PSYCHOLOGY, HEALTH AND STATISTICAL SCIENCES

University of Pavia

Quantifying the genetic component of the metabolic syndrome using a novel proposal score and SNP-based heritability

Ph.D candidate
Francesca Graziano

Tutor:
Prof. Grassi Mario

A. A. 2015-2016 – XXIX ciclo

ABSTRACT

Introduction. Metabolic syndrome (MetS) is a complex, multifactorial disease that poses a major public health problem. MetS increases the risk of coronary heart disease (CHD), atherosclerotic cardiovascular diseases (ASCVD), type 2 diabetes mellitus (T2DM), and all-cause mortality. Currently, there are a many different criteria that define MetS but the physiopathology is not completely understood both in terms of clinical progression and genetic contribution.

Aims. The present work characterizes MetS components (obesity, hypertension, glucose, etc.) as one continuous phenotype and genetic components of the proposed MetS score were estimated using both family-based samples and population-based samples.

Methods. In the first step, Confirmatory Factor Analysis (CFA) was used to select a model with the best fit. After the selection of the best factor structure and development an algorithm to calculate the score, heritability was performed in both pedigrees and SNPs/markers data. For the first sample, SOLAR (Sequential Oligogenic Linkage Analysis Routines) software was used to obtain the estimates. For the second sample, genetic variance components were calculated by fitting a linear mixed model (LMM) using two types of genetic relatedness matrices (Identity-By-Descend, IBD and Genome-Wide Complex Trait Analysis, GCTA), different levels of Linkage Disequilibrium (LD) pruning (0.20 – 0.80 and no LD pruning), and suggestive Genome-Wide Association Study (GWAS) SNPs.

Results. According to the analyses, the best CFA model was the bifactor model; estimated coefficients were used to calculate the MetS score. The score showed good performance and good agreement compared to the International Diabetes Federation (IDF) criteria, the gold standard used for clinical diagnosis.

With regards to the estimation of genetic variance, heritability was significant and ranged from 0.1 to 0.4 in whole samples and in all models. The heterogeneity of the results was due to the different samples and different types of matrix inputs into the LMMs. Heritability obtained using the GCTA matrix was significantly increased compared to the IBD matrix. No significant differences between family data and genetic data (markers) in Sardinia samples were observed using an LD threshold of 0.80 with no pruning.

Conclusions. Evidence of complex interactions in metabolic syndrome and significant genetic contributions were obtained from these analyses. Increased knowledge of the environmental and genetic components could allow for better assessment and identification of patients with this syndrome.

CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

1. INTRODUCTION

1.1 Background

Metabolic syndrome (MetS) is a multi-component human disease that is gradually increasing worldwide, particularly in countries with increasing obesity trends, sedentary lifestyle, and high consumption of calories. MetS represents a major health problem due to the increased the risk of coronary heart disease (CHD), atherosclerotic cardiovascular diseases (ASCVD), and type 2 diabetes mellitus (T2DM), as well as the risk of all-cause mortality (Kassi, Pervanidou, Kaltsas, & Chrousos, 2011). MetS is characterized by chronic low grade inflammation as a consequence of the complex interplay between genetic and environmental factors.

A cluster of interconnected risk factors defines MetS. The core components of metabolic syndrome include the following features: abnormal body fat distribution (high value of waist circumference or BMI>30), insulin resistance (diabetes and elevated glucose levels), atherogenic dyslipidemia (TGR, LDL, HDL), and elevated blood pressure (Systolic and Diastolic Blood Pressure, SBP and DBP, respectively).

However, the predominant risk factors appear to be abdominal obesity and, most importantly, insulin resistance.

Due to the multiple components and clinical implications, there is currently no universally accepted pathogenic mechanism or clearly defined diagnostic criteria for MetS. Additionally, there is no standardized or validated method to assess the severity of aggregated Metabolic Syndrome risk factors and there are no studies with replicated and validated results that examine the genetic contribution of MetS.

1.2 Research questions and thesis outline

Metabolic syndrome and the underlying components reflect a complex polygenic background, interactions of which are not completely understood. For this is reason, the focus of my project is to answer the following research questions concerning the definition of MetS and the genetic influences:

- Which is the best model to describe the cluster of Metabolic syndrome? Is it possible to have a score to determine the degree of pathology?

- What is the genetic variance using the newly proposed score and different types of samples?
- Is the proportion of variation in MetS that is captured by genotyped SNPs compared to one captured by family information?
- Which genetic variants or genes are associated with MetS? Are these results demonstrated in the literature or something new?

2. DEFINITION OF METABOLIC SYNDROME

2.1 Terms and criteria for the diagnosis

There is still debate as to whether this entity represents a specific syndrome or is a surrogate of combined risk factors that put the individual at particular risk.

Several terms have been proposed to describe this clustering: metabolic syndrome, metabolic disorder, syndrome X, insulin-resistance syndrome, etc. The most commonly used term to define this pathology is the first one, metabolic syndrome.

In clinical practice, many organizations attempt to define criteria for diagnosis. Currently, several definitions exist and, due to which factors are emphasized, different components, and relationships are used to describe the pathology **(Table 1**).

The first analysis was performed in 1998 by the World Health Organization (Alberti & Zimmet, 1998; Zimmet, Alberti, & Shaw, 2005). This organization focused the attention on insulin resistance as the major risk factor for diagnosis. One year later, the European Group for Study of Insulin Resistance (EGIR) proposed additional changes to the previous definition, but the focus remained the same (Balkau & Charles, 1999).

In 2001, the National Cholesterol Education Program-Third Adult Treatment Panel (NCEP-ATPIII) introduced alternative clinical criteria that did not identify insulin resistance as the most important evidence (due to laborious measurements), but instead identified the higher long-term risk of ASCVD ("Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report," 2002).

In 2005, and later in 2009, the International Diabetes Federation (IDF) modified the previous criteria set by the ATPIII. A novel feature, abdominal obesity, was introduced as a required characteristic of MetS (Alberti et al., 2009). A summary of the four most commonly used MetS definitions is shown in **Table 1**.

For international comparisons and to facilitate research on MetS etiology, a commonly agreed set of criteria exists that define MetS (Alberti et al., 2009).

According to the most recent consensus statement, patients are diagnosed with MetS if they have three or more of the following features (Grundy, Brewer, Cleeman, Smith, & Lenfant, 2004; Grundy et al., 2005):

- Elevated serum **triglycerides** (≥1.7 mmol/L) or drug treatment for elevated triglycerides.

- Reduced serum **HDL cholesterol** (in men: <1.0 mmol/L; in women: <1.3 mmol/L) or drug treatment for reduced HDL cholesterol.
- Increased **blood pressure** (systolic ≥130 and/or diastolic ≥85 mm Hg) or antihypertensive drug treatment.
- Increased fasting plasma **glucose** (>5.6 mmol/L) or glucose lowering drug treatment.
- Increased **waist circumference** (containing an ethnic specific cut-off point). However, recommendations on cut-off points for Caucasians differ considerably; a waist circumference ≥94 cm for men and ≥80 cm for women, corresponding with the **BMI** cut-off point for overweight, or a waist circumference ≥102 cm for men and ≥88 cm for women, corresponding with the BMI cut-off point for obesity, is recommended (Alberti, Zimmet, & Shaw, 2005).

Table 1. Definition of Metabolic syndrome

	NCEP-ATPIII	WHO	EGIR	IDF
Absolutely required	None	Insulin resistance (IGT, IFG, T2D or other evidence of IR)	Hyperinsulinemia (plasma insulin >75th percentile	Central obesity (waist circumference[≈]): ≥94 cm (M), ≥80 cm (F)
Criteria	Any three of the five below	Insulin resistance or diabetes, plus two of the five criteria below	Hyperinsulinemia, plus two of the four criteria below	Obesity, plus two of the four criteria below
Obesity	Waist circumference: >40 inches (M), >35 inches (F)	Waist/hip ratio: >0.90 (M), >0.85 (F); or BMI >30kg/m^2	Waist circumference: >0.94 cm (M), >0.80 cm (F)	Central obesity already required
Hyperglicemia	Fasting glucose ≥100 mg/dl or Rx	Insulin resistance already required	Insulin resistance already required	Fasting glucose ≥ 100 mg/dl
Dyslipidemia	TG ≥150 mg/dl or Rx	TG ≥150 mg/dl or HDL-C: <35 mg/dl (M), <39 mg/dl (F)	TG ≥177 mg/dl or HDL-C: <39 mg/dl	TG ≥ 150 mg/dl or Rx
Dyslipidemia(second, separate criteria)	HDL cholesterol: <40 mg/dl (M), <50 mg/dl (F)			HDL cholesterol: <40 mg/dl (M), <50 mg/dl (F); or Rx
Hypertension	>130 mmHg systolic or >85 mmHg diastolic or Rx	≥140/90 mmHg	≥140/90 mmHg or Rx	>130 mmHg systolic or >85 mmHg diastolic or Rx
Other criteria		Microalbuminuria		

[≈] Criteria for central obesity are specific for each population; values are given for European men and women

2.2 Prevalence of Metabolic Syndrome

Prevalence of MetS varies and depends on the criteria that are used and the types of populations. Sex, age, race, and ethnicity influence the syndrome in different ways.

Prevalence is relatively high in all populations and is rising worldwide (Aguilar, Bhuket, Torres, Liu, & Wong, 2015). The cause of these increases are related to a number of factors such as demographics, type of lifestyle, diet, and physical activity. MetS prevalence increases dramatically with BMI increment according to gender (van Vliet-Ostaptchouk et al., 2014).

MetS becomes more prevalent with increasing age and is associated with a rise in age-associated diseases and disabilities. Some studies have found that MetS prevalence increases with age through approximately 60-75 years. The plateau in prevalence estimates after the sixth and seventh decade is likely due to survival effect (Cameron, Magliano, Zimmet, Welborn, & Shaw, 2007; Cornier et al., 2008; Ford, Giles, & Mokdad, 2004; Lechleitner, 2008).

Other population studies have confirmed this trend. For example, in the NHANES 2003-2006 cohort, the prevalence estimate of MetS was equal to 20% and 16% (male and female, respectively) in people under the age of 40, 41% and 37% in people aged 40-59 years, and 52% and 54% in people aged 60 years and older (Ford et al., 2004; Kassi et al., 2011).

Prevalence of MetS is also dependent on the definition used, according to race and ethnicity. Values vary dramatically between countries. Differences may be due to cultural differences, different waist circumference thresholds, or different combinations of individual components of MetS used in different populations (Scuteri et al., 2015). For this is the reason, it is currently not possible to estimate only one value of prevalence.

2.3 Risk factors

Many middle-aged people with MetS are at risk of developing ASCVD in the near future (e.g., 10-year risk), have a two-fold increased risk of developing coronary heart disease (CHD), and have a five-fold increased risk of developing T2DM in the next five to ten years (Ford, 2005).

To prevent or delay the onset of ASCVD and diabetes, underlying risk factors need to be modified or removed. There are multiple underlying risk factors for MetS including genetic factors, physical inactivity, over nutrition, and abdominal obesity.

For people with both MetS and abdominal obesity, weight reduction is the first priority. Both weight reduction and maintenance of a lower weight are the best way to prevent metabolic risk factors. Increasing physical activity has beneficial effects on both metabolic and ASCVD risks. Also, reduction of total calories and diet with fruits, vegetables, and grains is encouraged to decrease the risk of MetS, ASCVD, and diabetes (Ervin, 2009; Grundy et al., 2005).

2.4 Clinical manifestations and treatments

Due to the complex pathology, the clinical manifestations are a cluster of conditions with no immediate physical symptoms. Usually people with MetS display central obesity, a strong family history of diabetes mellitus, and insulin resistance.

Currently, no defined therapies are available. The best approach to clinical management is to consider different treatments for each component (e.g., obesity, hypertension, etc.). However, prevention of MetS is important because it increases the risk of several health complications, such as cardiovascular disease, T2DM, non-alcoholic fatty liver disease, youth MetS, and the risk of adult outcomes. Promotion of a healthy lifestyle and family intervention is effective at reducing the incidence of T2DM compared with placebo (Vattikuti, Guo, & Chow, 2012). In addition, three-generation family histories provide evidence of conditions associated with MetS.

2.5 Clustering of MetS

There are multiple interrelated causal mechanisms that underlie MetS development. Knowledge about the mechanisms and the degree of association between MetS components will help the community determine prevention and intervention of cardiovascular disease and diabetes (Cornier et al., 2008). Most likely the clustering of these features is caused by multiple underlying, interrelated causal mechanisms.

Although MetS is believed to have multifactorial causes, the most accepted and unified hypothesis to describe the pathophysiological basis of MetS is insulin resistance and abdominal obesity (Eckel, Grundy, & Zimmet, 2005; Reaven, 1988).

In addition, the pathogenesis of hypertension as a condition of MetS is only partially understood and not considered to be one of the primary causes (Laaksonen et al., 2008). **Figure 1** shows the global mechanisms and the causal connections between MetS features. As shown in **Figure 1**, the core of this disease is defined by obesity and insulin resistance, which work together and are considered by most criteria to be the causes of the syndrome.

Next, hypertension, dyslipidemia, glucose intolerance, and microalbuminuria are recognized as MetS components. Additional risks factors are included in the complex mechanism but they usually are not used in the criteria. Together, these components collaborate to increase the risk of cardiovascular disease and T2DM.

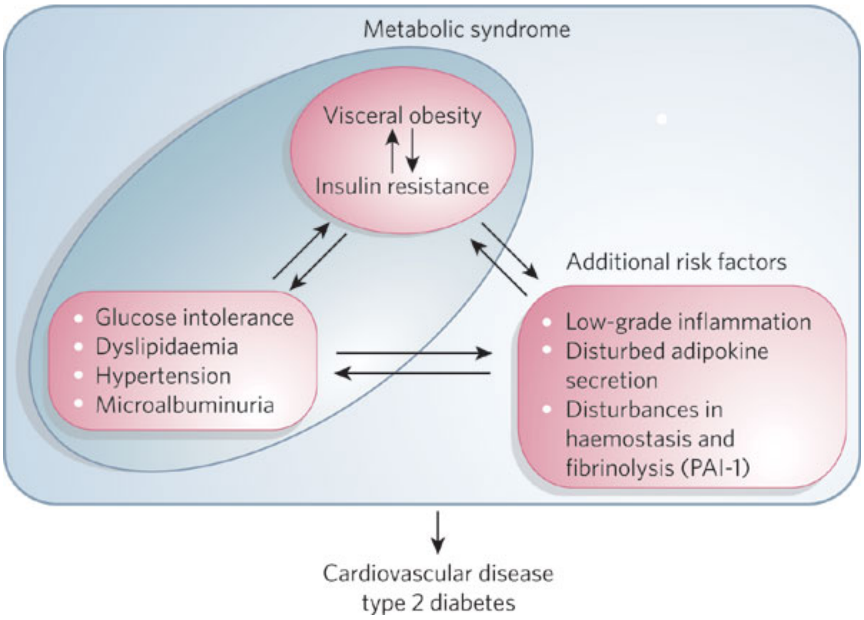


Figure 1 - Mechanisms and the causal connection between its features

2.6 MetS models

In recent years, to examine the pattern of the MetS, several studies used factor analysis, a method that explains the correlation among a set of variables in terms of a smaller set of unobserved "factors". The following techniques are commonly used: Confirmatory Factor Analysis (CFA), which is hypothesis driven data reduction technique, and Exploratory Factor Analysis (EFA), which is a data-driven technique.

In most cases, owing to the explorative and subjective nature of EFA, results of EFA studies on MetS are inconsistent. By contrast, conclusions of CFA studies have thus far been quite consistent, suggesting that MetS features included in the most widely accepted definitions represent a unified disease construct (Shen, Goldberg, Llabre, & Schneiderman, 2006).

Published CFA studies have tested various hypothetical models, including single-factor model, 2-factor, 3-factor, 4-factor (called Correlated CFA models), bifactor CFA models, and hierarchical CFA models, to determine which model best represents the factor structure underlying MetS (Babyak & Green, 2010; Li & Ford, 2007; Martinez-Vizcaino et al., 2010; Pladevall et al., 2006; Shen et al., 2006).

Briefly, as mentioned above, there are different characteristics that describe the CFA models.

The single-factor model is the simplest CFA model. A one-factor model specifies a single dimension underlying a set of measures and, thus, provides a parsimonious explanation for the responses on these measures. **Figure 2** is a graphical presentation of a model with a single factor (F_1) and a number of variables (X_1, X_2, X_3).

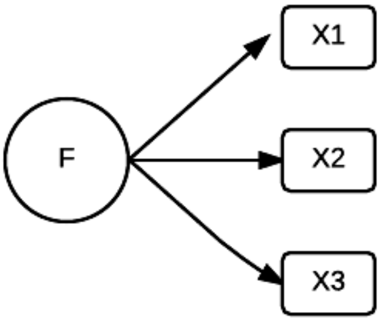


Figure 2 - Single-factor model

The correlated CFA model specifies two or more factors underlie a set of measured variables and that these factors are correlated. **Figure 3** presents a model for five variables with two correlated factors.

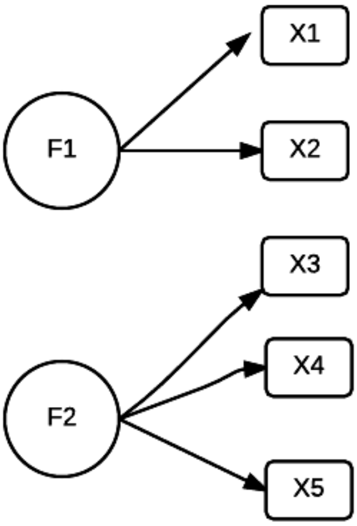


Figure 3 - Two-factor model

The bifactor model may include a general factor associated with all variables and one or more groups of factors associated with a limited number of measures (**Figure 4**).

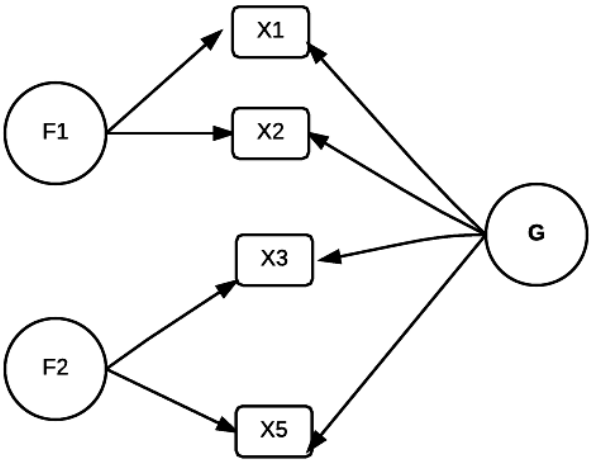


Figure 4 – Bifactor model

The last model, the hierarchical CFA model, usually contains two to four first-order factors and one second-order factor underlying the first-order one (**Figure 5**).

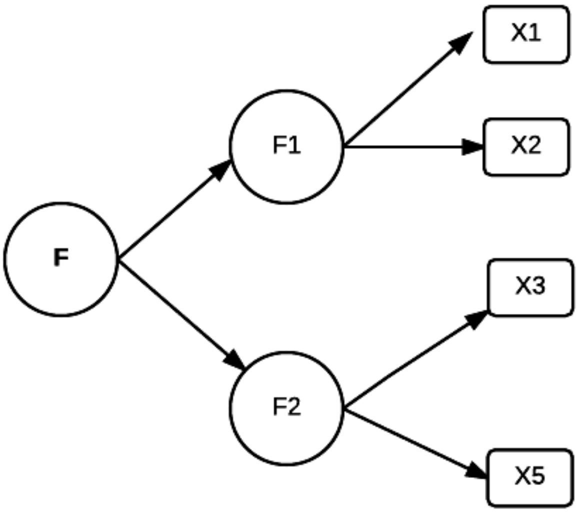


Figure 5 – Hierarchical factor model

In addition, to consider all standard MetS components, several changes to the current MetS definition have been suggested in the scientific literature.

In order to increase the predictive ability of MetS for T2DM and cardiovascular disease (CVD), some studies have proposed to add features to the definition of MetS (Shen et al., 2006). These features include sex and age, which have important roles in the definition of MetS. Among others, some studies suggest adding circulating adiponectin, C-reactive protein (CRP), albumin, APOB, and free fatty acid levels (FFA) or fatty liver (Povel et al., 2013).

Currently, it is unclear if MetS represents one statistical entity after addition of one or more of these features.

Another important point is that in the current binary MetS definition, part of the information is lost. For example, a minor change in triglyceride levels from 1.70 mmol/L to 1.64 mmol/L, could result in an individual no longer being classified as having MetS (Hillier et al., 2006). However, this change in triglyceride levels has only a minor effect on the metabolic profile and the risk for T2DM and ASCVD of this individual. Furthermore, when plotted against the number of positive features, the risk for ASCVD increases continuously, with no suggestion of a threshold effect (Woodward & Tunstall-Pedoe, 2009).

Currently, some groups are working on a new definition that considers metabolic syndrome as a continuous trait (Graziano et al., 2015; Janghorbani & Amini, 2016; Soldatovic, Vukovic, Culafic, Gajic, & Dimitrijevic-Sreckovic, 2016; Wiley & Carrington, 2016).

For example, recently, several authors have developed and validated a continuous MetS score to clustering its components in different ways and with different results (Gurka, Ice, Sun, & Deboer, 2012; Gurka, Lilly, Oliver, & DeBoer, 2014; Ragland, 1992; Wijndaele et al., 2006).

For epidemiological analyses, there are many advantages to using continuous traits instead of binary ones. For example, the binary definition has lower statistical power than the continuous definition; cardiovascular and diabetes risks increase progressively with increasing number of MetS risk factors, whereas using the continuous trait, a cut-off point for the components could be removed. Therefore, the continuous score is less error prone than the binary score (Ragland, 1992).

Table 2 illustrates all recent studies that implement MetS score as continuous trait in adult populations.

As shown in **Table 2**, some characteristics are common for all studies (e.g., waist circumference, HDL, etc.). Only a few studies have added new elements in the definition (e.g., sex or age).

Table 2 - Summary of approaches used to calculate the continuous MetS in adult

STUDY	OBESITY	LIPIDS	GLUCOSE INSULINE	BP	OTHER	STATISTICAL APPROACH
NHANES 1999-2010 (Gurka et al., 2014)	WC	HDL, TGR	-	SBP	-	CFA; one-factor model
D.E.S.I.R. cohort (Hillier et al., 2006)	WC	HDL, TGR	glucose	SBP	-	Principal component analysis
PANIC study, KIHD, DR's EXTRA study (Viitasalo et al., 2014)	WC	TGR, HDL	Glucose, insuline	BP	-	z-scores
Healty Hearts study (Wiley & Carrington, 2016)	WC	TGR, HDL	glucose	SBP, DBP	Sex	Standard deviations and weigth from PCA
Flemish study (Wijndaele et al., 2006)	WC	TGR, HDL	glucose	BP		Summing individual PC scores , each weighted for the relative contribution PC1 and PC2 In the explained variance
Ghana, Nigeria and Kenya (Tekola-Ayele et al., 2015)	WC, BMI	TGR, HDL	glucose	SBP, DBP	-	Sum of standardized residuals of MetS component traits

3. GENETIC EPIDEMIOLOGY OF METS

3.1 Complex genetic disease

When a complex pathology is considered as a whole, both the clinical and genetics aspects must be considered in order to describe it. When diseases are at least partially or mostly heritable, the methods of genetic epidemiology are used to identify phenotypic variability. Some complex, multifactorial diseases can be caused by a combination of genetic and environmental factors.

Genetic epidemiology focuses on genetic predisposition to disease and the joint effects of genetic and non-genetic (environmental) effects on disease risk. This type of research seeks to identify links between disease and genetic factors that increase the risk of disease. Depending on which kinds of scientific questions the researchers want to answer, different study approaches can be used (Cichon et al., 2009). For example, heritability, candidate genes study, and association analysis can be performed to discover new mechanisms.

Table 3 shows some of the common term definitions used in the following chapters. Details of the analysis and definition are illustrated in the Materials and Methods chapter.

Table 3 – Definition of terms

TERM	DEFINITION
Heritability	Proportion of the variance of a trait that is due to genes
Complex disease	Disease caused by of multiple genetic and/or factors
SNPs	Single Nucleotide Polymorphisms; specific position (among 3.2 billion in the genome) where chromosomes carry different nucleic acids
Common SNPs	≥5% frequency. Approximately 10 million in the genome These SNPs are targeted in GWAS
Linkage Disequilibrium	Correlation between SNPS that are close together
Genome Wide association study (GWAS)	A systematic search for common SNPs that influence a disease or traits
Genome Wide SNP chip (array)	A system for assaying 300.000 to 1.000.000 SNPs for an individual subjects, using an array of bead-based or hybridation assay on a glass slide

3.2 Genetic aspects

Briefly, a list of the most important features of MetS (i.e., characteristics of T2DM, dyslipidemia, and obesity) are described below, in terms of both definition and heritability.

3.2.1 Type 2 diabetes

The risk of developing T2DM is approximately 3-4 times higher among first degree relatives of diabetic subjects compared to subjects without a family history of diabetes (Rich, 1990). Similar numbers have been calculated from the offspring of diabetic subjects. If one parent has diabetes, the risk that the offspring will develop the disease is about 40%, and if both parents have diabetes the risk is approximately 70%. This supports the hypothesis that there are familial factors that contribute to the disease and suggests that these factors, to some extent, are additive. Very high concordance rates of T2DM have been reported in monozygotic twins. These studies most likely have overestimated concordance by ascertaining twins based upon disease status, which does not distinguish between familial genetic and non-genetic components. One population-based twin study suggested concordance rates of 34% among monozygotic and 16% among dizygotic twin pairs. Thus, approximately 40% of variability of the diabetic phenotype may be heritable (familial genetic). In one recent study, heritability seemed to be higher for diabetes (0.60), than for diabetes alone (0.26). There are also a few monogenic forms of diabetes with similarities to adult T2DM but that generally develop at earlier ages (American Diabetes, 2009). Maturity onset diabetes of the young (MODY) represents insulin deficient/insulin sensitive forms of T2DM that make up about 5% of all diabetic cases (Olokoba, Obateru, & Olokoba, 2012). MODY is caused by defects in β-cells that eventually lead to insulin deficiency. MODY1 is caused by mutations in the hepatocyte nuclear factor 4α gene (chromosome 20q12-q13.1), MODY2 by mutations in the glucokinase gene (chromosome 7p15-p13), MODY3 by mutations in the hepatocyte nuclear factor 1α gene (chromosome 12q24.2), MODY4 by mutations in the insulin promoter factor 1 (chromosome 13q12.1) and MODY5 by mutations in the hepatocyte nuclear factor 1β gene (chromosome 17cen-q21.3). Diabetes can also develop as a consequence of mutations in the insulin receptor gene (chromosome 19p13.2) or in mitochondrial DNA (tRNALeu). In addition, familial forms of adipose tissue deficiency (partial and congenital lipodystrophy) are associated with diabetes (American Diabetes, 2009).

3.2.2 Dyslipidemia – Lipid metabolism

The lipid metabolism pathway has been shown to play an important role in the genetic background of MetS. Heritability estimates for plasma triglyceride and HDL cholesterol levels range from 20 to 87%. In a recent study that includes twins reared apart, genetic factors contributed to one third of the variability of plasma triglycerides and nearly half of the variability of HDL cholesterol levels. In particular, triglyceride levels appear to be highly influenced by individual-specific environmental factors (Shirali et al., 2016). Several studies have shown that the most important variants associated with lipid metabolism are present in LPL, CETP, ZNF259 genes.

Lipoprotein lipase is encoded by the LPL gene and is expressed in the myocardium, adipose tissue, and skeletal muscle. The LPL gene is located in the short arm (p) of chromosome 8 at position 22 (Mirhafez et al., 2016).

Cholesterylester transferase protein (CETP), encoded by the CETP gene, plays a key role in cholesteryl ester transfer from HDL-C to TG-rich lipoprotein but its role in MetS pathogenesis is not clear. CETP has been reported to play a role in CVD pathogenesis and CETP polymorphisms are associated with MetS. Studies have demonstrated a relationship between CETP polymorphisms and increased risk of Coronary Artery Disease (CAD) (Frosst et al., 1995). The CETP gene is located on the long arm (q) of chromosome 16 at position 21.

The ZNF259 gene is located in the long arm of chromosome 11 (q) at position 23. Zinc finger ZPR1 protein, encoded by the ZNF259 gene, has been shown to affect lipid levels in the blood and ZNF259 polymorphisms have been associated with increased risk of coronary heart disease (Waterworth et al., 2010).

3.2.3 Obesity

Pathogenesis of obesity involves multiple interactions between environmental and genetic factors (Srivastava, Srivastava, & Mittal, 2016). Heritability estimates range between 20%-90% for obesity and between 30%-50% for abdominal obesity. Many of the available estimates include non-genetic familial factors, thus reflecting household effects. In Pima Indians, heritability was 80% for body fat and waist circumference and 50% for BMI (Thompson, Ravussin, Bennett, & Bogardus, 1997). Most studies agree on a heritability of BMI around 50%, and the remaining variability of BMI seems to be largely attributed to shared environmental factors (Dubois et al., 2012). Studies of twins have shown that the propensity to gain weight in response to overfeeding is largely heritable (Garver et al., 2013). In the largest review conducted by Elks et al., the median heritability in siblings was estimated to equal 0.75

(Elks et al., 2012). Monogenic obesity often develops in childhood and progresses over time. Obesity cases due to single mutations have been reported in 11 different genes including leptin, leptin receptor, POMC and MC4R genes (O'Rahilly & Farooqi, 2006). However, several genome-wide association studies (GWAS) have revealed numerous genetic susceptibility loci for obesity risk and some GWAS results have been replicated in different populations (Visscher P , Brown M , McCarthy M , & Yang, 2012).

Evidence also suggests that obesity is influenced by genes that are regulated by other genes. For example, SNPs in the FTO gene that are associated with obesity could be due to linkage disequilibrium between FTO and other genes (Fawcett & Barroso, 2010).

3.3 MetS heritability

One tool that accounts for genetic effects is the estimation of the heritability coefficient, which quantitatively evaluates how much of the phenotypic variance is compatible with a genetic transmission across generations (for more details see chapter 4.4).

In this case, MetS is a complex polygenic disease and the genetic basis of the syndrome is under investigation. Several lines of evidence support a genetic basis for the disease.

Recent studies suggest that complex networks of metabolic pathways modulated by interactions between genetic and environmental factors underlie MetS.

Due to the complexity of the pathology and the many criteria that define MetS status, several studies have evaluated heritability, yielding different results.

Most of the results were obtained through estimation of MetS components independently (e.g., Body Mass Index, Blood pressure, HDL, etc.) (Bosy-Westphal et al., 2007; van Dongen, Willemsen, Chen, de Geus, & Boomsma, 2013). Globally, authors have discovered a moderate to high heritability for all traits (Teran-Garcia & Bouchard, 2007) and significant differences across age and gender have been found.

Other studies have used metabolic syndrome as a binary phenotype. One study conducted by the Jackson Heart Study used ATPIII criteria (Khan et al., 2015) and another study used Dutch data (Henneman et al., 2008) to carry out heritability estimations considering MetS as a binary trait. These studies found significantly different results ranging from 19-38% (Bellia et al., 2009; Henneman et al., 2008; Khan et al., 2015; Lin et al., 2005).

To summarize, studies have shown that genetic effects influence the variability of MetS and indicate that, in representative population-based samples, metabolic syndrome and its components are moderately to highly heritable. Even if several heritability values exist in the literature, no single study has obtained results using Mets as continuous trait.

3.4 MetS GWAS

Representative and significant heritability estimates can be used to obtain an estimate of the total genetic variation of traits but are not informative about loci or associations with particular SNPs. Earlier genome-wide linkage studies reported links between MetS and several chromosomal regions including 1p34.1, 10p11.2, and 19q13.4 (Loos et al., 2003).

The hypothesis that MetS has a genetic component is also supported by GWAS and post-GWAS results that have found significant SNPs and pathways associated with MetS (Kristiansson, 2012; Pollex & Hegele, 2006; Povel, Boer, Reiling, & Feskens, 2011; Wu et al., 2015).

Due to the availability of GWAS catalogues and many GWAS MetS results, systematic reviews have been conducted during the last years (Fall & Ingelsson, 2014; Povel et al., 2011).

Povel et al. (Povel et al., 2011) conducted a systematic review on genes associated with MetS. Using HuGE Navigator and eligibility criteria, they selected 87 articles, including a total of 125 associated genes. At the end of the analysis, authors found evidence for an association with MetS and eight SNPs.

All of these SNPs were also associated with an individual MetS feature, with most SNPs being associated with dyslipidemia. This result suggests that lipid metabolism plays a central role in MetS development.

Comparable results were obtained in a Finnish study. Significant genes from lipid metabolism pathways were found to play a key role in the genetic background of MetS. The authors also found little evidence for pleiotropy linking dyslipidemia and obesity to other MetS traits such as hypertension and glucose intolerance (Kristiansson, 2012).

As in heritability analysis, some GWASs have been conducted by considering each component independently while considering MetS as the main trait.

For example, results published by the STAMPEED consortium carried out GWAS using MetS components individually. SNPs in or near 15 genes were significantly associated with at least one of the 11 traits studied (e.g., BMI, DBP, SBP, and HDL).

Furthermore, MetS was associated with several variants in genes including BUD13, ZNF259, APOA5, LPL, and CETP. This GWAS was conducted using 7 independent studies, comprising 22,161 participants from European ancestry. In this study, five SNPs (BUD13 rs10790162, ZNF259 rs2075290, APOA5 rs2266788, LPL rs295, and CETP rs173539) were associated with MetS (Kraja et al., 2011).

Some GWASs have confirmed and replicated these published results in different populations, while other studies tried to identify new loci associated with the syndrome. This is the case for a GWAS study conducted in 2011 that considered five different populations. Three new loci associated with metabolic disorder were identified in this study (Avery et al., 2011). These loci (APOC1, BRAP and PLCG1) were in or near genes associated with atherogenic dyslipidemia, vascular inflammation, type I diabetes, and central adiposity.

These examples demonstrate that not all genetic variants that explain part of the clustering of MetS features are also associated with MetS itself. In addition, some results have been replicated in different populations for validation, whereas other results appear to reveal new loci.

Currently, due to the variability of the results, it is unclear how and which set of genes contribute to the development of MetS. The genetics of MetS involves a large number of genes with weak effects, however, they may interact with each other and work synergistically with environmental factors (e.g., diet, physical activity, alcohol intake, and smoking) in the pathogenesis of the MetS (Andreassi & Botto, 2003).

No standard genetic test is available that may be used for diagnosis of MetS. The lack of confirmed associations is likely due to the complex interplay between genes and environmental factors that are necessary for expression of this phenotype (Joy, Lahiry, Pollex, & Hegele, 2008). Finally, more analysis considering MetS as a unique pathology that is not divided by each component is needed to better understand the complexity and the interrelationships of this syndrome.

4. MATERIALS AND METHODS

4.1 Populations

Three populations were considered to carry out the following analyses:

- **The Gubbio population** (external validation and heritability using family data)
- **The Sardinia population** (MetS score, heritability, and GWAS)
- **The ARIC sample** (heritability using SNP information)

Gubbio Population Study is a prospective epidemiological investigation on blood pressure and cardiovascular risk factors that began in 1983 and concluded in 2007 in Gubbio, a town in central Italy. Three surveys (between 1983-1985, 1988–1992, and 2001–2007) were conducted over the course of 25 years (Cirillo et al., 2014; Menotti et al., 2009).

These surveys targeted patients aged 5 years or older, living within medieval walls with their close relatives living outside. Among the 6,831 participants, 51.78% were residents within the medieval city (Cirillo et al., 2014).

Information on demographic, clinical, anthropometric, and environmental variables was collected. All participants provided their informed written consent. At each survey, genealogical information was also registered and updated through a structured interview administered to each participant. From these data, nuclear and extended pedigrees were drawn. Nuclear pedigrees are two-generation families with first-degree relationships, that is, parent–offspring and/or siblings. Drawings of extended pedigrees were carried out to include: three generations when applicable, the nuclear family of spouses in the second generation, and, consequently, all first cousins (maternal and paternal) in the third generation (Khoury, Beaty, & Cohen, 1993). Data from the last survey carried out in 2001-2007 have been considered in this analysis due to comparability of data collected in the Sardinia population in terms of span of years.

The Sardinia Population Study was a large population-based epidemiologic survey carried out in villages of the Ogliastra region in Sardinia, Italy, between 2002-2008 (Biino et al., 2011; Cappello et al., 1996).

People were invited to participate by means of public advertisement and letters sent to every family. Samples of blood, anthropometric and blood pressure measurements, and bioelectrical impedance analyses were collected. In addition, a standardized interview including sociodemographic, lifestyle, medical, and pharmacological history was obtained (Cappello et al., 1996). All participants provided informed written consent.

Among the 12,517 subjects, 8,102 (3,485 men and 4,617 women older than 18 years) were included in the analysis due to complete information on their MetS components. MetS score algorithm implementation was conducted using this sample size.

Some of the whole phenotype data, equal to 1,270 subjects, also contained genetics (SNPs) and pedigree information.

Analyses were performed using these types of samples because both of these populations have particular characteristics: no immigration and isolated populations (details are included in the Discussion chapter).

The Atherosclerosis Risk in Communities Study (ARIC), sponsored by the National Heart, Lung and Blood Institute (NHLBI), was a prospective epidemiologic study conducted in four U.S. communities (Investigators, 1989). ARIC was designed to investigate the causes of atherosclerosis and its clinical outcomes, and variations in cardiovascular risk factors, medical care, and disease by race, gender, location, and date. It consisted of a large sample of unrelated individuals and some families across North America. Specifically, the population was recruited from four centers across the United States: Forsyth County, North Carolina; Jackson, Mississippi; Minneapolis, Minnesota; and Washington County, Maryland. For this study, a restricted subgroup of European-Americans was considered. The ARIC population consisted of 8,592 unrelated subjects.

4.2 Outcomes

4.2.1 MetS components

Weight was determined on a portable electronic scale to the nearest 0.1 kg, height was measured to the nearest 0.5 cm with a stadiometer, and waist circumference was measured at a mid-distance between iliac crest and rib cage and was rounded to the nearest 0.1 cm. Body mass index (BMI) was calculated as kg/m^2. Bioelectrical impedance (BIA 101, RjL/Akern Systems, Detroit, MI), measuring the resistance and reactance, was used to determine many body composition parameters including fat mass percentage. SBP and DBP were measured in both arms with a standard mercury sphygmomanometer (Miniatur 300 B, Speidel & Keller) according to the ESH guidelines ("2003 European Society of Hypertension-European Society of Cardiology guidelines for the management of arterial hypertension," 2003). The arm with the higher pressure was used subsequently and the average value was obtained. Biochemical analyses, including HDL, TGR and fasting blood glucose levels, were obtained in a central laboratory.

4.2.2 MetS scores

After selection of the CFA model, a MetS continuous score including gender was calculated using a newly proposed equation.

The MetS binary variable was created using the IDF criteria to compare our results with the gold standard.

4.2.3 Genotyped data

DNA samples were isolated from blood of Sardinia and white American participants and genotyped using 500K Affymetrix Genome-Wide Human SNP 6.0 Array.

Experiments were performed using the following recommended protocol as described in the Affymetrix manual. Briefly, total genomic DNA (500 ng) was digested with Nsp I and Sty I restriction enzymes, ligated to adaptors, and amplified using a primer that recognizes the adaptor sequence. Amplified DNA was then fragmented, labeled, and hybridized to oligonucleotide probes attached to the surface of an array in a GeneChip Hybridization Oven 640 (Affymetrix, Inc. Santa Clara, CA, USA), followed by washing and staining procedures performed on a GeneChip Fluidics Station 450. Arrays were finally scanned using the GeneChip Scanner 3000 7G (Affymetrix, Inc.) (LaFramboise, 2009).

The obtained 321 CEL intensity files were analyzed with executables included in the Affymetrix Power Tools package (APT version 1.12.0). Quality Control was performed using the Contrast QC algorithm.

4.3 Statistical analysis

A diagram of the analytical process is presented in **Figure 6**. Analyses were divided in two macro areas, including the clinical and genetic aspects.

Specifically, the first step was to analyze the clinical aspect of the syndrome. Using the CFA model and comparisons between models, the best model that described the syndrome globally was chosen. Subsequently, using the CFA results, a newly proposed equation was used to obtain a score. External validation was performed to guarantee the effectiveness of the newly proposed score.

As shown in the flow-chart, the second part of this study included the estimation of heritability due to the availability of pedigree and GWAS data. Finally, previous GWAS analysis of the new quantitative traits was carried out. All data analyses were analyzed using R (v. 3.2.1) software (Team, 2005).

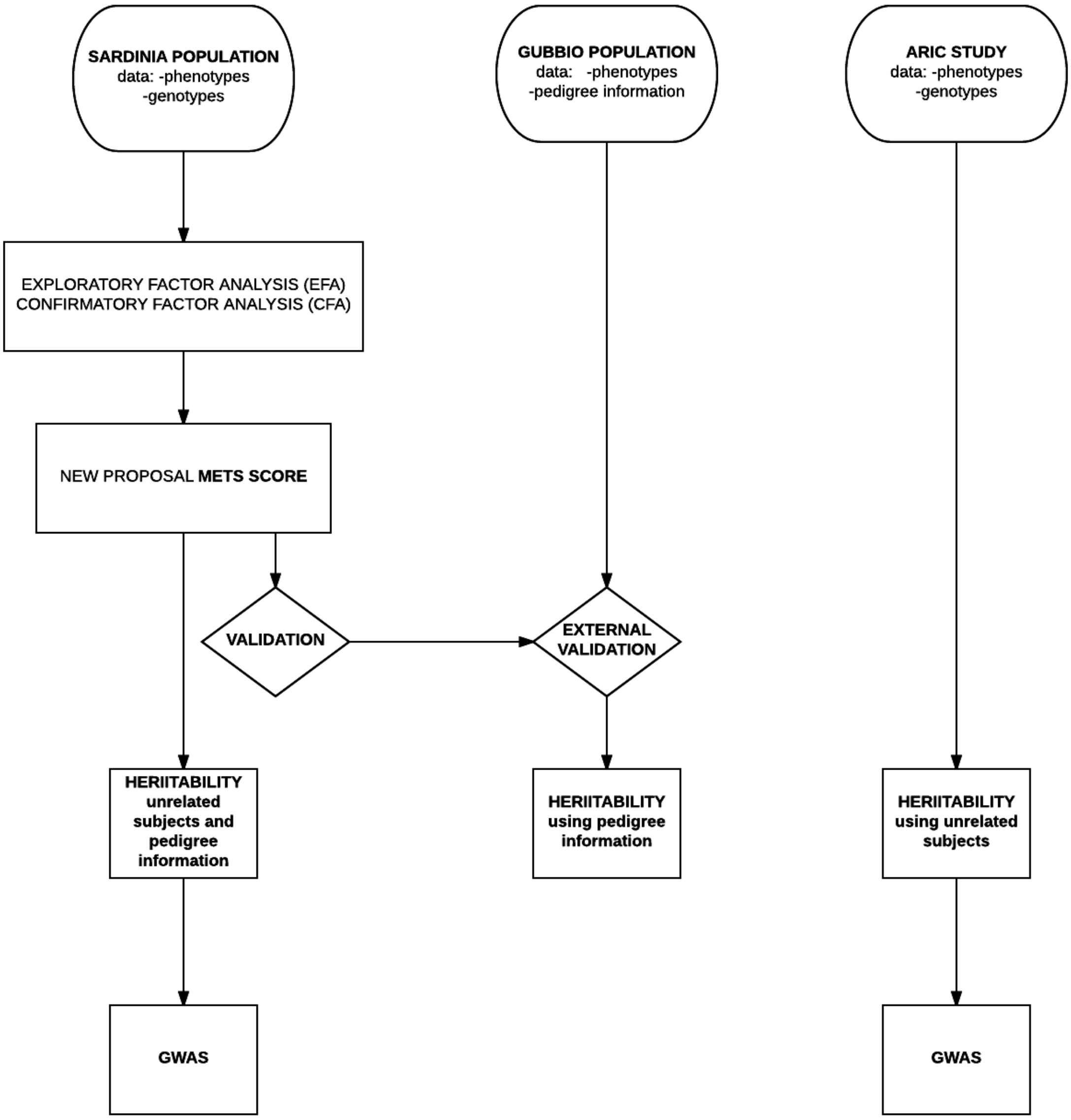

SARDINIA POPULATION
data: -phenotypes
-genotypes
GUBBIO POPULATION
data: -phenotypes
-pedigree information
ARIC STUDY
data: -phenotypes
-genotypes
EXPLORATORY FACTOR ANALYSIS (EFA)
CONFIRMATORY FACTOR ANALYSIS (CFA)
NEW PROPOSAL METS SCORE
VALIDATION
EXTERNAL VALIDATION
HERIITABILITY unrelated subjects and pedigree information
HERIITABILITY using pedigree information
HERIITABILITY using unrelated subjects
GWAS
GWAS

Figure 6 - Flow chart of steps taken in statistical analysis

4.3.1 Descriptive analysis

The normal distribution was determined using Kolmogorov-Smirnov test. Descriptive statistics were presented as the mean ± standard deviation. In addition, numbers and percentages were determined for all variables. Frequencies of MetS components were summarized for each population.

For normally distributed variables, Student's *t*-test was used to compare gender differences. If needed, data are given for men and women, separately.

4.3.2 MetS score from the Sardinia population

As described in the Introduction chapter, EFA and CFA were performed in the Sardinia population. Four-factor model, bifactor, and hierarchical models were analyzed and compared to select for the best one.

Using results from the selected model, an algorithm used to compute MetS as a continuous variable that summarizes clinical parameters could then be proposed (Graziano et al., 2015). This algorithm summarizes waist circumference, BMI, blood pressure, blood glucose, HDL-cholesterol, and triglycerides into one quantitative phenotype. In this way, the syndrome could become clinically interpretable and useful for evaluation of the association with cardiovascular diseases and for investigating genetic components.

After fitting different CFA models (CFA with a single factor, correlated CFA, bifactor CFA, and hierarchical CFA), several goodness-of-fit criteria were used to choose the best one (Kline, 2015):

- comparative fit index (CFI) >0.9,
- standardized mean square residual (SMSR) closer value to 0,
- root mean square of approximation (RMSEA) value < 0.08 and,
- the smallest AIC/BIC.

Using results of the CFA model, a newly proposed equation was used to obtain a score that summarizes its components. If differences between gender were discovered, two equations were reported.

4.3.3 External validation

After proposing a new score, validation analysis using ROC curve and external validation were carried out to evaluate the performance of the score. A new cut-off point using Youden's Index was also proposed to dichotomize the trait (Weng, 2001).

Characteristics of the population score values were analyzed and compared with the gold standard. In particular, the following score characteristics were verified in the sample:

- If the score computed using the proposed algorithm ranged between the predicted 0 and 100.
- If it is normally distributed.

4.4 Theories about the heritability analysis

4.4.1 Definition of heritability (h^2)

To facilitate the estimation of genetic components and to understand the architecture of the disease, heritability of the newly proposed score was carried out using both pedigrees and marker/SNPs information.

The concept of heritability was introduced by Fisher and Wright (Norton & Pearson, 1976; S. Wright, 1921) to refer to variance for phenotypes that is explained by sharing of genomic regions. Based on this theory, twin studies are useful because twins are exposed to the same environmental factors. This reduces environmental variability and genetics can therefore be better quantified.

Many definitions of heritability have been proposed but in general, heritability represents the amount of variation in a phenotype that is influenced by genetic variation. Mathematically, heritability is defined as the proportion of variance for a phenotype that is explained by sharing genomic regions. Specifically, it is defined as the proportion of total variance in a population for a particular measurement, taken at a particular time or age, that is attributable to variation in additive genetic or total genetic values (termed the narrow-sense heritability or just heritability, h^2), or the broad-sense heritability (H^2), respectively (Visscher, Hill, & Wray, 2008).

In statistical models, observed phenotypes take into account the contribution of unobserved genotype (G) and unobserved environmental factors (E):

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)}$$

The variance of the observed phenotypes (σ_P^2) can be expressed as a sum of unobserved underlying variances, σ_G^2 and σ_E^2.

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

Broad-sense heritability (H^2) is defined as $H^2 = \sigma_G^2/\sigma_P^2$, which is the proportion of genotypic variance that is responsible for the proportion of phenotypic variance.

However, genetic variance can be partitioned into the variance of additive genetic effects (breeding values; σ_A^2), of dominance genetic effects (interactions between alleles at the same locus; σ_D^2), and of epistatic genetic effects (interactions between alleles at different loci; σ_I^2) (Vinkhuyzen, Wray, Yang, Goddard, & Visscher, 2013):

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

If the narrow-sense heritability does not include epistatic and dominance effects, then the final formula is $h^2 = \sigma_A^2/\sigma_P^2$.

h^2 is the parameter usually used to indicate heritability because dominance and other non-additive genetic effects that are based on sharing two copies do not contribute to phenotypic resemblance. This is because individuals transmit only one copy of each gene to their offspring and then relatives share only one or no copy that are "identical-by-descent" (IBD). Identical twins and sibs are the most important exceptions.

In the equation $h^2 = \sigma_A^2/\sigma_P^2$, the numerator and denominator need attention for a correct assignment. The denominator contains the total observed variation, excluding variation that is due to known fixed factors and covariates, such as sex, age, and cohort. The numerator contains variation that is due to genetic additive values in the population. They are defined as the sum of the average effects of parental genes that give rise to the mean genotypic value of their progeny. Narrow sense heritability is time and population specific.

Traditionally, heritability is estimated from simple designs, such as simple functions of the regression of offspring on parental phenotypes, the correlation of full or half sibs, and the difference in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs. Recently, availability of genotype information on a large number of loci has made it possible to estimate genetic contribution using genetic relatedness among unrelated data.

This is useful because instead of testing the effect of each SNP independently on the trait like a GWAS (Genome-wide association study) or CVAS (common variant association study), a total variance explained by fitting all SNPs simultaneously can be obtained.

4.4.2 Heritability using LMM

Briefly, the classical method to estimate heritability is based on a simple assumption about correlations between relatives such as: $cor\left(Y^{(relative\ i)}, Y^{(relative\ j)}\right) = A_{ij}\, h^2$, where A_{ij} is a coefficient that depends on the pedigree relationship. For example, the coefficient equals unity if the relatives are monozygotic twins, ½ if they are parents and offspring, ¼ if they are uncle (aunt) and nephew (niece), and so on. The use of the empirical correlation to estimate h^2 does not take into account or resemble each other, not only for genetic reasons, but also for environmental ones (James J. Lee, Vattikuti, & Chow, 2016; Zuk, Hechter, Sunyaev, & Lander, 2012).

Alternative methods for estimating the total heritability attributable to addittive common variants (i.e., the narrow-sense heritability) is via the Linear Mixed Model (LMM), also called Mixed Linear Models (MLMs) (Bonnet, Gassiat, & Lévy-Leduc, 2014; de Los Campos, Sorensen, & Gianola, 2015; David Golan & Rosset, 2011; Hall & Bush, 2016; Hu & Yang, 2014; J. Yang et al., 2010; J. Yang, Manolio, et al., 2011). The model is mixed because it jointly accounts for fixed (*β*) and random (g_a and *e*) effects in the equation:

$$y = X\beta + g_a + e$$

where y is the vector $(nx1)$ of phenotypic trait, *X* is the matrix (nxp) of observed covariates (i.e., sex, age, principal component of genetic substructure, etc.) corresponding to the fixed effect in *β*, g_a is the vector $(nx1)$ of additive random effects (the degree of genomic individual sharing), $g_a \sim MVN(0, \sigma_A^2 G)$, and *e* is the vector $(nx1)$ of residual random effects (representing environmental, non-genetic effects), $e \sim MVN(0, \sigma_E^2 I)$.

In particular, elements in the vector g_a are correlated because they include the known sharing gene information (*G*), the genetic variance is assumed equal to the sum of the squares-effect sizes of *S* loci: $\sigma_A^2 = \sum_{s \in S} a_s^2$, and the environmental contribution to phenotype is assumed equal to σ_E^2 for all the subjects. Therefore, the mean vector and the covariance matrix of *y* are:

$$E(y) = X\beta \quad \& \quad V(y) = \sigma_A^2 G + \sigma_E^2 I$$

The recently developed LMM approach (J. Yang et al., 2010) defines a polygenic additive model with many markers of small, null, and outlier markers with large effects. The key assumption is that we are interested in the value $\sum_{s \in S} a_s^2$, rather than each of the individual effects size a_s , these effect sizes may be regarded as "nuisance" offset parameters.

The variance of the random effects (σ_A^2 and σ_E^2, or the heritability: $h^2 = \sigma_A^2 / \sigma_P^2$, and the total variance: $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$) are typically fit using Maximum Likelihood (ML), or REstricted Maximum Likelihood estimation (REML), which are iterative methods that find the best fit for the model. A likelihood ratio test (LRT) is performed, examining the significance of the genetic random effects on the fit model, yielding *P*-values.

Depending on which type of data is available for the analysis, differences in estimation of the genetic relationship matrix, *G* is reflected in the evaluation of heritability. Thus, for pedigree data where relationships are known, elements of G are the coefficient of actual genetic relatedness derived from probabilities of Identity-By-Descent (IBD), $G = K_{PED}$. In the population data where dense marker data are known but family information is unknown, elements of G are the expected IBD based on the Identity-By-State (IBS) coefficients, $G = K_{IBD}$, or the Genetic Relationship Matrix (GRM) between pairs of subjects that are captured by observed markers, $G = K_{GRM}$. The heritability using GWAS data is sometimes called "SNP-heritability".

4.4.3 Estimating Identity-By-Descend (IBD) matrices

To identify the degree of relatedness between each pair of study samples, Identity-By-Descend IBD estimation can be done using either the method of moments (MoM) (Purcell et al., 2007) or maximum likelihood estimation (MLE) (Milligan, 2003).

Likelihood estimators (MLE) are based on a probability models of the sampled data. In this case, the unit of sampling is a pair of individuals, each one of which has been assayed genetically at L loci. The estimator is based on the assumption of independently segregating marker loci. The likelihood for the overall sample, therefore, is simply the product of the likelihoods across the loci.

In MoM, a correction factor based on allele counts is used to adjust for sampling. However, if allele frequencies are specified, no correction factor is conducted since the specified allele frequencies are assumed to be known without sampling. In particular, Pr(IBD=0)=k0, Pr(IBD=1)=k1, = Pr(IBD=2)=k2, and each IBD coefficient pair is calculated from $0.25k_1 + 0.5(1 - k_0 - k_1)$.

Although MLE estimates are more reliable than MoM, the IBD MoM method is computationally more efficient relative to MLE. For these computational reasons, only IBD MoM was used in the analysis.

4.4.4 Estimating genetic relationship matrices (GRMs)

Estimates of genetic sharing across study samples using GWAS datasets is often represented as a *genetic relationship matrix* (GRM).

If the information captured by the GWAS dataset represented 100% of all genetic variation, this analysis would yield a perfectly accurate estimate of trait heritability. Because genotyping technologies do not capture all genetic variants, the estimation of shared genetic variations are limited to the genotyped information. Thus, when properly adjusted for factors, the variance explained by GWAS-genotyped SNPs can be considered a surrogate of the heritability due to additive genetic effects (narrow sense heritability).

In this way, the most common GRM is implemented by Yang et al., called Genome-wide Complex Trait Analysis (GCTA) (J. Yang, Lee, Goddard, & Visscher, 2011). This tool has been followed to estimate heritability using marker information. Due to large-scale, genome-wide single nucleotide polymorphism (SNP) genotyping sample sizes, the computational process can be very intensive.

As mentioned previously, one of the core functions of GCTA is to estimate the genetic relationships between individuals from the GWAS SNPs. In practice, the genetic relationship between the individuals j and k is typically based on the additive sharing of alleles across all (N) genotyped SNPs, according to the following equation:

$$A_{ik} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

where x_{ij} is the number of copies of the reference allele for the i^{th} SNP on the j^{th} individual, and p_i is the frequency of the reference allele. Importantly, this model of relatedness assumes an additive effect for SNPs and is typically created using autosomal SNPs only.

To avoid the possibility of including the non-genetic effect, authors suggest to exclude closed subjects (one individual of a pair whose relationship is greater than a specified cut-off value, e.g., 0.025).

GCTA is also sensitive to linkage disequilibrium (LD). Heritability can be under or overestimated in influential regions with high or low LD. This correction is currently under debate, (S. Lee et al., 2011) suggesting that LD has a relatively minimal effect. (Hill & Maki-Tanila, 2015) However, other authors argue that in region where there is high LD near causal variants, heritability is overestimated, which is the opposite for areas of low LD.

Thus, LD pruning before estimation of GRM can be used to filter SNPs using an LD threshold. In this way, heritability can be estimated by decreasing the potential for confounding due to LD.

4.4.5 Phenotype Correlation-Genotype Correlation (PCGC)

Recently, another ingenious method to estimate heritability was developed by Golan (D. Golan, Lander, & Rosset, 2014). It is called phenotype correlation-genotype correlation (PCGC) regression and it is derived from the well-known Haseman-Elston regression method (Haseman & Elston, 1972).

Under the previous additive LMM for convenience without fixed effects, and using "normalized" phenotypes y_i and genotypes x_i, where the values y_i and x_i have been centered to have mean 0 and standardized to have variance 1, we have the following relationship:

$$cor(y_i; y_j) = E(y_i; y_j) = h^2 G_{ij}$$

where G_{ij} is the genetic correlation between individuals i and j; given by:

$$G_{ij} = cor(x_i; x_j) = \frac{1}{N} \sum_{k=1}^{N} x_{ik} x_{jk}$$

Thus, under additive model with random effects, the estimated slope of the regression of the empirical phenotypic correlation $(y_i; y_j)$ onto the genetic correlation $(x_i; x_j)$ is the heritability of the trait, h^2.

The PCGC regression approach for estimating heritability is easy to understand and implement; it produces unbiased estimators. More precisely, the PCGC regression estimator is a moments-based estimator; it looks at pairs of individuals at a time (Mitchell et al., 2015) and the covariates can be considered using the phenotype residuals before the PCGC.

4.4.6 Beyond heritability analysis

Each of the heritability estimation methods described above make different assumptions about the model generating phenotype, discussed in detail by Zaitlen et al., Lee et al., and Yang et al. (Zaitlen & Kraft, 2012) (J. J. Lee & Chow, 2014; S. Yang et al., 2014). Here, we reported some hot-spots beyond heritability.

A. Choose unrelated subjects

One of the core functions of the GTCA method is to estimate the genetic relationship matrix, leaving out closer relatives (e.g., 3rd cousins or closer; cut-off = 0.025). The reason was to avoid the possibility that the resemblance between close relatives could be due to non-genetic effects (shared environment). In this case, heritability includes not only genetic effects but also environmental components. The estimate of heritability from an analysis with many close relatives would be similar to the estimate using only those relatives and fitting an AE model, excluding the common environment in the ACE model (v. par. 4.5). Such an analysis would not tell us something new and would not be informative with respect to variation due to causal variants that are in LD with common SNPs (Visscher, Yang, & Goddard, 2010). Kumar et al. claim that this filtering is subject to a lot of error, and does not resolve the cryptic relatedness in the observed GTCA (Kumar, Feldman, Rehkopf, & Tuljapurkar, 2016).

B. Population structure

The problem of confounding by population structure, family structure, and cryptic relatedness in heritability analysis is widely appreciated. Statistical methods for correcting these confounders include linear mixed models (LMMs), genomic control, family-based association tests, structured association, and eigenstrat (Zhou & Stephens, 2012). Compared to other methods, LMMs can capture all of these confounders simultaneously, without knowledge of which are present and without the need to tease them apart.

Shin and Lee suggest that the mixed model methodology was useful to reduce spurious genetic associations produced by population stratification, even with a high degree of admixture (Shin & Lee, 2015). To achieve these goals, the authors simulate datasets based on the HapMap data under various scenarios. Results indicate that the mixed-model approach performs well in controlling for population structure/admixture. It has a similar performance as that based on Principal Component Analysis (PCA). However, the approach combining mixed-model and principal component analysis does not perform as well as either method itself (Liu, Zhao, Patki, Limdi, & Allison, 2011). Correction for population structure and selection of unrelated subjects to obtain an unbiased estimation of heritability is still under debate.

C. Missing heritability

To summarize, different values of variance explained can be obtain based on availability of sample size and type of population (family or population-based samples). The discrepancy between estimates of h^2 from studies of pedigrees and the percentage of the variance ascribable to phenotype-associated SNPs identified with high confidence in GWAS is called “missing heritability”. The differences between two types of estimation are still unknown but have been studied previously (Blanco-Gomez et al., 2016; Manolio et al., 2009; Zuk et al., 2012). Zaitlen and Kraft define “bottom-up” heritability, the h^2 computed given a GWAS using the effect size estimates from the markers with a pre-specified genome-wide significance level ($P < 10^{-5}$, or $P < 10^{-6}$), and “top-down” heritability, the h^2 computed with all the markers, and their ratio is a crude relative measure of missing heritability (Zaitlen & Kraft, 2012).

D. h^2 dissemination

Regarding interpretation of heritability values, in general it ranges from 0 to 1, from no heritability trait (0%), which means no genetic effect, to full heritability trait (100%), which means strong genetic contribution. Even if there are many heritability studies, it is difficult to compare the results and make conclusions, but it is very useful to the investigation of the architecture and for discovering specific genetic component of the complex traits (James J. Lee et al., 2016). The GTCA method is quite robust. It should remain a valuable tool in quantitative genetics for some time to come (J. J. Lee & Chow, 2014).

4.5 Heritability analysis in the three study samples

Using family data (Gubbio and Sardinia) and two GWAS datasets (ARIC and Sardinia), heritability of MetS scores were estimated in two ways:

- using pedigree design (both nuclear and extended families) and
- using population design.

4.5.1 Family data (Gubbio and Sardinia)

Due to the availability of pedigree data in Gubbio and Sardinia samples, Maximum-likelihood heritability estimates were performed by classical ACE models.

In these models, parameters are estimated under the assumption that the variance of the trait is attributable to a combination of non-shared environmental factors (E), common environmental variance (C), and additive genetic variance (A). Combinations of this parameters are denoted as follows:

E model = null model,
AE model = additive + environmental,
CE model = no genetic component, and
ACE model = full model.

Each hypothesis was tested against the null model by use of the likelihood ratio test computed as $-2[lnL_{hyp} - lnL_{null}]$, where lnL is the log-likelihood. The likelihood ratio has a χ^2 distribution with the degrees of freedom equal to the number of parameters in the null model minus the number of parameters in the ACE/AE/CE models.

All models were adjusted for age, but not for gender because it is included in the MetS equation. Sibs-household (C^2) and household effects (C^2) were considered in nuclear and extended pedigrees, respectively.

4.5.2 Unrelated individual data (Sardinia and ARIC)

Estimation of heritability was carried out under a number of different scenarios. In particular, for both Sardinia and ARIC samples, two types of matrices were considered to impute the relationship of unrelated subjects: the IBD and GCTA matrices.

All models were adjusted for age and principal components (PCs) if necessary.

LD pruning with two different thresholds were utilized to estimate both IBD and GCTA matrices. Pruning with LD equal to 0.20, 0.80 and no threshold were used to capture the differences between heritability and linkage disequilibrium.

Also, these analyses, considering three LD thresholds, were repeated with Phenotype Correlation-Genotype Correlation (PCGC) using the same GRM GCTA as impute to estimate heritability.

4.5.3 Genome-Wide Association Study (GWAS) (Sardinia and ARIC)

Estimation of "bottom-up" heritability in Sardinia and ARIC samples was computed by classical GWAS analysis (Barsh, Copenhaver, Gibson, & Williams, 2012; H. Zhang, Liu, Wang, & Gruen, 2007). In brief, GWAS analysis of genotyped data was composed by the following essential steps:

- data pre-processing and quality control (QC),
- principal component analysis (PCA), and
- association analysis for typed data.

A. Data pre-processing

Several Quality Control (QC) procedures were performed on the genotype data. This step involves both SNPs, sample, and family filtering. In this step, selected SNPs and individuals were excluded from analysis (Wang, Barratt, Clayton, & Todd, 2005; W. Zhang et al., 2008).

In the SNP-level filtering, the call rate for a given SNP is defined as the proportion of individuals in the sample for which the corresponding SNP information is not missing. In the Sardinia population, we used a call rate filter equal to 95%; the minor allele frequency (MAF) refers to the frequency at which the second most common allele occurs in a given population. SNPs for which the MAF was less than 1% were removed. Another pre-processing filter applied to this population was the Hardy-Weinberg Equilibrium (HWE). This type of equilibrium says that alleles and genotype frequencies in a population remain constant from generation to generation in the absence of other evolutionary influences. HWE is generally measured at a given SNP using a chi-squared goodness-of-fit test between the observed and expected genotypes. SNPs which had a HWE test statistic corresponding to a *P*-value less than $1x10^{-6}$ were removed.

Another important concept is Linkage Disequilibrium (LD). It refers to a nonrandom assortment of alleles at two loci. LD pruning was performed using a several threshold values (0.20, 0.50, and 0.80) to eliminate a large degree of redundancy in the data and reduce the influence of chromosomal artifacts.

In the sample-level filtering, call rate, similar to SNP-level filtering, refers to exclusion of individuals who are missing genotype data across more than a pre-defined percentage of typed SNPs. The threshold for Sardinia population was equal to 95%.

For family-based data only, families with more than 5% Mendel errors (considering all SNPs) and SNPs (i.e., based on the number of trios) with more than 10% Mendel error rate were discarded.

B. Principal Component Analysis (PCA)

Another parameter evaluated was ancestry, which was carried out using Principal Component Analysis (PCA). PCA is one approach used to visualize and classify individuals into ancestry groups based on reference panels (Liu et al., 2011; Reich, Price, & Patterson, 2008). Sardinia and ARIC populations were compared with CEPH (Utah residents with ancestry from northern and western Europe) (abbreviation: CEU in black) and other population, Yoruba in Ibadan, Nigeria (abbreviation: YRI); Japanese in Tokyo, Japan (abbreviation: JPT); and Han Chinese in Beijing, China (abbreviation: CHB). The first PCs capture information of latent population substructure and usually the first 10 PCs are considered as possible confounders (conservative approach). For Sardinia and ARIC, we selected the first PC because no evident sub-structures were revealed (see par. 5.6).

C. Model used to test association

Association analysis for the MetS score was conducted by linear regression analysis, adjusting for age, PCs, and for family information (using IBD matrix) in Sardinia and ARIC samples. A single additive model was selected; each SNP was represented as the corresponding number of minor alleles (0, 1, 2).

The fixed threshold for genome-wide significance was set at the consensus level of $P < 5\times10^{-8}$ and a genome-wide suggestive P-values if $1\times10^{-5} > P > 5\times10^{-8}$ because of the larger number of tests conducted in a genome-wide survey. These thresholds help to avoid false positives and ensure that reported associations in other samples from the same population. The typed SNPs identified as genome-wide significant ($P < 10^{-4}$) were used to estimate the "bottom-up" heritability using both IBD and GCTA matrices.

All data analyses were analyzed using R (v. 3.2.1) software (Team, 2005). ACE models were performed by function implemented in SOLAR (Sequential Oligogenic Linkage Analysis Routines) software package Eclipse version 7.6.4 (Almasy & Blangero, 1998).

SNP-based heritability estimation, collection whole genetic/frequencies information, estimation of matrices, GWAS analysis, and linear mixed models were performed using R packages as SNPRelate (Zheng et al., 2012), SNPstats, GenABEL (Karssen, van Duijn, & Aulchenko, 2016), and PLINK 1.9 (Purcell et al., 2007).

5. RESULTS

5.1 Characteristics of the study populations

Table 4, below, describes the clinical characteristics and phenotypic details of Gubbio, Sardinia, and ARIC participating cohorts, regarding the MetS score components (BMI, SBP, SBP, WC, glucose, TRG, HDL). Sample size, mean, standard deviation (SD) and range are reported by gender.

In particular, characteristics of the Gubbio population are described in the first part. As reported in the Materials and Methods chapter, only the third survey was considered. A sample of 4,111 subjects were analyzed from this cohort (1,852 males and 2,259 females).

Summary characteristics for the Sardinia population are reported in the middle section of **Table 4**; a total of 8,102 subjects were enrolled for the implementation of the MetS score equation. In total, 3,485 males and 4,617 females were enrolled in the Sardinia population.

Finally, ARIC contained the largest sample size in this analysis. Sample size was equal to 8,592 subjects, including 4,419 females and 4,173 males (**Table 4**, last section).

Table 4 - Descriptive features of Gubbio, Sardinia and ARIC samples

	GUBBIO				**SARDEGNA**				**ARIC**			
	Men, N = 1852		**Women, N = 2259**		**Men, N = 3485**		**Women, N = 4617**		**Men, N = 4173**		**Women, N = 4419**	
	Mean ± SD	**Range**	**Mean ± SD**	**Range**	**Mean ± SD**	**Range**	**Mean ± SD**	**Range**	**Mean ± SD**	**Range**	**Mean ± SD**	**Range**
BMI (kg/m^2)	27.15 ± 3.77	13.78 - 45.11	26.14 ± 4.83	15.79 – 50.37	26.5 ± 3.9	15.8-48.9	25.2 ± 4.9	14.3-50.7	27.42 ± 3.97	16.10 – 56.26	26.65 ± 5.48	14.91 – 55.20
Waist circ. (cm)	93.05 ± 11.12	56–140	85.08 ± 13.96	52–135	93.0 ± 10.5	60.2-135.7	84.7 ± 13.4	52.7-146.5	99.60 ± 10.36	66 - 171	93.14 ± 14.81	52 - 169
HDL (mg/dl)	49.40 ± 12.47	22–109	61.37 ± 14.40	24–117	47.7 ± 11.5	19.1-121.0	56.6 ± 12.3	17.7-120.4	43.78 ± 12.19	9.63 – 128.01	57.42 ± 17.01	11.55 – 134. 82
TRG (mg/dl)	139.4 ± 96.46	17–1377	107.4 ± 61.51	23–630	125.6 ± 96.3	21.5-1456.5	94.8 ± 55.8	15.7-870.7	147.30 ± 97.9	24 – 1876	128.87 ± 82.58	26 - 1563
Glucose (mg/dl)	94.29 ± 23.30	59–346	89.82 ± 22.08	56–388	99.7 ± 25.0	56.4-435.9	92.6 ± 23.4	54.2-426.9	105.05 ± 30.6	53.88 – 517.67	102.86 ± 30.85	36.97 – 446. 47
SBP (mmHg)	130.88 ± 15.79	87–221	128.6 ± 19.12	87–217	130.9 ± 16.3	83-220	125.2 ± 18.3	80-200	120.16 ± 16.03	72- 206	117.05 ± 17.45	61 - 203
DBP (mmHg)	77.93 ± 8.92	45–115	75.48 ± 9.53	47–111	82.9 ± 10.0	50-130	79.1 ± 10.2	40-150	73.60 ± 10.01	12-130	69.88 ± 9.66	27 - 129
Age (years)	53.44 ± 16.08	11–92	55.95 ± 16.57	8–93	49.1 ± 17.5	18-100	49.4 ± 17.7	18-98	54. 69 ± 5.68	44 - 66	54.01 ± 5.67	44 - 66

Variables are expressed as means ± standard deviation

5.2 MetS scoring

As illustrated in the flow chart at the beginning of the chapter (**Figure 6**), the first step was to carry out Confirmatory Factor Analysis to select the best model that describes metabolic syndrome. Comparing three different CFA models, the best CFA fitting, represented by the lowest AIC and BIC scores and adequate summary indices, was found to be the bifactor model with one general factor (g) and three specific factors (Graziano et al., 2015).

As shown in **Figure 7**, the bifactor CFA model was the best way to summarize MetS components. In detail, the general factor (g) was the MetS syndrome and was used to estimate the newly proposed score. The other three factors summarized f1 “the obesity trait”, f2 the “blood pressure trait” and f3 the “lipid trait”. Only the glucose feature was independent from the other factors.

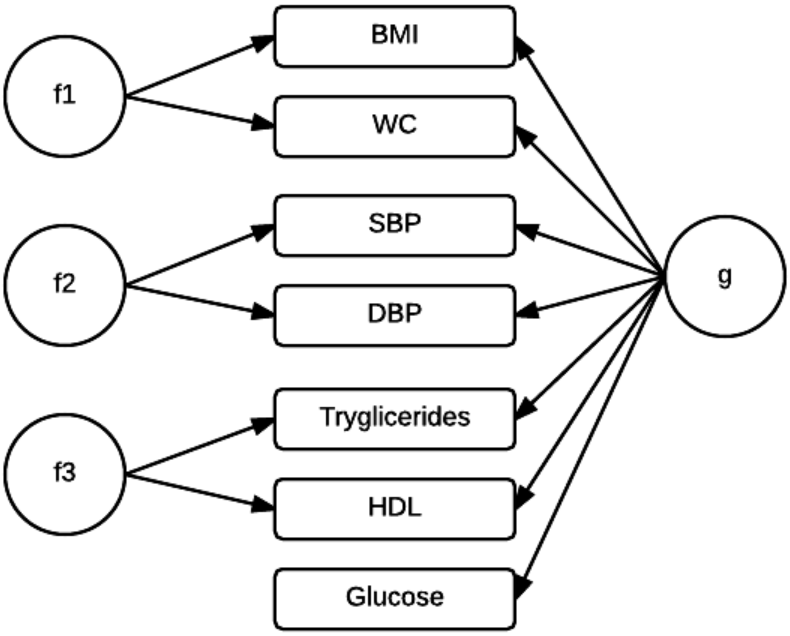


Figure 7. Bifactor CFA model

The indexes from this model and its results were used to propose the equation.

Algorithms to calculate MetS score (*g*) were based on bifactor results and were defined by the sum of each centered and scaled MetS variable (*x*’s) weighted with the corresponding ratio between factor loading (λ’s) and residual variance (ϑ’s) of the general factor *g* of bifactor CFA model (McDonald, 2013):

$$g = \sum_{j=1}^{p} \frac{\lambda_j}{\vartheta_j} \cdot \frac{x_j - mean(x_j)}{sd(x_j)}$$

Metabolic syndrome correlation structure was modeled by gender group (two-group CFA) because some components were statistically different. When significant group differences in factor loadings/residual variances were detected, factor score estimates were considered separately for each gender group (Viitasalo et al., 2014). Finally, g scores were rescaled in the range of 0 to 100 with: $100 * (g - \min)/(max - \min)$.

Results from this equation were used to calculate the MetS score for each individual. A continuous trait, normally distributed, that summarizes common components of MetS was obtain for each sample.

As shown in detail below, due to statistically significant differences in gender, two algorithms were performed:

For males:

Gm = 0.645*WC + 0.933*BMI + 0.059*SBP + 0.087*DBP + 0.011*GLU -0.022*HDL+ 0.003* TRIG - 63.0

For females:

Gf = 0.342*WC + 0.636*BMI + 0.133*SBP + 0.146*DBP + 0.021*GLU -0.027*HDL + 0.009*TRIG - 44.4

5.3 Comparison the proposal score with other criteria

In **Figure 8**, three ROC curves are illustrated using the proposal score calculated in the Sardinia population compared with the three most common criteria (IDF, harmonized, and ATPIII).

As shown in **Table 5**, the IDF criteria had the best performance compared with the others criteria (sensitivity and specificity were equal to 0.802 and 0.803, respectively).

Table 5. Sensitivity and Specificity of each criteria

Criteria (overall cut-off)	Sensitivity (SE)	Specificity (SE)
IDF criteria (36.3)	0.802 ± 0.02	0.801 ± 0.01
ATPIII criteria (35.05)	0.769 ± 0.02	0.738 ± 0.01
Harmonized criteria (34.3)	0.738 ± 0.02	0.801 ± 0.01

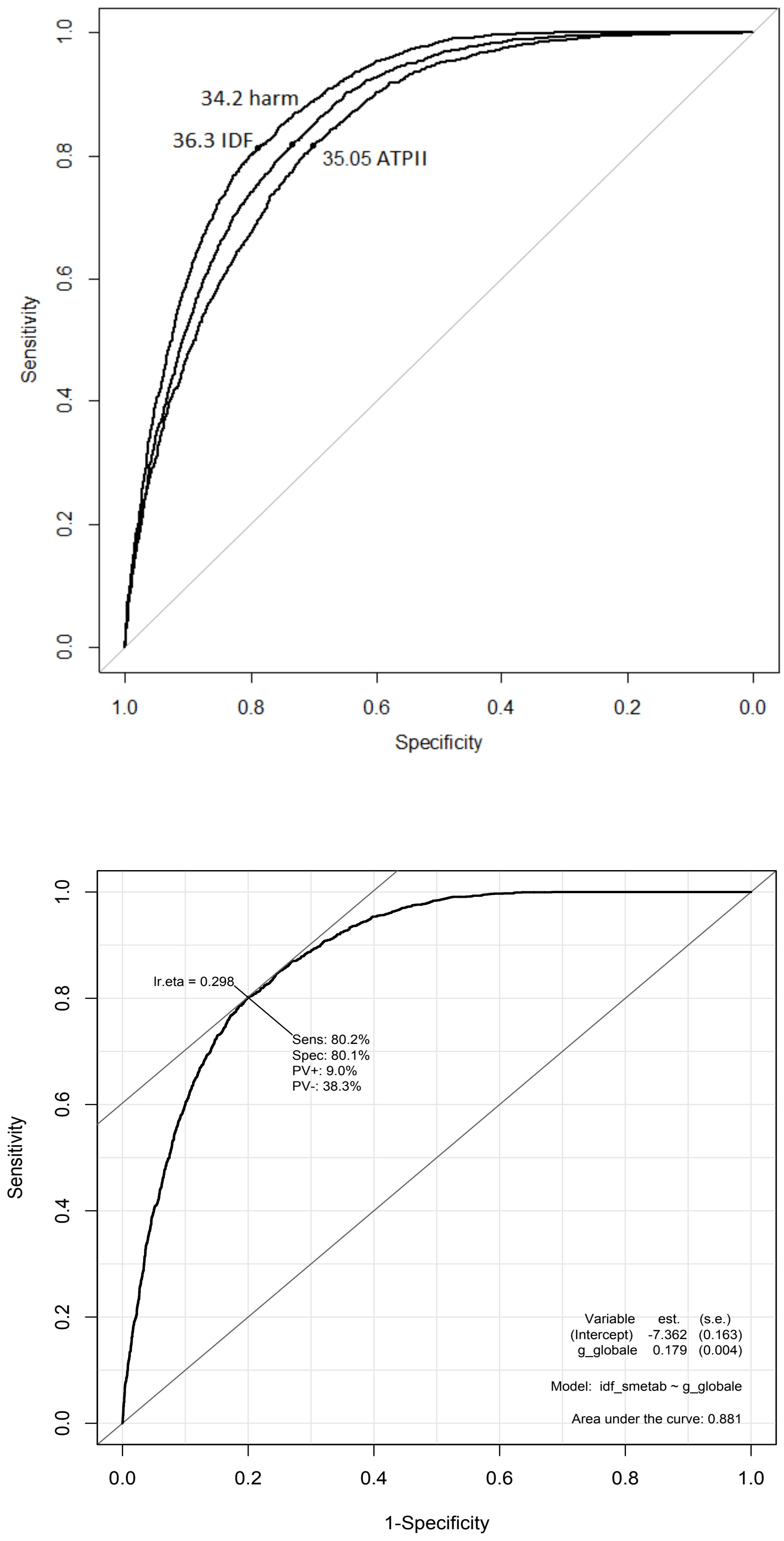


Figure 8 a) three compared ROC curves using Harmonized, IDF and ATPIII criteria b) ROC curve of MetS score for identifying Metabolic Syndrome (using IDF criteria)

In detail, MetS Score effectiveness, compared with IDF criteria, was evaluated using ROC (Receiver Operating Characteristics) curve (Hajian-Tilaki, 2013). The optimal cut-off was also calculated to determine the presence or absence of metabolic syndrome. Using Youden's Index (*J*), the optimal cut-point (*c*) was equal to 36.33 with 0.80 specificity and 0.80 sensitivity (**Figure 8**).

A good concordance measured using Cohen's Kappa and 36.33 as cut-off point was obtained (coefficient of agreement was equal to 0.55).

5.4 External validation in the Gubbio population

This score was validated using another Italian population. External validation was performed on data from the Gubbio Population Study (Cirillo et al., 2014; Graziano, Grassi, Bonati, Zanchetti, & Biino, 2016).

Data from the last survey, carried out in 2001-2007 (sample size equal to 4,111 subjects), was considered for the validation due to the availability of all seven phenotypes and the comparability with data collected in the Sardinia population in terms of span of years.

ROC curve analysis was conducted for assessing the performance of the MetS score as a binary classifier, both in the whole sample and stratified by sex (**Table 6**).

Table 6. Metabolic Syndrome Score in Gubbio population

	MEN		**WOMEN**	
	Mean (SD)	**Range**	**Mean (SD)**	**Range**
MetS score	37.26 (10.95)	0.13–89.59	30.59 (10.07)	5.03–73.23

In addition, Cohen's Kappa was computed to measure agreement between the MetS diagnosis, obtained by applying the proposed cut-off of 36.33 to the MetS score in the Gubbio sample, and the gold standard diagnosis using IDF criteria.

Performance evaluation of the MetS score revealed AUC equal to 0.89, specificity equal to 0.75, and sensitivity equal to 0.86, thus confirming the good predictive accuracy of the score as a binary classifier (**Figure 9**).

Furthermore, a Cohen's Kappa of 0.80, 0.77, and 0.82 in the whole sample, in men, and in women, respectively, shows a good agreement between the MetS diagnosis through the proposed cut-off and the gold standard, IDF criteria (Graziano et al., 2016).

The MetS score calculated in this population has a value of 37.26 (SD = 10.95) in men and 30.59 (SD = 10.07).

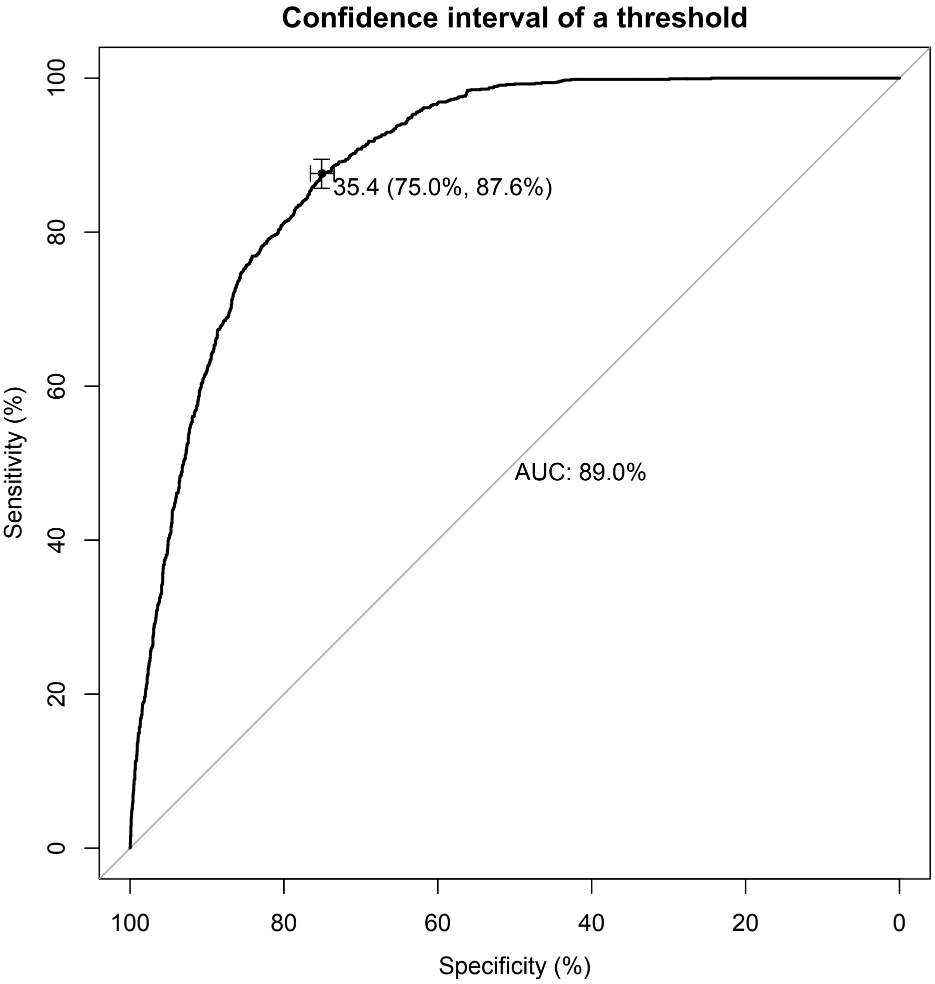
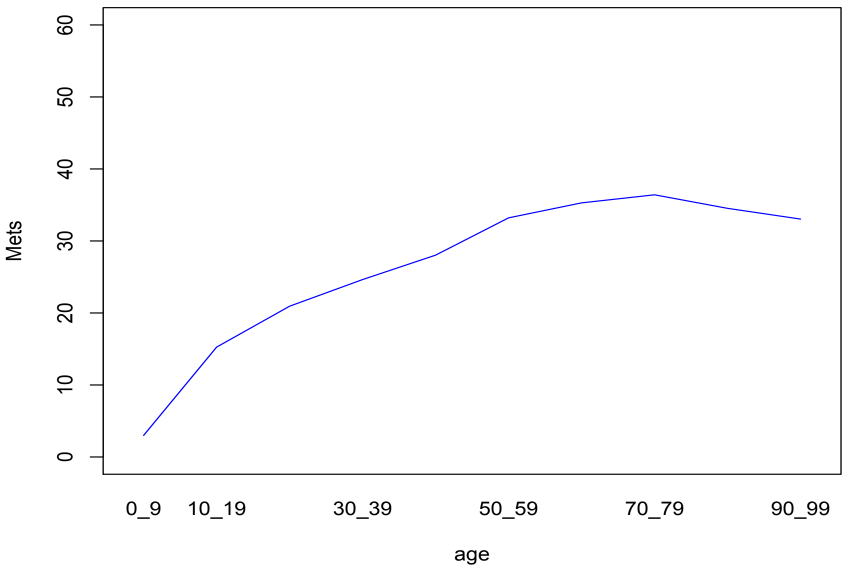


Figure 9 - Receiver operating characteristic (ROC) curves for MetS score

Due to the availability of a twenty-five years span life, collection of individual ages was considered and plotted to compare with the MetS score. As the literature suggests, prevalence of MetS increases with age. As shown in **Figure 10**, a positive trend is observed until the fourth decade and a plateau of MetS score value is obtained in the fifth decade. Due to missing values and deaths of older people, a decrease in the MetS score is observed in the last decade. Considering the cut-off point equal to 36.33, founded in ROC curve analysis, MetS variable increases dramatically with age in the fifth decade. Due to missing values, the last decade in the graph is not representative of the MetS trend.

Results of the external validation support the applicability of the model in clinical practice for diagnosis and screening. It also supports the possibility of using the model to evaluate the genetic component of this complex disease and to discover genomic regions involved in the clinical presentation of this disorder.



AGE decades	METS
0-9	3.0 - NA
10-19	15.26
20-29	20.94
30-39	24.64
40-49	28.03
50-59	33.21
60-69	35.29
70-79	36.41
80-89	34.53
90-99	33.04

Figure 10 - MetS over time, and in the table, the values corresponded each decade

5.5 Heritability using the family data

5.5.1 Gubbio population study

A total of 711 nuclear pedigrees were available for heritability analysis in the Gubbio population. All models were adjusted for age. AE and ACE (both households and sibs-household effects) models were performed. As shown in **Table 7**, heritability is statistically significant in each model. h^2 is estimated to equal 35% under the AE model assumption.

Table 7. Heritability analysis in Gubbio population (n=2620, 711 pedigrees)

MODEL	h^2 (SE)	c^2 (SE)	Variance due to covariates
AE	0.354*** (0.051)	-	0.199
ACE (Household effects)	0.129* (0.118)	0.128* (0.060)	0.199
ACE (Sibs-household effects)	0.300*** (0.061)	0.094*** (0.049)	0.194

*p<0.05; *** p<0.0001

If familial environmental factors are considered, ACE (sibs-household and household effects) model estimations were statistically significant. In the first case, h^2 and c^2 (household effect) were both equal to 13%; in the second case, h^2 was higher than the first one and equal to 30% and c^2 (sibs-households) was lower and equal to 9%.

5.5.2 Sardinia population study

Due to the availability of pedigree information, MetS heritability was also estimated in the Sardinia population. A total of 589 pedigrees, n = 8,096 subjects, was collected and analyzed to calculate additive and environmental factors. Again, all models were adjusted only for age because gender is included yet in the definition of the score. h^2 was statistically significant and equal to 34% in the AE model.

When c^2 (household effects) were considered, both c^2 and h^2 were significant, but h^2 was lower than before and equal to 29% and c^2 was equal to 3%. Results are summarized in **Table 8**.

Table. 8 Heritability analysis in Sardinia population (n=8096, 589 pedigrees)

MODEL	h^2 (SE)	c^2 (SE)	Variance due to covariates
AE	0.340*** (0.0224)	-	0.221
ACE (Household effects)	0.292*** (0.0225)	0.027* (0.050)	0.221

*p<0.05; *** p<0.0001

Sibs-household effects were not calculated because only extended pedigrees were analyzed in this sample.

5.6 Heritability using marker information and LD thresholds

5.6.1 Sardinia population study

A total of 1,163 subjects were considered to estimate heritability using IBD and GCTA. Subjects from the Sardinia samples demonstrated, in part, relatedness in the within-family design.

Figure 11 shows the relatedness inference from IBD estimates. Specifically, estimates of the IBD coefficients, k0 and k1, were used to infer relatedness.

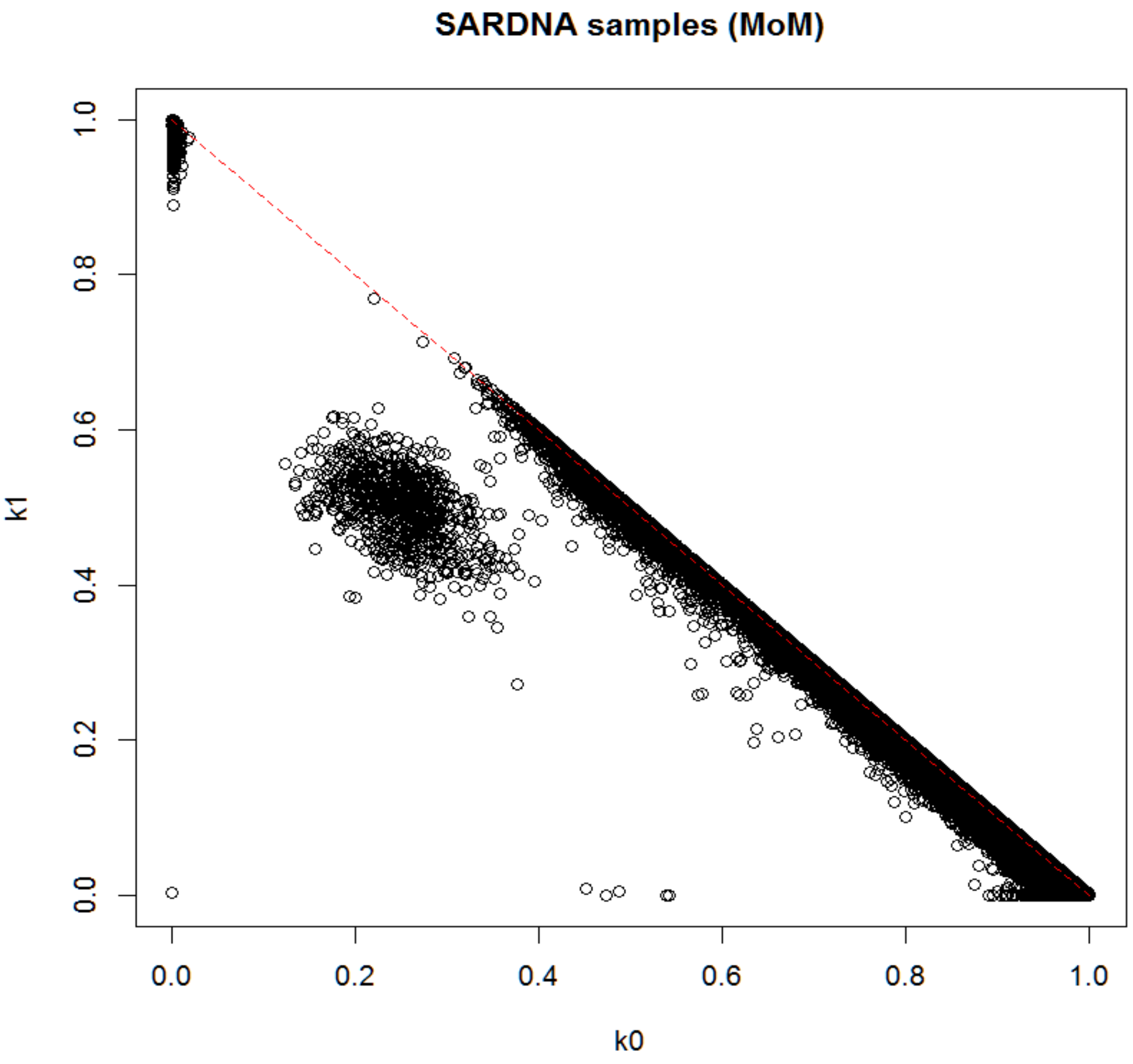


Figure 11 - Relatedness inference from IBD estimates

Each point is for a pair of samples and the diagonal line (red line) represents k0 + k1 equal to 1. Parent-offspring pairs are expected to occur at k1 = 1 and k0=0 and duplicates (or identical twins) at k0 = k1 = 0. As illustrated in **Figure 11**, a duplicate subject is present, also pairs of samples with kinship coefficient estimates for full sibs (k0 = 0.25 ± 0.08, k1 = 0.50 ± 0.10), half sibs (k1 = 0.50 ± 0.08, k0 = 1 - k1), and first cousins (k1 = 0.25 ± 0.08, k0 = 1 - k1) are present. As shown in **Figure 11**, many supposed "unrelated" subjects in the Sardinia samples (1/3 of the total samples) as expected are close or very closely related subjects. PCA confirms this common ancestral origin displaying no population subgroups (substructure) using the first four PCs (**Figure 12**).

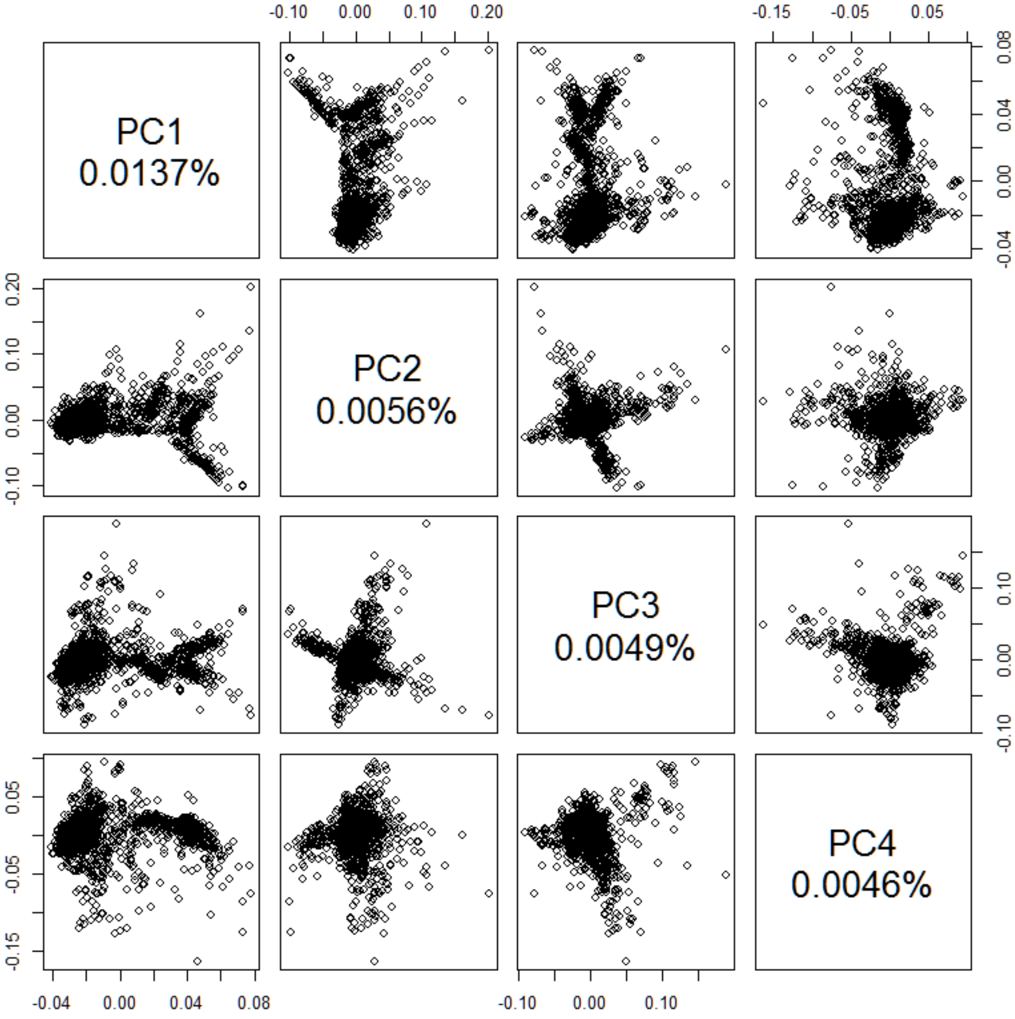


Figure 12. Plot the principal component pairs for the first four PCs of the Sardinia sample

The PCs number K=4 is found at the eigenvalue in which an elbow of an eigenvalue decay is observed on the scree plot (Cattell, 1966), and on the diagonal of scatter matrix of **Figure 12**, the percent of variation accounted by the first four principal components is plotted.

The implication is that the population structure may be accounted for by using the kinship matrices (IBD or GTCA), without the need of fitting PCs as fixed effects. As previously described, for calculating heritability, two types of matrices and three levels of LD were considered. The number of SNPs (m) selected for the analysis are shown in **Table 9** at the top. In detail, m = 65,298, 22,6771 and 46,1015 were considered for heritability estimation after LD pruning equal to 0.20, 0.80 and no LD, respectively.

All heritability estimates of MetS (adjusted for the fixed effects of age and PC1), both using IBD and GCTA matrix, were significant. In particular, considering the IBD matrix (**Table 9**, top left) and applying an LD threshold equal to 0.20, h^2 of metabolic syndrome was equal to

0.243 (0.179 - 0.307); using an LD threshold equal to 0.80, heritability was estimated to be 0.273 (0.120 - 0.347), and using no LD pruning, h^2 estimate was equal to 0.262 (0.189 - 0.334). Using the GCTA matrix (**Table 9**, top right) and different levels of LD threshold, for LD equal to 0.20, h^2 was estimated to be 0.408 (0.308 - 0.508); applying an LD threshold of 0.80, h^2 was equal to 0.355 (0.268 - 0.441), and without LD pruning, h^2 was equal to 0.332 (0.250 - 0.413).

5.6.2 ARIC population

Similar to the Sardinia study, **Figure 13** shows the relatedness inference from IBD estimates. Only pairs with kinship coefficient estimates >1/32 are plotted. At the bottom of the figure, the red line represents the independent subjects with k0 + k1 equal to 1, whereas the red line at the top of the figure represents the parent-offspring pairs. Only 619 pairwise on 8592 x 8591/2 were offline from the diagonal line, suggesting a negligible hidden relatedness.

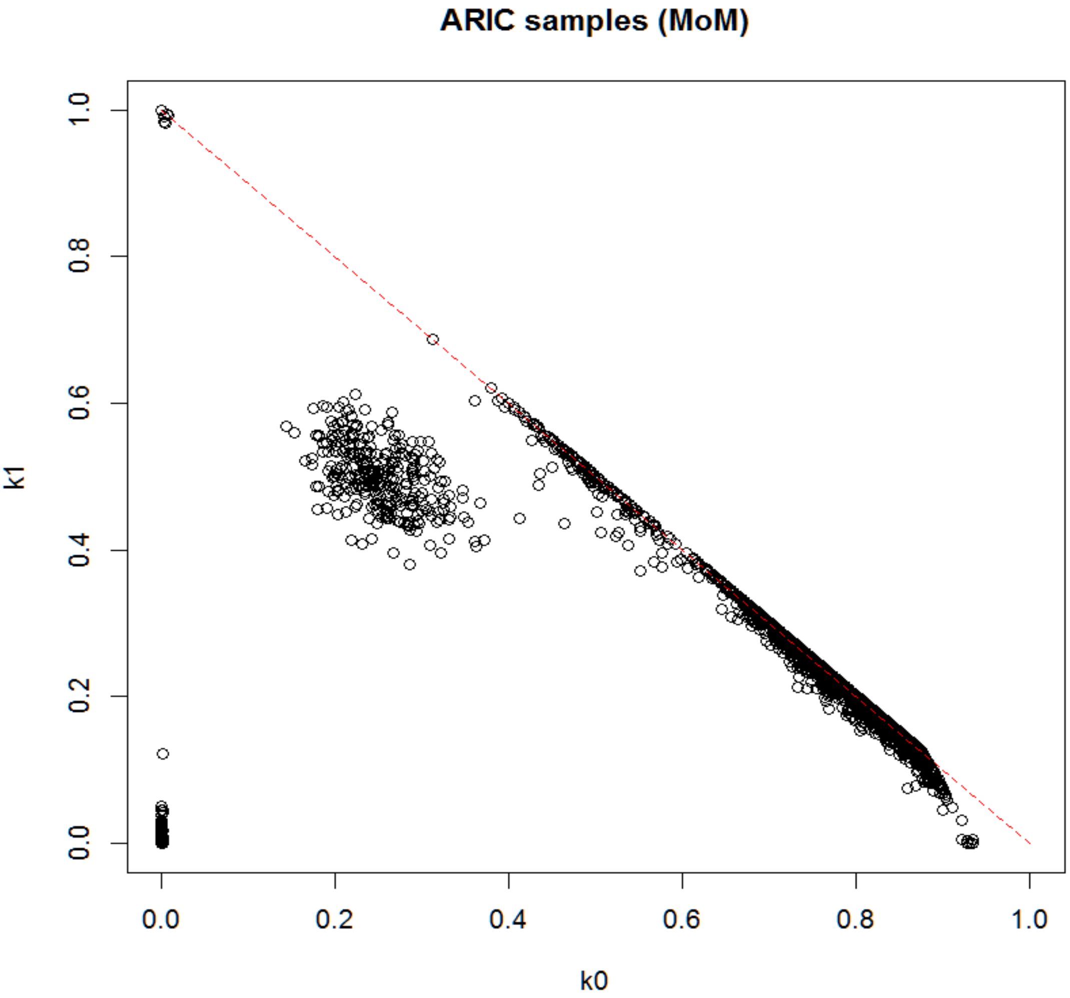


Figure - 13 Relatedness inference from IBD estimates

Next, PCA was performed to differentiate if there is a latent population substructure. As shown in **Figure 14**, the scatter matrix of the first four PCs selected using the scree plot illustrates that there are no significant population structure differences.

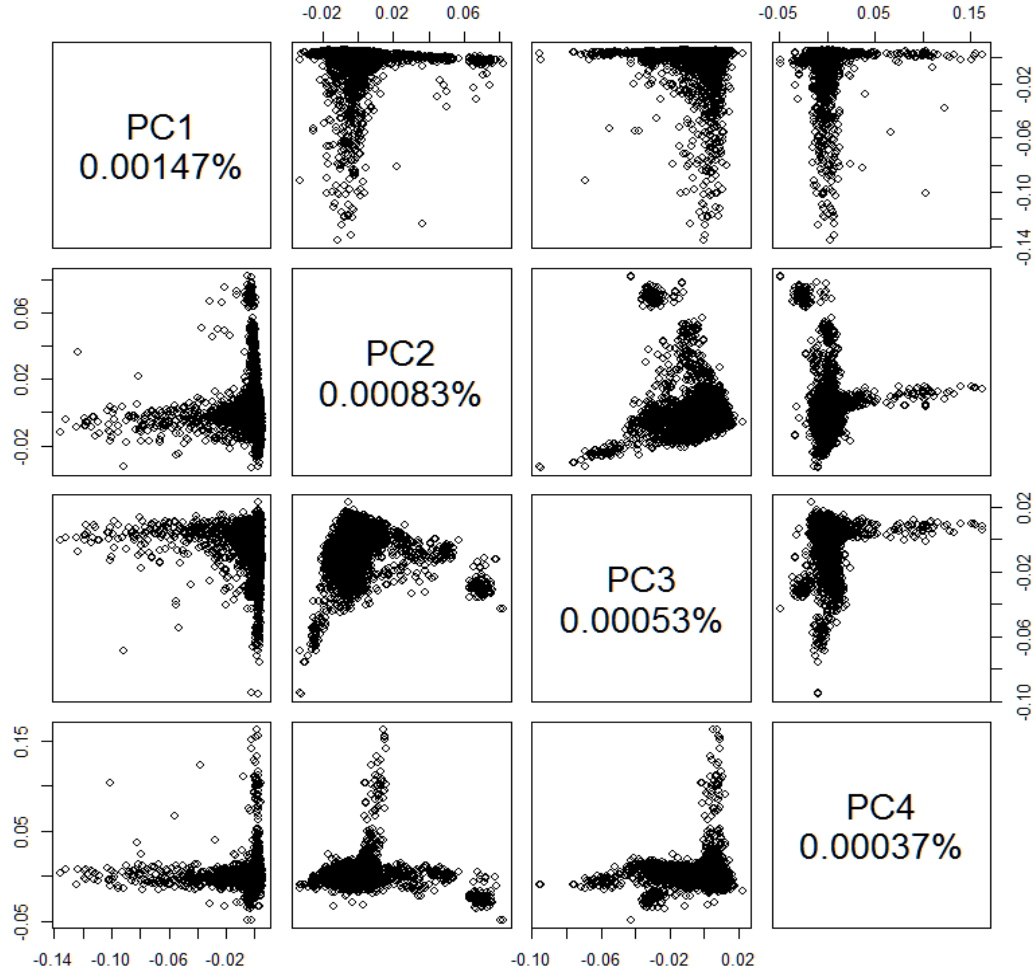


Figure 14. Plot the principal component pairs for the first four PCs of the ARIC sample

Therefore, as in the Sardinia samples, heritability analysis was performed using IBD or GCTA matrices and applying three levels of LD pruning, without the need of fitting PCs as fixed effects. In each case, heritability was significant. The number of SNPs (m) selected for the ARIC heritability analysis are shown in **Table 9,** at the bottom. Specifically, the number of subjects (n) was equal to 8,592 and the number of SNPs equal to 70,594, 273,341, and 571,466 were considered for heritability estimation after LD pruning at 0.20, 0.80 and no LD, respectively.

Using the IBD matrix (**Table 9**, bottom left) and LD = 0.20, h^2 was equal to 0.101 (0.059 - 0.150); using an LD threshold equal to 0.80, h^2 was estimated to be 0.152 (0.089 - 0.215).

Finally, using no threshold LD, h^2 was equal to 0.117 (0.065 - 0.168). In models using the GCTA matrix (**Table 9**, bottom right), heritability was higher than the IBD matrix and statistically significant using all three of the LD thresholds. Specifically, for LD = 0.20, h^2 was equal to 0.206 (0.135 - 0.277); using LD = 0.80, h^2 was equal to 0.227 (0.148 - 0.305); and finally, without LD pruning, h^2 was estimated equal to 0.195 (0.126 - 0.264).

Figure 15 shows a scatter plot matrix comparing the three threshold levels, illustrating that matrices imputed using an LD threshold equal to 0.80 or without LD pruning have the same trend, demonstrating similar pairwise genetic relatedness results across LD pruning.

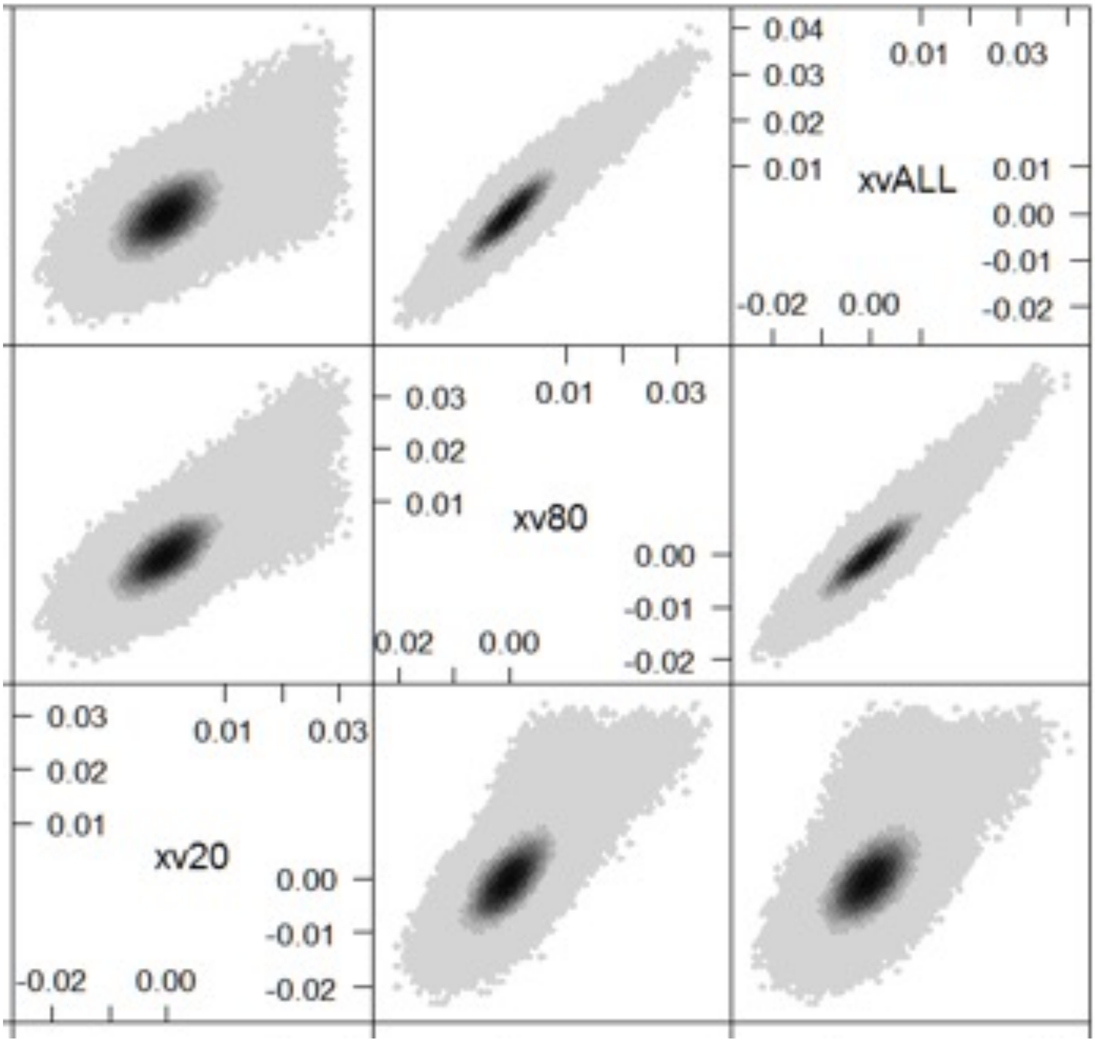


Figure 15 - Scatter plot matrix comparison (LD=0.20,0.80, none) in ARIC sample

Table 9 – Summary results from Sardinia and ARIC samples

		IBD			GCTA		
SARDNA	**LD cut-off**	**0.20**	**0.80**	**ALL**	**0.20**	**0.80**	**ALL**
n=1163	**n. of SNPs**	65298	226771	461015	65298	226771	461015
	heritability	0.243*	0.273*	0.262*	0.408*	0.355*	0.332*
	95%CI	0.179 - 0.307	0.120 - 0.347	0.188 - 0.337	0.308 - 0.508	0.268 - 0.441	0.250 - 0.413
	h2$_{PCGC}$	-	-	-	0.203	0.158	0.145
ARIC **n=8592**	**n. of SNPs**	70594	273341	571466	70594	273341	571466
	heritability	0.1015*	0.152*	0.117*	0.206*	0.227*	0.195*
	95%CI	0.031 - 0.149	0.089 - 0.215	0.065 - 0.168	0.135 - 0.277	0.148 - 0.305	0.126 - 0.264
	h2$_{PCGC}$	-	-	-	0.176	0.203	0.174

heritability estimation using IBD and GCTA matrices at LD=0.20, 0.80 and no LD pruning; n=sample size, *=p-value <0.0001, 95%CI= 95% approximate symmetric confidence interval.

5.7 Heritability using suggestive genome-wide association

5.7.1 Sardinia population

Genotype data were filtered on the basis of quality control measures (see Chapter 4.5). After Quality Control, GWAS was carried out using a sample size equal to 1,163 subjects with a total of 361,504 SNPs. A single marker linear regression with age as a covariate and 52 sub-groups obtained by Hierarchical Average Clustering on the Identity By State (IBS) matrix for controlling hidden family structure of the “unrelated” subjects (see par. 5.6.1) was performed. The number of clusters K = 52 was selected cutting the dendrogram at the default threshold, h = 15.

Results of typed SNPs are shown in **Table 10** and displayed in a Manhattan plot of **Figure 16**, where the two lines indicate the Bonferroni threshold and the less stringent *P*-value.

Table 10 - SNPs with suggestive P-values ($P<10^{-4}$) in the Sardinia sample

SNP.id	Chromosome	Position (BP)	MAF	P-value
rs4862188	**4**	**184355370**	**0.305291**	**5.02E-25**
rs3883013	**15**	**85088657**	**0.305681**	**1.08E-24**
rs2880301	**13**	**20100534**	**0.304878**	**4.01E-24**
rs3883014	**15**	**85088729**	**0.292813**	**1.96E-22**
rs3013384	**1**	**243087578**	**0.369584**	**5.66E-13**
rs7184960	16	89961661	0.011608	9.69E-08
rs1819043	1	202432002	0.416594	1.72E-07
rs2270459	16	89979851	0.012038	1.97E-07
rs11158185	14	58431087	0.094234	6.87E-07
rs6586608	8	17369652	0.471195	9.68E-07
rs10955346	8	105311364	0.242882	1.69E-06
rs4448239	8	105309928	0.242033	1.90E-06
rs2730268	7	158759451	0.047496	2.02E-06
rs12699098	7	71277486	0.310181	2.03E-06
rs11076654	16	90074085	0.011082	2.23E-06
rs1867523	15	35551374	0.2513	2.43E-06
rs17567007	15	28201539	0.086414	2.48E-06
rs10486666	7	35547956	0.098404	2.58E-06
rs17332419	5	60136626	0.053833	3.61E-06
rs10904949	10	17439939	0.130603	4.06E-06
rs2494356	1	42823366	0.494411	4.12E-06
rs2730276	7	158758611	0.047474	4.54E-06
rs2622759	15	35549647	0.25631	4.61E-06
rs1536651	1	42823768	0.492685	5.35E-06
rs10509940	10	113120504	0.365633	6.34E-06
rs1259724	12	32029577	0.271815	6.88E-06
rs1550744	15	35550795	0.253478	6.91E-06
rs2392369	7	35537950	0.106466	8.12E-06

rs318331	15	35543852	0.254091	8.72E-06
rs17094008	14	58428106	0.090278	9.13E-06
rs874599	7	158077350	0.416021	1.03E-05
rs11973244	7	158075769	0.459895	1.16E-05
rs4244290	10	113127090	0.389276	1.17E-05
rs4723436	7	35647449	0.282515	1.20E-05
rs373634	16	24164724	0.230735	1.34E-05
rs373690	14	58021286	0.168966	1.35E-05
rs2765315	13	101672045	0.047414	1.54E-05
rs2788223	6	720465	0.313149	1.62E-05
rs17331746	5	60116613	0.055751	1.77E-05
rs10141903	14	82018202	0.433995	1.83E-05
rs2281273	6	1590446	0.029336	1.91E-05
rs7112116	11	96686871	0.486547	1.94E-05
rs2896871	11	96672691	0.476611	1.95E-05
rs9838604	3	181346790	0.329321	2.26E-05
rs738370	22	35982672	0.29199	2.29E-05
rs1997883	22	36103709	0.263113	2.34E-05
rs7001464	8	1220024	0.312608	2.39E-05
rs17606892	4	37965518	0.101563	2.68E-05
rs381901	16	24152467	0.243327	2.70E-05
rs12287911	11	96670957	0.476273	2.71E-05
rs11253637	10	15810354	0.028497	2.75E-05
rs996604	11	96672579	0.475494	2.81E-05
rs1499968	3	117699039	0.068847	2.96E-05
rs1499097	3	1227382	0.239583	3.26E-05
rs1190100	14	58015648	0.166233	3.40E-05
rs6651244	8	128217462	0.0162	3.44E-05
rs5755921	22	36104730	0.267212	3.46E-05
rs5995155	22	36102700	0.263997	3.51E-05
rs6707241	2	46584852	0.341516	3.58E-05
rs10836637	11	36937056	0.313523	3.61E-05
rs268821	14	58033552	0.179406	3.65E-05
rs10509941	10	113127257	0.409716	3.83E-05
rs7101678	11	96686388	0.470406	3.91E-05
rs10950200	7	69865264	0.243668	4.01E-05
rs1383349	11	36938773	0.314064	4.05E-05
rs17423790	7	71273602	0.231441	4.16E-05
rs2781007	9	86266068	0.239424	4.25E-05
rs1681946	6	64216909	0.477528	4.37E-05
rs7116990	11	96680330	0.475779	4.51E-05
rs4665522	2	23040960	0.109809	4.54E-05
rs11860279	16	61140529	0.106989	4.84E-05
rs17101549	14	75096934	0.02972	5.03E-05
rs13236867	7	54896159	0.17962	5.05E-05
rs11144596	9	78339320	0.129273	5.11E-05
rs17221776	10	109414982	0.062769	5.16E-05
rs11860196	16	61139382	0.105517	5.26E-05
rs228556	1	79650782	0.177603	5.29E-05

rs9319845	18	70283365	0.049308	5.42E-05
rs6759518	2	27486595	0.067863	5.43E-05
rs6597610	9	136091096	0.114789	5.43E-05
rs1452967	11	96669797	0.47745	5.44E-05
rs10885145	10	113140675	0.28821	5.63E-05
rs9667859	11	96679894	0.473386	5.77E-05
rs4716356	6	169824254	0.25	5.98E-05
rs10053787	5	128568188	0.116269	6.02E-05
rs17490471	4	11449278	0.168966	6.02E-05
rs3000891	1	12699337	0.061039	6.29E-05
rs10430541	10	56824247	0.211952	6.54E-05
rs17008750	2	119287185	0.050347	6.56E-05
rs755542	8	1846758	0.155885	6.57E-05
rs2025214	14	81949014	0.377846	6.57E-05
rs2677822	6	133923532	0.153184	6.77E-05
rs1544954	16	62866857	0.340536	6.81E-05
rs2274914	9	86500979	0.471195	6.95E-05
rs431718	14	56531186	0.278979	7.36E-05
rs1247489	10	78003580	0.316423	7.36E-05
rs12932136	16	7611973	0.40203	7.44E-05
rs2677821	6	133922011	0.153114	7.48E-05
rs10802680	1	238536388	0.467326	7.56E-05
rs10124390	9	86549939	0.444107	7.85E-05
rs12122035	1	50726052	0.030095	8.24E-05
rs800562	8	116709020	0.314703	8.35E-05
rs11942525	4	48205942	0.026667	8.41E-05
rs2225378	14	83583841	0.323401	8.51E-05
rs10520433	4	180669418	0.054206	8.54E-05
rs1681947	6	64217440	0.480224	8.65E-05
rs2511606	8	105295695	0.182958	8.75E-05
rs4590798	10	113194854	0.351852	8.77E-05
rs823608	8	16787703	0.018487	8.80E-05
rs7207189	17	51537585	0.340069	8.86E-05
rs228564	1	79646081	0.172978	8.92E-05
rs693420	1	238055618	0.195652	8.96E-05
rs17072059	13	49255972	0.042132	9.31E-05
rs2298100	1	238048325	0.195482	9.33E-05
rs513000	16	1092871	0.076425	9.42E-05
rs17061497	6	132921520	0.087927	9.63E-05
rs7920368	10	113201886	0.351082	9.63E-05
rs11726451	4	59852381	0.480503	9.65E-05
rs10505391	8	122175826	0.092943	9.73E-05
rs5999861	22	35923192	0.414422	9.74E-05
rs6807027	3	60092789	0.465176	9.82E-05
rs10504891	8	91171216	0.025541	9.97E-05
rs7302568	12	55869990	0.054819	9.97E-05
rs1541967	4	54382119	0.14298	9.98E-05

A total of **124** selected SNPs (*P*-values <10^{-4}) were used to calculate IBD and GCTA matrices. Results show that five Bonferroni significant SNPs are located in four chromosomes:
2 SNPs, rs3883013 and 3883014 (gene mapped, UBE2Q2P1), in chromosome 15. Other SNPs in chromosome 4 (rs4862188 in WWC2/CDKN2AIP Gene), chromosome 13 (rs2880301, in TPTE2/MPHOSPH8) and chromosome 1 (rs3013384).

Heritability was significant (*P* = 1.5e-40 and *P* = 1.1e-109) and equal to **0.158** (approximate symmetric 95% CI: 0.135 to 0.182) and **0.293** (approximate symmetric 95% CI: 0.268 to 0.319) using IBD and GCTA matrices, respectively.

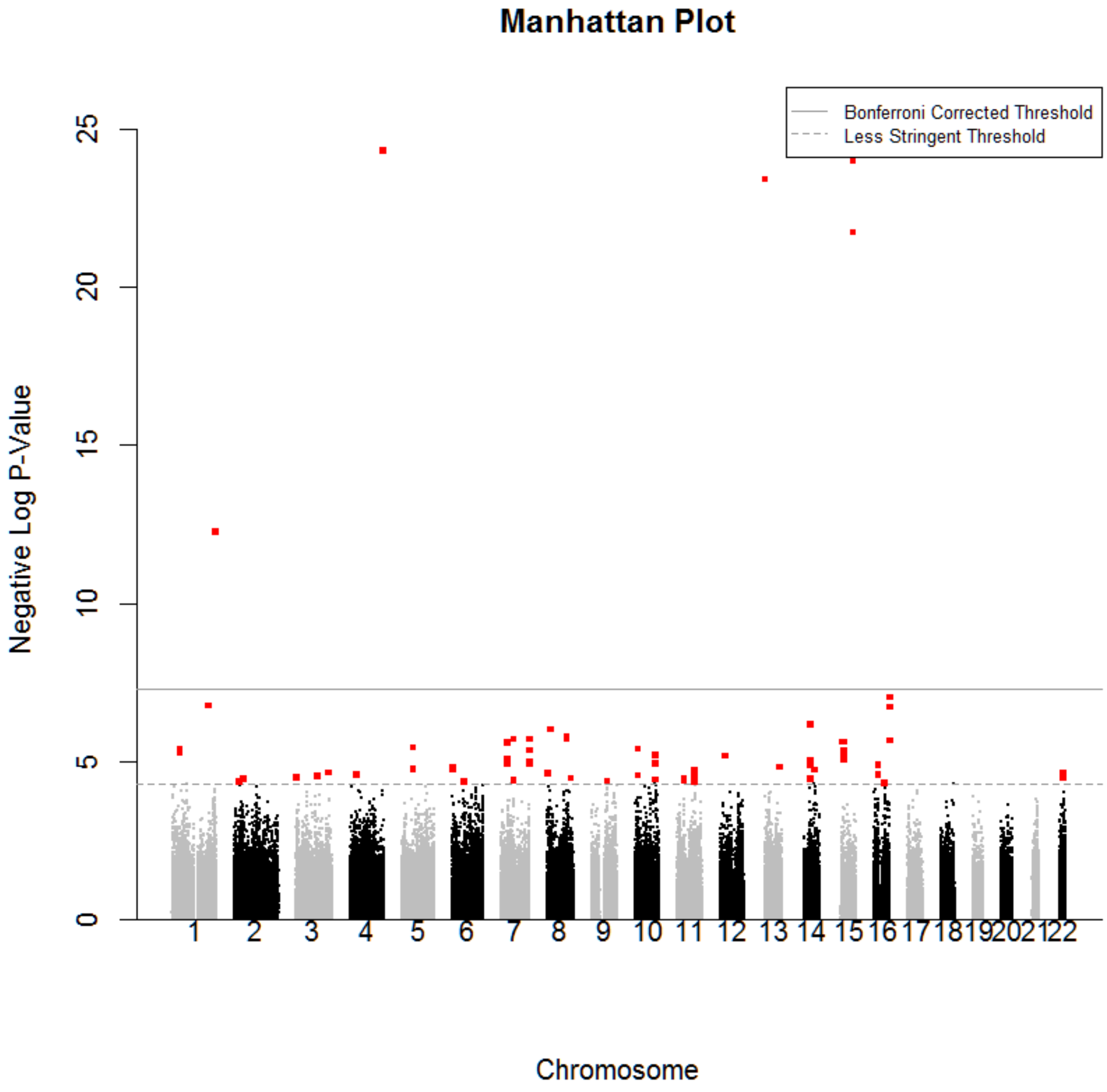


Figure 16 – Manhattan plot from Sardinia typed SNPs

Figure 17 displays a heatmap of pairwise linkage disequilibrium (LD) measurements calculated for selected SNPs. The upper triangle represents R^2 measures between pairs of SNPs. R^2 can ranged from 0 (SNPs in Linkage Equilibrium) to 1 (Linkage Disequilibrium).

Here, results showed that only a few SNPs are in LD. This means that almost all of the selected SNPs have an independent contribution into the estimated genetic effect on MetS.

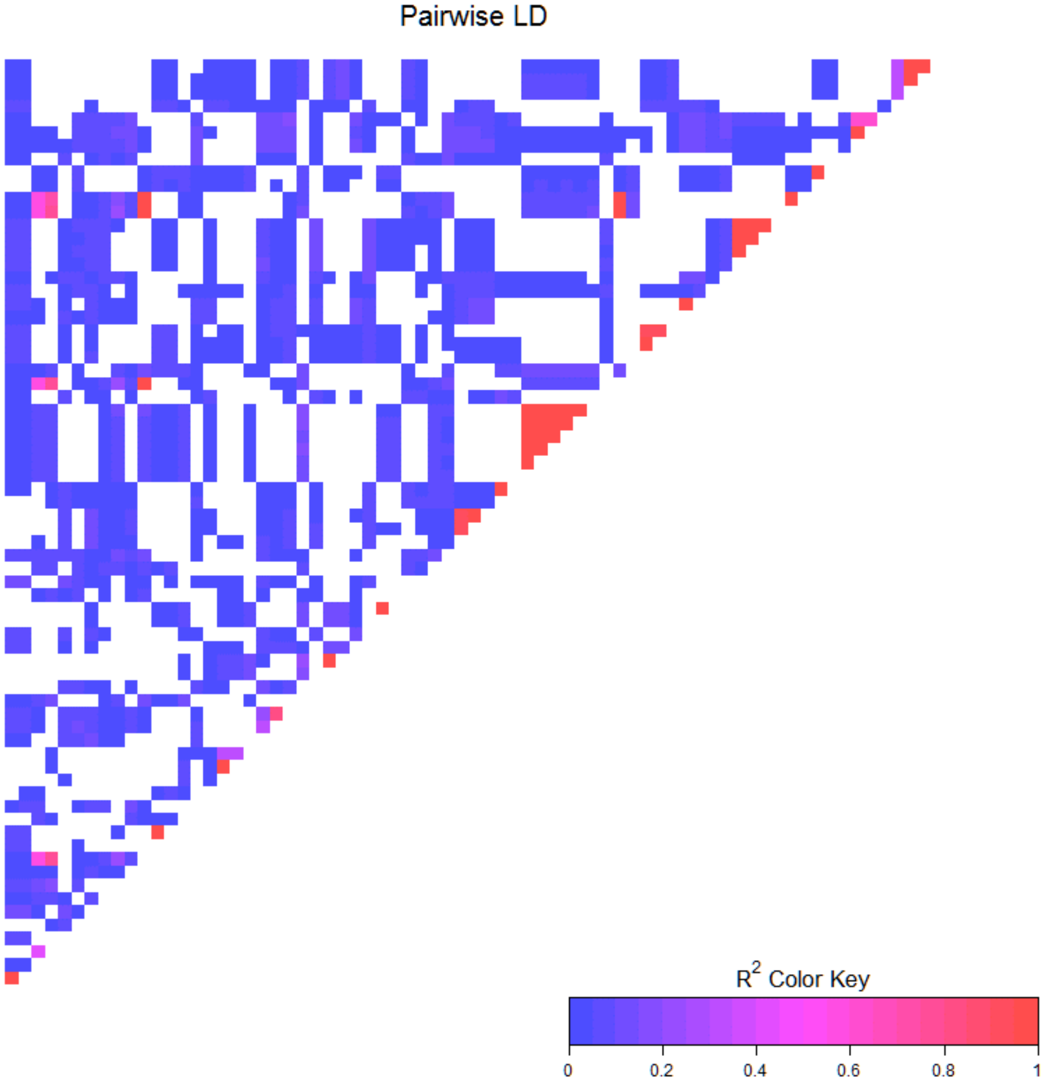


Figure 17 – LD matrix from selected SNPs (Sardinia)

5.7.2 ARIC population

GWAS from genotyped ARIC data was carry out using a sample size equal to 8,592 subjects with a total of 570,390 SNPs, after QC. A single marker linear regression with covariates of age and 10 PCs was performed.

GWAS results are displayed in the Manhattan plot (**Figure 18)** and selected SNPs are collected in **Table 11**.

Table 11. SNPs with suggestive P-values (P<10^{-4}) in the ARIC sample

SNP.id	Chromosome	Position (BP)	MAF	P-value
rs8050136	**16**	**53816275**	**0.406472**	**6.52E-09**
rs9941349	**16**	**53825488**	**0.42004**	**1.28E-08**
rs9940128	**16**	**53800754**	**0.432514**	**1.29E-08**
rs9939973	**16**	**53800568**	**0.432204**	**1.65E-08**
rs1121980	**16**	**53809247**	**0.432554**	**2.00E-08**
rs9930506	16	53830465	0.441803	2.31E-07
rs7782904	7	1.37E+08	0.275082	1.04E-06
rs1350146	3	36516326	0.084955	3.40E-06
rs1521252	13	63424494	0.201004	3.41E-06
rs3930017	7	76720582	0.423323	4.76E-06
rs9838604	3	1.81E+08	0.365398	4.88E-06
rs9838053	3	36494163	0.074181	6.06E-06
rs6435822	2	2.15E+08	0.152098	8.10E-06
rs2013012	1	64128209	0.257276	8.49E-06
rs17527820	1	2.2E+08	0.103472	9.86E-06
rs9397778	6	1.55E+08	0.247265	1.74E-05
rs4678415	3	1.38E+08	0.349744	1.79E-05
rs9730331	1	2.32E+08	0.302047	1.86E-05
rs731632	3	1.38E+08	0.349942	2.00E-05
rs4678260	3	1.38E+08	0.35025	2.04E-05
rs16877320	6	15923026	0.045492	2.24E-05
rs9419105	10	1.35E+08	0.169912	2.66E-05
rs7185307	16	12046899	0.236732	2.76E-05
rs7637666	3	1.38E+08	0.350839	3.18E-05
rs41375446	6	28024408	0.051334	3.19E-05
rs16893505	8	1.21E+08	0.019167	3.34E-05
rs7615658	3	1.38E+08	0.349959	3.46E-05
rs3773752	3	1.38E+08	0.347118	3.48E-05
rs6782181	3	1.38E+08	0.354632	3.50E-05
rs12683176	9	1.06E+08	0.020474	3.58E-05
rs6769261	3	1.38E+08	0.349511	3.60E-05
rs9951002	18	73426923	0.231188	3.90E-05
rs1718123	12	1.22E+08	0.497586	3.90E-05
rs4678409	3	1.38E+08	0.34986	3.96E-05

rs6914640	6	1.55E+08	0.230513	3.99E-05
rs10014286	4	1.66E+08	0.185871	4.35E-05
rs2054468	3	1.38E+08	0.350268	4.47E-05
rs12759915	1	1.99E+08	0.425628	4.48E-05
rs7704854	5	57416499	0.187776	4.54E-05
rs6087024	20	1031043	0.378536	4.57E-05
rs7116004	11	81273554	0.340713	4.78E-05
rs9888962	16	73749154	0.022292	5.06E-05
rs6499646	16	53843533	0.079211	5.09E-05
rs17709097	6	28002963	0.052206	5.18E-05
rs4678408	3	1.38E+08	0.374738	5.18E-05
rs9842371	3	1.81E+08	0.346199	5.37E-05
rs4895708	6	1.48E+08	0.037673	5.43E-05
rs2306589	17	34848874	0.475326	5.45E-05
rs1952836	14	28576698	0.291078	5.59E-05
rs7950435	11	81276907	0.340142	5.71E-05
rs7761864	6	15994826	0.046682	5.76E-05
rs6082455	20	21552504	0.36035	5.77E-05
rs6496903	15	92736994	0.102778	5.78E-05
rs17046025	4	1.66E+08	0.203628	5.89E-05
rs10509483	10	85708756	0.016142	6.03E-05
rs7183436	15	92737310	0.106132	6.11E-05
rs4428477	6	88633009	0.112488	6.14E-05
rs17032807	4	1.56E+08	0.01904	6.43E-05
rs1924338	13	1.09E+08	0.399267	6.54E-05
rs6485456	11	43766902	0.301293	6.60E-05
rs12292013	11	7696338	0.18116	6.62E-05
rs4412595	1	5830815	0.168161	6.77E-05
rs2993123	1	42354111	0.267691	6.79E-05
rs9393879	6	28018944	0.050848	6.88E-05
rs867382	9	1.3E+08	0.049465	7.00E-05
rs4072521	1	5830248	0.168645	7.03E-05
rs1191378	14	28597909	0.226246	7.28E-05
rs1191381	14	28602041	0.289192	7.51E-05
rs9308491	1	2.32E+08	0.302933	7.54E-05
rs7688470	4	22171998	0.204826	7.59E-05
rs17044860	3	6461402	0.012903	7.60E-05
rs12053372	2	99810055	0.244255	7.76E-05
rs855349	1	64127189	0.323237	7.91E-05
rs9725346	1	2.32E+08	0.332322	8.07E-05
rs1180187	6	83967454	0.361945	8.08E-05
rs2187539	11	81264443	0.340203	8.18E-05
rs3935073	1	5830967	0.167831	8.29E-05
rs1191379	14	28598952	0.294401	8.41E-05
rs10982499	9	1.18E+08	0.048842	8.58E-05
rs676160	11	1.21E+08	0.089687	8.67E-05
rs9450829	6	88624437	0.041565	8.70E-05
rs11023974	11	16541291	0.302247	9.28E-05
rs1157836	9	1.18E+08	0.445539	9.31E-05

rs10825738	10	57804722	0.136131	9.37E-05
rs11937241	4	1.56E+08	0.018973	9.54E-05
rs1679178	3	1.38E+08	0.159681	9.78E-05
rs2030600	15	26344449	0.308928	9.79E-05

A total of **87** selected SNPs ($P < 10^{-4}$) were used to calculate IBD and GCTA matrices. Results show that five Bonferroni significant SNPs are all located in chromosome 16 (rs8050136, rs9941349, rs9940128, rs9939973, rs1121980), whole mapped in FTO gene.

Heritability was significant (P = 1.1e-62 and P = 1.8e-137) and equal to **0.027** (approximate symmetric 95%CI: 0.024 to 0.031) and **0.086** (approximate symmetric 95%CI: 0.080 to 0.093) using IBD and GCTA approaches, respectively.

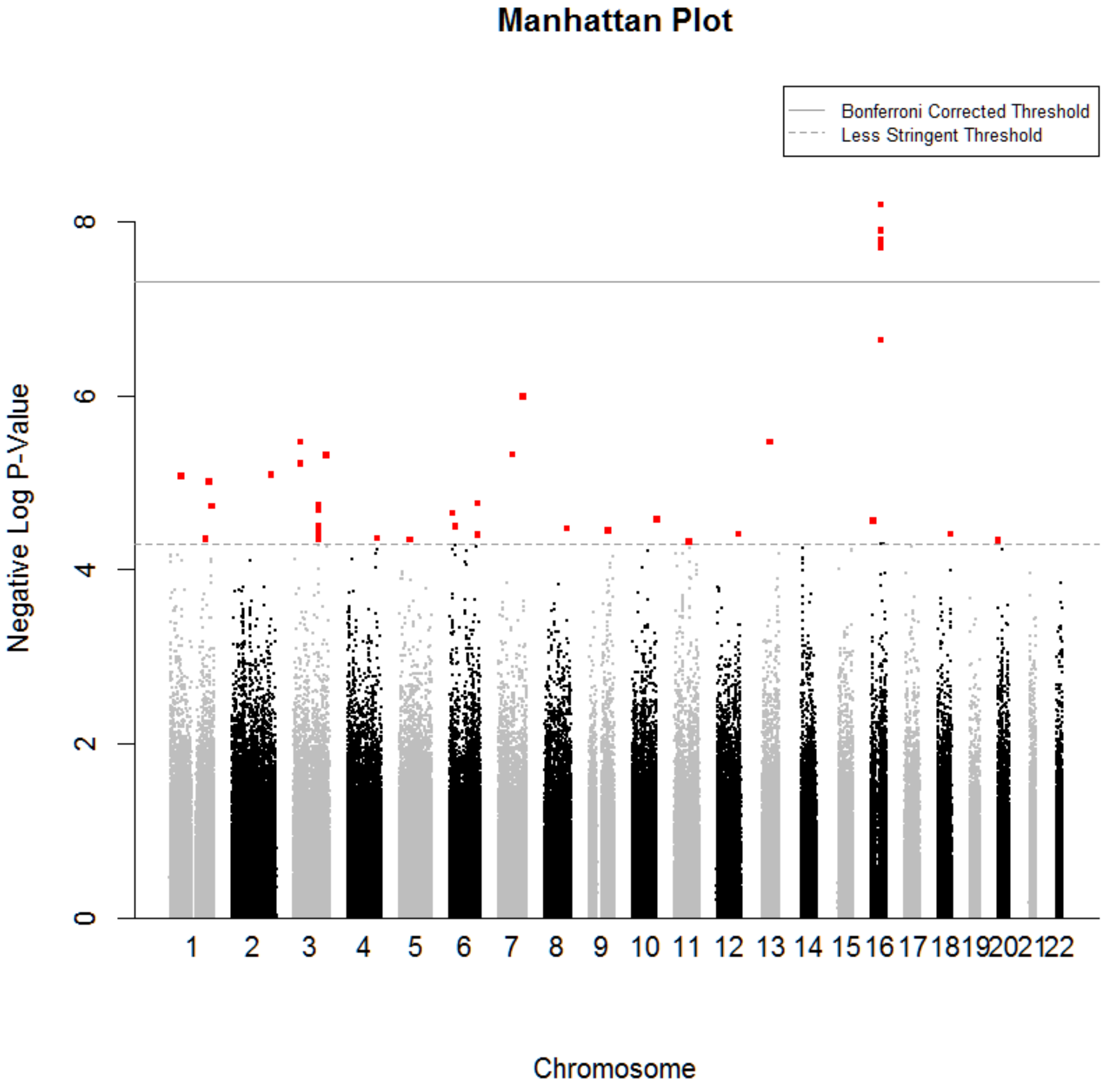


Figure 18 – Manhattan plot from ARIC typed SNPs

LD heatmap was calculated to evaluate LD for each pair of significant typed SNPs. Results are displayed in **Figure 19**. Similar to the Sardinia study results, few SNPs are in Linkage Disequilibrium, demonstrating that independent SNPs that contribute to estimate the genetic effect on MetS.

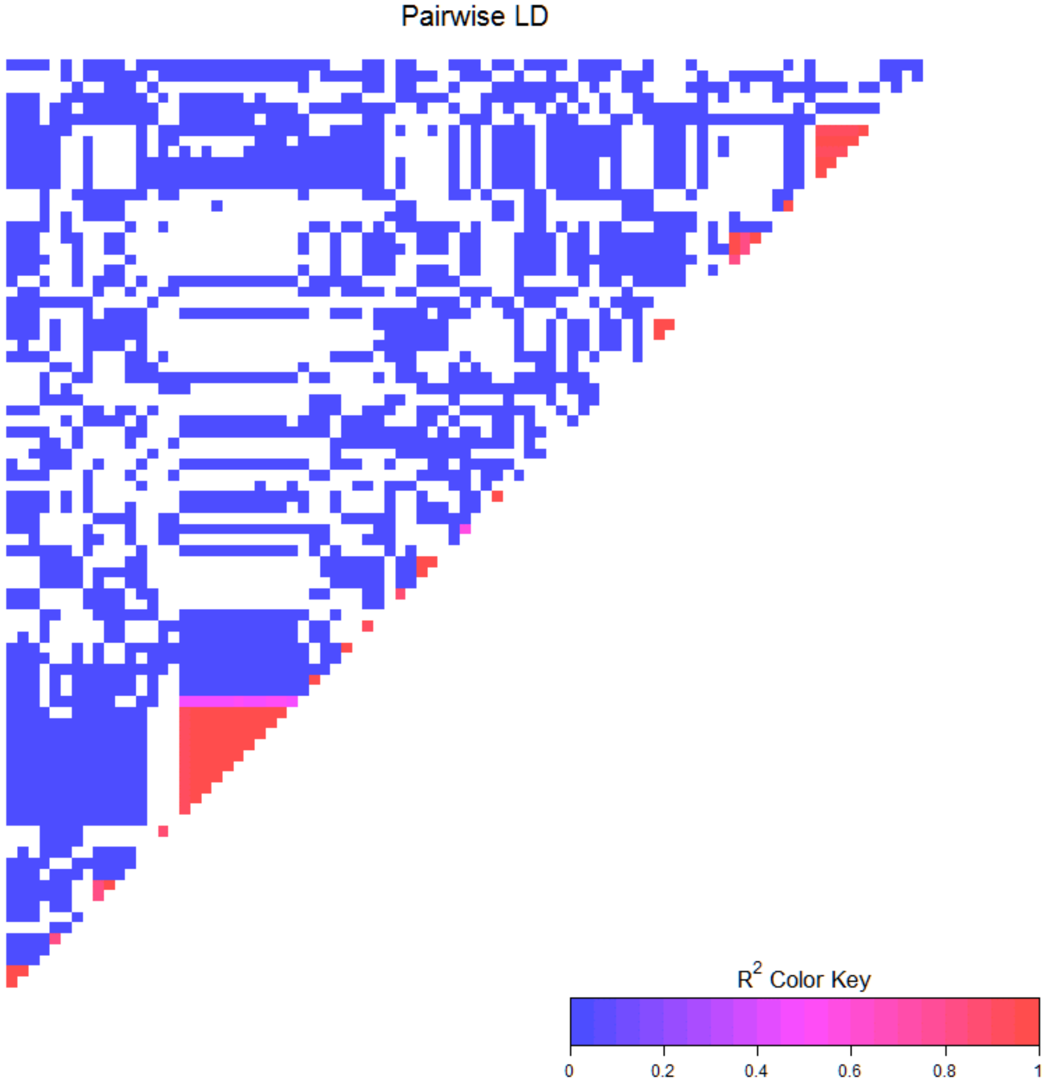


Figure 19 – LD matrix from selected SNPs (ARIC)

6. DISCUSSION

MetS is a complex disease and knowledge of the underlying mechanisms may contribute to a better understanding of MetS pathogenesis. Currently, information about each component is available, however, few studies have evaluated both clinical and genetic aspects. In this thesis, the focus was to understand interactions between components that are known to contribute to the syndrome and to fill in gaps with regards to the genetic aspects.

First, a model that summarizes the components will help us to understand in which factors are more relevant and in what way these factors contribute to MetS. Results have shown that the bifactor model is the best model to describe what happens when all components are taken into account.

Three factors (lipids, fat, and BP) and a general variable (the syndrome) were identified when inter-correlation was considered. This result confirms that these factors are the focus of the physiopathology.

Using a proposal algorithm derived from the bifactor model could have different advantages both in diagnosis and in discovery by taking account for continuous component values with the level of gravity, not binary information yes/no as has been done previously. The same reasoning could be applied for MetS itself, a value of gravity was indicated for each subject dependent on each of the seven components through a score (quantitative trait). In fact, when a continuous outcome variable is dichotomized, some of the information contained in the underlying distribution is discarded. In this case, whole MetS information through a continuous variable was taken into account.

Another advantage in the scoring analysis was inclusion of the possibility to define scores for this syndrome using isolated populations. The Sardinia population has been geographically isolated for centuries, has undergone low immigration and slow population growth, and is characterized by a great deal of homogeneity in their genetic pool, in life style, and eating habits. Such genetic, demographic, and environmental isolation represents an ideal condition for studying complex diseases because of a reduction in background variability due to unpredictable factors, and this approach has proven to be extremely cost and time effective (Varilo & Peltonen, 2004; A. F. Wright, Carothers, & Pirastu, 1999). To permit the comparison with other types of samples (unrelated and not isolated samples), external validation was carried out. Nevertheless, scoring analysis was performed in other populations (Gubbio and ARIC) and results confirmed good agreement and good performance.

After validation of the MetS score using different samples, the thesis was focused on estimation of genetic components using different approaches by the use of pedigree information and SNP data (Chen et al., 2015; Shetty, Qin, Namkung, Elston, & Zhu, 2011).

Using GWAS data and consortium, genome-scale sequencing data provide an opportunity to estimate relatedness of individuals using marker information and then using this relatedness to infer heritability from the proportion of phenotypic variance explained by genotyped SNPs.

These approaches have many advantages than the traditional ones. First, GWAS data allows empirical estimates of genomic sharing rather than relying on theoretical distributions used in family-based study designs. In addition, population-based datasets reduce variability and provide more precise estimates of heritability. Finally, collection of large twin or family-based cohorts is difficult and not cost-effective.

However, pedigree and within-family design also have some advantages. For example, the coefficients of relatedness are blind to allele frequencies of causal variants. Moreover, the proportion of genetic variance explained by SNPs depends on the LD of measured SNPs and unknown causal variants.

To reduce bias and to increase the precision of heritability estimates, LD pruning at different levels was conducted (one conservative level at 0.2, one less conservative at 0.8, and full information without LD pruning). Results demonstrated no significant differences between h^2 calculated using LD at 0.20, 0.80, and no LD and statistical significance in all of the cases.

In general, bias in heritability estimations may also come from environmental factors that are not modeled in an adequate way. If individuals who share SNP genotypes more often than the average also tend to share a common environment, then the heritability explained by SNPs will be overestimated.

This would be expected in the Sardinia population where closely related people were included in the sample without adjustment. In this case, the estimation of additive genetic variance could not be free of confounding by environmental factors. It is this reason why the estimates in **Table 9** are higher than heritability in family-based data adjusted for household effects (Gubbio and Sardinia) and the ARIC population adjusted for hidden population substructure.

Generally, close relatives give more precision but potentially more bias, whereas distant relatives give less precision and less bias. In addition, precision in parameter estimates depends on the total number of individuals with a phenotype. In this case, a comparison between different types of populations, including relatives and with distant relationships, allows an unbiased estimation of heritability without much error.

Beyond the methodological approaches, the key finding is that the heritability of MetS using the proposed score attributable to common genetic variants is high and significant.

When focusing on the results of heritability, in general the estimates are significant in the entire samples and ranged between 0.1 and 0.4, with the combined value at about 0.2, which indicates the presence of marked genetic components in the phenotype.

When pruned SNPs at LD = 0.80 were considered to calculate the matrices, h^2 increased in both samples; from 0.243 to 0.273 in the Sardinia population and from 0.101 to 0.152 in the ARIC population, but decreased using all SNPs. Comparing the different approaches, results from the GCTA matrix appeared higher and more stable than the IBD matrix.

The estimated heritability using the PCGC approach (**Table 9** and Chapter 4.4.5) in ARIC samples was lower but similar compared to the GCTA approach, and equal to 0.176, 0.203, and 0.174, when pruning SNPs with LD thresholds at 0.20, 0.80, and no LD, respectively. If the Sardinia samples were analyzed removing closely related family members, a similar heritability was obtained (0.203, 0.158, and 0.145 for LD at 0.20, 0.80, and no LD, respectively), indicating an overestimated heritability due to non-genetic factors.

Heritability with suggestive GWAS results (at a significance threshold $P < 10^{-4}$: 124 SNPs and 84 SNPS in Sardinia and ARIC, respectively) suggest that significant genetic components combined with environmental factors were present in the syndrome. After prior selection of variants based on association with MetS, genetic variants and genetic interaction or family environment could make important contributions to estimate unbiased heritability.

In particular, using ARIC GWAS results, the total fraction of the MetS variation explained in the identified SNPs remains small considering the total number of SNPs (after LD pruning) as expected from the literature on the “missing” heritability (Manolio et al, 2009). By comparison, in the Sardinia samples selected from an isolated population in which individuals were closely related to each other, heritability using both approaches had significant and similar values (GTCA with suggestive SNPs: 0.29, and GTCA with all SNPs: 0.33), which was also similar to heritability using a pedigree matrix without genetic markers (ACE model: 0.29 and 0.30 in Sardinia and Gubbio family samples, respectively).

As reported in the literature, LD between markers and the unknown Quantitative Trait Loci (QTL) plays a central role in determination of genomic variances, (de Los Campos et al., 2015) especially when a large proportion of the markers used in the analysis are in LE with QTL, models can be incorrectly specified. However, complex traits (e.g., MetS) are possibly affected by large numbers of small-effect QTL. Close relatives share long chromosome segments and, under these assumptions, the patterns of allele sharing at markers and at QTL are very similar.

Thus, high accuracy and very small bias in heritability estimates using SNP information can be obtained.

On the other hand, considering distantly related subjects, the addition of a large number of markers that are in LE with QTL can lead to incorrect specification of relatedness and potential inconsistencies in estimates of heritability.

One of the major limitations in this study is that this analysis does not consider newly emerging risk factors for MetS, such as the inflammatory state or procoagulant variables, that could interact with the other considered factors. Moreover, improvements could be taken into account with other risk factors. Another limitation is that results showed in this work considered only one genetic variance component. Future analyses will take into account dominance and epistatic models using new and more efficient algorithms for LMMs, for example the R package software (Covarrubias-Pazaran, 2016), or other Genomic Selection (GS)-based resources.

To summarize, these results show that a significant genetic component is present in MetS. However, future analysis could be conducted to include more interpretation that fully understands the architecture of the pathology.

Finally, GWAS analyses combining typed and imputed data were not used, because the focus of this project was to estimate the genetic component and to compare the results between different populations and Genetic Relationship Matrices. However, a meta-analysis GWAS using typed and imputed data to individual novel loci, gene pathways, and heritability, assuming that if causal QTL are in contained within the large marker panel there should be no missing heritability, will be performed in the future.

In the coming years, adding important features listed above, a fuller picture of the genetic architecture of MetS will be obtained.

7. REFERENCES

2003 European Society of Hypertension-European Society of Cardiology guidelines for the management of arterial hypertension. (2003). *J Hypertens, 21*(6), 1011-1053. doi:10.1097/01.hjh.0000059051.65882.32

Aguilar, M., Bhuket, T., Torres, S., Liu, B., & Wong, R. J. (2015). Prevalence of the metabolic syndrome in the United States, 2003-2012. *Jama, 313*(19), 1973-1974. doi:10.1001/jama.2015.4260

Alberti, K. G., Eckel, R. H., Grundy, S. M., Zimmet, P. Z., Cleeman, J. I., Donato, K. A., . . . Smith, S. C., Jr. (2009). Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation, 120*(16), 1640-1645. doi:10.1161/circulationaha.109.192644

Alberti, K. G., Zimmet, P., & Shaw, J. (2005). The metabolic syndrome--a new worldwide definition. *Lancet, 366*(9491), 1059-1062. doi:10.1016/s0140-6736(05)67402-8

Alberti, K. G., & Zimmet, P. Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet Med, 15*(7), 539-553. doi:10.1002/(sici)1096-9136(199807)15:7<539::aid-dia668>3.0.co;2-s

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet, 62*(5), 1198-1211. doi:10.1086/301844

American Diabetes, A. (2009). Diagnosis and Classification of Diabetes Mellitus. In *Diabetes Care* (Vol. 32, pp. S62-67).

Andreassi, M. G., & Botto, N. (2003). DNA damage as a new emerging risk factor in atherosclerosis. *Trends Cardiovasc Med, 13*(7), 270-275.

Avery, C. L., He, Q., North, K. E., Ambite, J. L., Boerwinkle, E., Fornage, M., . . . Lin, D. Y. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet, 7*(10), e1002322. doi:10.1371/journal.pgen.1002322

Babyak, M. A., & Green, S. B. (2010). Confirmatory factor analysis: an introduction for psychosomatic medicine researchers. *Psychosom Med, 72*(6), 587-597. doi:10.1097/PSY.0b013e3181de3f8a

Balkau, B., & Charles, M. A. (1999). Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). *Diabet Med, 16*(5), 442-443.

Barsh, G. S., Copenhaver, G. P., Gibson, G., & Williams, S. M. (2012). Guidelines for Genome-Wide Association Studies. In *PLoS Genet* (Vol. 8). San Francisco, USA.

Bellia, A., Giardina, E., Lauro, D., Tesauro, M., Di Fede, G., Cusumano, G., . . . Sbraccia, P. (2009). "The Linosa Study": epidemiological and heritability data of the metabolic syndrome in a Caucasian genetic isolate. *Nutr Metab Cardiovasc Dis, 19*(7), 455-461. doi:10.1016/j.numecd.2008.11.002

Biino, G., Balduini, C. L., Casula, L., Cavallo, P., Vaccargiu, S., Parracciani, D., . . . Pirastu, M. (2011). Analysis of 12,517 inhabitants of a Sardinian geographic isolate reveals that predispositions to thrombocytopenia and thrombocytosis are inherited traits. *Haematologica, 96*(1), 96-101. doi:10.3324/haematol.2010.029934

Blanco-Gomez, A., Castillo-Lluva, S., Del Mar Saez-Freire, M., Hontecillas-Prieto, L., Mao, J. H., Castellanos-Martin, A., & Perez-Losada, J. (2016). Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes. *Bioessays, 38*(7), 664-673. doi:10.1002/bies.201600084

Bonnet, A., Gassiat, E., & Lévy-Leduc, C. (2014). Heritability estimation in high dimensional linear mixed models. *arXiv preprint arXiv:1404.3397*.

Bosy-Westphal, A., Onur, S., Geisler, C., Wolf, A., Korth, O., Pfeuffer, M., . . . Muller, M. J. (2007). Common familial influences on clustering of metabolic syndrome traits with central obesity and insulin resistance: the Kiel obesity prevention study. *Int J Obes (Lond), 31*(5), 784-790. doi:10.1038/sj.ijo.0803481

Cameron, A. J., Magliano, D. J., Zimmet, P. Z., Welborn, T., & Shaw, J. E. (2007). The metabolic syndrome in Australia: prevalence using four definitions. *Diabetes Res Clin Pract, 77*(3), 471-478. doi:10.1016/j.diabres.2007.02.002

Cappello, N., Rendine, S., Griffo, R., Mameli, G. E., Succa, V., Vona, G., & Piazza, A. (1996). Genetic analysis of Sardinia: I. data on 12 polymorphisms in 21 linguistic domains. *Ann Hum Genet, 60*(Pt 2), 125-141.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research, 1*(2), 245-276.

Chen, F., He, J., Zhang, J., Chen, G. K., Thomas, V., Ambrosone, C. B., . . . Stram, D. O. (2015). Methodological Considerations in Estimation of Phenotype Heritability Using Genome-Wide SNP Data, Illustrated by an Analysis of the Heritability of Height in a

Large Sample of African Ancestry Adults. *PLoS One, 10*(6), e0131106. doi:10.1371/journal.pone.0131106

Cichon, S., Craddock, N., Daly, M., Faraone, S. V., Gejman, P. V., Kelsoe, J., . . . Sullivan, P. F. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry, 166*(5), 540-556. doi:10.1176/appi.ajp.2008.08091354

Cirillo, M., Terradura-Vagnarelli, O., Mancini, M., Menotti, A., Zanchetti, A., & Laurenzi, M. (2014). Cohort profile: The Gubbio Population Study. *Int J Epidemiol, 43*(3), 713-720. doi:10.1093/ije/dyt025

Cornier, M. A., Dabelea, D., Hernandez, T. L., Lindstrom, R. C., Steig, A. J., Stob, N. R., . . . Eckel, R. H. (2008). The metabolic syndrome. *Endocr Rev, 29*(7), 777-822. doi:10.1210/er.2008-0024

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS One, 11*(6), e0156744. doi:10.1371/journal.pone.0156744

de Los Campos, G., Sorensen, D., & Gianola, D. (2015). Genomic heritability: what is it? *PLoS Genet, 11*(5), e1005048. doi:10.1371/journal.pgen.1005048

Dubois, L., Ohm Kyvik, K., Girard, M., Tatone-Tokuda, F., Pérusse, D., Hjelmborg, J., . . . Martin, N. G. (2012). Genetic and Environmental Contributions to Weight, Height, and BMI from Birth to 19 Years of Age: An International Study of Over 12,000 Twin Pairs. In G. Wang (Ed.), *PLoS One* (Vol. 7). San Francisco, USA.

Eckel, R. H., Grundy, S. M., & Zimmet, P. Z. (2005). The metabolic syndrome. *Lancet, 365*(9468), 1415-1428. doi:10.1016/s0140-6736(05)66378-7

Elks, C. E., den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J. F., & Ong, K. K. (2012). Variability in the Heritability of Body Mass Index: A Systematic Review and Meta-Regression. *Front Endocrinol (Lausanne), 3*. doi:10.3389/fendo.2012.00029

Ervin, R. B. (2009). Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. *Natl Health Stat Report*(13), 1-7.

Fall, T., & Ingelsson, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. *Mol Cell Endocrinol, 382*(1), 740-757. doi:10.1016/j.mce.2012.08.018

Fawcett, K. A., & Barroso, I. (2010). The genetics of obesity: FTO leads the way. In *Trends Genet* (Vol. 26, pp. 266-274).

Ford, E. S. (2005). Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome: a summary of the evidence. *Diabetes Care, 28*(7), 1769-1778.

Ford, E. S., Giles, W. H., & Mokdad, A. H. (2004). Increasing prevalence of the metabolic syndrome among u.s. Adults. *Diabetes Care, 27*(10), 2444-2449.

Frosst, P., Blom, H. J., Milos, R., Goyette, P., Sheppard, C. A., Matthews, R. G., . . . et al. (1995). A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat Genet, 10*(1), 111-113. doi:10.1038/ng0595-111

Garver, W. S., Newman, S. B., Gonzales-Pacheco, D. M., Castillo, J. J., Jelinek, D., Heidenreich, R. A., & Orlando, R. A. (2013). The genetics of childhood obesity and interaction with dietary macronutrients. In *Genes Nutr* (Vol. 8, pp. 271-287). Berlin/Heidelberg.

Golan, D., Lander, E. S., & Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci U S A, 111*(49), E5272-5281. doi:10.1073/pnas.1419064111

Golan, D., & Rosset, S. (2011). Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics, 27*(13), i317-i323.

Graziano, F., Grassi, M., Bonati, M. T., Zanchetti, A., & Biino, G. (2016). External validation of the MetS score, a prediction tool for metabolic syndrome. *Nutr Metab Cardiovasc Dis, 26*(4), 359-360. doi:10.1016/j.numecd.2015.12.014

Graziano, F., Grassi, M., Sacco, S., Concas, M. P., Vaccargiu, S., Pirastu, M., & Biino, G. (2015). Probing the factor structure of metabolic syndrome in Sardinian genetic isolates. *Nutr Metab Cardiovasc Dis, 25*(6), 548-555. doi:10.1016/j.numecd.2015.02.004

Grundy, S. M., Brewer, H. B., Jr., Cleeman, J. I., Smith, S. C., Jr., & Lenfant, C. (2004). Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Arterioscler Thromb Vasc Biol, 24*(2), e13-18. doi:10.1161/01.atv.0000111245.75752.c6

Grundy, S. M., Cleeman, J. I., Daniels, S. R., Donato, K. A., Eckel, R. H., Franklin, B. A., . . . Costa, F. (2005). Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation, 112*(17), 2735-2752. doi:10.1161/circulationaha.105.169404

Gurka, M. J., Ice, C. L., Sun, S. S., & Deboer, M. D. (2012). A confirmatory factor analysis of the metabolic syndrome in adolescents: an examination of sex and racial/ethnic differences. *Cardiovasc Diabetol, 11*, 128. doi:10.1186/1475-2840-11-128

Gurka, M. J., Lilly, C. L., Oliver, M. N., & DeBoer, M. D. (2014). An examination of sex and racial/ethnic differences in the metabolic syndrome among adults: a confirmatory factor analysis and a resulting continuous severity score. *Metabolism, 63*(2), 218-225. doi:10.1016/j.metabol.2013.10.006

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. In *Caspian J Intern Med* (Vol. 4, pp. 627-635). Babol, Iran.

Hall, J. B., & Bush, W. S. (2016). Analysis of Heritability Using Genome-Wide Data. *Curr Protoc Hum Genet, 91*, 1.30.31-31.30.10. doi:10.1002/cphg.25

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics, 2*(1), 3-19. doi:10.1007/bf01066731

Henneman, P., Aulchenko, Y. S., Frants, R. R., van Dijk, K. W., Oostra, B. A., & van Duijn, C. M. (2008). Prevalence and heritability of the metabolic syndrome and its individual components in a Dutch isolate: the Erasmus Rucphen Family study. *J Med Genet, 45*(9), 572-577. doi:10.1136/jmg.2008.058388

Hill, W. G., & Maki-Tanila, A. (2015). Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *J Anim Breed Genet, 132*(2), 176-186. doi:10.1111/jbg.12140

Hillier, T. A., Rousseau, A., Lange, C., Lepinay, P., Cailleau, M., Novak, M., . . . Balkau, B. (2006). Practical way to assess metabolic syndrome using a continuous score obtained from principal components analysis. *Diabetologia, 49*(7), 1528-1535. doi:10.1007/s00125-006-0266-8

Hu, Z., & Yang, R. C. (2014). Marker-Based Estimation of Genetic Parameters in Genomics. In X. Cai (Ed.), *PLoS One* (Vol. 9). San Francisco, USA.

Investigators, A. (1989). The atherosclerosis risk in communities (ARIC) study: Design and objectives. *American journal of epidemiology, 129*(4), 687-702.

Janghorbani, M., & Amini, M. (2016). Utility of Continuous Metabolic Syndrome Score in Assessing Risk of Type 2 Diabetes: The Isfahan Diabetes Prevention Study. *Ann Nutr Metab, 68*(1), 19-25. doi:10.1159/000441851

Joy, T., Lahiry, P., Pollex, R. L., & Hegele, R. A. (2008). Genetics of metabolic syndrome. *Curr Diab Rep, 8*(2), 141-148.

Karssen, L. C., van Duijn, C. M., & Aulchenko, Y. S. (2016). The GenABEL Project for statistical genomics. *F1000Res, 5*. doi:10.12688/f1000research.8733.1

Kassi, E., Pervanidou, P., Kaltsas, G., & Chrousos, G. (2011). Metabolic syndrome: definitions and controversies. *BMC Med, 9*, 48. doi:10.1186/1741-7015-9-48

Khan, R. J., Gebreab, S. Y., Sims, M., Riestra, P., Xu, R., & Davis, S. K. (2015). Prevalence, associated factors and heritabilities of metabolic syndrome and its individual components in African Americans: the Jackson Heart Study. *BMJ Open, 5*(10), e008675. doi:10.1136/bmjopen-2015-008675

Khoury, M. J., Beaty, T. H., & Cohen, B. H. (1993). *Fundamentals of genetic epidemiology* (Vol. 22): Oxford University Press, USA.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*: Guilford publications.

Kraja, A. T., Vaidya, D., Pankow, J. S., Goodarzi, M. O., Assimes, T. L., Kullo, I. J., . . . Borecki, I. B. (2011). A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. *Diabetes, 60*(4), 1329-1339. doi:10.2337/db10-1011

Kristiansson, K. (2012). Genome-Wide Screen for Metabolic Syndrome Susceptibility Loci Reveals Strong Lipid Gene Contribution but No Evidence for Common Genetic Basis for Clustering of Metabolic Syndrome Traits. *5*(2), 242-249. doi:10.1161/circgenetics.111.961482

Kumar, S. K., Feldman, M. W., Rehkopf, D. H., & Tuljapurkar, S. (2016). Response to Commentary on "Limitations of GCTA as a solution to the missing heritability problem". *bioRxiv*. doi:10.1101/039594

Laaksonen, D. E., Niskanen, L., Nyyssönen, K., Lakka, T. A., Laukkanen, J. A., & Salonen, J. T. (2008). Dyslipidaemia as a predictor of hypertension in middle-aged men. In *Eur Heart J* (Vol. 29, pp. 2561-2568).

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. In *Nucleic Acids Res* (Vol. 37, pp. 4181-4193).

Lechleitner, M. (2008). Obesity and the metabolic syndrome in the elderly--a mini-review. *Gerontology, 54*(5), 253-259. doi:10.1159/000161734

Lee, J. J., & Chow, C. C. (2014). Conditions for the validity of SNP-based heritability estimation. *Hum Genet, 133*(8), 1011-1022. doi:10.1007/s00439-014-1441-5

Lee, J. J., Vattikuti, S., & Chow, C. C. (2016). Uncovering the Genetic Architectures of Quantitative Traits. *Computational and Structural Biotechnology Journal, 14*, 28-34. doi:http://dx.doi.org/10.1016/j.csbj.2015.10.002

Lee, S., Cho, Y., Lim, D., Kim, H., Choi, B., Park, H., . . . Yoon, D. (2011). Linkage disequilibrium and effective population size in hanwoo korean cattle. *Asian-Australasian Journal of Animal Sciences, 24*(12), 1660-1665.

Li, C., & Ford, E. S. (2007). Is there a single underlying factor for the metabolic syndrome in adolescents? A confirmatory factor analysis. *Diabetes Care, 30*(6), 1556-1561. doi:10.2337/dc06-2481

Lin, H. F., Boden-Albala, B., Juo, S. H., Park, N., Rundek, T., & Sacco, R. L. (2005). Heritabilities of the metabolic syndrome and its components in the Northern Manhattan Family Study. *Diabetologia, 48*(10), 2006-2012. doi:10.1007/s00125-005-1892-2

Liu, N., Zhao, H., Patki, A., Limdi, N. A., & Allison, D. B. (2011). Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals. *Stat Interface, 4*(3), 317-326.

Loos, R. J., Katzmarzyk, P. T., Rao, D. C., Rice, T., Leon, A. S., Skinner, J. S., . . . Bouchard, C. (2003). Genome-wide linkage scan for the metabolic syndrome in the HERITAGE Family Study. *J Clin Endocrinol Metab, 88*(12), 5935-5943. doi:10.1210/jc.2003-030553

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

Martinez-Vizcaino, V., Martinez, M. S., Aguilar, F. S., Martinez, S. S., Gutierrez, R. F., Lopez, M. S., . . . Rodriguez-Artalejo, F. (2010). Validity of a single-factor model underlying the metabolic syndrome in children: a confirmatory factor analysis. *Diabetes Care, 33*(6), 1370-1372. doi:10.2337/dc09-2049

McDonald, R. P. (2013). *Test theory: A unified treatment*: Psychology Press.

Menotti, A., Lanti, M., Angeletti, M., Botta, G., Cirillo, M., Laurenzi, M., . . . Zanchetti, A. (2009). Twenty-year cardiovascular and all-cause mortality trends and changes in cardiovascular risk factors in Gubbio, Italy: the role of blood pressure changes. *J Hypertens, 27*(2), 266-274. doi:10.1097/HJH.0b013e32831cbb0b

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics, 163*(3), 1153-1167.

Mirhafez, S. R., Avan, A., Pasdar, A., Khatamianfar, S., Hosseinzadeh, L., Ganjali, S., . . . Ghayour-Mobarhan, M. (2016). Zinc Finger 259 Gene Polymorphism rs964184 is Associated with Serum Triglyceride Levels and Metabolic Syndrome. In *Int J Mol Cell Med* (Vol. 5, pp. 8-18). Babol, Iran.

Mitchell, J. S., Johnson, D. C., Litchfield, K., Broderick, P., Weinhold, N., Davies, F. E., . . . Houlston, R. S. (2015). Implementation of genome-wide complex trait analysis to quantify the heritability in multiple myeloma. *Scientific Reports, 5*, 12473. doi:10.1038/srep12473

http://www.nature.com/articles/srep12473 - supplementary-information

Norton, B., & Pearson, E. S. (1976). A note on the background to, and refereeing of, R. A. Fisher's 1918 paper 'On the correlation between relatives on the supposition of Mendelian inheritance'. *Notes Rec R Soc Lond, 31*(1), 151-162.

O'Rahilly, S., & Farooqi, I. S. (2006). Genetics of obesity. In *Philos Trans R Soc Lond B Biol Sci* (Vol. 361, pp. 1095-1105). London.

Olokoba, A. B., Obateru, O. A., & Olokoba, L. B. (2012). Type 2 Diabetes Mellitus: A Review of Current Trends. In *Oman Med J* (Vol. 27, pp. 269-273).

Pladevall, M., Singal, B., Williams, L. K., Brotons, C., Guyer, H., Sadurni, J., . . . Haffner, S. (2006). A single factor underlies the metabolic syndrome: a confirmatory factor analysis. *Diabetes Care, 29*(1), 113-122.

Pollex, R. L., & Hegele, R. A. (2006). Genetic determinants of the metabolic syndrome. *Nat Clin Pract Cardiovasc Med, 3*(9), 482-489. doi:10.1038/ncpcardio0638

Povel, C. M., Beulens, J. W., van der Schouw, Y. T., Dolle, M. E., Spijkerman, A. M., Verschuren, W. M., . . . Boer, J. M. (2013). Metabolic syndrome model definitions predicting type 2 diabetes and cardiovascular disease. *Diabetes Care, 36*(2), 362-368. doi:10.2337/dc11-2546

Povel, C. M., Boer, J. M., Reiling, E., & Feskens, E. J. (2011). Genetic variants and the metabolic syndrome: a systematic review. *Obes Rev, 12*(11), 952-967. doi:10.1111/j.1467-789X.2011.00907.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Ragland, D. R. (1992). Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology, 3*(5), 434-440.

Reaven, G. M. (1988). Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes, 37*(12), 1595-1607.

Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. In *Nat Genet* (Vol. 40, pp. 491-492). United States.

Rich, S. S. (1990). Mapping genes in diabetes. Genetic epidemiological perspective. *Diabetes, 39*(11), 1315-1319.

Scuteri, A., Laurent, S., Cucca, F., Cockcroft, J., Cunha, P. G., Manas, L. R., . . . Nilsson, P. M. (2015). Metabolic syndrome across Europe: different clusters of risk factors. *Eur J Prev Cardiol, 22*(4), 486-491. doi:10.1177/2047487314525529

Shen, B. J., Goldberg, R. B., Llabre, M. M., & Schneiderman, N. (2006). Is the factor structure of the metabolic syndrome comparable between men and women and across three ethnic groups: the Miami Community Health Study. *Ann Epidemiol, 16*(2), 131-137. doi:10.1016/j.annepidem.2005.06.049

Shetty, P. B., Qin, H., Namkung, J., Elston, R. C., & Zhu, X. (2011). Estimating heritability using family and unrelated individuals data. In *BMC Proc* (Vol. 5, pp. S34).

Shin, J., & Lee, C. (2015). A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics, 105*(4), 191-196. doi:10.1016/j.ygeno.2015.01.006

Shirali, M., Pong-Wong, R., Navarro, P., Knott, S., Hayward, C., Vitart, V., . . . Haley, C. S. (2016). Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. *Heredity (Edinb), 116*(3), 333-338. doi:10.1038/hdy.2015.107

Soldatovic, I., Vukovic, R., Culafic, D., Gajic, M., & Dimitrijevic-Sreckovic, V. (2016). siMS Score: Simple Method for Quantifying Metabolic Syndrome. *PLoS One, 11*(1), e0146143. doi:10.1371/journal.pone.0146143

Srivastava, A., Srivastava, N., & Mittal, B. (2016). Genetics of Obesity. *Indian Journal of Clinical Biochemistry, 31*(4), 361-371. doi:10.1007/s12291-015-0541-x

Team, R. C. D. (2005). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. In: Vienna Austria.

Tekola-Ayele, F., Doumatey, A. P., Shriner, D., Bentley, A. R., Chen, G., Zhou, J., . . . Rotimi, C. N. (2015). Genome-wide association study identifies African-ancestry specific variants for metabolic syndrome. *Mol Genet Metab, 116*(4), 305-313. doi:10.1016/j.ymgme.2015.10.008

Teran-Garcia, M., & Bouchard, C. (2007). Genetics of the metabolic syndrome. *Appl Physiol Nutr Metab, 32*(1), 89-114. doi:10.1139/h06-102

Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. (2002). *Circulation, 106*(25), 3143-3421.

Thompson, D. B., Ravussin, E., Bennett, P. H., & Bogardus, C. (1997). Structure and sequence variation at the human leptin receptor gene in lean and obese Pima Indians. *Human Molecular Genetics, 6*(5), 675-679.

van Dongen, J., Willemsen, G., Chen, W. M., de Geus, E. J., & Boomsma, D. I. (2013). Heritability of metabolic syndrome traits in a large population-based sample. *J Lipid Res, 54*(10), 2914-2923. doi:10.1194/jlr.P041673

van Vliet-Ostaptchouk, J. V., Nuotio, M.-L., Slagter, S. N., Doiron, D., Fischer, K., Foco, L., . . . Hiekkalinna, T. (2014). The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. *BMC endocrine disorders, 14*(1), 1.

Varilo, T., & Peltonen, L. (2004). Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev, 14*(3), 316-323. doi:10.1016/j.gde.2004.04.008

Vattikuti, S., Guo, J., & Chow, C. C. (2012). Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. In P. M. Visscher (Ed.), *PLoS Genet* (Vol. 8). San Francisco, USA.

Viitasalo, A., Lakka, T. A., Laaksonen, D. E., Savonen, K., Lakka, H. M., Hassinen, M., . . . Rauramaa, R. (2014). Validation of metabolic syndrome score by confirmatory factor analysis in children and adults and prediction of cardiometabolic outcomes in adults. *Diabetologia, 57*(5), 940-949. doi:10.1007/s00125-014-3172-5

Vinkhuyzen, A. A., Wray, N. R., Yang, J., Goddard, M. E., & Visscher, P. M. (2013). Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet, 47*, 75-95. doi:10.1146/annurev-genet-111212-133258

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet, 9*(4), 255-266. doi:10.1038/nrg2322

Visscher, P. M., Yang, J., & Goddard, M. E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res Hum Genet, 13*(6), 517-524. doi:10.1375/twin.13.6.517

Visscher P , M., Brown M , A., McCarthy M , I., & Yang, J. (2012). Five Years of GWAS Discovery. In *Am J Hum Genet* (Vol. 90, pp. 7-24).

Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet, 6*(2), 109-118. doi:10.1038/nrg1522

Waterworth, D. M., Ricketts, S. L., Song, K., Chen, L., Zhao, J. H., Ripatti, S., . . . Sandhu, M. S. (2010). Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol, 30*(11), 2264-2276. doi:10.1161/atvbaha.109.201020

Weng, T. (2001). Evaluation of diagnostic tests: measuring degree of agreement and beyond. *Drug information journal, 35*(2), 577-588.

Wijndaele, K., Beunen, G., Duvigneaud, N., Matton, L., Duquet, W., Thomis, M., . . . Philippaerts, R. M. (2006). A continuous metabolic syndrome risk score: utility for epidemiological analyses. In *Diabetes Care* (Vol. 29, pp. 2329). United States.

Wiley, J. F., & Carrington, M. J. (2016). A metabolic syndrome severity score: A tool to quantify cardio-metabolic risk factors. *Preventive medicine, 88*, 189-195.

Woodward, M., & Tunstall-Pedoe, H. (2009). The metabolic syndrome is not a sensible tool for predicting the risk of coronary heart disease. *Eur J Cardiovasc Prev Rehabil, 16*(2), 210-214. doi:10.1097/HJR.0b013e3283282f8d

Wright, A. F., Carothers, A. D., & Pirastu, M. (1999). Population choice in mapping genes for complex diseases. *Nat Genet, 23*(4), 397-404. doi:10.1038/70501

Wright, S. (1921). Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics, 6*(2), 111-123.

Wu, Y., Yu, S., Wang, S., Shi, J., Xu, Z., Zhang, Q., . . . Yaqin, Y. (2015). Zinc Finger Protein 259 (ZNF259) Polymorphisms are Associated with the Risk of Metabolic Syndrome in a Han Chinese Population. *Clin Lab, 61*(5-6), 615-621.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet, 42*(7), 565-569. doi:10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. In *Am J Hum Genet* (Vol. 88, pp. 76-82).

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., . . . Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet, 43*(6), 519-525. doi:10.1038/ng.823

Yang, S., Liu, Y., Jiang, N., Chen, J., Leach, L., Luo, Z., & Wang, M. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics, 15*, 13. doi:10.1186/1471-2164-15-13

Zaitlen, N., & Kraft, P. (2012). Heritability in the genome-wide association era. *Hum Genet, 131*(10), 1655-1664. doi:10.1007/s00439-012-1199-6

Zhang, H., Liu, L., Wang, X., & Gruen, J. R. (2007). Guideline for data analysis of genomewide association studies. *Cancer Genomics Proteomics, 4*(1), 27-34.

Zhang, W., Duan, S., Kistner, E. O., Bleibel, W. K., Huang, R. S., Clark, T. A., . . . Cox, N. J. (2008). Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet, 82*. doi:10.1016/j.ajhg.2007.12.015

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*(24), 3326-3328. doi:10.1093/bioinformatics/bts606

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet, 44*(7), 821-824. doi:10.1038/ng.2310

Zimmet, P., Alberti, G., & Shaw, J. (2005). A new IDF worldwide definition of the metabolic syndrome: the rationale and the results. *Diabetes Voice, 50*(3), 31.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A, 109*(4), 1193-1198. doi:10.1073/pnas.1119675109

8. ACKNOWLEDGMENT

I would like to thank my PhD supervisor professor Mario Grassi for the support he has given me across these years. I also want to thank the head of the department, professor Luisa Bernardinelli.

Sincerely thank to Prof. Benjamin Neale and to Prof. Marcel De Hoed for the opportunity to work with them in their Labs, teams and resources.

I would like to thank my colleagues, Teresa, Emilia, Gabriele, Fernando, for the help and friendship across these years in Pavia.

On the personal site, I would like to thank my fiancé, Francesco for everything (almost), and of course my family.