovich et al (Nature 2010)		
rs12328675	2	165249046	HDL	c	t	0.133	*COBLL1*	Teslovich et al (Nature 2010)		**
rs2943645	2	226807424	TG	t	c	0.658	*IRS1*	Teslovich et al (Nature 2010)		
rs1515100	2	226837161	HDL	c	a	0.333	*IRS1*	Teslovich et al (Nature 2010)		
rs11563251	2	234344123	TC	t	c	0.1	*UGT1A1*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs2290159	3	12603920	TC	g	c	0.8	*RAF1*	Teslovich et al (Nature 2010)		
rs645040	3	137409312	TG	t	g	0.8	*MSL2L1*	Teslovich et al (Nature 2010)		
rs442177	4	88249285	TG	t	g	0.608	*AFF1/KLHL8*	Teslovich et al (Nature 2010)		
rs13107325	4	103407732	HDL	c	t	0.908	*SLC39A8*	Teslovich et al (Nature 2010)		
rs6450176	5	53333782	HDL	g	a	0.783	*ARL15*	Teslovich et al (Nature 2010)		
rs9686661	5	55897543	TG	t	c	0.15	*ANKRD55/MAP3K1*	Teslovich et al (Nature 2010)		
rs12916	5	74692295	LDL	c	t	0.392	*HMGCR*	Teslovich et al (Nature 2010)		**
rs12916	5	74692295	TC	c	t	0.392	*HMGCR*	Teslovich et al (Nature 2010)		**
rs6882076	5	156322875	LDL	c	t	0.7	*TIMD4/HAVCR1*	Teslovich et al (Nature 2010)		**
rs6882076	5	156322875	TC	c	t	0.7	*TIMD4/HAVCR1*	Teslovich et al (Nature 2010)		
rs1553318	5	156411901	TG	c	g	0.708	*TIMD4/HAVCR1*	Teslovich et al (Nature 2010)		
rs3757354	6	16235386	LDL	c	t	0.817	*MYLIP*	Teslovich et al (Nature 2010)		
rs3757354	6	16235386	TC	c	t	0.817	*MYLIP*	Teslovich et al (Nature 2010)		
rs1800562	6	26201120	LDL	g	a	0.967	*HFE/HIST1H4C*	Teslovich et al (Nature 2010)		
rs1800562	6	26201120	TC	g	a	0.967	*HFE/HIST1H4C*	Teslovich et al (Nature 2010)		
rs2247056	6	31373469	TG	c	t	0.703	*HLA*	Teslovich et al (Nature 2010)		
rs389883	6	32055439	TG	t	g	0.7	*HLA*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs3177928	6	32520413	LDL	a	g	0.181	*HLA*	Teslovich et al (Nature 2010)		
rs3177928	6	32520413	TC	a	g	0.181	*HLA*	Teslovich et al (Nature 2010)		
rs2814982	6	34654538	TC	c	t	0.858	*C6orf106*	Teslovich et al (Nature 2010)		
rs2814944	6	34660775	HDL	g	a	0.858	*C6orf106*	Teslovich et al (Nature 2010)		
rs9488822	6	116419586	TC	a	t	0.692	*FRK*	Teslovich et al (Nature 2010)		
rs11153594	6	116461284	LDL	c	t	0.625	*FRK*	Teslovich et al (Nature 2010)		
rs605066	6	139871359	HDL	t	c	0.6	*CITED2*	Teslovich et al (Nature 2010)		
rs1564348	6	160498850	LDL	c	t	0.208	*LPA*	Teslovich et al (Nature 2010)		
rs1564348	6	160498850	TC	c	t	0.208	*LPA*	Teslovich et al (Nature 2010)		
rs3123629	6	160826076	TG	a	g	0.367	*LPA*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs1084651	6	161009807	HDL	g	a	0.908	*LPA*	Teslovich et al (Nature 2010)		
rs2285942	7	21549442	TC	t	c	0.142	*DNAH11*	Teslovich et al (Nature 2010)		
rs12670798	7	21573877	LDL	c	t	0.208	*DNAH11*	Teslovich et al (Nature 2010)		**
rs2072183	7	44545705	TC	c	g	0.283	*NPC1L1*	Teslovich et al (Nature 2010)		
rs217386	7	44567220	LDL	g	a	0.608	*NPC1L1*	Teslovich et al (Nature 2010)		
rs13238203	7	71767603	TG	c	t	0.975	*TYW1B*	Teslovich et al (Nature 2010)		
rs7811265	7	72572446	TG	a	g	0.833	*MLXIPL*	Teslovich et al (Nature 2010)		**
rs17145738	7	72620810	HDL	t	c	0.133	*MLXIPL*	Teslovich et al (Nature 2010)		
rs4731702	7	130083924	HDL	t	c	0.45	*KLF14*	Teslovich et al (Nature 2010)		
rs9987289	8	9220768	HDL	g	a	0.925	*PPP1R3B*	Teslovich et al (Nature 2010)		
rs2126259	8	9222556	LDL	c	t	0.908	*PPP1R3B*	Teslovich et al (Nature 2010)		
rs2126259	8	9222556	TC	c	t	0.908	*PPP1R3B*	Teslovich et al (Nature 2010)		
rs11776767	8	10721339	TG	c	g	0.375	*PINX1/XKR6*	Teslovich et al (Nature 2010)		**
rs6983129	8	11628545	TG	a	c	0.508	*GATA4*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs1961456	8	18299989	TC	g	a	0.317	*NAT2*	Teslovich et al (Nature 2010)		
rs1495743	8	18317580	TG	g	c	0.258	*NAT2*	Teslovich et al (Nature 2010)		
rs12679834	8	19864713	HDL_2ndary	c	t	0.125	*LPL*	Asselbergs et al (The American Journal of Human Genetics 2012)	Secondary indipendent signal	
rs12678919	8	19888502	HDL	g	a	0.125	*LPL*	Teslovich et al (Nature 2010)		**
rs12678919	8	19888502	TG	a	g	0.875	*LPL*	Teslovich et al (Nature 2010)		**
rs1030431	8	59474251	LDL	a	g	0.3	*CYP7A1*	Teslovich et al (Nature 2010)		
rs1030431	8	59474251	TC	a	g	0.3	*CYP7A1*	Teslovich et al (Nature 2010)		
rs2293889	8	116668374	HDL	g	t	0.642	*TRPS1*	Teslovich et al (Nature 2010)		
rs2737229	8	116717740	TC	a	c	0.717	*TRPS1*	Teslovich et al (Nature 2010)		
rs2954022	8	126551803	LDL	c	a	0.583	*TRIB1*	Teslovich et al (Nature 2010)		
rs2954022	8	126551803	TC	c	a	0.583	*TRIB1*	Teslovich et al (Nature 2010)		**
rs2954029	8	126560154	TG	a	t	0.575	*TRIB1*	Teslovich et al (Nature 2010)		**
rs10808546	8	126565000	HDL	t	c	0.425	*TRIB1*	Teslovich et al (Nature 2010)		
rs7388248	8	144376728	HDL	c	g	0.242	*GPIHBP1*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs11136341	8	145115531	LDL	g	a	0.392	*PLEC1*	Teslovich et al (Nature 2010)		
rs11136341	8	145115531	TC	g	a	0.392	*PLEC1*	Teslovich et al (Nature 2010)		
rs643531	9	15286034	HDL	a	c	0.85	*TTC39B*	Teslovich et al (Nature 2010)		**
rs581080	9	15295378	TC	c	g	0.8	*TTC39B*	Teslovich et al (Nature 2010)		
rs1883025	9	106704122	HDL	c	t	0.783	*ABCA1*	Teslovich et al (Nature 2010)		**

(** indicates that this is not the first report of association, but the source of the info reported here)

Appendix table 5: *Lipids G-W significant SNPs reported from published GWAS (before October 2012). Continue. PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES	
rs1883025	9	106704122	TC	c	t	0.783	*ABCA1*	Teslovich et al (Nature 2010)		
rs651007	9	135143696	TC	t	c	0.233	*ABO*	Teslovich et al (Nature 2010)		
rs649129	9	135144125	LDL	t	c	0.233	*ABO*	Teslovich et al (Nature 2010)		
rs10761731	10	64697616	TG	a	t	0.533	*JMJD1C*	Teslovich et al (Nature 2010)		
rs2068888	10	94829632	TG	g	a	0.525	*CYP26A1*	Teslovich et al (Nature 2010)		
rs11597086	10	101943695	TC	c	a	0.442	*CHUK*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs1129555	10	113900711	LDL	a	g	0.275	*GPAM*	Teslovich et al (Nature 2010)		
rs2255141	10	113923876	TC	a	g	0.267	*GPAM*	Teslovich et al (Nature 2010)		
rs2923084	11	10345358	HDL	a	g	0.867	*ADM/AMPD3*	Teslovich et al (Nature 2010)		
rs11024739	11	18602419	LDL	a	c	0.708	*SPTY2D1*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs10832963	11	18620817	TC	g	t	0.708	*SPTY2D1*	Teslovich et al (Nature 2010)		
rs3136441	11	46699823	HDL	c	t	0.083	*LRP4/NR1H3*	Teslovich et al (Nature 2010)		
rs174546	11	61326406	TG	t	c	0.383	*FADS1–2–3*	Teslovich et al (Nature 2010)		**
rs174550	11	61328054	TC	t	c	0.625	*FADS1–2–3*	Teslovich et al (Nature 2010)		**
rs174583	11	61366326	LDL	c	t	0.617	*FADS1–2–3*	Teslovich et al (Nature 2010)		**
rs174601	11	61379716	HDL	c	t	0.625	*FADS1–2–3*	Teslovich et al (Nature 2010)		**
rs11236530	11	75167052	HDL	c	a	0.883	*DGAT2*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs964184	11	116154127	HDL	c	g	0.85	*APOA1–C3–A4–A5*	Teslovich et al (Nature 2010)		**
rs964184	11	116154127	LDL	g	c	0.15	*APOA1–C3–A4–A5*	Teslovich et al (Nature 2010)		
rs964184	11	116154127	TC	g	c	0.15	*APOA1–C3–A4–A5*	Teslovich et al (Nature 2010)		
rs964184	11	116154127	TG	g	c	0.15	*APOA1–C3–A4–A5*	Teslovich et al (Nature 2010)		**
rs9804646	11	116170289	HDL_2ndary	t	c	0.058	*BUD13/APOA1*	Asselbergs et al (The American Journal of Human Genetics 2012)	Secondary indipendent signal	
rs12225230	11	116233840	HDL_2ndary	c	g	0.15	*BUD13/APOA1*	Asselbergs et al (The American Journal of Human Genetics 2012)	Secondary indipendent signal	
rs7941030	11	122027585	TC	c	t	0.425	*UBASH3B*	Teslovich et al (Nature 2010)		
rs7115089	11	122035801	HDL	g	c	0.392	*UBASH3B*	Teslovich et al (Nature 2010)		
rs11220462	11	125749162	LDL	a	g	0.158	*ST3GAL4*	Teslovich et al (Nature 2010)		
rs11220463	11	125753421	TC	t	a	0.15	*ST3GAL4*	Teslovich et al (Nature 2010)		
rs7134375	12	20365025	HDL	a	c	0.392	*PDE3A*	Teslovich et al (Nature 2010)		
rs11613352	12	56078847	TG	c	t	0.717	*LRP1*	Teslovich et al (Nature 2010)		
rs3741414	12	56130316	HDL	t	c	0.275	*LRP1*	Teslovich et al (Nature 2010)		
rs7134594	12	108484576	HDL	t	c	0.5	*MMAB/MVK*	Teslovich et al (Nature 2010)		**
rs11065987	12	110556807	LDL	a	g	0.617	*BRAP*	Teslovich et al (Nature 2010)		
rs11065987	12	110556807	TC	a	g	0.617	*BRAP*	Teslovich et al (Nature 2010)		
rs1169288	12	119901033	LDL	c	a	0.292	*HNF1A*	Teslovich et al (Nature 2010)		**
rs1169288	12	119901033	TC	c	a	0.292	*HNF1A*	Teslovich et al (Nature 2010)		
rs4759361	12	121744233	HDL	a	t	0.158	*HCAR2*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs4759375	12	122362191	HDL	t	c	0.108	*SBNO1*	Teslovich et al (Nature 2010)		
rs4765127	12	123026120	HDL	t	g	0.392	*CCDC92/ZNF664*	Teslovich et al (Nature 2010)		
rs12310367	12	123052631	TG	a	g	0.617	*CCDC92/ZNF664*	Teslovich et al (Nature 2010)		
rs838880	12	123827546	HDL	c	t	0.25	*SCARB1*	Teslovich et al (Nature 2010)		
rs9534275	13	31838345	LDL	c	a	0.525	*BRCA2*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs2332328	14	23952898	LDL	t	c	0.533	*CBLN3/KIAA1305*	Teslovich et al (Nature 2010)		
rs2412710	15	40471079	TG	a	g	0.009	*CAPN3*	Teslovich et al (Nature 2010)		
rs2929282	15	42033223	TG	t	a	0.017	*FRMD5*	Teslovich et al (Nature 2010)		
rs4775041	15	56461987	HDL_2ndary	c	g	0.292	*LIPC*	Asselbergs et al (The American Journal of Human Genetics 2012)	Secondary indipendent signal	
rs1532085	15	56470658	HDL	a	g	0.392	*LIPC*	Teslovich et al (Nature 2010)		**
rs1532085	15	56470658	TC	a	g	0.392	*LIPC*	Teslovich et al (Nature 2010)		
rs261342	15	56518445	TG	g	c	0.25	*LIPC*	Teslovich et al (Nature 2010)		
rs2652834	15	61183920	HDL	g	a	0.8	*LACTB*	Teslovich et al (Nature 2010)		
rs11649653	16	30825988	TG	c	g	0.55	*CTF1*	Teslovich et al (Nature 2010)		
rs1421085	16	52358455	HDL	t	c	0.542	*FTO*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs247616	16	55547091	LDL	c	t	0.683	*CETP*	Teslovich et al (Nature 2010)		
rs3764261	16	55550825	HDL	a	c	0.308	*CETP*	Teslovich et al (Nature 2010)		**
rs3764261	16	55550825	TC	a	c	0.308	*CETP*	Teslovich et al (Nature 2010)		
rs4783961	16	55552395	HDL_2ndary	a	g	0.458	*CETP*	Asselbergs et al (The American Journal of Human Genetics 2012)	Secondary indipendent signal	
rs7205804	16	55562390	TG	g	a	0.55	*CETP*	Teslovich et al (Nature 2010)		
rs16942887	16	66485543	HDL	a	g	0.108	*LCAT*	Teslovich et al (Nature 2010)		**
rs2000999	16	70665594	LDL	a	g	0.175	*HP/HPR/DHX38*	Teslovich et al (Nature 2010)		
rs2000999	16	70665594	TC	a	g	0.175	*HP/HPR/DHX38*	Teslovich et al (Nature 2010)		
rs2925979	16	80092291	HDL	c	t	0.7	*CMIP*	Teslovich et al (Nature 2010)		
rs881844	17	35063744	HDL	g	c	0.717	*STARD3*	Teslovich et al (Nature 2010)		
rs7225700	17	42746803	LDL	c	t	0.608	*OSBPL7*	Teslovich et al (Nature 2010)		
rs7206971	17	42780114	TC	a	g	0.5	*OSBPL7*	Teslovich et al (Nature 2010)		
rs1801689	17	61641042	LDL	c	a	0.017	*APOH*	Asselbergs et al (The American Journal of Human Genetics 2012)		
rs4148008	17	64386889	HDL	c	g	0.75	*ABCA8*	Teslovich et al (Nature 2010)		
rs4082919	17	73889077	HDL	t	g	0.542	*PGS1*	Teslovich et al (Nature 2010)		
rs7241918	18	45414951	HDL	t	g	0.825	*LIPG*	Teslovich et al (Nature 2010)		**
rs7239867	18	45418715	TC	g	a	0.825	*LIPG*	Teslovich et al (Nature 2010)		
rs12967135	18	56000003	HDL	g	a	0.758	*RPS3A/MC4R*	Teslovich et al (Nature 2010)		
rs7255436	19	8339196	HDL	a	c	0.642	*ANGPTL4*	Teslovich et al (Nature 2010)		**
rs6511720	19	11063306	LDL	g	t	0.917	*LDLR*	Teslovich et al (Nature 2010)		**
rs6511720	19	11063306	TC	g	t	0.917	*LDLR*	Teslovich et al (Nature 2010)		**
rs737337	19	11208493	HDL	t	c	0.958	*DOCK6/LOC55908*	Teslovich et al (Nature 2010)		
rs10401969	19	19268718	LDL	t	c	0.908	*CSPG3/CILP2/PBX4*	Teslovich et al (Nature 2010)		**
rs10401969	19	19268718	TC	t	c	0.908	*CSPG3/CILP2/PBX4*	Teslovich et al (Nature 2010)		**
rs10401969	19	19268718	TG	t	c	0.908	*CSPG3/CILP2/PBX4*	Teslovich et al (Nature 2010)		**
rs439401	19	50106291	TG	c	t	0.608	*APOE–C1–C2*	Teslovich et al (Nature 2010)		**
rs4420638	19	50114786	HDL	a	g	0.817	*APOE–C1–C2*	Teslovich et al (Nature 2010)		
rs4420638	19	50114786	LDL	g	a	0.183	*APOE–C1–C2*	Teslovich et al (Nature 2010)		**
rs4420638	19	50114786	TC	g	a	0.183	*APOE–C1–C2*	Teslovich et al (Nature 2010)		**
rs492602	19	53898229	TC	g	a	0.542	*FUT2/FLJ36070*	Teslovich et al (Nature 2010)		
rs386000	19	59484573	HDL	c	g	0.183	*LILRA3/LILRB2*	Teslovich et al (Nature 2010)		
rs2277862	20	33616196	TC	c	t	0.908	*ERGIC3*	Teslovich et al (Nature 2010)		
rs2902940	20	38524901	TC	a	g	0.742	*MAFB*	Teslovich et al (Nature 2010)		
rs2902941	20	38524928	LDL	a	g	0.742	*MAFB*	Teslovich et al (Nature 2010)		**
rs4297946	20	39244689	TC	c	g	0.433	*TOP1*	Teslovich et al (Nature 2010)		
rs909802	20	39370229	LDL	t	c	0.45	*TOP1*	Teslovich et al (Nature 2010)		
rs1800961	20	42475778	HDL	c	t	0.958	*HNF4A*	Teslovich et al (Nature 2010)		**
rs1800961	20	42475778	TC	c	t	0.958	*HNF4A*	Teslovich et al (Nature 2010)		
rs4810479	20	43978455	TG	c	t	0.275	*PLTP*	Teslovich et al (Nature 2010)		**
rs6065906	20	43987422	HDL	t	c	0.808	*PLTP*	Teslovich et al (Nature 2010)		**
rs181362	22	20262068	HDL	c	t	0.842	*UBE2L3*	Teslovich et al (Nature 2010)		
rs5756931	22	36875979	TG	t	c	0.583	*PLA2G6*	Teslovich et al (Nature 2010)		

(** indicates that this is not the first report of association, but the source of the info reported here)

Appendix table 5: *Continuation.*

SNP	CHR	Position HG18 (UCSC Build36)	PHENO	EA	NEA	EAF	LOCUS	CITATION	NOTES
rs880315	1	10719453	DBP	c	t	0.358	*CASZ1*	Takeuchi et al (Circulation 2010)	Japanese
rs4846049	1	11772952	DBP	g	t	0.642	*MTHFR/NPPB*	Johnson et al (The American Journal of Human Genetics 2011)	
rs17367504	1	11785365	SBP	a	g	0.817	*MTHFR-NPPB*	Newton-Cheh et al (Nature Genetics 2009)	
rs13306560	1	11788770	DBP	g	a	0.358	*MTHFR/CLCN6*	Tomazewski et al (Hypertension 2010)	
rs17030613	1	112992330	DBP	c	a	0.175	*ST7L/CAPZA1*	Kato et al (Nature Genetics 2011)	East Asian
rs2932538	1	113018066	SBP/DBP	g	a	0.7	*MOV10*	Ehret et al (Nature 2011)	
rs2004776	1	228915325	HTN	t	c	0.242	*AGT*	Johnson et al (The American Journal of Human Genetics 2011)	
rs16849225	2	164615066	SBP	c	t	0.75	*FIGN/GRB14*	Kato et al (Nature Genetics 2011)	East Asian
rs13002573	2	164623454	PP	a	g	0.75	*FIGN*	Wain et al (Nature Genetics 2011)	
rs1446468	2	164671732	MAP/SBP/DBP	c	t	0.517	*FIGN*	Wain et al (Nature Genetics 2011)	
rs13082711	3	27512913	DBP	c	t	0.225	*SLC4A7*	Ehret et al (Nature 2011)	
rs3774372	3	41852418	DBP	c	t	0.217	*ULK4*	Ehret et al (Nature 2011)	
rs9815354	3	41887655	DBP	a	g	0.217	*ULK4*	Levy et al (Nature Genetics 2009)	
rs319690	3	47902488	MAP	t	c	0.525	*MAP4_intron*	Wain et al (Nature Genetics 2011)	
rs419076	3	170583580	SBP/DBP	t	c	0.467	*MECOM*	Ehret et al (Nature 2011)	
rs871606	4	54494002	PP	t	c	0.883	*CHIC2*	Wain et al (Nature Genetics 2011)	
rs1458038	4	81383747	DBP/SBP	t	c	0.267	*FGF5*	Ehret et al (Nature 2011)	
rs16998073	4	81403365	DBP/HTN	t	c	0.358	*FGF5*	Newton-Cheh et al (Nature Genetics 2009), Takeuchi et al (Circulation 2010)	in Janapenese population also with HTN
rs13107325	4	103407732	DBP/SBP	c	t	0.908	*SLC39A8*	Ehret et al (Nature 2011)	
rs6825911	4	111601087	DBP	c	t	0.217	*ENPEP*	Kato et al (Nature Genetics 2011)	East Asian
rs13139571	4	156864963	DBP	c	a	0.717	*GUCY1A3/GUCY1B3*	Ehret et al (Nature 2011)	
rs1173766	5	32840285	SBP	c	t	0.45	*NPR3*	Kato et al (Nature Genetics 2011)	East Asian
rs1173771	5	32850785	SBP/DBP/HTN	g	a	0.475	*NPR3-C5orf23*	Ehret et al (Nature 2011)	
rs11953630	5	157777980	SBP/DBP	c	t	0.658	*EBF1*	Ehret et al (Nature 2011)	
rs1799945	6	26199158	SBP/DBP/HTN	g	c	0.125	*HFE*	Ehret et al (Nature 2011)	
rs805303	6	31724345	SBP/DBP/HTN	g	a	0.698	*BAT2-BAT5*	Ehret et al (Nature 2011)	
rs17477177	7	106199094	PP/SBP	c	t	0.242	*PIK3CG*	Wain et al (Nature Genetics 2011)	
rs3918226	7	150321109	DBP	t	c	0.092	*NOS3*	Johnson et al (The American Journal of Human Genetics 2011)	
rs2898290	8	11471318	SBP	c	t	0.525	*GATA4*	Ho et al (Journal of Hypertension 2010)	
rs2071518	8	120504993	PP	t	c	0.208	*NOV_3UTR*	Wain et al (Nature Genetics 2011)	
rs4373814	10	18459978	SBP/DBP	c	g	0.333	*CACNB2_5UTR*	Ehret et al (Nature 2011)	
rs1813353	10	18747454	SBP/DBP/HTN	t	c	0.633	*CACNB2_3UTR*	Ehret et al (Nature 2011)	
rs11014166	10	18748804	DBP	a	t	0.633	*CACNB2*	Levy et al (Nature Genetics 2009)	
rs4590817	10	63137559	SBP/DBP/HTN	g	c	0.817	*C10orf107*	Ehret et al (Nature 2011)	
rs1530440	10	63194597	DBP	c	t	0.817	*C10orf107*	Newton-Cheh et al (Nature Genetics 2009)	
rs932764	10	95885930	SBP/HTN	g	a	0.425	*PLCE1*	Ehret et al (Nature 2011)	
rs1004467	10	104584497	SBP	a	g	0.908	*CYP17A1*	Levy et al (Nature Genetics 2009)	
rs12413409	10	104709086	DBP/SBP/HTN	g	a	0.917	*CYP17A1/CNNM2*	Takeuchi et al (Circulation 2010)	Japanese
rs11191548	10	104836168	SBP	t	c	0.917	*CYP17A1/NT5C2*	Newton-Cheh et al (Nature Genetics 2009)	
rs2782980	10	115771517	MAP	c	t	0.817	*ADRB1*	Wain et al (Nature Genetics 2011)	
rs661348	11	1861868	MAP	c	t	0.4	*LSP1/TNNT3*	Johnson et al (The American Journal of Human Genetics 2011)	
rs7129220	11	10307114	SBP	a	g	0.117	*ADM*	Ehret et al (Nature 2011)	
rs381815	11	16858844	SBP	t	c	0.317	*PLEKHA7*	Levy et al (Nature Genetics 2009)	
rs633185	11	100098748	SBP/DBP/HTN	c	g	0.708	*FLJ32810/TMEM133*	Ehret et al (Nature 2011)	
rs11222084	11	129778440	PP	t	a	0.35	*ADAMTS8*	Wain et al (Nature Genetics 2011)	
rs2681472	12	88533090	DBP/HTN	a	g	0.9	*ATP2B1*	Levy et al (Nature Genetics 2009)	
rs2681492	12	88537220	SBP	t	c	0.892	*ATP2B1*	Levy et al (Nature Genetics 2009)	
rs11105354	12	88550654	HTN	a	g	0.9	*ATP2B1*	Johnson et al (The American Journal of Human Genetics 2011)	
rs17249754	12	88584717	SBP/DBP/HTN	g	a	0.9	*ATP2B1*	Ehret et al (Nature 2011)	
rs3184504	12	110368991	DBP/SBP	t	c	0.45	*SH2B3*	Ehret et al (Nature 2011), Levy et al (Nature Genetics 2009)	
rs653178	12	110492139	DBP	c	t	0.417	*SH2B3*	Newton-Cheh et al (Nature Genetics 2009)	
rs11066280	12	111302166	DBP/SBP	t	a	1	*ALDH2/RPL6/PTPN11*	Kato et al (Nature Genetics 2011)	East Asian
rs2384550	12	113837114	DBP	g	a	0.642	*TBX5/TBX3*	Levy et al (Nature Genetics 2009)	
rs10850411	12	113872179	DBP	t	c	0.683	*TBX5/TBX3*	Ehret et al (Nature 2011)	
rs35444	12	114036820	DBP	a	g	0.625	*TBX3*	Kato et al (Nature Genetics 2011)	East Asian
rs1378942	15	72864420	DBP	c	a	0.3	*CYP1A1-ULK3*	Newton-Cheh et al (Nature Genetics 2009)	
rs6495122	15	72912698	DBP	a	c	0.358	*CSK/ULK3*	Levy et al (Nature Genetics 2009)	
rs2521501	15	89238392	SBP/DBP	t	a	0.383	*FURIN/FES*	Ehret et al (Nature 2011)	
rs13333226	16	20273155	HTN	a	g	0.808	*UMOD*	Padmanabhan et al (Plos Genetics2010)	
rs12946454	17	40563647	SBP	t	a	0.242	*PLCD3*	Newton-Cheh et al (Nature Genetics 2009)	
rs17608766	17	42368270	SBP	c	t	0.092	*GOSR2*	Ehret et al (Nature 2011)	
rs12940887	17	44757806	DBP/SBP	t	c	0.375	*ZNF652*	Ehret et al (Nature 2011)	
rs16948048	17	44795465	DBP	g	a	0.375	*ZNF652*	Newton-Cheh et al (Nature Genetics 2009)	
rs1327235	20	10917030	DBP/SBP	g	a	0.508	*JAG1*	Ehret et al (Nature 2011)	
rs6015450	20	57184512	SBP/DBP/HTN	g	a	0.058	*GNAS/EDN3*	Ehret et al (Nature 2011)	

Appendix table 6: *Blood pressure and HTN G-W significant SNPs reported from published GWAS (before October 2012). PHENO: phenotype; EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency in CEU population (from 1000G data, pilot 1).*

SNP ID	NEAR LOCUS	ORIGINAL ASSOCIATION	SUB-CLUSTER NAMES			DAPPLE SIGNIFICANCE			STRING SIGNIFICANCE			GeneMANIA SIGNIFICANCE			GOrilla SIGNIFICANCE			MULTIPLE EFFECTS DEFINITIONS		
			CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3
rs1714573	MLXIPL	HDL	H25_1	H20_1	H15_1	1			3.65E-10			1			1			LIPIDS		
rs7811265	MLXIPL	TG																		
rs1267983	LPL	HDL																		
rs964184	APOA1–C3–A4–A5	HDL/LDL/TC/TG	H25_2	H20_2	H15_3															
rs1260326	GCKR	2hGlu/TC/TG/T2D/FGlu/Fins	H25_3	H20_3	H15_4															
rs4420638	APOE–C1–C2	HDL/LDL/TC	H25_4	H20_4	H15_5	0.000999			4.25E-08			2.4517E-07			1			LIPIDS		
rs629301	CELSR2/PSRC1/SORT1	LDL/TC																		
rs6511720	LDLR	LDL/TC			H15_6															
rs1367117	APOB	LDL/TC																		
rs1083096	MTNR1B	T2D/Fglu	H25_5	H20_5	H15_7	1			1			4.42764E-05			1			GLYCAEMIC		
rs1387153	MTNR1B	T2D/HbA1C																		
rs560887	G6PC2	Fglu																		
rs6892884	FBXW11	Height	H25_6	H20_6	H15_9	0	1	1	0.01	1	1	1	1	1	1	1	1	MIXED		METS
rs3792752	NPR3	Height																		
rs1001032	TET2	Height																		
rs974801	TET2	FinsadjBMI/Fins																		
rs1778008	LRRC37B	Height																		
rs1018706	CCDC108/IHH	Height																		
rs1087474	RPL5	Height																		
rs7515577	GVI1/EVI5	TC																		
rs35767	IGF1	Fins																		
rs1167800	HIP1	Fins																		
rs1703632	PPARG	FinsadjBMI/T2D			H15_10			1			1			1			1			HOUL
rs1308138	PPARG	T2D																		
rs731839	PEPD	Fins/FinsadjBMI																		
rs6450176	ARL15	HDL																		
rs4865796	ARL15	Fins/FinsadjBMI																		
rs3786897	PEPD	T2D																		
rs459193	ANKRD55	T2D																		
rs7944584	MADD	FGlu		H20_7	H15_11															
rs1050132	MADD	FastingPro-insulin																		
rs4846567	LYPLAL1	WHR	H25_7	H20_8	H15_12	0.01	1		0	0		1	1		1	1		HOUL		
rs2785980	LYPLAL1	Fins/FinsadjBMI																		
rs2820436	LYPLAL1	Fins/WHR																		
rs9491696	RSPO3	WHR/Fins																		
rs4765127	CCDC92/ZNF664	HDL/TG			H15_13			1			0			1			1			
rs9686661	ANKRD55/MAP3K1	TG																		
rs605066	CITED2	HDL																		
rs1055144	NFE2L3	WHR																		
rs1011731	DNM3/PIGC	WHR																		
rs718314	ITPR2-SSPN	WHR																		
rs6784615	NISCH/STAB1	WHR																		
rs4823006	ZNRF3-KREMEN1	WHR																		
rs1443512	HOXC13	WHR																		
rs6905288	VEGFA	WHR																		
rs984222	TBX15/WARS2	WHR																		
rs2943641	IRS1	T2D/Fins/TG/HDL		H20_9	H15_14		0.006			7.83E-14			6.3965E-05			1				
rs7578326	IRS1	T2D/Fins																		
rs4731702	KLF14	HDL/T2D																		
rs3822072	FAM13A	FinsadjBMI																		
rs1232867	COBLL1/GRB14	HDL/Fins																		
rs4691380	PDGFC	Fins/FinsadjBMI																		
rs1019525	GRB14	WHR/Fins/FinsadjBMI/TG/T2D																		
rs3923113	GRB14	T2D																		
rs439401	APOE–C1–C2	TG	H25_8	H20_10	H15_15	0.05	0.01	1	0	0	0	0	0	0	1	1	1	LIPIDS		
rs1040196	CILP2	T2D																		
rs2954022	TRIB1	LDL/TC/TG/HDL			H15_16			1			1			0			1			
rs1042034	APOB	HDL/TG																		
rs2000999	HP/HPR/DHX38	LDL/TC		H20_11	H15_17		1			0.017			7.9525E-05			1				
rs4299376	ABCG5/8	LDL/TC																		
rs1564348	LPA	LDL/TC																		
rs6882076	TIMD4/HAVCR1	LDL/TC/TG																		
rs1122046	ST3GAL4	LDL/TC																		
rs2479409	PCSK9	LDL/TC																		
rs217386	NPC1L1	LDL																		
rs3757354	MYLIP	LDL/TC																		
rs2332328	CBLN3/KIAA1305	LDL																		
rs2072183	NPC1L1	TC																		
rs1113634	PLEC1	LDL/TC																		
rs1421085	FTO	Fins/HDL/BMI/T2D	H25_9	H20_12	H15_18															
rs7241918	LIPG	HDL/TC	H25_10	H20_13	H15_19	0.05	1	1	0	0	0.06	0	0	0.04	1	1	1	MIXED	LIPIDS	LIPIDS STRANGE
rs1883025	ABCA1	HDL/TC																		
rs386000	LILRA3/LILRB2	HDL																		
rs7134594	MMAB/MVK	HDL																		
rs1694288	LCAT	HDL																		
rs9804646	BUD13/APOA1	HDL			H15_20			1			0			0			1			LIPIDS
rs2290159	RAF1	TC																		
rs514230	IRF2BP2/TOMM20	LDL/TC																		
rs1202713	TMEM57/LDLRAP1	LDL/TC																		
rs1267079	DNAH11	LDL																		
rs2285942	DNAH11	TC																		
rs2902940	MAFB	TC/LDL																		
rs1030431	CYP7A1	LDL/TC																		
rs2737229	TRPS1	TC																		
rs1169288	HNF1A	LDL/TC																		
rs651007	ABO	TC/LDL																		
rs1106598	BRAP	LDL/TC		H20_14	H15_21		1	1		0.03	0		0.05	1		1	1		BLOOD PRESSURE	BMI
rs3184504	SH2B3	DBP/SBP																		
rs2287019	QPCTL	BMI																		
rs1042392	GIPR	2hGlu																		
rs29941	KCTD15	BMI																		
rs1514175	TNNI3K	BMI																		
rs2890652	LRP1B	BMI																		
rs9816226	ETV5	BMI																		
rs1736750	MTHFR-NPPB	SBP			H15_22			1			0			0.04			1			BLOOD PRESSURE
rs4846049	MTHFR/NPPB	DBP																		
rs7129220	ADM	SBP																		
rs2681472	ATP2B1	SBP/DBP/HTN																		
rs1122208	ADAMTS8	PP																		
rs1530440	C10orf107	DBP																		
rs4590817	C10orf107	SBP/DBP/HTN																		
rs1313957	GUCY1A3/GUCY1B3	DBP																		
rs4986172	ACBD4	Height																		
rs1294645	PLCD3	SBP																		
rs2521501	FURIN/FES	SBP/DBP																		
rs6699417	PKN2	Height																		
rs2932538	MOV10	SBP/DBP																		
rs1703061	ST7L/CAPZA1	DBP																		
rs1333322	UMOD	HTN																		
rs1327235	JAG1	DBP/SBP																		
rs633185	FLJ32810/TMEM133	SBP/DBP/HTN																		
rs6495122	CSK/ULK3	DBP																		
rs1378942	CYP1A1-ULK3	DBP																		
rs1800562	HFE	HbA1C/LDL/TC		H20_15	H15_23		0			0			0.01			0.05			HEIGHT	
rs1799945	HFE	SBP/DBP/HTN																		
rs1490384	C6orf173	Height			H15_24			1			1			1			1			HEIGHT
rs720390	IGF2BP2	Height																		
rs849134	JAZF1	T2D																		
rs1708299	JAZF1	Height																		
rs1156325	UGT1A1	TC			H15_25			0.01			0.02			0.01			1			LIPIDS STRANGE
rs1801274	FCGR2A	TC																		
rs1102473	SPTY2D1	LDL/TC																		
rs1159708	CHUK	TC																		
rs1159975	CPN1	Height																		
rs581080	TTC39B	TC																		
rs643531	TTC39B	HDL																		
rs1222523	BUD13/APOA1	HDL																		
rs1800961	HNF4A	HDL/TC																		
rs2277862	ERGIC3	TC																		
rs7532866	LIN28	Height																		
rs1115359	FRK	LDL																		
rs9488822	FRK	TC																		
rs7941030	UBASH3B	TC/HDL																		
rs1129555	GPAM	LDL/TC			H15_26			0.04			0			1			1			LIPIDS STRANGE
rs2807834	MOSC1	LDL/TC																		
rs7206971	OSBPL7	TC																		
rs7225700	OSBPL7	LDL																		
rs6759321	RAB3GAP1	TC																		
rs1530559	YSK4	FIns																		
rs2814944	C6orf106	HDL/Height																		
rs2814982	C6orf106	TC																		
rs2256183	MICA	Height																		
rs7507204	NFIC	Height			H15_27			0.05			1			1			1			STRANGE
rs9727115	SNX7	FastingPro-insulin																		
rs1004467	CYP17A1	DBP/SBP/HTN																		
rs805303	BAT2-BAT5	SBP/DBP/HTN																		
rs2796441	TLE1	T2D																		
rs243088	BCL11A	T2D																		
rs4790333	SGSM2	FastingPro-insulin																		
rs391300	SRR	T2D																		
rs306549	DDX31	FastingPro-insulin																		
rs7460090	SDR16C5	Height																		
rs1582931	CEP120	Height																		
rs1659127	MKL2	Height																		
rs1106306	CCND2	T2D																		
rs1325598	PAPPA2	Height																		
rs2154319	SCMH1	Height																		
rs2110001	TMEM176A	Height			H15_28			1			0.02			1			1			HEIGHT
rs4282339	SLIT3	Height																		
rs1074812	FRS2	Height																		
rs7849585	QSOX2	Height																		
rs1043515	PIP4K2B	Height																		
rs3129109	OR2J3	Height																		
rs1731859	ATP5SL	Height																		
rs1257763	PTPDC1	Height																		
rs1780688	SUCLG2	Height																		
rs7155279	TRIP11	Height																		
rs526896	PITX1	Height																		
rs494459	TREH	Height																		
rs6959212	STARD3NL	Height																		
rs1708193	POLR2B	Height																		
rs2070776	CSH1/GH1	Height																		
rs1095847	SDR16C5	Height																		
rs9456307	TULP4	Height																		
rs1351164	TNS1	Height																		
rs1694234	ACAN	Height																		
rs1298274	DOT1L	Height																		
rs2353398	HHIP	Height																		
rs2279008	MYO9B	Height																		
rs7971536	CCDC53/GNPTAB	Height																		
rs1085956	SOCS2	Height																		
rs6714546	LTBP1	Height																		
rs237743	ZNFX1	Height																		
rs1317771	FER	Height																		
rs7916441	PPIF	Height																		
rs1741344	SMOX	Height																		
rs1047014	ID4	Height																		
rs7864648	BNC2	Height																		
rs2071518	NOV	3UTR																		
rs4470914	TWISTNB	Height																		
rs1290242	MYO9A	Height																		
rs4072910	ADAMTS10	Height																		
rs7652177	GHSR	Height																		
rs310405	FAM46A	Height																		
rs1215339	FBXW11	Height																		
rs1183010	SBNO1	Height																		
rs3829109	DNLZ	FGlu																		
rs7697556	ADAMTS3	Height																		
rs8052560	CTU2/GALNS	Height																		
rs1308846	DOCK3	Height																		
rs1195877	SLC38A9	Height																		
rs2629046	SERPINE2	Height																		
rs7274811	ZNF341	Height																		
rs2597513	HDAC11	Height																		
rs1253409	IGF2BP3	Height																		
rs1046896	FN3K	HbA1C		H20_16	H15_29		1			1			6.0491E-06			1				
rs1692624	HK1	HbA1C																		

***Appendix table 7:**.Summary of groups of shared multiple effects and of pathway analysis results for cardiometabolic SNPs. Continue.*

SNP ID	NEAR LOCUS	ORIGINAL ASSOCIATION	SUB-CLUSTER NAMES			DAPPLE SIGNIFICANCE			STRING SIGNIFICANCE			GeneMANIA SIGNIFICANCE			GOrilla SIGNIFICANCE			MULTIPLE EFFECTS DEFINITIONS		
			CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3	CI set 1	CI set 2	CI set 3
rs2112347	FLJ35779	BMI	H25_11	H20_17	H15_30	0.000999		1	1		0	2.3765E-08		0	1		1	HOUL		
rs12916	HMGCR	LDL/TC																		
rs2131925	ANGPTL3/DOCK7	TG/LDL/TC			H15_31															
rs389883	C4B	TG			H15_32			0			0.02			0			1			
rs2247056	HLA	TG																		
rs6457620	HLA	Height																		
rs4297946	TOP1	TC/LDL																		
rs3177928	HLA	LDL/TC																		
rs4965598	ADAMTS17	Height	H25_12	H20_18	H15_33	0.000999		0.05	4.72E-07		1	4.9367E-07		1	2.99E-10		0.01	HEIGHT/METS		HEIGHT/METS
rs862034	LTBP2	Height																		
rs1950500	NFATC4	Height																		
rs572169	GHSR	Height																		
rs3782089	SSSCA1	Height																		
rs1468758	LPAR1	Height																		
rs6457821	PPARD/FANCE	Height																		
rs1751110	CDC42EP3	Height																		
rs1401796	NOG	Height																		
rs750460	PML	Height																		
rs7759938	LIN28B	Height																		
rs6684205	TGFB2	Height																		
rs2780226	HMGA1	Height																		
rs2145272	BMP2	Height																		
rs2079795	TBX2	Height																		
rs1013209	ADAM28	Height																		
rs822552	PDIA4	Height																		
rs9835332	C3orf63	Height																		
rs3764419	ATAD5/RNF135	Height																		
rs7319045	GPC5	Height																		
rs7466269	FUBP3	Height																		
rs788867	PRKG2/BMP3	Height																		
rs798489	GNA12	Height																		
rs3791675	EFEMP1	Height																		
rs1046934	TSEN15	Height																		
rs4800452	CABLES1	Height																		
rs1125993	ADAMTSL3	Height																		
rs1173771	NPR3-C5orf23	SBP/DBP/HTN/Height																		
rs1173766	NPR3	SBP																		
rs2280470	ACAN	Height																		
rs1079944	JMJD4	Height																		
rs1110711	SOCS2	Height																		
rs7763064	GPR126	Height																		
rs7027110	ZNF462	Height																		
rs7689420	HHIP	Height			H15_34			0			0			0			0			HEIGHT
rs1120527	SF3B4	Height																		
rs143384	GDF5	Height																		
rs1351394	HMGA2	Height																		
rs806794	Histone	cluster																		
rs724016	ZBTB38	Height																		
rs1171719	ADCY5	2hGlu	H25_13	H20_19	H15_35	1	1	1	0.01	0.01	0.01	0.07	0.04	0	1	1	1	GLYCAEMIC STRANGE		
rs1170806	ADCY5	T2D/Fglu																		
rs1096525	CDKN2A/B	T2D/Fglu																		
rs7041847	GLIS3	T2D/Fglu																		
rs516946	ANK1	T2D																		
rs1294421	LY86	WHR			H15_36															
rs1161931	PDX1	Fglu/FastingPro-insulin			H15_37			0.06			0.01			0.02			1			
rs1160592	CRY2	FGlu																		
rs6048205	FOXA2	FGlu																		
rs6113722	FOXA2	FGlu																		
rs1192009	SLC2A2	FGlu																		
rs1716848	DGKB	T2D																		
rs4869272	PCSK1	Fglu/FastingPro-insulin																		
rs1160333	ARAP1	Fglu/FastingPro-insulin/T2D																		
rs174546	FADS1–2–3	TG/Fglu/TC/LDL/HDL		H20_20	H15_38															
rs1532085	LIPC	HDL/TC	H25_14	H20_21	H15_39	1			0.01			0.01			1			STRANGE		
rs4775041	LIPC	HDL																		
rs261342	LIPC	TG																		
rs4783961	CETP	HDL			H15_40															
rs9987289	PPP1R3B	HDL/Fins/Fglu/FInsadjBMI/LDL/TC		H20_22	H15_41															
rs983309	PPP1R3B	Fglu/Fins																		
rs1178238	PPP1R3B	2hGlu																		
rs7138803	FAIM2	BMI	H25_15	H20_23	H15_42	0.02397602		0.07	0.00098		0	1		1	1		1	MIXED		METS
rs1307880	CADM2	BMI																		
rs4836133	ZNF608	BMI																		
rs887912	FANCL	BMI																		
rs1244497	GPRC5B	BMI																		
rs1096857	LRRN6C	BMI																		
rs1184769	PRKD1	BMI																		
rs2815752	NEGR1	BMI																		
rs1187330	MC4R	T2D																		
rs1015033	NRXN3	BMI/WC																		
rs7359397	SH2B1	BMI																		
rs1093839	GNPDA2	BMI																		
rs4929949	RPL27A	BMI																		
rs987237	TFAP2B	BMI/WC																		
rs1076766	BDNF	BMI																		
rs543874	SEC16B	BMI																		
rs2867125	TMEM18	BMI																		
rs1297013	MC4R	WC/T2D																		
rs489693	MC4R	WC																		
rs571312	MC4R	BMI/HDL/Height																		
rs3817334	MTCH2	BMI			H15_43			0.02			0			1			1			HOUL
rs1310732	SLC39A8	BMI/DBP/SBP/HDL																		
rs4771122	MTIF3	BMI																		
rs206936	NUDT3	BMI																		
rs2241423	MAP2K5	BMI																		
rs1555543	PTBP2	BMI																		
rs3810291	TMEM160	BMI																		
rs1684922	FIGN/GRB14	SBP/PP																		
rs5215	KCNJ11	T2D																		
rs4665736	DNAJC27	Height																		
rs713586	RBJ	BMI																		
rs7332115	PDS5B/BRCA2	Height																		

Appendix table 7:*.Continuation.*

SNP ID	NEAR LOCUS	ORIGINAL ASSOCIATION	SUB-CLUSTER NAMES			DAPPLE SIGNIFICANCE			STRING SIGNIFICANCE			GeneMANIA SIGNIFICANCE			GOrilla SIGNIFICANCE			MULTIPLE EFFECTS DEFINITIONS		
			Cl set 1	Cl set 2	Cl set 3	Cl set 1	Cl set 2	Cl set 3	Cl set 1	Cl set 2	Cl set 3	Cl set 1	Cl set 2	Cl set 3	Cl set 1	Cl set 2	Cl set 3	Cl set 1	Cl set 2	Cl set 3
rs1747717	PIK3CG	PP/SBP	H25_16	H20_24	H15_44	1	1	1	0.03	1	1	1	1	1	1	0.07	1	MIXED	MIXED	METS
rs6825911	ENPEP	DBP																		
rs1246435	INSIG2	TC																		
rs2302593	GIPR	FGlu																		
rs3783347	WARS	FGlu																		
rs855791	TMPRSS6	HbA1C																		
rs1164965	CTF1	TG																		
rs645040	MSL2L1	TG																		
rs442177	AFF1/KLHL8	TG																		
rs7255436	ANGPTL4	HDL																		
rs4759361	HCAR2	HDL																		
rs2925979	CMIP	HDL																		
rs6983129	GATA4	TG																		
rs1177676	PINX1/XKR6	TG																		
rs2293889	TRPS1	HDL																		
rs4660293	MACF1/PABPC4	HDL																		
rs1495743	NAT2	TG			H15_45			1			1			1			1			METS
rs1961456	NAT2	TC																		
rs5756931	PLA2G6	TG																		
rs492602	FUT2/FLJ36070	TC																		
rs2657879	GLS2	FgluadjBMI																		
rs2068888	CYP26A1	TG																		
rs3123629	LPAL2	TG																		
rs1367226	EFEMP1	Height			H15_46			1			0			1			1			MIXED
rs1330656	MTHFR/CLCN6	DBP																		
rs9470794	ZFAND3	T2D																		
rs1074708	P2RX2	FGlu																		
rs871606	CHIC2	PP																		
rs576674	KL	FGlu																		
rs1277979	CDC123/CAMK1D	T2D																		
rs1760876	GOSR2	SBP																		
rs319690	MAP4	intron																		
rs4737009	ANK1	HbA1C																		
rs6474359	ANK1	HbA1C																		
rs7998202	ATP11A/TUBGCP3	HbA1C																		
rs2779116	SPTA1	HbA1C																		
rs1308271	SLC4A7	DBP																		
rs880315	CASZ1	DBP																		
rs1123653	DGAT2	HDL																		
rs1535500	KCNK16	T2D																		
rs1092393	NOTCH2	T2D																		
rs7957197	HNF1A/TCF1	T2D																		
rs1163439	ZFAND6	T2D																		
rs1329213	CHCHD9/TLE4	T2D																		
rs661348	LSP1/TNNT3	MAP																		
rs1758449	PTPRD	T2D																		
rs4812829	HNF4A	T2D																		
rs6017317	FITM2/R3HDML/HNF4A	T2D																		
rs381815	PLEKHA7	SBP																		
rs419076	MECOM	SBP/DBP																		
rs6467136	GCC1/PAXA4	T2D																		
rs1085041	TBX5/TBX3	DBP																		
rs2384550	TBX5/TBX3	DBP																		
rs1549318	LARP6	FastingPro-insulin																		
rs6015450	GNAS/EDN3	SBP/DBP/HTN																		
rs1195363	EBF1	SBP/DBP																		
rs1446468	FIGN	MAP/SBP/DBP																		
rs1813353	CACNB2	3UTR																		
rs1458038	FGF5	DBP/SBP/HTN																		
rs1802295	VPS26A	T2D																		
rs932764	PLCE1	SBP/HTN																		
rs3774372	ULK4	DBP																		
rs6861681	CPEB4	WHR																		
rs2145998	PPIF	Height		H20_25	H15_47		0.01798202			1			1			0.0431			GLYCAEMIC	GLYCAEMIC
rs961764	VGLL2	Height																		
rs1003751	MEF2C	Height																		
rs6943153	GRB10	FGlu																		
rs1171591	AMT	FGlu																		
rs1801689	APOH	LDL																		
rs9534275	BRCA2	LDL																		
rs4430796	HNF1B/TCF2	T2D																		
rs1359790	SPRY2	T2D																		
rs2782980	ADRB1	MAP																		
rs4373814	CACNB2	5UTR																		
rs654723	FLI1	Height																		
rs2898290	GATA4	SBP																		
rs6912327	C6orf107/UHRF1BP1	FInsadjBMI/Fins																		
rs9967417	DYM	Height		H20_26	H15_48		1	1		0.04	0		0.01	0.05		1	1		MIXED	HOUL
rs2871865	IGF1R	Height																		
rs6449353	LCORL	Height																		
rs2665838	CSH1/GH1	Height																		
rs6470764	GSDMC	Height																		
rs6439167	C3orf47	Height																		
rs2336725	RTF1	Height																		
rs1046943	ZBTB24	Height																		
rs1269499	Sep-02	Height																		
rs879882	MICA	Height																		
rs42235	CDK6	Height																		
rs9844666	PCCB	Height																		
rs2580816	NPPC	Height																		
rs4640244	KCNJ12	Height																		
rs3118905	DLEU7	Height																		
rs1111834	LYPLAL1	Height																		
rs6569648	L3MBTL3	Height																		
rs751543	PAPPA	Height																		
rs1268065	ZFAT	Height																		
rs1739169	GIPC2	Height																		
rs4605213	NME2	Height			H15_49			1			0.05			1			1			HEIGHT
rs3110496	ANKRD13B	Height																		
rs2247341	SLBP/FGFR3	Height																		
rs634552	SERPINH1	Height																		
rs955748	WWC2	Height																		
rs227724	NOG	Height																		
rs2856321	ETV6	Height																		
rs889014	BOD1	Height																		
rs1077070	SLCO1C1	Height																		
rs7112925	RHOD	Height																		
rs4821083	SYN3	Height																		
rs7909670	CCDC3	Height																		
rs2237886	KCNQ1	Height																		
rs2421992	DNM3	Height																		
rs2778031	SPIN1	Height																		
rs1015259	TLE3	Height																		
rs6473015	PEX2	Height																		
rs2834442	KCNE2	Height																		
rs274546	SLC22A5	Height																		
rs26868	CASKIN1	Height																		
rs7567288	NCKAP5	Height																		
rs1570106	RAD51L1	Height																		
rs9472414	SUPT3H/RUNX2	Height																		
rs1247420	SOCS5	Height																		
rs1686132	ST6GAL1	T2D																		
rs891088	INSR	Height																		
rs543650	ESR1	Height																		
rs2028299	AP3S2	T2D																		
rs2004776	AGT	HTN																		
rs545854	MSRA	WC																		
rs7612463	UBE2E2	T2D																		
rs35444	TBX3	DBP																		
rs2334499	DUSP8	T2D																		
rs8042680	PRC1	T2D																		
rs231361	KCNQ1	T2D																		
rs2093210	SIX6	Height																		
rs1738475	HTR1D	Height																		
rs7853377	C9orf64	Height																		
rs1814175	FOLH1	Height																		
rs1083880	PTPRJ/SLC39A13	Height																		
rs5017948	OR4A5	Height																		
rs422421	FGFR4/NSD1	Height																		
rs9863706	RYBP	Height																		
rs1114468	PCSK5	Height																		
rs1330	NUCB2	Height																		
rs7926971	TEAD1	Height																		
rs4601530	CLIC4	Height																		
rs3812163	BMP6	Height																		
rs2638953	CCDC91	Height																		
rs1168440	EIF2AK3	Height																		
rs9360921	SENP6	Height																		
rs1186747	KCNJ16/KCNJ2	Height																		
rs8181166	ZCCHC6	Height																		
rs1696421	CYP19A1	Height																		
rs2066807	STAT2	Height																		
rs9969804	IPPK	Height																		
rs1086393	DTL	Height																		
rs1734645	DNM3	Height																		
rs6879260	GFPT2	Height																		
rs9428104	SPAG17	Height																		
rs473902	PTCH1/FANCC	Height																		
rs1164879	NARFL	Height																		
rs4711336	HMGA1	Height																		
rs5742915	PML	Height																		
rs1247050	CCDC108/IHH	Height																		
rs2341459	C2orf34	Height																		
rs2724475	LCORL	Height																		
rs2284746	MFAP2	Height																		
rs6065906	PLTP	HDL			H15_50			1			0.03			1			1			LIPIDS
rs4810479	PLTP	TG																		
rs4846914	GALNT2	HDL/TG																		
rs2929282	FRMD5	TG																		
rs2412710	CAPN3	TG																		
rs1161335	LRP1	TG/HDL																		
rs4759375	SBNO1	HDL																		
rs7134375	PDE3A	HDL																		
rs4148008	ABCA8	HDL																		
rs2923084	ADM/AMPD3	HDL																		
rs1689800	ZNF648	HDL																		
rs7388248	GP1HBP1	HDL																		
rs737337	DOCK6/LOC55908	HDL																		
rs1084651	LPA	HDL																		
rs2652834	LACTB	HDL																		
rs4082919	PGS1	HDL																		
rs1076173	JMJD1C	TG																		
rs1083868	MADD	FastingPro-insulin																		
rs3136441	LRP4/NR1H3	HDL																		
rs6072275	TOP1	FGlu																		
rs225694	GPR126	Height																		
rs425277	PRKCZ	Height																		
rs7567851	PDE11A	Height																		
rs1323820	TYW1B	TG																		
rs1294088	ZNF652	DBP/SBP																		
rs2072153	ZNF652	Height																		
rs1691369	IKBKAP	FGlu		H20_27	H15_51		1	1		0.06	0		0	1		1	1		GLYCAEMIC STRANGE	GLICAEMIC STRANGE
rs1776245	RREB1	FGlu																		
rs1189986	THADA	T2D																		
rs1088512	ADRA2A	FGlu																		
rs340874	PROX1	T2D/Fglu																		
rs6960043	DGKB	T2D																		
rs1483121	OR4S1	FGlu																		
rs1019503	ERAP2	2hGlu																		
rs1107165	FAM148B/VPS13C/C2CD4A	FGlu																		
rs4502156	FAM148B/VPS13C/C2CD4A	FastingPro-insulin/T2D																		
rs7178424	C2CD4A	Height																		
rs1727130	FAM148B/VPS13C/C2CD4A	2hGlu			H15_52			1			0			1			1			GLYCAEMIC STRANGE
rs1111875	HHEX/IDE	T2D																		
rs7651090	IGF2BP2	FGlu/2hGlu/T2D																		
rs9368222	CDKAL1	Fglu/T2D																		
rs7754840	CDKAL1	T2D			H15_53			1			0			0.03			1			T2D/METS
rs7202877	BCAR1	T2D																		
rs4760790	TSPAN8/LGR5	T2D																		
rs1084299	KLHDC5	T2D																		
rs1257175	ZMIZ1	T2D																		
rs8108269	GIPR	T2D																		
rs944801	CDKN2A/B	T2D																		
rs831571	PSMD6	T2D																		
rs1531343	HMGA2	T2D																		
rs163184	KCNQ1	T2D/T2D																		
rs896854	TP53INP1	T2D																		
rs7593730	RBMS1	T2D																		
rs7178572	HMG20A	T2D/T2D																		
rs1001013	WFS1	T2D																		
rs231362	KCNQ1	T2D																		
rs4457053	ZBED3	T2D/FGluadjBMI																		
rs4607103	ADAMTS9	T2D																		
rs6795735	ADAMTS9	WHR/T2D																		
rs881844	STARD3	HDL			H15_54			0.07			1			1			1			GLYCAEMIC STRANGE
rs838880	SCARB1	HDL																		
rs181362	UBE2L3	HDL																		
rs1371614	DPYSL5	FGlu																		
rs7205804	CETP	TG	H25_17	H20_28	H15_55															
rs247616	CETP	LDL/HDL/TC																		
rs1799884	GCK	HbA1C/2hGlu/T2D/Fglu	H25_18	H20_29	H15_56	1			1			1			1			GLYCAEMIC STRANGE		
rs552976	G6PC2	HbA1C																		
rs1224332	TCF7L2	2hGlu	H25_19	H20_30	H15_57	1			1			1			1			STRANGE		
rs4506565	TCF7L2	Fglu/T2D/Fins																		
rs1326663	SLC30A8	T2D/Fglu/FastingPro-insulin																		

Appendix table 7:.*Continuation.*

6 References

1 Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature reviews. Genetics* **12**, 204-213, doi:10.1038/nrg2949 (2011).

2 Allison, D. B. *et al.* Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *American journal of human genetics* **63**, 1190-1201, doi:10.1086/302038 (1998).

3 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).

4 Dumitrescu, L. *et al.* No evidence of interaction between known lipid-associated genetic variants and smoking in the multi-ethnic PAGE population. *Human genetics*, doi:10.1007/s00439-013-1375-3 (2013).

5 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356-369, doi:10.1038/nrg2344 (2008).

6 Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics* **14**, 483-495, doi:10.1038/nrg3461 (2013).

7 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

8 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

9 Pendergrass, S. A. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic epidemiology* **35**, 410-422, doi:10.1002/gepi.20589 (2011).

10 Kim, S., Sohn, K. A. & Xing, E. P. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* **25**, i204-212, doi:10.1093/bioinformatics/btp218 (2009).

11 Tyler, A. L., Asselbergs, F. W., Williams, S. M. & Moore, J. H. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays : news and reviews in molecular, cellular and developmental biology* **31**, 220-227, doi:10.1002/bies.200800022 (2009).

12 Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. Major depression and generalized anxiety disorder. Same genes, (partly) different environments? *Archives of general psychiatry* **49**, 716-722 (1992).

13 Criswell, L. A. *et al.* Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *American journal of human genetics* **76**, 561-571, doi:10.1086/429096 (2005).

14 Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nature genetics* **40**, 310-315, doi:10.1038/ng.91 (2008).

15 Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988, doi:10.1038/ng2085 (2007).

16 Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* **42**, 937-948, doi:10.1038/ng.686 (2010).

17 Iles, M. M. *et al.* A variant in FTO shows association with melanoma risk not due to BMI. *Nature genetics* **45**, 428-432, 432e421, doi:10.1038/ng.2571 (2013).

18 Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics* **44**, 991-1005, doi:10.1038/ng.2385 (2012).

19 Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* **42**, 579-589, doi:10.1038/ng.609 (2010).

20 Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *American journal of human genetics* **89**, 607-618, doi:10.1016/j.ajhg.2011.10.004 (2011).

21 Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-1241, doi:10.1016/j.cell.2012.02.039 (2012).

22 Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS biology* **11**, e1001661, doi:10.1371/journal.pbio.1001661 (2013).

23 Plate, L. in *Festschrift zum sechzigsten Geburtstag Richard Hertwigs* (1910).

24 Haecker, V. in *Bibliographia Genetica* Vol. 1 1-314 (1925).

25 Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767-773, doi:10.1534/genetics.110.122549 (2010).

26 Fisher, R. A. *The genetical theory of natural selection*. 2d rev. edn, (Dover ; Constable, 1958).

27 Mayr, E. *Animal species and evolution*. (Belknap Press of Harvard University Press ;
Oxford University Press, 1963).

28 Beadle, G. W. & Tatum, E. L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* **27**, 499-506 (1941).

29 Hadorn, E. in *Methuen and Company* (London, 1961).

30 Williams, G. C. Pleiotropy, natural selection, and evolution of senescence. *Evolution; international journal of organic evolution* **11**, 398-411 (1957).

31 Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695 (1977).

32 Weber, J., Jelinek, W. & Darnell, J. E., Jr. The definition of a large viral transcription unit late in Ad2 infection of HeLa cells: mapping of nascent RNA molecules labeled in isolated nuclei. *Cell* **10**, 611-616 (1977).

33 Benne, R. *et al.* Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819-826 (1986).

34 Waxman, D. & Peck, J. R. Pleiotropy and the preservation of perfection. *Science* **279**, 1210-1213 (1998).

35 Hodgkin, J. Seven types of pleiotropy. *The International journal of developmental biology* **42**, 501-505 (1998).

36 Orr, H. A. Adaptation and the cost of complexity. *Evolution; international journal of organic evolution* **54**, 13-20 (2000).

37 Welch, J. J. & Waxman, D. Modularity and the cost of complexity. *Evolution; international journal of organic evolution* **57**, 1723-1734 (2003).

38 Wagner, G. P. *et al.* Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* **452**, 470-472, doi:10.1038/nature06756 (2008).

39 Strassmann, J. E., Zhu, Y. & Queller, D. C. Altruism and social cheating in the social amoeba Dictyostelium discoideum. *Nature* **408**, 965-967, doi:10.1038/35050087 (2000).

40 Su, Z., Zeng, Y. & Gu, X. A preliminary analysis of gene pleiotropy estimated from protein sequences. *Journal of experimental zoology. Part B, Molecular and developmental evolution* **314**, 115-122, doi:10.1002/jez.b.21315 (2010).

41 Li, R. *et al.* Structural model analysis of multiple quantitative traits. *PLoS genetics* **2**, e114, doi:10.1371/journal.pgen.0020114 (2006).

42 Wang, Z., Liao, B. Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 18034-18039, doi:10.1073/pnas.1004666107 (2010).

43 He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885-1891, doi:10.1534/genetics.106.060269 (2006).

44 Shriner, D. Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. *Frontiers in genetics* **3**, 1, doi:10.3389/fgene.2012.00001 (2012).

45 Ioannidis, J. P., Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nature reviews. Genetics* **10**, 318-329, doi:10.1038/nrg2544 (2009).

46 Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429-435, doi:10.1038/nature06757 (2008).

47 Dodds, C. & Allison, J. Postoperative cognitive deficit in the elderly surgical patient. *British journal of anaesthesia* **81**, 449-462 (1998).

48 Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nature reviews. Genetics* **14**, 379-389, doi:10.1038/nrg3472 (2013).

49 Fisher, R. A. *Statistical methods for research workers*. (Oliver & Boyd, 1925).

50 Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics* **7**, e1002254, doi:10.1371/journal.pgen.1002254 (2011).

51 Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *American journal of human genetics* **90**, 821-835, doi:10.1016/j.ajhg.2012.03.015 (2012).

52 O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087 (1984).

53 Yang, Q., Wu, H., Guo, C. Y. & Fox, C. S. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology* **34**, 444-454, doi:10.1002/gepi.20497 (2010).

54 van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics* **9**, e1003235, doi:10.1371/journal.pgen.1003235 (2013).

55 Huang, J., Johnson, A. D. & O'Donnell, C. J. PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics* **27**, 1201-1206, doi:10.1093/bioinformatics/btr116 (2011).

56 Weller, J. I., Wiggans, G. R., Vanraden, P. M. & Ron, M. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **92**, 998-1002, doi:10.1007/BF00224040 (1996).

57 Korol, A. B., Ronin, Y. I., Itskovich, A. M., Peng, J. & Nevo, E. Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics* **157**, 1789-1803 (2001).

58 Klei, L., Luca, D., Devlin, B. & Roeder, K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic epidemiology* **32**, 9-19, doi:10.1002/gepi.20257 (2008).

59 Ott, J. & Rabinowitz, D. A principal-components approach based on heritability for combining phenotype information. *Human heredity* **49**, 106-111, doi:22854 (1999).

60 Wang, Y., Fang, Y. & Jin, M. A ridge penalized principal-components approach based on heritability for high-dimensional data. *Human heredity* **64**, 182-191, doi:10.1159/000102991 (2007).

61 Ferreira, M. A. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**, 132-133, doi:10.1093/bioinformatics/btn563 (2009).

62 O'Reilly, P. F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* **7**, e34861, doi:10.1371/journal.pone.0034861 (2012).

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

64 Amos, C. I., Elston, R. C., Bonney, G. E., Keats, B. J. & Berenson, G. S. A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *American journal of human genetics* **47**, 247-254 (1990).

65 Eaves, L. J., Neale, M. C. & Maes, H. Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior genetics* **26**, 519-525 (1996).

66 Mardia, K. V., Bibby, J. M. & Kent, J. T. *Multivariate analysis*. (Academic Press, 1979).

67 Yang, F., Tang, Z. & Deng, H. Bivariate association analysis for quantitative traits using generalized estimation equation. *Journal of genetics and genomics = Yi chuan xue bao* **36**, 733-743, doi:10.1016/S1673-8527(08)60166-6 (2009).

68 Lee, P. H. *et al.* Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework. *Human heredity* **72**, 10-20, doi:10.1159/000327158 (2011).

69 Hartley, S. W. & Sebastiani, P. PleioGRiP: genetic risk prediction with pleiotropy. *Bioinformatics* **29**, 1086-1088, doi:10.1093/bioinformatics/btt081 (2013).

70 Hartley, S. W., Monti, S., Liu, C. T., Steinberg, M. H. & Sebastiani, P. Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. *Frontiers in genetics* **3**, 176, doi:10.3389/fgene.2012.00176 (2012).

71 Williams, J. T., Van Eerdewegh, P., Almasy, L. & Blangero, J. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. *American journal of human genetics* **65**, 1134-1147, doi:10.1086/302570 (1999).

72 Zeger, S. L. & Liang, K. Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130 (1986).

73 Zhang, L., Pei, Y. F., Li, J., Papasian, C. J. & Deng, H. W. Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PloS one* **4**, e6502, doi:10.1371/journal.pone.0006502 (2009).

74 Heijnen, C. J., van der Meer, J. W. & Zegers, B. J. Altered antigen-presentation in the induction of the in-vitro antigen-specific T helper cell function in patients with chronic granulomatous disease. *Clinical and experimental immunology* **66**, 111-117 (1986).

75 Liu, J., Pei, Y., Papasian, C. J. & Deng, H. W. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic epidemiology* **33**, 217-227, doi:10.1002/gepi.20372 (2009).

76 Zhang, H., Liu, C. T. & Wang, X. An Association Test for Multiple Traits Based on the Generalized Kendall's Tau. *Journal of the American Statistical Association* **105**, 473-481, doi:10.1198/jasa.2009.ap08387 (2010).

77 Chen, C. H., Chang, C. J., Yang, W. S., Chen, C. L. & Fann, C. S. A genome-wide scan using tree-based association analysis for candidate loci related to fasting plasma glucose levels. *BMC genetics* **4 Suppl 1**, S65, doi:10.1186/1471-2156-4-S1-S65 (2003).

78 Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PloS one* **8**, e65245, doi:10.1371/journal.pone.0065245 (2013).

79 International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).

80 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).

81 Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2542, doi:10.1093/bioinformatics/bts474 (2012).

82 Henderson, N. C. *et al.* Galectin-3 regulates myofibroblast activation and hepatic fibrosis. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5060-5065, doi:10.1073/pnas.0511167103 (2006).

83 Honjo, Y., Nangia-Makker, P., Inohara, H. & Raz, A. Down-regulation of galectin-3 suppresses tumorigenicity of human breast carcinoma cells. *Clinical cancer research : an official journal of the American Association for Cancer Research* **7**, 661-668 (2001).

84 Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular systems biology* **1**, 2005 0001, doi:10.1038/msb4100004 (2005).

85 Sandhu, M. S., Debenham, S. L., Barroso, I. & Loos, R. J. Mendelian randomisation studies of type 2 diabetes: future prospects. *Diabetologia* **51**, 211-213, doi:10.1007/s00125-007-0903-x (2008).

86 Katan, M. B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **1**, 507-508 (1986).

87 Thomas, D. C. & Conti, D. V. Commentary: the concept of 'Mendelian Randomization'. *International journal of epidemiology* **33**, 21-25, doi:10.1093/ije/dyh048 (2004).

88 Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572-580, doi:10.1016/S0140-6736(12)60312-2 (2012).

89 Freathy, R. M. *et al.* Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes* **57**, 1419-1426, doi:10.2337/db07-1466 (2008).

90 Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a mendelian randomization analysis. *PLoS medicine* **10**, e1001474, doi:10.1371/journal.pmed.1001474 (2013).

91 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

92 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

93 Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

94 Baker, M. Biorepositories: Building better biobanks. *Nature* **486**, 141-146, doi:10.1038/486141a (2012).

95 Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* **25**, 309-316, doi:10.1038/nbt1295 (2007).

96 Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a pathway resource list. *Nucleic acids research* **34**, D504-506, doi:10.1093/nar/gkj126 (2006).

97 WHO World Health Organisation. *WHO Global Infobase*, <https://apps.who.int/infobase/> (2010).

98 McCarthy, M. I. Genomics, type 2 diabetes, and obesity. *The New England journal of medicine* **363**, 2339-2350, doi:10.1056/NEJMra0906948 (2010).

99 Travers, M. E. & McCarthy, M. I. Type 2 diabetes and obesity: genomics and the clinic. *Human genetics* **130**, 41-58, doi:10.1007/s00439-011-1023-8 (2011).

100 Prokopenko, I., McCarthy, M. I. & Lindgren, C. M. Type 2 diabetes: new genes, new understanding. *Trends in genetics : TIG* **24**, 613-621, doi:10.1016/j.tig.2008.09.004 (2008).

101 Diabetes Genetics Initiative of Broad Institute of Harvard Mit *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336, doi:10.1126/science.1142358 (2007).

102 Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-1345, doi:10.1126/science.1142382 (2007).

103 Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885, doi:10.1038/nature05616 (2007).

104 Steinthorsdottir, V. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature genetics* **39**, 770-775, doi:10.1038/ng2043 (2007).

105 Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).

106 Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* **40**, 638-645, doi:10.1038/ng.120 (2008).

107 Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* **8**, e1002793, doi:10.1371/journal.pgen.1002793 (2012).

108 Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981-990, doi:10.1038/ng.2383 (2012).

109 Yamauchi, T. *et al.* A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nature genetics* **42**, 864-868, doi:10.1038/ng.660 (2010).

110 Tsai, F. J. *et al.* A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS genetics* **6**, e1000847, doi:10.1371/journal.pgen.1000847 (2010).

111 Unoki, H. *et al.* SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature genetics* **40**, 1098-1102, doi:10.1038/ng.208 (2008).

112 Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature genetics* **43**, 984-989, doi:10.1038/ng.921 (2011).

113 Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature genetics* **42**, 255-259, doi:10.1038/ng.530 (2010).

114 Lyssenko, V. *et al.* Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *The Journal of clinical investigation* **117**, 2155-2163, doi:10.1172/JCI30706 (2007).

115 Pearson, E. R. *et al.* Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. *Diabetes* **56**, 2178-2182, doi:10.2337/db07-0440 (2007).

116 Prokopenko, I. *et al.* Variants in MTNR1B influence fasting glucose levels. *Nature genetics* **41**, 77-81, doi:10.1038/ng.290 (2009).

117 Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics* **42**, 105-116, doi:10.1038/ng.520 (2010).

118 Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nature genetics* **42**, 142-148, doi:10.1038/ng.521 (2010).

119 Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics* **44**, 659-669, doi:10.1038/ng.2274 (2012).

120 Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229-3239, doi:10.2337/db10-0502 (2010).

121 Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624-2634, doi:10.2337/db11-0415 (2011).

122 Ingelsson, E. *et al.* Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes* **59**, 1266-1275, doi:10.2337/db09-1568 (2010).

123 Lyssenko, V. *et al.* Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nature genetics* **41**, 82-88, doi:10.1038/ng.288 (2009).

124 Meigs, J. B. *et al.* Body mass index, metabolic syndrome, and risk of type 2 diabetes or cardiovascular disease. *The Journal of clinical endocrinology and metabolism* **91**, 2906-2912, doi:10.1210/jc.2006-0594 (2006).

125 Loos, R. J. *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics* **40**, 768-775, doi:10.1038/ng.140 (2008).

126 Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics* **42**, 949-960, doi:10.1038/ng.685 (2010).

127 Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics* **41**, 25-34, doi:10.1038/ng.287 (2009).

128 Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature genetics* **41**, 18-24, doi:10.1038/ng.274 (2009).

129 Chambers, J. C. *et al.* Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nature genetics* **40**, 716-718, doi:10.1038/ng.156 (2008).

130 Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS genetics* **5**, e1000508, doi:10.1371/journal.pgen.1000508 (2009).

131 Heard-Costa, N. L. *et al.* NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS genetics* **5**, e1000539, doi:10.1371/journal.pgen.1000539 (2009).

132 Kilpelainen, T. O. *et al.* Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nature genetics* **43**, 753-760, doi:10.1038/ng.866 (2011).

133 Wen, W. *et al.* Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature genetics* **44**, 307-311, doi:10.1038/ng.1087 (2012).

134 Okada, Y. *et al.* Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nature genetics* **44**, 302-306, doi:10.1038/ng.1086 (2012).

135 Church, C. *et al.* A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS genetics* **5**, e1000599, doi:10.1371/journal.pgen.1000599 (2009).

136 Ren, D. *et al.* Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *The Journal of clinical investigation* **117**, 397-406, doi:10.1172/JCI29417 (2007).

137 Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838, doi:10.1038/nature09410 (2010).

138 Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS genetics* **6**, e1001113, doi:10.1371/journal.pgen.1001113 (2010).

139 Asselbergs, F. W. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *American journal of human genetics* **91**, 823-838, doi:10.1016/j.ajhg.2012.08.032 (2012).

140 Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics* **40**, 161-169, doi:10.1038/ng.76 (2008).

141 Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics* **40**, 189-197, doi:10.1038/ng.75 (2008).

142 Kooner, J. S. *et al.* Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nature genetics* **40**, 149-151, doi:10.1038/ng.2007.61 (2008).

143 Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics* **41**, 47-55, doi:10.1038/ng.269 (2009).

144 Chasman, D. I. *et al.* Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS genetics* **5**, e1000730, doi:10.1371/journal.pgen.1000730 (2009).

145 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713, doi:10.1038/nature09270 (2010).

146 Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719, doi:10.1038/nature09266 (2010).

147 Padmanabhan, S. *et al.* Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS genetics* **6**, e1001177, doi:10.1371/journal.pgen.1001177 (2010).

148 Johnson, T. *et al.* Blood pressure loci identified with a gene-centric array. *American journal of human genetics* **89**, 688-700, doi:10.1016/j.ajhg.2011.10.013 (2011).

149 Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nature genetics* **41**, 677-687, doi:10.1038/ng.384 (2009).

150 International Consortium for Blood Pressure Genome-Wide Association, S. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103-109, doi:10.1038/nature10405 (2011).

151 Wain, L. V. *et al.* Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature genetics* **43**, 1005-1011, doi:10.1038/ng.922 (2011).

152 Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nature genetics* **41**, 666-676, doi:10.1038/ng.361 (2009).

153 Kato, N. *et al.* Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nature genetics* **43**, 531-538, doi:10.1038/ng.834 (2011).

154 Takeuchi, F. *et al.* Blood pressure and hypertension are associated with 7 loci in the Japanese population. *Circulation* **121**, 2302-2309, doi:10.1161/CIRCULATIONAHA.109.904664 (2010).

155 Odegaard, J. I. & Chawla, A. Pleiotropic actions of insulin resistance and inflammation in metabolic homeostasis. *Science* **339**, 172-177, doi:10.1126/science.1230721 (2013).

156 Beer, N. L. *et al.* The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Human molecular genetics* **18**, 4081-4088, doi:10.1093/hmg/ddp357 (2009).

157 Reaven, G. M. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes* **37**, 1595-1607 (1988).

158 Grundy, S. M. *et al.* Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* **109**, 433-438, doi:10.1161/01.CIR.0000111245.75752.C6 (2004).

159 Ruderman, N. B., Schneider, S. H. & Berchtold, P. The "metabolically-obese," normal-weight individual. *The American journal of clinical nutrition* **34**, 1617-1621 (1981).

160 Karelis, A. D. *et al.* The metabolically healthy but obese individual presents a favorable inflammation profile. *The Journal of clinical endocrinology and metabolism* **90**, 4145-4150, doi:10.1210/jc.2005-0482 (2005).

161 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).

162 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561-568, doi:10.1093/nar/gkq973 (2011).

163 Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**, D808-815, doi:10.1093/nar/gks1094 (2013).

164 Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-2939, doi:10.1093/bioinformatics/btn564 (2008).

165 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

166 Team, R. C. *R: A language and environment for statistical computing*, <http://www.R-project.org/> (2013).

167 Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning* Ch. 14.3.12, 520–528 (Springer, 2009).

168 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720, doi:10.1093/bioinformatics/btm563 (2008).

169 Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-208 (2002).

170 Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS genetics* **7**, e1001273, doi:10.1371/journal.pgen.1001273 (2011).

171 Montojo, J. *et al.* GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* **26**, 2927-2928, doi:10.1093/bioinformatics/btq562 (2010).

172 Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48, doi:10.1186/1471-2105-10-48 (2009).

173 Tomaszewski, M. *et al.* Genetic architecture of ambulatory blood pressure in the general population: insights from cardiovascular gene-centric array. *Hypertension* **56**, 1069-1076, doi:10.1161/HYPERTENSIONAHA.110.155721 (2010).

174 Ho, J. E. *et al.* Discovery and replication of novel blood pressure genetic loci in the Women's Genome Health Study. *Journal of hypertension* **29**, 62-69, doi:10.1097/HJH.0b013e3283406927 (2011).

175 Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human molecular genetics* **19**, 2706-2715, doi:10.1093/hmg/ddq156 (2010).

176 Cho, Y. S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature genetics* **44**, 67-72, doi:10.1038/ng.1019 (2012).

177 Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-874, doi:10.1038/nature08625 (2009).

178 Shu, X. O. *et al.* Identification of new genetic risk variants for type 2 diabetes. *PLoS genetics* **6**, e1001127, doi:10.1371/journal.pgen.1001127 (2010).

179 Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337, doi:10.1093/bioinformatics/btq419 (2010).

180 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).

181 Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**, 369-375, S361-363, doi:10.1038/ng.2213 (2012).

182 Paternoster, L. *et al.* Genome-wide population-based association study of extremely overweight young adults--the GOYA study. *PloS one* **6**, e24303, doi:10.1371/journal.pone.0024303 (2011).

183 Lind, L., Fors, N., Hall, J., Marttala, K. & Stenborg, A. A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study. *Arteriosclerosis, thrombosis, and vascular biology* **25**, 2368-2375, doi:10.1161/01.ATV.0000184769.22061.da (2005).

184 Eijgelsheim, M. *et al.* Genome-wide association analysis identifies multiple loci related to resting heart rate. *Human molecular genetics* **19**, 3885-3894, doi:10.1093/hmg/ddq303 (2010).

185 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124, doi:10.1038/nature11582 (2012).

186 Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118-1125, doi:10.1038/ng.717 (2010).

187 Strachan, D. P. *et al.* Lifecourse influences on health among British adults: effects of region of residence in childhood and adulthood. *International journal of epidemiology* **36**, 522-531, doi:10.1093/ije/dyl309 (2007).

188 Mit, D. G. I. o. B. I. o. H. a. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336, doi:10.1126/science.1142358 (2007).

189 Nelis, M. *et al.* Genetic structure of Europeans: a view from the North-East. *PloS one* **4**, e5472, doi:10.1371/journal.pone.0005472 (2009).

190 Metspalu, A., Kohler, F., Laschinski, G., Ganten, D. & Roots, I. [The Estonian Genome Project in the context of European genome research]. *Deutsche medizinische Wochenschrift* **129 Suppl 1**, S25-28, doi:10.1055/s-2004-824840 (2004).

191 Vartiainen, E. *et al.* Thirty-five-year trends in cardiovascular risk factors in Finland. *International journal of epidemiology* **39**, 504-518, doi:10.1093/ije/dyp330 (2010).

192 Kaprio, J., Pulkkinen, L. & Rose, R. J. Genetic and environmental factors in health-related behaviors: studies on Finnish twins and twin families. *Twin research : the official journal of the International Society for Twin Studies* **5**, 366-371, doi:10.1375/136905202320906101 (2002).

193 Pajunen, P. *et al.* The metabolic syndrome as a predictor of incident diabetes and cardiovascular events in the Health 2000 Study. *Diabetes & metabolism* **36**, 395-401, doi:10.1016/j.diabet.2010.04.003 (2010).

194 Yliharsila, H. *et al.* Body mass index during childhood and adult body composition in men and women aged 56-70 y. *The American journal of clinical nutrition* **87**, 1769-1775 (2008).

195 Eriksson, J. G., Forsen, T., Tuomilehto, J., Osmond, C. & Barker, D. J. Early growth and coronary heart disease in later life: longitudinal study. *Bmj* **322**, 949-953 (2001).

196 Wichmann, H. E., Gieger, C., Illig, T. & Group, M. K. S. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67 Suppl 1**, S26-30, doi:10.1055/s-2005-858226 (2005).

197 Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European journal of human genetics : EJHG* **14**, 79-84, doi:10.1038/sj.ejhg.5201508 (2006).

198 Rantakallio, P. Groups at risk in low birth weight infants and perinatal mortality. *Acta paediatrica Scandinavica* **193**, Suppl 193:191+ (1969).

199 Jarvelin, M. R. *et al.* Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension* **44**, 838-846, doi:10.1161/01.HYP.0000148304.33869.ee (2004).

200 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics* **41**, 35-46, doi:10.1038/ng.271 (2009).

201 Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin research and human genetics : the official journal of the International Society for Twin Studies* **13**, 231-245, doi:10.1375/twin.13.3.231 (2010).

202 Hedstrand, H. A study of middle-aged men with particular reference to risk factors for cardiovascular disease. *Upsala journal of medical sciences. Supplement* **19**, 1-61 (1975).

203 Raitakari, O. T. *et al.* Cohort profile: the cardiovascular risk in Young Finns Study. *International journal of epidemiology* **37**, 1220-1226, doi:10.1093/ije/dym225 (2008).

204 Magi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC bioinformatics* **11**, 288, doi:10.1186/1471-2105-11-288 (2010).

205 Global Lipids Genetics, C. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature genetics* **45**, 1274-1283, doi:10.1038/ng.2797 (2013).

206 Berglund, G. *et al.* Cardiovascular risk groups and mortality in an urban swedish male population: the Malmo Preventive Project. *Journal of internal medicine* **239**, 489-497 (1996).

207 Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *International journal of epidemiology* **35**, 34-41, doi:10.1093/ije/dyi183 (2006).

208 Liu, J. Z. *et al.* Genome-wide association study of height and body mass index in Australian twin families. *Twin research and human genetics : the official journal of the International Society for Twin Studies* **13**, 179-193, doi:10.1375/twin.13.2.179 (2010).

209 Whitfield, J. B., Zhu, G., Nestler, J. E., Heath, A. C. & Martin, N. G. Genetic covariation between serum gamma-glutamyltransferase activity and cardiovascular risk factors. *Clinical chemistry* **48**, 1426-1431 (2002).

210 Whitfield, J. B. *et al.* Measuring carbohydrate-deficient transferrin by direct immunoassay: factors affecting diagnostic sensitivity for excessive alcohol intake. *Clinical chemistry* **54**, 1158-1165, doi:10.1373/clinchem.2007.101733 (2008).

211 Penninx, B. W. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International journal of methods in psychiatric research* **17**, 121-140, doi:10.1002/mpr.256 (2008).

212 Isomaa, B. *et al.* A family history of diabetes is associated with reduced physical fitness in the Prevalence, Prediction and Prevention of Diabetes (PPP)-Botnia study. *Diabetologia* **53**, 1709-1713, doi:10.1007/s00125-010-1776-y (2010).

213 Hofman, A. *et al.* The Rotterdam Study: 2010 objectives and design update. *European journal of epidemiology* **24**, 553-572, doi:10.1007/s10654-009-9386-z (2009).

214 Spector, T. D. & Williams, F. M. The UK Adult Twin Registry (TwinsUK). *Twin research and human genetics : the official journal of the International Society for Twin Studies* **9**, 899-906, doi:10.1375/183242706779462462 (2006).

215 Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS genetics* **8**, e1002607, doi:10.1371/journal.pgen.1002607 (2012).

216 Okauchi, Y. *et al.* Changes in serum adiponectin concentrations correlate with changes in BMI, waist circumference, and estimated visceral fat area in middle-aged general population. *Diabetes care* **32**, e122, doi:10.2337/dc09-1130 (2009).

217 Arai, T. *et al.* Increased plasma cholesteryl ester transfer protein in obese subjects. A possible mechanism for the reduction of serum HDL cholesterol levels in obesity. *Arteriosclerosis and thrombosis : a journal of vascular biology / American Heart Association* **14**, 1129-1136 (1994).

218 Asayama, K. *et al.* Increased activity of plasma cholesteryl ester transfer protein in children with end-stage renal disease receiving continuous ambulatory peritoneal dialysis. *Nephron* **72**, 231-236 (1996).

219 Ebenbichler, C. F. *et al.* Relationship between cholesteryl ester transfer protein and atherogenic lipoprotein profile in morbidly obese women. *Arteriosclerosis, thrombosis, and vascular biology* **22**, 1465-1469 (2002).

220 Tzotzas, T. *et al.* Early decreases in plasma lipid transfer proteins during weight reduction. *Obesity* **14**, 1038-1045, doi:10.1038/oby.2006.119 (2006).

221 Herman, M. A. *et al.* A novel ChREBP isoform in adipose tissue regulates systemic glucose metabolism. *Nature* **484**, 333-338, doi:10.1038/nature10986 (2012).

222 Eissing, L. *et al.* De novo lipogenesis in human fat and liver is linked to ChREBP-beta and metabolic health. *Nature communications* **4**, 1528, doi:10.1038/ncomms2537 (2013).

223 Clemente-Postigo, M. *et al.* Adipose tissue gene expression of factors related to lipid processing in obesity. *PloS one* **6**, e24783, doi:10.1371/journal.pone.0024783 (2011).

224 Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research* **615**, 28-56, doi:10.1016/j.mrfmmm.2006.09.003 (2007).

225 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics* **91**, 224-237, doi:10.1016/j.ajhg.2012.06.007 (2012).

226 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics* **5**, e1000384, doi:10.1371/journal.pgen.1000384 (2009).

227 Casonato, A. *et al.* A common ancestor more than 10,000 years old for patients with R854Q-related type 2N von Willebrand's disease in Italy. *Haematologica* **98**, 147-152, doi:10.3324/haematol.2012.066019 (2013).