Università Degli Studi di Napoli "Federico II"

Dipartimento di Biologia e Patologia Cellulare e Molecolare

"L. Califano"

Tesi di Dottorato di Ricerca in Fisiopatologia e Patologia Molecolare

THREE METHODS TO INCREASE THE LIKELY TO IDENTIFY GENE INVOLVED IN COMPLEX DISEASE

Candidato: dott. **Michele Pinelli**

Docente Guida: prof. **Sergio Cocozza**

Coordinatore del Dottorato: prof. **Vittorio Enrico Avvedimento**

XXI Ciclo di Dottorato, anni 2005-2008

I greatly respect past successes of medical genetics, since they were achieved with limited means and crude methods. But medical genetics are not dead or finished disciplines – it is just starting now.

John P. A. Ioannidis

Table of Contents

Table of contents....III
Summary and thesis plan....V
Abbreviations....VI
Publications....VII
.
Chapter 1: Complex Disease....1
Introduction....1
Why to study the genetics of complex disease....6
Type 2 Diabetes Mellitus....7
Genetics DM....9
Obesity....10
Genetics of Obesity....14
Thrift genotype hypothesis....15
Aim....17
References....18

Chapter 2: Evolution-enhanced candidate gene method identified association of ACO1 gene with Type 2 Diabetes Mellitus....29
Introduction....29
Methods....34
Selective Pressure Information....34
Sequence- and Annotation-based Candidate Gene Selection....35
GeneSeeker....36
Analysis of candidate gene expression using eVOC annotation....37
Disease Gene Prediction (DGP)....38
PROSPECTR/SUSPECTS....39
G2D....41
POCUS....42
Merging candidate gene lists....45
Collecting case-control cohort for T2DM....45
Statistical Methods....46
Results....47
Conclusions....52
References....53

Chapter 3: Pro12Ala polymorphism of the PPARγ locus modulates the relationship between energy intake and body weight in type 2 diabetic patients....60
Introduction....60
Methods....62
Statistical analysis....63
RESULTS....65
CONCLUSIONS....70
References....73

Chapter 4: Statistical method to identify gene-environment interaction.....77
Introduction.....77
Methods.....80
Feature Selection Methods.....80
Univariate methods.....80
Backward stepwise logistic regression.....80
Multifactor Dimensionality Reduction.....81
Linear Discriminant Analysis.....83
Population simulator.....84
Analysis strategy.....90
Ensemble.....92
Real World sample.....92
Results.....94
Comparison of different FSM against simulated populations.....94
Real World Sample.....99
Conclusion.....104
References.....105

Summary and thesis plan

The large part of human pathology is composed by *complex disease*, such as heart disease, obesity, cancer, diabetes, and many common psychiatric and neurological conditions. The common feature of all these conditions is the multifactorial etiology that involves both genetic and environmental factors. The common disease-common variant (CDCV) hypothesis posits that common, interacting alleles underlie most common diseases, in association with environmental factors. Furthermore, according to the *thrift genotype,* such alleles have been subjected to selective pressure, mainly those involved in metabolic disease such as T2DM and obesity.

Although the concept of gene-environment interaction is central to ecogenetics, and has long been recognized by geneticists (Haldane 1946), there are relatively few detailed descriptions of gene–environment interaction in biomedical literature. This lacking may be explained by difficulties in collecting environmental information of enough quality and by great difficulties in analyze them. Indeed, when the number of factors to analyze is large, become overwhelming the course of dimensionality and the multiple testing problems.

In the present thesis the hypothesis that knowledge-driven approaches may improve the ability to identify genes involved in complex disease was checked. Three approaches have been presented, each of them leading to the identification of a factor or of a interaction of factors. As the study a complex disease is composed by three steps: (1) selection of candidate genes, (2) collecting of genetic and non-genetic information and (3) statistical analysis of data, it is showed that

each of these steps may be improved by consideration of the biological background.

The first study, regarded the possibility to exploit evolutionary information to identify genes involved in type 2 diabetes. This hypothesis was based on the thrifty genotype hypothesis. A gene was identified, ACO1, and was successfully associated to the disease.

In the second study, we analyses the case of a gene, PPAGγ that have been inconsistency associated with obesity. We hypothesized that the inconsistence of association may be due to its relationship with environment. Then we jointly analyzed the genotype of the gene and comprehensive nutritional information about a cohort and proved an interaction. The genotype of PPARγ modulated the response to the diet. Ala-carriers gained more weight than ProPro individuals when had the same caloric intake.

In the third study, we implemented a software tool to create simulated populations based on gene-environment interactions. The system was based on genetic information to simulate realistic populations. We used these simulated populations to collect information on statistical methods more frequently used to study case-controls samples. Afterward, we built an ensemble of these methods and applied it to a real sample. We showed that ensemble had better performances of each single methods in condition of small sample size.

Genetics of complex disease is becoming exclusive field of epidemiology and large consortia. In this scenario, studies are based on brute-force approaches,

using even-larger sample sizes and genotyping capabilities. However it may be difficult to imagine a consortium for each phenotype and that evidence-based approach may study complex genetics phenomena. Indeed, a more knowledge-driven approach may increase the likelihood to shed light on the genetics of complex disease.

Abbreviations

GWAS: Genome-Side Association Study
T2DM: Type 2 Diabetes Mellitus
CDCV: Common disease common variant hypothesis
CGM: Candidate Gene Selection Method
FSM: Feature Selection Method

Publications

Castaldo I, Pinelli M, Monticelli A, Acquaviva F, Giacchetti M, Filla A, Sacchetti S, Keller S, Avvedimento VE, Chiariotti L, Cocozza S. DNA methylation in intron 1 of the frataxin gene is related to GAA repeat length and age of onset in Friedreich's ataxia patients. J Med Genet. 2008 Aug 12. [Epub ahead of print] PMID: 18697824

Acquaviva F, Castaldo I, Filla A, Giacchetti M, Marmolino D, Monticelli A,Pinelli M, Saccà F, Cocozza S. Recombinant human erythropoietin increases frataxin protein expression without increasing mRNA expression. Cerebellum. 2008;7(3):360-5. PMID: 18581197

Ciaramella A, Cocozza S, Iorio F, Miele G, Napolitano F, Pinelli M, Raiconi G, Tagliaferri R.Interactive data analysis and clustering of genomic data. Neural Netw. 2008 Mar-Apr;21(2-3):368-78. Epub 2007 Dec 31. Erratum in: Neural Netw. 2008 May;21(4):698.PMID: 18255261

R Tagliaferri, A Ciaramella, S Cocozza, F Iorio, F Napolitano, M Pinelli, G Raiconi & G Miele. Clustering, Assessment and Validation: an application to gene expression data. 2007 International Joint Conference on Neural Networks (IJCNN), Orlando, Florida, USA, 2007

Vaccaro O, Lapice E, Monticelli A, Giacchetti M, Castaldo I, Galasso R, Pinelli M, Donnarumma G, Rivellese AA, Cocozza S, Riccardi G. Pro12Ala polymorphism of the PPARgamma2 locus modulates the relationship between energy intake and body weight in type 2 diabetic patients. Diabetes Care. 2007 May;30(5):1156-61. Epub 2007 Jan 26. PMID: 17259473

Pinelli M, Giacchetti M, Acquaviva F, Cocozza S, Donnarumma G, Lapice E, Riccardi G, Romano G, Vaccaro O, Monticelli A. Beta2-adrenergic receptor and UCP3 variants modulate the relationship between age and type 2 diabetes mellitus. BMC Med Genet. 2006 Dec 6;7:85. PMID: 17150099

Chapter 1: Complex Disease

Introduction

The large part of human pathology is composed by *complex disease*, such as heart disease, obesity, cancer, diabetes, and many common psychiatric and neurological conditions [1, 2]. The common feature of all these conditions is the multifactorial etiology that involves both genetic and environmental factors [3, 4].

The burden of common complex disease is rapidly increasing worldwide. It has been calculated that, in 2001, they contributed approximately 60% of the 56.5 million total reported deaths in the world and approximately 46% of the global burden of disease. It has also been projected that, by 2030, complex diseases will account for almost three-quarters of all deaths worldwide [5]. More than half of them are attributable to cardiovascular diseases, cancer, diabetes and obesity. These conditions are also showing worrying trends, not only because they already affect a large proportion of the population, but also because they have started to appear earlier in life. Furthermore for the ageing of populations in low- and middle-income countries, complex disease problem is far from being limited to the developed regions of the world [6].

For most complex diseases are well recognized a familiar predisposition. In many cases also a genetic predisposition has been proved by classical genetics methods, such as the analysis of concordance between twins, the evaluation of the frequency of disease between families and between populations. Although

complex diseases tend to cluster within families, they do not segregate in a mendelian fashion [7] and they are further caused by an interplay between genetic and environmental factors [3]. Environment and life-style are major contributors to the pathogenesis, nevertheless genetics background could be the necessary condition to allow the damaging effects of environmental factors. In fact, not all the people subject to an environmental exposure develop a disease (i.e. not all smokers develop a cancer). Moreover, a gene or a combination of genes might make an individual sensible to an environment and other combinations of genes might make him susceptible to a different environment [7]. For these reason different genetic backgrounds, different environmental susceptibilities, and resulting different gene-environment interactions could be present in different families. Therefore, when considered by a population level, most susceptibility alleles result conferring only a modest increase in risk and are neither necessary nor sufficient to cause disease. A popular model of the genetic architecture of common disease posits that the minor-allele frequencies (MAFs) of genetic variants influencing susceptibility are often also common (i.e.,≥ 1%) and that such alleles are therefore old and found in multiple populations, rather than being rare and population specific. This model is known as the common-variant/common-disease (CV/CD) hypothesis [8]. Under this model, disease susceptibility is suggested to result from the joint action of several common variants, and unrelated affected individuals share a significant proportion of disease alleles [9]. There is currently not enough empirical evidence to either prove or disprove the CD/CV hypothesis. However, a few prototypical examples of such common

variants are known, i.e Pro12Ala PPARγ in both T2DM [10] and obesity [11], rs7903146 TCF7L2 in T2DM [12] and rs9939609 FTO in obesity [13] that have been studied in various populations. Furthermore, a large meta-analyses suggested that disease causing alleles presently know are largely shared among ethnic groups [14]. Also simulation studies provided support for the common disease–common variant hypothesis [15].

Although the concept of gene-environment interaction is central to ecogenetics, and has long been recognized by geneticists [16], there are relatively few detailed descriptions of gene–environment interaction in biomedical literature [4].

To find genes involved in a disease is necessary to prove a significant association between disease and a functional polymorphism. Generally, this is achieved by comparing a random sample of unrelated affected individuals with a matched control group. This approach may reveal a polymorphic allele that is increased in frequency in the patient group and such a significant association might point towards a disease-susceptibility locus [17, 18]. Classically this approach can be applied to a selected loci (candidate gene) or to a set of markers along the genome (genome-wide approach). Both methods have points of strength and weakness. Candidate gene studies, being hypothesis-driven, allow a more specific description of a phenomenon, in this setting is possible to validate complex hypothesis and shed light on specific physiologic process. The main weakness of the candidate gene approach is the difficulty to make hypotheses because in complex disease the number of elements involved often are very large. Genome-Wide Studies (GWAS) allow the identification of loci associated with the disease,

with an unbias, brute-force approach. On the other hand, GWAS are expensive, difficult to organize, because often require a very large sample size, and have overwhelming statistical problems that allow only simple analyses. In fact, in the last years have been performed several Genome-Wide Association Studies (GWAS) allowing the imputation of a large number of loci in many complex disease [19, 20]. However in most of them there is no consideration of any environmental factors role [4, 17, 20]. It is likely that this lacking is in part caused by difficulties that are encountered both at sampling and analytical level. At the sampling level, just collecting enough environmental and clinical data, of a quality that can allow a gene-environment interaction could be a compelling task, especially when the sample size is large and the information are not natively of numeric or categorical type. At a statistical level, analyzing relationships between factors, even with few factors, could lead to overwhelming problems such as the course of dimensionality and the multiple testing problem [17]. The course of the dimensionality is when the number of possible categories is relatively larger than the sample size. I.e. in a genetic association study all the possible genotypes of three biallelic SNPs results in 81 combinations and, in such a situation, only with a large sample size is possible to have enough individuals in each combination to statistically evaluate the interaction [17, 21]. On the other hand, the multiple testing problem occurs when a researcher wants to study all the possible interactions between factors [17]. For example, in the case of 2-elements interaction among 100 factors there are 4950 possible combinations! Then if a p value threshold of 0.05 is imposed for each test, in other terms a probability to

have 1/20 of false positive results, and 4950 tests are performed, we could expect to have 247 results obtained just by chance. In such a situation, to keep a experiment-wise p-value threshold equal to 0.05 the researcher has to proportionally adjust the p-values of each test. The most conservative option is to multiply each test p-value by the number of performed tests and check if the new p-value is still lower the 0.05 value, the Bonferroni correction method [22]. In such a case, only very strong effect could be individuated and it is likely to reject many false negative results. Furthermore, the interactions between genetic and environmental factors could be in several cases of a complex non-linear nature. For this reason several statistical methods, first of all the Binary Logistic Regression could be not efficient to identify involved factors [21]. Although some further methods have been proposed, such as MDR, there is still lacking a proper method to study complex interactions. The lacking of a proper method to analyze complex interactions could be a further reason of the rarity of this type of study.

The difficulties in the identification of complex interactions is particularly high in GWAS, whereas utilizing a gene candidate approach could overcome some of these problems. This, mainly because in candidate gene approach all the study steps are tailored on specific hypotheses and a shorter number of variables have to be collected and analyzed. However, in this case the greatest difficulty is the appropriate selection of the genes to study. Classically the candidate gene process relied on information of gene function or on involvement in same o similar disease. A common critic is that we have only few information for most of genes and there is an important bias in favor of few *popular* genes. This kind of

approach should further increase the bias forward these popular genes. To avoid this problem, further candidate gene selection methods have been developed based on sequence analyses. These methods rely on the assumption that genes involved in diseases tend to share some sequence characteristics, as length, few paralogs, highly evolutionary conservations [23]. According to this method a geneome-wide scanning of the gene sequences could output a set of genes that has high likely to be involved in a disease. The major weakness of this method is that the large part of the criteria used to search for candidate genes are based on monogenic disease genes.

Why to study the genetics of complex disease

The difficulties to find genes involved in complex disease, the large amount of money invested in this searching, and the notion that in several cases the genetic risk is lower than most of the environmental risk have risen doubt on the overall utility of this researches [24].

There are at least three order of reason by which the identification of genetics of complex disease is important: at a population, individual and physiological level. By the population level, the understanding that in a specific population there is a genetic predisposition to a disease or to be particularly sensible to a environmental exposure could drive public health policy [4]. At individual levels, genomic information could be used to predict the future occurrence of disease for patients and their families, design interventions, and tailor therapeutic strategies to individual patients [25]. Furthermore, the notion that a discovery in human

genetics consists of identifying ‘the gene’ for a disease should be overcome. This effort serves also, and perhaps mainly, for discovery science. It will help to determine the mechanisms of gene function and how they are perturbed in different situations, ultimately providing insights into possible preventive or therapeutic strategies [26]. Large resources of information on biologic mechanism are generated without necessarily knowing in advance which pieces of information will prove most important for human health [27]. Nevertheless, these findings may not only discover “new genes”, but permit advances in our understanding of how human evolution has “used them” to develop the diseases that are common today [28].

Type 2 Diabetes Mellitus

The number of cases of diabetes worldwide in 2000 among adults older than 20 years of age is estimated to be 171 million with the vast majority being cases of Type 2 Diabetes Mellitus (T2DM). In USA have been estimated that more than one in three born in 2000 will develop type 2 diabetes [28]. Furthermore, even if the prevalence of obesity remains stable, which seems unlikely, it is anticipated that the worldwide number of people with diabetes will more than double until 2030 [29].

Italy is among the 10 countries with highest number of affected [29]. The diabetes prevalence is 8.4% in males and 6% in females [30] and it accounts for 3% of all the deaths nationwide [31].

T2DM often leads to a number of long-term complications, generally subdivided

into micro- and macrovascular complications. It is these long-term chronic complications that have the greatest impact on the health and quality of life of patients. The microvascular complications include retinopathy, neuropathy and nephropathy, with T2DM being one of the main causes of blindness, lower limb amputations, and renal failure in adults. The macrovascular complications mean that T2DM is a major risk factor for cardiovascular disease and stroke. Diabetic patients have 2-4 times higher rate to die for cardiovascular accidents [32]. The overall mortality rate is double in individuals affected by diabetes and is mainly linked to cardiovascular disease [33]. These chronic complications have a high socio-economic cost and put a heavy burden on public health services [32].

The T2DM is characterized by elevated plasma glucose levels. Normal glucose homeostasis depends on the balance between glucose production by the liver , and glucose uptake by the brain, muscle and adipose tissue. Insulin, the predominant anabolic hormone involved, increases the uptake of glucose from the blood, enhances its conversion to glycogen and triglyceride and also increases glucose oxidation. Plasma glucose levels are normally kept within a small range (4 to 6 mmol/l) by multiple mechanisms. After a meal, a small increase in plasma glucose will lead to an increased insulin secretion by the pancreatic β-cells.

Both insulin's inhibitory effect on liver glucose production and its stimulatory effect on peripheral glucose uptake are diminished. Although many T2DM patients have a basal hyperinsulinemia, elevations in plasma glucose have a characteristically reduced stimulatory effect on insulin secretion.

Many risk factors have been identified which influence the prevalence and the

incidence. Factors of particular importance are a family history of T2DM, age, overweight, increased abdominal fat, hypertension, lack of physical exercise, and ethnic background [34]. The familial predisposition could indicate for the involvement of genes in people's susceptibility for the disease.

Genetics DM

According to the multifactorial model, predisposition to the T2DM could be determined by many different combinations of genetic variants (genotypes) and environmental factors. The genetically predisposed individuals will not necessarily develop the overt syndrome unless they are also exposed to particular environmental factors [35]. It is well known that exogenous factors such as age, physical activity, diet, and obesity, play a major role in the disease aetiology of T2DM [36]. However, there are several evidence proving the genetic bases of T2DM [37, 38].

Evidence from family and twin studies. The common familial aggregation of T2DM is clearly consistent with a genetic component to disease susceptibility, although a shared environment may also contribute. The extent of familial aggregation is often summarized in terms of the sibling relative risk (ls, the ratio of disease prevalence in the siblings of affected individuals compared with that in the general population). Ls for T2DM in European populations is approximately 3.5 (35% versus 10%) [39]. The patterns of segregation in families with T2DM are (with rare exceptions, such as genetically determined maturity onset diabetes of the young – MODY) consistent with a complex, multifactorial inheritance [40].

Several studies have shown higher concordance rates in monozygotic (MZ) twins than in dizygotic (DZ) twins [41] for example, in a population-based cohort of twins in Finland, the concordance rate in MZ twins was 34% whereas in DZ twins it was 16% [42]. In a Japanese study these figures were 83% for MZ twins and 40% for DZ twins [43]. Such figures show the difference of environmental influences within populations (i.e. the difference between MZ and DZ twins). The large variation in concordance rates between populations may be due to bias or a different selection from the populations studied, but it may also indicate differences in genetic susceptibility between these populations [44, 45].

Evidence from population studies. The high prevalence of T2DM in some populations, such as Nauruan Islanders and Pima Indians, is also consistent with a genetic aetiology [46, 47]. Migration studies provide additional evidence in favor of the genetic basis of the disease. For example, individuals from the Indian subcontinent, for example, have high prevalence rates of T2D whether in urban India [48]or as migrants [49]. The prevalence of T2D in elderly Nauruans was reported to be 83% in full-blooded islanders but only 17% in those with (unsuspected) foreign genetic admixture [50]. Since there were no apparent cultural differences between the groups, this indicated a protective effect of foreign genotypes on diabetes risk. Similar findings have been reported in Pima Indians [33] and other Native American populations [51].

Obesity

Obesity can be described as an excess amount of fat tissue accumulated as a result

of imbalance between energy intake and energy expenditure. There are large differences between countries in the prevalence of obesity [52]. However, obesity has become increasingly prevalent both in Western societies and in developing countries [52-54]. The prevalence of obesity is high for example in Eastern Europe, Eastern Mediterranean, North, Central and South America (especially in the US, Argentina, Chile, Paraguay and Mexico), as well as in many Western European countries [52, 55, 56]. There are certain isolated Pacific Islands such as Samoa, Nauru, Tonga, the Cook Islands and French Polynesia where obesity is extremely common with a prevalence close to 75% [53]. Within some of these ethnic groups large physical size is still considered as a mark of beauty and social status.

The BMI is a crude measure of adiposity but correlates well with body fatness [57-59]. BMI is calculated by dividing person's weight in kilograms by square of person's height in meters. The cut-off points proposed by the World Health Organization (WHO) for defining obesity is 30 Kg/m^2, [60].

The marked increase in the prevalence of obesity in the US during the last 20 years is well documented in the reports of the Behavioral Risk Factor Surveillance System, conducted annually in the US by the Centers for Disease Control and Prevention (http://www.cdc.gov/nccdphp/dnpa/obesity/trend/) [61]. Data collected for 1999-2002 estimates that 30.4% of the adult US population has a BMI more than 30 kg/m2 and is thus considered obese [62]. This number has almost doubled when compared to the results from the same survey in 1976-1980.

In Italy, the data seem to confirm the worldwide trends. Recording of height and

weight in schools showed about 17 % of boys and 7 % of girls aged 15 years old to be nearing obesity, and 3 % of boys and 1 % of girls, obese. Italy has a ratio of pre-obese 15-year-olds that is even higher than the European average [63].

Body weight is the result of the complex interplay between genetic, environmental and psychosocial factors acting through the physiological mediators of energy intake and energy expenditure. Environmental factors must play a significant role in obesity, as evidenced by increasing prevalence of obesity in the last decade. A sedentary life style and low physical activity promote obesity [64]. Body weight also increases with age [65]. Of the dietary factors, high fat content and energy density have been associated with obesity [66-69]. An association between low socioeconomic status and obesity has also been reported [64, 65, 70, 71]. In addition, overweight individuals more often have difficulties controlling eating, have stronger feeling of hunger, and they tend to engage in emotional eating [66, 72]. However, in a similar, shared environment some people are likely to become obese, whereas others are not. Twin, family and adoption studies suggest a major genetic component in the determination of body weight [73-78]. Currently, obesity is thus seen as a complex disorder with an individual's genetic background affecting the susceptibility, but ultimately genetic, physiological and psychosocial factors acting together to determine the body composition.

The increased public and scientific attention to obesity is largely due to its health consequences. Total mortality associated with BMI shows a J shaped curve, meaning that overweight and obesity, as well as underweight, are associated with increased total mortality [79].

The increased death in obese individuals can be explained by chronic diseases that are more common in obese than in normal weight individuals [80]. T2DM [81, 82], CHD [83-85], hypertension [86], cholelithiasis [87], and cancer [88, 89] are the most common disorders associated with obesity.

Obese females with BMI > 31 kg/m^2 have about 40 times higher risk for T2DM compared to lean individuals with BMI < 22 kg/m^2 and more than 90 times higher risk when BMI exceeds 35 kg/m^2 [82]. In males, the association between obesity and T2DM has been detected, too; when BMI exceeds 35 kg/m^2, the risk for T2DM is more than 40 times greater when compared to lean individuals [81]. The significant increase in risk for T2DM can be seen even in normal weight people, especially in the case of women; the risk for T2DM is increased five times with BMI 24-25 kg/m^2 compared to women with BMI < 22 kg/m^2 [81, 82]. The weight change also affects the risk for T2DM; loss of approximately 10 kg decreases the risk 1.4 times, whereas gaining the same amount of weight increases the risk 2.2 times [90].

Because of the serious health consequences related to overweight and the large resources that the obesity-related diseases require on the health care system, it is of great interest to discover the mechanisms that predispose to obesity, as well as to create efficient prevention and treatment for obesity.

Obese individuals often have elevated insulin levels and are insulin resistant . A major contributor to the insulin resistance is excess free fatty acids, FFA, [91]. FFAs are derived from the triglyceride (TG) stores in the adipose tissue. Normally, insulin inhibits the lipolysis in adipose tissue. Thus, more fatty acids

are released from the adipose tissue when insulin resistance develops. The lipotoxic effects of excess FFAs may also affect beta-cell activity inducing an inhibition of the insulin signaling [92].

Genetics of Obesity

Obesity aggregates in families, but the pattern of inheritance does not in most cases follow any Mendelian segregation. This suggests a complex mode of inheritance, and the proportion of obesity due to genes is somewhat difficult to predict. The risk for obesity (defined as 90th BMI percentile or BMI > 30 kg/m^2) was two to three times higher for a person with family history of obesity and the risk increased with the severity of obesity. Studies on monozygotic and dizygotic twins or monozygotic twins reared apart give the highest heritability estimates, of the order of 70% [73, 77, 93]. Adoption studies suggest the lowest heritability with the values clustering around 30% [74-76, 94]. Results from the family studies are intermediate between the twin and adoption studies [78]. Certain diseases and traits that co-occur with obesity also show high heritability. Twin studies also suggest a considerable genetic component to eating behavior [95].

Multiple genome-wide scans have been performed for obesity and traits related to body composition [96]. Multiple genes have been associated with common forms of obesity, although only some of them have been replicated in other studies [96]. The genes for which at least five different studies found association with obesity or obesity related phenotypes include Adiponectin, Adrenergic, beta-2- and beta-3- receptors (ADRB2 and ADRB3), Guanine nucleotide binding protein (G

protein), beta polypeptide 3 (GNB3), Interleukin 6 (interferon, beta 2) (IL6), Insulin, Leptin (LEP), Leptin receptor (LEPR), Lamin A/C (LIPE), Nuclear receptor subfamily 3, group C, member 1 (NR3C1), PPARG, Tumor necrosis factor TNF superfamily, member 2 (TNF), as well as Uncoupling protein proteins 1, 2 and 3 (mitochondrial, proton carrier) (UCP1, UCP2 and UCP3) [96].

Most patients with T2DM are obese, which led to the finding that obesity is associated with diminished insulin action both in the liver and in the periphery. The association between T2DM and obesity is probably due to multiple mechanisms, including elevations in plasma free fatty acids [97] and tumour necrosis factor-alpha (TNFα) released from "full" adipocytes [98].

Thrift genotype hypothesis

The "thrifty gene" hypothesis was initially introduced by James Neel in 1962 as an attempt to explain the increase in T2DM prevalence [99]. He suggested that genes or genotypes responsible for improved energy storage during famine and starvation provided a survival advantage at a time when humans were hunter-gatherers, and there have been periods of time when the food supply was plentiful followed by periods of famine. All food in the stone-age was obtained via extensive physical activity. Thus, the lives of our ancestors alternated between shortage and abundance, the latter possibly occurring after successful hunting tending to lead to reduced physical activity. Excessive energy consumed during this period was stored as TGs in adipose tissue (and glycogen in muscles) referred as thrifty storage. When this was followed by decreased amounts of food available

with possible famine, considerable physical activity was needed to provide food again. Individuals with maximal energy storing capabilities during time of abundance, combined with their economical usage of the stored energy during famine, were probably the most capable of surviving the physical rigors of life. Genes or genetic variations enhancing these features were restored in the human genome during the evolution. Dramatic changes have occurred in the process of food supply within the last thousands of years, which is still a short period in evolutionary terms. Nowadays food supply is constant and plentiful, obtainable with minimal physical effort, consequently creating a so-called "obesogenic" environment. Most of the job descriptions of people in Western societies do not include physical labor, as neither do leisure time activities. Therefore, the body of the human being designed to 34 function in the cycles of abundance and shortage have stalled to the abundance step prepared to take on the next shortage. When the shortage or extensive physical activity never arrives, the properties of efficient energy storage become detrimental, predisposing to diseases typical to Western societies, such as overweight, obesity and T2DM. Thus, genes previously beneficial are now causing diseases. In accordance with the thrifty gene hypothesis, it has been suggested that excess fat should not be considered as a disease, i.e., a biological abnormality of an individual, but instead as a collective adaptation to the pathological pressure of the environment to eat too much and exercise too little [100].

Aim

The aim of my work was to identify and develop methods to overcome some of the major difficulties imposed by complex disease. Firstly, by developing a method, based on evolution, to prioritize candidate gene (chapter 2). Secondly, by proving that when jointly analysed it is possible to identify gene-environment interactions (chapter 3). Thirdly by developing a statistical method that allows the identification of factors involved in complex disease, overcoming the major problems of such diseases, as the sample size (chapter 4).

References

1. The GAIN Collaborative Research Group New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat. Genet. 39, 1045-51(2007).
2. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. 33, 177-82(2003).
3. Hunter, D.J. Gene-environment interactions in human diseases. Nat. Rev. Genet. 6, 287-98(2005).
4. Khoury, M.J., Davis, R., Gwinn, M., Lindegren, M.L. & Yoon, P. Do we need genomic research for the prevention of common diseases with environmental causes?. Am. J. Epidemiol. 161, 799-805(2005).
5. Joint WHO/FAO Expert Consultation Diet, nutrition and the prevention of chronic diseases. , (2003).
6. WHO The global burden of disease: 2004 update. , (2004).
7. Peltonen, L. & McKusick, V.A. Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291, 1224-9(2001).
8. Guthery, S.L., Salisbury, B.A., Pungliya, M.S., Stephens, J.C. & Bamshad, M. The structure of common genetic variation in United States populations. Am. J. Hum. Genet. 81, 1221-31(2007).
9. Wang, W.Y.S., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. 6, 109-18(2005).
10. Ludovico, O., Pellegrini, F., Di Paola, R., Minenna, A., Mastroianno, S., Cardellini, M. et al. Heterogeneous effect of peroxisome proliferator-activated receptor gamma2 Ala12 variant on type 2 diabetes risk. Obesity (Silver

Spring) 15, 1076-81(2007).

11. Vaccaro, O., Lapice, E., Monticelli, A., Giacchetti, M., Castaldo, I., Galasso, R. et al. Pro12Ala polymorphism of the PPARgamma2 locus modulates the relationship between energy intake and body weight in type 2 diabetic patients. Diabetes Care 30, 1156-61(2007).
12. Florez, J.C., Jablonski, K.A., Bayley, N., Pollin, T.I., de Bakker, P.I.W., Shuldiner, A.R. et al. TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. N. Engl. J. Med. 355, 241-50(2006).
13. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M. et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316, 889-94(2007).
14. Lohmueller, K.E., Mauney, M.M., Reich, D. & Braverman, J.M. Variants associated with common disease are not unusually differentiated in frequency across populations. Am. J. Hum. Genet. 78, 130-6(2006).
15. Peng, B. & Kimmel, M. Simulations provide support for the common disease-common variant hypothesis. Genetics 175, 763-76(2007).
16. Haldane, J. The interaction of nature and nurture. Ann Eugen 13, 197-205(1946).
17. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat. Rev. Genet. 4, 701-9(2003).
18. Tabor, H.K., Risch, N.J. & Myers, R.M. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. 3, 391-7(2002).
19. Hemminki, K., Försti, A. & Bermejo, J.L. The 'common disease-common variant' hypothesis and familial risks. PLoS ONE 3, e2504(2008).
20. The Wellcme Trust Case Control Consortium Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature

447, 661-78(2007).

21. Hahn, L.W., Ritchie, M.D. & Moore, J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376-82(2003).

22. Bland, J.M. & Altman, D.G. Multiple significance tests: the Bonferroni method. BMJ 310, 170(1995).

23. Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M. et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res. 34, 3067-81(2006).

24. Buchanan, A.V., Weiss, K.M. & Fullerton, S.M. Dissecting complex disease: the quest for the Philosopher's Stone?. Int J Epidemiol 35, 562-71(2006).

25. Varmus, H. Getting ready for gene-based medicine. N. Engl. J. Med. 347, 1526-7(2002).

26. Manolio, T.A., Brooks, L.D. & Collins, F.S. A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. 118, 1590-605(2008).

27. Millikan, R.C. Commentary: the Human Genome: philosopher's stone or magic wand?. Int J Epidemiol 35, 578-81; discussion 593-6(2006).

28. Freimer, N.B. & Sabatti, C. Human genetics: variants in common diseases. Nature 445, 828-30(2007).

29. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 27, 1047-53(2004).

30. Pilotto, L., Gaggioli, A., Noce, C.L., Dima, F. & Palmieri, L. Il diabete in Italia: un problema di sanità pubblica. Ital Heart J Suppl 5, 480-486(2004).

31. http://www.mortalita.iss.it/ . , ().

32. WHO Definition, Diagnosis and Classification of Diabetes Mellitus and its

complications.. , (1999).

33. Knowler, W.C., Bennett, P.H., Hamman, R.F. & Miller, M. Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. Am. J. Epidemiol. 108, 497-505(1978).

34. Kanaya, A.M. & Narayan, K.M. Prevention of type 2 diabetes: data from recent trials. Prim. Care 30, 511-26(2003).

35. Valsania, P. & Micossi, P. Genetic epidemiology of non-insulin-dependent diabetes. Diabetes Metab Rev 10, 385-405(1994).

36. Gerich, J.E. The genetic basis of type 2 diabetes mellitus: impaired insulin secretion versus impaired insulin sensitivity. Endocr. Rev. 19, 491-503(1998).

37. Gloyn, A.L. & McCarthy, M.I. The genetics of type 2 diabetes. Best Pract. Res. Clin. Endocrinol. Metab. 15, 293-308(2001).

38. Diamond, J. The double puzzle of diabetes. Nature 423, 599-602(2003).

39. Kobberling J, T.H. The Gnetics of Diabetes Mellitus (ed.) (Academic Press, London, 1982).

40. Rich, S.S. Mapping genes in diabetes. Genetic epidemiological perspective. Diabetes 39, 1315-9(1990).

41. Medici, F., Hawa, M., Ianari, A., Pyke, D.A. & Leslie, R.D. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis. Diabetologia 42, 146-50(1999).

42. Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J. et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35, 1060-7(1992).

43. Japan Diabetes Society Diabetes mellitus in twins: a cooperative study in Japan. Committee on Diabetic Twins. Diabetes Res. Clin. Pract. 5, 271-80(1988).

44. Hamman, R.F. Genetic and environmental determinants of non-insulin-dependent diabetes mellitus (NIDDM). Diabetes Metab Rev 8, 287-338(1992).

45. MacGregor, A.J., Snieder, H., Schork, N.J. & Spector, T.D. Twins. Novel uses to study complex traits and genetic diseases. Trends Genet. 16, 131-4(2000).

46. Zimmet, P., Dowse, G., Finch, C., Serjeantson, S. & King, H. The epidemiology and natural history of NIDDM--lessons from the South Pacific. Diabetes Metab Rev 6, 91-124(1990).

47. Rushforth, N.B., Bennett, P.H., Steinberg, A.G., Burch, T.A. & Miller, M. Diabetes in the Pima Indians. Evidence of bimodality in glucose tolerance distributions. Diabetes 20, 756-65(1971).

48. Mather, H.M. & Keen, H. The Southall Diabetes Survey: prevalence of known diabetes in Asians and Europeans. Br Med J (Clin Res Ed) 291, 1081-4(1985).

49. Ramachandran, A., Snehalatha, C., Dharmaraj, D. & Viswanathan, M. Prevalence of glucose intolerance in Asian Indians. Urban-rural difference and significance of upper body adiposity. Diabetes Care 15, 1348-55(1992).

50. Serjeantson, S.W., Owerbach, D., Zimmet, P., Nerup, J. & Thoma, K. Genetics of diabetes in Nauru: effects of foreign admixture, HLA antigens and the insulin-gene-linked polymorphism. Diabetologia 25, 13-7(1983).

51. Gardner, L.I.J., Stern, M.P., Haffner, S.M., Gaskill, S.P., Hazuda, H.P., Relethford, J.H. et al. Prevalence of diabetes in Mexican Americans. Relationship to percent of gene pool derived from native American sources. Diabetes 33, 86-92(1984).

52. Kopelman, P.G. Obesity as a medical problem. Nature 404, 635-43(2000).

53. Björntorp, P. Obesity. Lancet 350, 423-6(1997).

54. Flegal, K.M., Carroll, M.D., Ogden, C.L. & Johnson, C.L. Prevalence and trends in obesity among US adults, 1999-2000. JAMA 288, 1723-7(2002).

55. James, P.T. Obesity: the worldwide epidemic. Clin. Dermatol. 22, 276-

80(2004).

56. James, P.T., Leach, R., Kalamara, E. & Shayeghi, M. The worldwide obesity epidemic. Obes. Res. 9 Suppl 4, 228S-233S(2001).

57. Gray, D.S. & Fujioka, K. Use of relative weight and Body Mass Index for the determination of adiposity. J Clin Epidemiol 44, 545-50(1991).

58. Strain, G.W. & Zumoff, B. The relationship of weight-height indices of obesity to body fat content. J Am Coll Nutr 11, 715-8(1992).

59. Steinberger, J., Jacobs, D.R., Raatz, S., Moran, A., Hong, C. & Sinaiko, A.R. Comparison of body fatness measurements by BMI and skinfolds vs dual energy X-ray absorptiometry and their relation to cardiovascular risk factors in adolescents. Int J Obes (Lond) 29, 1346-52(2005).

60. WHO Obesity: preventing and managing the global epidemic. , (1997).

61. Li, F., Fisher, K.J. & Harmer, P. Prevalence of overweight and obesity in older U.S. adults: estimates from the 2003 Behavioral Risk Factor Surveillance System survey. J Am Geriatr Soc 53, 737-9(2005).

62. Hedley, A.A., Ogden, C.L., Johnson, C.L., Carroll, M.D., Curtin, L.R. & Flegal, K.M. Prevalence of overweight and obesity among US children, adolescents, and adults, 1999-2002. JAMA 291, 2847-50(2004).

63. WHO Obesity: preventing and managing the global epidemic. , (2004).

64. Rissanen, A.M., Heliövaara, M., Knekt, P., Reunanen, A. & Aromaa, A. Determinants of weight gain and overweight in adult Finns. Eur J Clin Nutr 45, 419-30(1991).

65. Rahkonen, O., Lundberg, O., Lahelma, E. & Huuhka, M. Body mass and social class: a comparison of Finland and Sweden in the 1990s. J Public Health Policy 19, 88-105(1998).

66. Lindroos, A.K., Lissner, L., Mathiassen, M.E., Karlsson, J., Sullivan, M., Bengtsson, C. et al. Dietary intake in relation to restrained eating, disinhibition, and hunger in obese and nonobese Swedish women. Obes. Res.

5, 175-82(1997).

67. Bray, G.A., Paeratakul, S. & Popkin, B.M. Dietary fat and obesity: a review of animal, clinical and epidemiological studies. Physiol. Behav. 83, 549-55(2004).

68. Bray, G.A. & Popkin, B.M. Dietary fat intake does affect obesity!. Am. J. Clin. Nutr. 68, 1157-73(1998).

69. McCrory, M.A., Fuss, P.J., Saltzman, E. & Roberts, S.B. Dietary determinants of energy intake and weight regulation in healthy adults. J. Nutr. 130, 276S-279S(2000).

70. Kahn, H.S. & Williamson, D.F. The contributions of income, education and changing marital status to weight change among US men. Int J Obes 14, 1057-68(1990).

71. Sarlio-Lähteenkorva, S., Silventoinen, K. & Lahelma, E. Relative weight and income at different levels of socioeconomic status. Am J Public Health 94, 468-72(2004).

72. Hakala, P., Rissanen, A., Koskenvuo, M., Kaprio, J. & Rönnemaa, T. Environmental factors in the development of obesity in identical twins. Int. J. Obes. Relat. Metab. Disord. 23, 746-53(1999).

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

74. Stunkard, A.J., Sørensen, T.I., Hanis, C., Teasdale, T.W., Chakraborty, R., Schull, W.J. et al. An adoption study of human obesity. N. Engl. J. Med. 314, 193-8(1986).

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

76. Vogler, G.P., Sørensen, T.I., Stunkard, A.J., Srinivasan, M.R. & Rao, D.C.

Influences of genes and shared family environment on adult body mass index assessed in an adoption study by a comprehensive path model. Int. J. Obes. Relat. Metab. Disord. 19, 40-5(1995).

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

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

79. Manson, J.E., Willett, W.C., Stampfer, M.J., Colditz, G.A., Hunter, D.J., Hankinson, S.E. et al. Body weight and mortality among women. N. Engl. J. Med. 333, 677-85(1995).

80. Willett, W.C., Dietz, W.H. & Colditz, G.A. Guidelines for healthy weight. N. Engl. J. Med. 341, 427-34(1999).

81. Chan, J.M., Rimm, E.B., Colditz, G.A., Stampfer, M.J. & Willett, W.C. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. Diabetes Care 17, 961-9(1994).

82. Colditz, G.A., Willett, W.C., Rotnitzky, A. & Manson, J.E. Weight gain as a risk factor for clinical diabetes mellitus in women. Ann. Intern. Med. 122, 481-6(1995).

83. Hubert, H.B., Feinleib, M., McNamara, P.M. & Castelli, W.P. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation 67, 968-77(1983).

84. Willett, W.C., Manson, J.E., Stampfer, M.J., Colditz, G.A., Rosner, B., Speizer, F.E. et al. Weight, weight change, and coronary heart disease in women. Risk within the 'normal' weight range. JAMA 273, 461-5(1995).

85. Hu, G., Tuomilehto, J., Silventoinen, K., Barengo, N.C., Peltonen, M. &

Jousilahti, P. The effects of physical activity and body mass index on cardiovascular, cancer and all-cause mortality among 47 212 middle-aged Finnish men and women. Int J Obes (Lond) 29, 894-902(2005).

86. Huang, Z., Willett, W.C., Manson, J.E., Rosner, B., Stampfer, M.J., Speizer, F.E. et al. Body weight, weight change, and risk for hypertension in women. Ann. Intern. Med. 128, 81-8(1998).

87. Maclure, K.M., Hayes, K.C., Colditz, G.A., Stampfer, M.J., Speizer, F.E. & Willett, W.C. Weight, diet, and the risk of symptomatic gallstones in middle-aged women. N. Engl. J. Med. 321, 563-9(1989).

88. Calle, E.E., Rodriguez, C., Walker-Thurmond, K. & Thun, M.J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N. Engl. J. Med. 348, 1625-38(2003).

89. Calle, E.E. & Kaaks, R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nat. Rev. Cancer 4, 579-91(2004).

90. Tuomilehto, J., Lindström, J., Eriksson, J.G., Valle, T.T., Hämäläinen, H., Ilanne-Parikka, P. et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N. Engl. J. Med. 344, 1343-50(2001).

91. Eckel, R.H., Grundy, S.M. & Zimmet, P.Z. The metabolic syndrome. Lancet 365, 1415-28(2005).

92. Dresner, A., Laurent, D., Marcucci, M., Griffin, M.E., Dufour, S., Cline, G.W. et al. Effects of free fatty acids on glucose transport and IRS-1-associated phosphatidylinositol 3-kinase activity. J. Clin. Invest. 103, 253-9(1999).

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

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

95. Tholin, S., Rasmussen, F., Tynelius, P. & Karlsson, J. Genetic and environmental influences on eating behavior: the Swedish Young Male Twins Study. Am. J. Clin. Nutr. 81, 564-9(2005).

96. Pérusse, L., Rankinen, T., Zuberi, A., Chagnon, Y.C., Weisnagel, S.J., Argyropoulos, G. et al. The human obesity gene map: the 2004 update. Obes. Res. 13, 381-490(2005).

97. Uysal, K.T., Wiesbrock, S.M., Marino, M.W. & Hotamisligil, G.S. Protection from obesity-induced insulin resistance in mice lacking TNF-alpha function. Nature 389, 610-4(1997).

98. Hotamisligil, G.S., Arner, P., Caro, J.F., Atkinson, R.L. & Spiegelman, B.M. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. J. Clin. Invest. 95, 2409-15(1995).

99. NEEL, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?. Am. J. Hum. Genet. 14, 353-62(1962).

100. Bell, C.G., Benzinou, M., Siddiq, A., Lecoeur, C., Dina, C., Lemainque, A. et al. Genome-wide linkage analysis for severe obesity in french caucasians finds significant susceptibility locus on chromosome 19q. Diabetes 53, 1857-65(2004).

Chapter 2: Evolution-enhanced candidate gene method identified association of ACO1 gene with Type 2 Diabetes Mellitus

Introduction

The common disease-common variant (CDCV) hypothesis posits that common, interacting disease alleles underlie most common diseases, perhaps in association with environmental factors [1, 2]. This hypothesis has been the scientific paradigm for association studies that have been or are being conducted on many common diseases. Numbers of new susceptibility loci are being identified. For example, the recent study by the Wellcome Trust Case Control Consortium detected 24 independent association signals for 7 major diseases [3, 4].

It is possible that the recent increase in life expectancy uncovered the latent genetic susceptibility to diseases with post-reproductive age of onset. Hence, ancestral disease risk variants might be expected not to have had fitness consequences; this scenario, in which ancestral or derived alleles are equally likely to increase disease risk, might easily fit within the common disease–common variant (CDCV) hypothesis. However, the evidence for natural selection favoring derived protective alleles might require more complex models for the evolution of the genes influencing the susceptibility to some common diseases. More specifically, it suggests not only that these genes did not evolve neutrally, but also that the environmental pressures acting on them changed during human

evolution. Indeed, it was hypothesized [5-7] that disease-susceptibility genotypes conferred a selective advantage in ancestral human populations. Also simulation studies provided support for the common disease–common variant hypothesis [8].

The most difficult aspect of the CVCD hypothesis is that detrimental genetic variants should be negatively selected during the evolution, leading to a small frequency. However, it is possible to argue that our species evolved in a different environment (i.e. hunter-gather scenario) that shaped our characteristics. After the industrial revolution, the environment in which we are living radically changed causing a conditions of un-adaptation. In its simplest incarnation, the thrifty genotype hypothesis [5] posits that the genetic predisposition to type 2 diabetes is the consequence of metabolic adaptations to an ancient lifestyle characterized by fluctuating and unpredictable food supply and high levels of physical activity. With the switch to a sedentary lifestyle and energy-dense diets, the thrifty genotype is no longer advantageous and gives rise to disease phenotypes, such as type 2 diabetes and obesity [7]. An analogous evolutionary framework, sometimes referred to as the sodium-retention hypothesis, was proposed to explain the increased prevalence of essential hypertension in some ethnic groups [6]. Briefly, ancient human populations living in hot, humid areas adapted to their environment by retaining salt, whereas populations in cooler, temperate climates adapted to conditions of higher sodium levels. This hypothesis recently received support from the finding of a latitudinal cline of allele frequency for variants likely to influence sodium homeostasis and/or hypertension, including those in AGT, CYP3A5 , angiotensin I converting enzyme (ACE), ENACa (also known as

SCNN1A) and ENACg, (also known as SCNN1G), which encode sodium channels [9, 10].

What these two hypotheses have in common is a radical and relatively recent change in the selective pressures acting on biological processes responsible for maintaining the correct balance between the organism and its environment. The recent environmental change disrupts this balance leading, in turn, to new detrimental phenotypes. Thus, these hypotheses, originally based only on disease physiology and epidemiology, can be translated into testable population genetics models of disease susceptibility. One such model could envision that the ancestral versions of genes that affect susceptibility to common diseases today reflect ancient adaptations [11, 12]. With the switch in lifestyle and environmental conditions, the ancestral alleles no longer confer a selective advantage and increase disease risk, whereas some derived alleles protect against disease and become either neutral or advantageous. Whether the selective advantage inferred for some of the derived variants is the direct consequence of disease protection or results from pleiotropic effects is currently a matter of speculation [9]. Despite intense research, only a relatively small number of regions and genes have been directly implicated as targets of selection in the human genome [13-25].

To date rarely the evolutionary and natural selection information have been used to aid the identification of gene involved in complex disease. One of the limitations to this approach is that regions under selective pressure can have been selected for several reasons. Knowing that a region is under selective pressure do not imply straightforwardly its association with a specific disease. To overcome

this limit it is possible to cross information form selective pressure with others that increases the likely that the region is involved in the studied disease. Furthermore, this approach may be further valuable for metabolic disease where a strong selective pressure can have been acting.

To candidate gene for the involvement in a disease many computational methods have been developed, mainly mining several different data sources containing, such as sequence data, biological information, functional information and expression data for candidate genes. [26]. Because it is currently impossible to prove that a predicted candidate gene is not associated with the disease of interest, it is not possible to select one of them as the best. An alternative strategy may be to identify the genes most commonly selected by several methods.

Mainly the candidate gene selection methods (CGMs) may be divided in two type: annotation-based and sequence-based. The annotation-base rely on the information previously collected about a gene to guess if it can be implicated in the disease. In many cases this candidature process also extend to paralogs and hortologs that in other species have been implicated in he same disease/phenotype. The weakness of this approach is that if a gene have been never studied it never will be selected. To overcome this weakness the selection-based CGMs have been developed. These methods rely on the assumption that genes that have largest chances to cause disease share some sequence characteristics. In particular, genes involved in disease have been harvested to identify common characteristics, and if a novel genes comply with them is considered a candidate gene. The major weakness of this method is that the large

majority of them are mendelian disease, thus is not straightforward that genes in complex disease have the same characteristics.

Therefore to select genes to study we crossed information about selective pressure with those regarding the sequence and the annotation of a candidate gene. In details, we collected evolutionary information by a genome-wide study conducted by Akey et al in 2002 [27] and a sequence- and annotation-based candidate genes from a meta-study conducted by Tiffin et al. In 2006 [26]. By an ensemble method we put together these results and obtained an ordered list of genes to empirically test in a case-control cohort of T2DM. The first gene in our list was ACO1, SNP rs1041321. When analyzed in the cohort there was a significant difference in frequencies between cases and controls.

Methods

The main strategies to identify genes in complex disease are genome-wide association study and candidate genes. The candidate gene approach is more suitable to study complex disease because overcome the multiple testing problem and allows the testing of gene-gene and gene-environment interactions (see first chapter). To select candidate genes we used an evolution-enhanced strategy that combined classical and evolutionary criteria.

Selective Pressure Information

In 2002 Akey et al. conducted a geneme-wide study to identify regions under selective pressure. Although, one of the specific aim was to "guide selection of loci for inclusion in population genetic studies" [27], to date any association study have used these information to candidate genes. Their strategy was to examine the variation in SNP allele frequencies between populations and quantify them by the statistic FST [14, 16, 23-25, 28]. Under selective neutrality, FST is determined by genetic drift, which will affect all loci across the genome in a similar and predictable fashion. On the other hand, natural selection is a locus-specific force that can cause systematic deviations in FST values for a selected gene and nearby genetic markers. For example, geographically restricted directional selection may lead to an increase in FST of a selected locus, whereas balancing or species-wide directional selection may lead to a decrease in FST compared with neutrally evolving loci [29-31]. Previous studies that have attempted to identify natural selection based on patterns of population differentiation relied on simulations to

obtain the expected distribution of FST under selective neutrality [28, 31, 32]. In brief, they studied the allele frequencies of 26,530 SNPs in three populations: African-American, East Asian, and European-American. The density of this SNP allele frequency map were the highest at that time, today overcame only by the HapMap data [33]. To examine interlocus variation in allele frequencies, they constructed the empirical genome-wide distribution of FST for all autosomal markers. The average FST for the 25,549 autosomal SNPs was 0.123, which was consisent with previous estimates [21, 31]. A high proportion of markers are located in the tails of the distribution and 6% of SNPs have FST > 0.40 [27].

Sequence- and Annotation-based Candidate Gene Selection

To increase the likelihood that a gene is involved in a disease and to candidate it for followig empirical analyses several Candidate Gene selection Methods (CGM) have been constructed. Because it is impossible to ascertain whether a CGM has good performances and because each CGM have specific peculiarity one soluion may be to use all the information that come from each CGM.

To this aim, Tiffin et al. filtered a list of 9556 candidate gene for T2DM by using 5 CGMs. The starting set of 9556 were derived by a bibliographic analysis of all linkage study for T2DM and resulted in the loci reported in box 2.1. The CGMs used are GeneSeeker, eVOC-based, DGP, PROSPECTR/SUSPECTS, G2D. In following sections are reported a brief description of the CGMs and the parameters used to identify T2DM.

Box 2.1: genomic region associated by linkage with T2DM
1q21–25, 1p31, 2p11, 2p22–2p13, 2p25, 2q12, 2q24, 2q33–2q37, 3p12–3p13, 3p24-22, 3p26, 3q11, 3q27–29, 4q27–4q28, 4q32–34, 5q13, 5q31–5q32, 6p21–6p22, 6q12, 6q15–6q27, 7p15, 7p21–7p22,7q22, 7q36, 8p21–8p22, 8p11–8p12, 8q11, 8q24, 9p13–p24, 9q31, 9q33,10p13, 10q23, 10q26, 11p12–p14
For reference list see [26]

GeneSeeker

GeneSeeker [34, 35] is a web tool that filters positional candidate disease genes based on expression and phenotypic data from both human and mouse. It queries several online databases directly through the web, guaranteeing that the most recent data are used at all times and removing the need for local repositories. In a test using 10 syndromes, GeneSeeker reduced the candidate gene lists from an average of 163 position-based candidate genes to an average of 22 candidates based on position and expression or phenotype. Though particularly well suited for syndromes in which the disease gene shows altered expression patterns in the affected tissues, it can also be applied to more complex diseases.

In the search for T2DM genes, GeneSeeker was run in batch mode with the following expression/phenotypic profiles '(insulin OR glucose OR pancreas OR fat OR adipose OR liver OR kidney OR gut) OR (muscle AND glucose) NOT (eye OR bone OR skin OR hair)'. The term 'brain' was intentionally left out of the 'NOT' section of both queries to avoid spuriously excluding valid genes that may also be expressed in the brain, given the broad expression profile of this organ

(e.g. the muscle glucose transporter GLUT4 is also expressed in the brain). House-keeping genes were not filtered out since glucose metabolism is a fundamental cellular process. The remaining settings were left at their defaults (10 cM maximum Oxford-grid distance, no databases excluded). The resulting candidate genes were validated against their respective loci using Ensembl BioMart, since GeneSeeker uses (Oxford-grid) chromosomal synteny for orthology determination and not per-gene orthology.

Analysis of candidate gene expression using eVOC annotation

This method performs candidate disease gene selection using the eVOC (a controlled vocabulary for unifying gene expression data) anatomy ontology. It selects candidate disease genes according to their expression profiles, using the eVOC anatomical system ontology as a bridging vocabulary to integrate clinical and molecular data through a combination of text- and data-mining [35]. The method first makes an association between each eVOC anatomy term and the disease name according to their co-occurrence in PubMed abstracts, and then ranks the identified anatomy terms and selects candidate genes annotated with the top-ranking terms. Candidate disease genes are thus selected according to their expression profiles within tissues associated with the disease of interest. In a test of 20 known disease associated genes, the gene was present in the selected subset of candidate genes for 19/20 cases (95%), with an average reduction in size of the candidate gene set to 64.2% (±10.7%) of the original set size. Thus, genes selected as most likely candidates from the candidate gene list are those annotated with at least three eVOC terms that match the four top-scoring disease-associated eVOC

terms.

Disease Gene Prediction (DGP)

The genes that are already known to be involved in monogenic hereditary disease have been shown to follow specific sequence property patterns that would make them more likely to suffer pathogenic mutations. Based on these patterns, DGP is able to assign probabilities to all the genes that indicate their likelihood to mutate solely based on their sequence properties [36]. In particular, the properties analyzed by DGP are protein length, degree of conservation, phylogenetic extent and paralogy pattern. The performance of this method has been assessed previously on a test dataset by building a model with a part of the data (learning set: 75%) and testing with the rest (test set: 25%). On average 70% of the disease genes in the test set were predicted correctly with 67% precision [36]. Genes involved in complex diseases, similarly to monogenic disease genes, need to have mutations or variations in the gene sequence that impair or modify the function or expression of the protein they encode, leading to a disease phenotype. Thus, we believe that, although DGP has been designed for the prediction of mendelian diseases, it can also be useful for the identification of complex-disease genes as it will identify those genes with higher likelihood of suffering mutations.

A decision tree-based model was built based on sequence properties (i.e. protein length, phylogenetic extent, degree of conservation and paralogy). This model was then applied to all the genes in the disease loci analyzed in order to obtain a probability score for these proteins to be involved in hereditary disease. Note that

this probability score is indicative of the probability of the genes to suffer mutations that impair the functionality of the protein encoded to cause a disease phenotype. It does not assume any particular phenotype and it does not account for specific phenotype features.

PROSPECTR/SUSPECTS

It can be shown that genes implicated in disease share certain patterns of sequence based features like larger gene lengths and broader conservation through evolution. PROSPECTR is an alternating decision tree which has been trained to differentiate between genes likely to be involved in disease and genes unlikely to be involved in disease [37]. By using sequence-based features like gene length, protein length and the percent identity of homologs in other species as input a score (ranging from 0 to 1) can be obtained for any gene of interest. Genes with scores over a certain threshold, 0.5, are classified as likely to be involved in some form of human hereditary disease while genes with scores under that threshold are classified as unlikely to be involved in disease. The score itself is a measure of confidence in the classification. PROSPECTR requires only basic sequence information to classify genes as likely or unlikely to be involved in disease. SUSPECTS builds on this by incorporating annotation data from Gene Ontology (GO), InterPro and expression libraries [38]. Candidate genes are scored using PROSPECTR and also on how significantly similar their annotation is to a set of genes already implicated in the same disorder (the 'training set'). This enables SUSPECTS to rank genes according to the likelihood that they are involved in a particular disorder rather than human hereditary disease in general. SUSPECTS

leverages the structure of the GO, requiring GO terms to be closely enough related semantically speaking to be considered significant [39]. As a rank-based system, it requires potential candidates to share GO terms with other disease genes to a greater extent than the other genes in the same region of interest. Performance of both PROSPECTR and SUSPECTS was tested separately with a set of oligogenic and complex disorders including Alzheimer's disease, hypertension, autism and systemic lupus erythematosus. At least two implicated genes for each disease were available. For each implicated gene, a region of interest was created containing the implicated gene itself (the 'target gene') and every gene within 7.5 Mb on either side. On average each region of interest contained 155 genes. Associated training sets were then created for SUSPECTS containing the remaining implicated genes for each disorder. Using PROSPECTR, on average the target gene was in the top 31.23% of the resulting ranked lists of candidates and in the top 5% of those lists 20 times out of 156 (13%). In comparison, on average the target gene was in the top 12.93% of the ranked list from SUSPECTS, which took both the region of interest and the relevant training set as input in each case. The target gene was in the top 5% of the ranked list 87 times out of 156 (56%) [37, 38].

In this study, the genes in each locus were scored by SUSPECTS first using a training set made up of genes already implicated in T2DM. The training set of genes implicated in T2DM was composed by: PPARG, GYS1, IRS1, INS, KCNJ11, ABCC8, CAPN10, SLC2A1, and PPARGC1 [40]. The top 10th percentile of each results set was then taken to represent a group of genes enriched

for good candidates. This proportion, providing a balance of sensitivity and specificity, was chosen on the basis of tests using positive controls as described in Adie et al. [38]. All genes were also scored by PROSPECTR based on their sequence features. Genes with PROSPECTR scores >0.65 (8~% of the total) were selected as possible candidates.

G2D

This system scores all terms in GO according to their relevance to each disease starting from MEDLINE queries featuring the name of the disease. This is done by relating symptoms to GO terms through chemical compounds, combining fuzzy binary relations between them previously inferred from the whole MEDLINE and RefSeq databases. Then, to identify candidate genes in a given a chromosomal region, G2D (genes to diseases) performs BLASTX searches [41] of the region against all the (GO annotated) genes in RefSeq. All hits in the region with an E-value $<10^{-10}$ are registered and sorted according to the GO-score of the RefSeq gene they hit (the average of the scores of their GO annotations). Note that hits in the genome might correspond to known or unknown genes, or to a pseudogene. In a test with 100 dis- eases chosen at random from OMIM (Online Mendelian Inheritance in Man) [42], using bands of 30 Mb [the average size of linkage regions [43]], G2D detected the disease gene in 87 cases. In 39% of these it was among the best three candidates, and in 47% among the best 8 candidates [43, 44].

G2D makes predictions of candidates on chromosomal regions by defining and

scoring a number of BLASTX matches of that region against a scored database of genes. For the sake of the comparison presented in this work, the results had to be mapped to genes and not genomic locations, therefore the BLASTX hits that did not overlap with any current ENSEMBL gene prediction were filtered out (these can be obtained using the G2D web server [43, 44] for a particular genomic region and disease). The final result is an ordered list of candidates for each chromosomal region and disease with a score that depends on their GO annotation. A second score have been added to the candidates, the R-score. This is the relative score of a sequence according to the distribution of GO scores of the RefSeq set used to characterize the region (the sequence ranking according to its GO-score minus one divided by the total number of sequences in the RefSeq set). R-score values close to zero indicate a strong possible relation of the sequence to the disease under consideration according to the current knowledge. The R-score allows comparing candidates for a given disease across different genomic regions linked to it; that is, one can see for which of the multiple genomic regions analyzed G2D obtained better candidates [44].

POCUS

POCUS [45] exploits the tendency for genes predisposing to the same disease to have identifiable similarities, such as shared GO annotation, shared InterPro domains or a similar expression profile. Therefore where genes within different susceptibility regions for the same disease share GO or InterPro annotation and/or are co-expressed, these genes may be considered good candidates. Although genes may be selected as candidates on the basis of sharing only a single GO term or

InterPro domain, genes lacking this annotation completely will not be selected. Some polygenic/complex diseases may be caused by different genes that are not functionally related. In such cases this method would not be expected to select the disease genes as candidates, but may still, by chance, find functional similarities between some other genes in the regions (especially where there are many regions or the regions contain many genes). Each observed similarity between genes in different regions is given a score. The score is based on the probability of seeing such specific (or more specific) similarities between genes in different randomly chosen regions of the genome containing many genes. Where such a specific (or more specific) similarity would not be seen by chance in >5% of sets of randomly chosen region analyzed, the similar genes are considered to be good candidates. Therefore in cases where disease genes are not functionally related (or where there is no data to suggest the disease genes are functionally related) POCUS will select no candidate genes in 95% of cases. This means that POCUS is far more conservative than the other methods discussed. Where many large regions are analyzed almost any similarity between genes in different regions will have a considerable probability of being seen by chance. Therefore this method is not likely to be successful when many large regions are analyzed, so analysis should be restricted to the most tightly defined and best-supported regions available.

The performance of POCUS was tested by using it to look for known disease genes. Test susceptibility regions were created containing known disease genes and the surrounding genes [45]. Test susceptibility regions were created for 120 diseases for which more than one associated gene appears in the OMIM database.

POCUS was then used to analyses the set of test regions corresponding to each disease. The performance was measured by the percentage of known disease genes selected as candidates from the test regions. The enrichment for disease genes in the selected genes compared to the whole susceptibility region was also considered. Enrichment was calculated as Enrichment = (disease genes selected/non-disease genes selected) / (disease genes in region/all genes in region). Where the test regions contained 20 genes in total the percentage of disease genes found was 41.7% and enrichment was 10.5-fold. For 100 genes the equivalent figures were 25.8% and 36.9, respectively, and for 200 genes 14.9% and 46.3. It is important to note that these results were obtained with no prior knowledge of disease pathogenesis. However, POCUS can also take into account prior knowledge of the disease, either in the form of known disease genes or preferred genes that are weighted during the analysis. Preferred genes could be genes expressed in the affected tissue or genes selected by other programs as being likely candidates.

The POCUS method is sensitive to noise. The inclusion of poorly defined susceptibility regions, or regions with a questionable association with the disease can result in failure to select similarities between disease genes, as such similaritiesare obscured by the background noise. Therefore the analysis was confined to the best supported and most tightly defined regions. These were 3p22–p24, 3q27–q terminal, 10q26, 11p12–p14, 14q11–14q13, 15q13 and 18q22–p23 [26]. Genes scoring above a threshold of 0.95 were considered good candidates. This stringent threshold is a direct reflection of the degree of statistical support for

the candidate genes returned by POCUS, according to performance on positive controls (known disease genes) unrelated to the present data. At this threshold, spurious, non- disease genes are expected to be nominated as candidates for <5% of diseases analysed. Using more liberal thresholds results in only a small increase in true positives (correctly identified disease genes) but an accompanying large increase in false positives (non-disease genes) returned as candidates by POCUS [45].

Merging candidate gene lists

The result CGMs study was a list of genes that were identified at least by a method. We joined this information with those about selective pressure. In particular genes that had a high likelihood to have been subject of selective pressure had an FST > 0.40. Then genes that were present in both lists were selected. A score was assigned at each list, calculated by adding the FST value to 0.2 for each CGM that identified it (we called the latter score as CGM*). In this way FST, that have range from 0 to 1, had broadly the same weight of the sum of the 5 methods. Finally, we obtained an ordered gene list, where at top positions were genes that have the higher likelihood to be involved in the disease. Because the evolutionary information were on SNPs and not on gene, in the following step we genotyped that SNP.

Collecting case-control cohort for T2DM

We analyzed a dataset of a case-control study on T2DM. The sample was randomly selected by the non-obese diabetic patients (BMI < 30 Kg/m^2) viewed at

the Diabetes Outpatient Clinic of the University “Federico II”, Naples. The selection of non-obese patients increased the likelihood to detect the diabetes caused by genetic predisposition. A cohort of unrelated non-diabetic non-obese glucose-tolerant control subjects were randomly selected among telephone company employees taking part in a company sponsored health screening. All study participants were Caucasians of Italian origin, unrelated and residents of the same geographical area. The study was approved by the local ethics committee; informed consent was obtained from all participants. No intervention was implemented; the prescribed medications were not modified by the study investigators. Weight and height were measured with participants wearing light clothing and no shoes. BMI was calculated as body weight in kilograms divided by the square of height in meters. DNA were extracted by peripheral blood cells by using standard laboratory techniques, and genotyping were performed by a genotyping service by the Kbioscience laboratory in Hoddesdon Herts, UK. The genetic variant analyzed was ACO1 (rs1041321).

Statistical Methods

Data are given as means and SDs or percentages. Proportions were compared by contingency tables and χ^2 analysis. The χ^2 goodness-of-fit test was used to assess deviation from Hardy-Weinberg equilibrium of the genotypic frequency by calculating expected frequencies of genotypes. A p value < 0.05 (two tailed) was considered significant. All statistical analyses were conducted using R for Windows ver 2.7.

Results

Twenty-seven genes were selected for both their sequence-annotation characteristics and for their evolutionary history (table 2.1). In that list the genes were ordered on the basis of their potential involvement in T2DM. Any of the genes in the list have ever been associated with T2DM. Only the SNP C-634G of VEGFC was associated with with development of diabetic macular edema and correlated with macular retinal thickness in type 2 diabetes [46].

Tab 2.1: gene at high probability to be involved in T2DM, on the basis of evolutionary and sequence-annotation information.

Genomic Region	Gene Symbol	GCM*	GCMs	FST	Score
9p13-p24	**ACO1**	**0,8**	**g2d eVOC suspects geneseeker**	**0,73**	**1,53**
11q23	PCSK7	1	g2d eVOC suspects dgp65 prospectr	0,51	1,51
7p15	OSBPL3	0,6	g2d eVOC geneseeker	0,8	1,4
14q23-14q24	ACTN1	0,8	g2d eVOC dgp65 geneseeker	0,57	1,37
8p21-8p22	DPYSL2	0,4	g2d eVOC	0,84	1,24
14q23-14q24	PRKCH	0,2	g2d	0,96	1,16
4q32-34	VEGFC	0,6	g2d suspects geneseeker	0,5	1,1
9q31	SMC2L1	0,6	g2d eVOC dgp65	0,48	1,08
12q15-12q21	PAWR	0,4	g2d eVOC	0,65	1,05
18q21-18q23	CDH19	0,4	g2d geneseeker	0,65	1,05
Xq23-27.3	CAPN6	0,4	g2d suspects	0,6	1
20q12-13	MYBL2	0,4	g2d eVOC	0,58	0,98
1p31	IL12RB2	0,4	g2d geneseeker	0,57	0,97
7q22	SRPK2	0,4	g2d eVOC	0,49	0,89
14q23-14q24	PPP2R5E	0,4	g2d eVOC	0,49	0,89
17p13-17q22	CHAD	0,2	g2d	0,68	0,88
9q31	INVS	0,4	dgp65 geneseeker	0,47	0,87
20q12-13	BMP7	0,2	g2d	0,67	0,87
6q15-6q27	VIL2	0,4	g2d eVOC	0,46	0,86
1p31	PDE4B	0,4	g2d eVOC	0,45	0,85
6p216p22	RNF8	0,2	eVOC	0,5	0,7
9q31	CDW92	0,2	eVOC	0,48	0,68
8q11	OPRK1	0,2	g2d	0,48	0,68
10p13	NMT2	0,2	g2d	0,47	0,67
6q15-6q27	IL20RA	0,2	eVOC	0,46	0,66
19q13	RPS11	0,2	g2d	0,46	0,66
2q24	GALNT5	0,2	g2d	0,45	0,65

Genomic region: genomic region where the gene is mapped, Gene Symbol: HUGO gene symbol, GCM*: arbitrary score derived by the sequence-annotation candidature process, GCMs: list of CGM that identified the gene, FST score of the SNP, Score: the arbitrary score of likelihood of been involved in the disease, calculated as the sum of GCM*+FST

The first gene in the list was Aconitase Cytoplasmic 1 (ACO1). ACO1 has two functions: iron responsive element (IRE)-binding protein and cytoplasmic isoform of aconitase [47]. Aconitase Cytoplasmic 1, have never been associated with diabetes, nor with obesity or metabolic syndrome. However, the role of iron in the pathogenesis of diabetes is suggested by two evidence, an increased incidence of T2DM in diverse causes of iron overload and a reversal or improvement in diabetes (glycemic control) with a reduction in iron load achieved using either phlebotomy or iron chelation therapy [48]. The SNP rs1041321 of ACO1 showed an FST of 0,73 implying that there are large differences in genotipic frequencies among ethnic groups. The CGM* score measured 0.8, implying that 4 CGMs have selected it, namely G2D, eVOC, Prospectr/Suspects and GeneSeeker. The SNP rs1041321 of ACO1 was in a non-coding region, in the first intron, however it was in linkage with two non-synonymous SNPs, one (rs41313772) in the second and (rs41304757) one in the fourth exon. Nevertheless, most SNPs associated to disease are in non-coding region [3].

Characteristics of the population studied are reported in the table 2.2 and genotypic frequency of ACO1 SNP in table 2.3. The genotypic frequency of ACO1 was in accord with the Hardy-Weinberg equilibrium.

Table 2.2: characteristics of the sample.			
	Cases	**Controls**	**t-test p-value**
N	292	227	
Age (y)	58 ± 7.1	55 ± 6.2	< 0.001
Gender (% F)	31	39	0.31
BMI	26.1 ± 2.3	25.1 ± 2.8	< 0.001
N, number of individuals in the category, BMI, body mass index ($m \cdot kg^{-2}$)			

Table 2.3: genotypic frequencies of SNP rs1041321 of ACO1 in cases and controls.				
		TT	TC	CC
ACO1	Cases	135 (67%)	119 (57%)	24 (57%)
	Controls	67 (33%)	89 (43%)	18 (43%)
Counts of individuals with that genotype are reported. In parenthesis are reported the percentage of individuals in that genotype that is affected or not-affected.				

While there was an equal distribution of CC and CT in cases and controls, the difference of these with TT was consistent (table 2.3 and figure 2.1). Indeed a χ^2 analyses of disease frequency between these two groups was significant ($p < 0.05$). Furthermore, also considering the role of age, gender and BMI, the association of ACO1 with the T2DM was still significant.

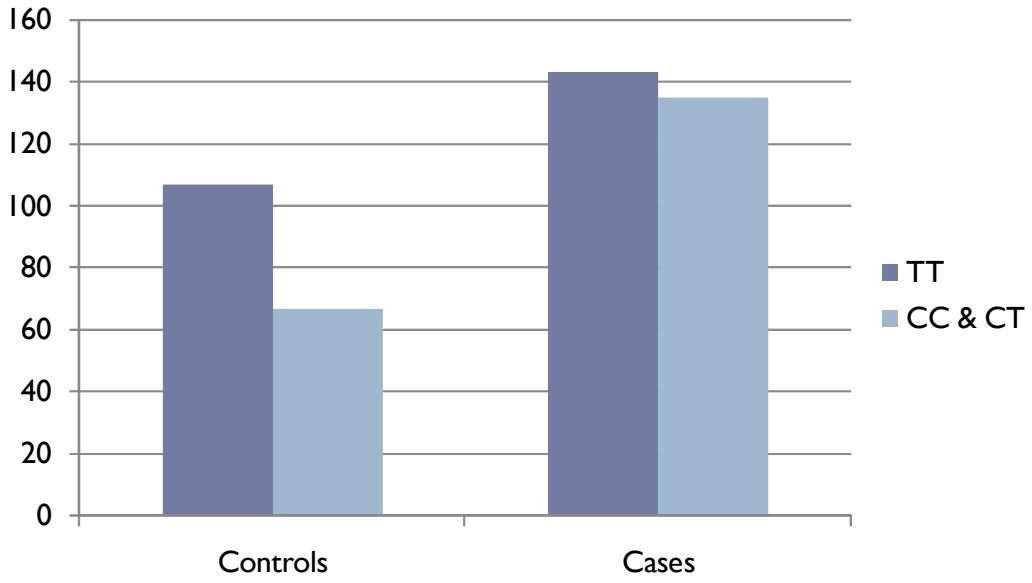


Figure 2.1: the frequencies of TT and CT&CC in cases and controls.

Conclusions

Common disease-common variant (CDCV) hypothesis suggested that common alleles underlie most common diseases [1, 2]. Furthermore, according to the *thrift genotype,* such alleles have been subjected to selective pressure, mainly those involved in metabolic disease such as T2DM and obesity [5, 7]. Therefore, we defined an algorithm that prioritized the candidature of gene for T2DM on the basis of their evolutionary history. Indeed, after the prioritizing process, the gene with the highest probability, ACO1, resulted significantly associated with T2DM in a case-control cohort. Nevertheless, although this proof of principle, it is needed to replicate the results on other high priority genes and in other populations.

References

1. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. Trends Genet. 17, 502-10(2001).
2. Wang, W.Y.S., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. 6, 109-18(2005).
3. Hemminki, K., Försti, A. & Bermejo, J.L. The 'common disease-common variant' hypothesis and familial risks. PLoS ONE 3, e2504(2008).
4. The Wellcme Trust Case Control Consortium Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-78(2007).
5. NEEL, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?. Am. J. Hum. Genet. 14, 353-62(1962).
6. Gleibermann, L. Blood pressure and dietary salt in human populations. Ecol Food Nutr 2, 143-156(1973).
7. Diamond, J. The double puzzle of diabetes. Nature 423, 599-602(2003).
8. Peng, B. & Kimmel, M. Simulations provide support for the common disease-common variant hypothesis. Genetics 175, 763-76(2007).
9. Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A. & Di Rienzo, A. CYP3A variation and the evolution of salt-sensitivity variants. Am. J. Hum. Genet. 75, 1059-69(2004).
10. Di Rienzo, A. & Hudson, R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet. 21, 596-601(2005).
11. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. 22, 239-47(1999).

12. Leabman, M.K., Huang, C.C., DeYoung, J., Carlson, E.J., Taylor, T.R., de la Cruz, M. et al. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc. Natl. Acad. Sci. U.S.A. 100, 5896-901(2003).

13. Kitano, T. & Saitou, N. Evolution of Rh blood group genes have experienced gene conversions and positive selection. J. Mol. Evol. 49, 615-26(1999).

14. Rana, B.K., Hewett-Emmett, D., Jin, L., Chang, B.H., Sambuughin, N., Lin, M. et al. High polymorphism at the human melanocortin 1 receptor locus. Genetics 151, 1547-57(1999).

15. Huttley, G.A., Easteal, S., Southey, M.C., Tesoriero, A., Giles, G.G., McCredie, M.R. et al. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. Nat. Genet. 25, 410-3(2000).

16. Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T. et al. Lactase haplotype diversity in the Old World. Am. J. Hum. Genet. 68, 160-172(2001).

17. Hull, J., Ackerman, H., Isles, K., Usen, S., Pinder, M., Thomson, A. et al. Unusual haplotypic structure of IL8, a susceptibility locus for a common respiratory virus. Am. J. Hum. Genet. 69, 413-9(2001).

18. Hurst, L.D. & Pál, C. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. 17, 62-5(2001).

19. Koda, Y., Tachida, H., Pang, H., Liu, Y., Soejima, M., Ghaderi, A.A. et al. Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma alpha(1,3)fucosyltransferase gene (FUT6) in human populations. Genetics 158, 747-56(2001).

20. Sullivan, A.D., Wigginton, J. & Kirschner, D. The coreceptor mutation CCR5Delta32 influences the dynamics of HIV epidemics and is selected for by HIV. Proc. Natl. Acad. Sci. U.S.A. 98, 10214-9(2001).

21. Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G. et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293, 455-62(2001).

22. Baum, J., Ward, R.H. & Conway, D.J. Natural selection on the erythrocyte surface. Mol. Biol. Evol. 19, 223-9(2002).

23. Fullerton, S.M., Bartoszewicz, A., Ybazeta, G., Horikawa, Y., Bell, G.I., Kidd, K.K. et al. Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. Am. J. Hum. Genet. 70, 1096-106(2002).

24. Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. & Skorecki, K. Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl. Acad. Sci. U.S.A. 99, 862-7(2002).

25. Hamblin, M.T., Thompson, E.E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. 70, 369-83(2002).

26. Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M. et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res. 34, 3067-81(2006).

27. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12, 1805-14(2002).

28. Lewontin, R.C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175-95(1973).

29. Cavalli-Sforza, L.L. Population structure and human evolution. Proc. R. Soc. Lond., B, Biol. Sci. 164, 362-79(1966).

30. Andolfatto, P. Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. 11, 635-41(2001).

31. Bowcock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K. et al. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc. Natl. Acad. Sci. U.S.A. 88, 839-43(1991).

32. Beaumont, M. & Nichols, R. Evaluating loci for use in the genetic analysis of population structure. 263, 1619-1626(1996).

33. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C. et al. Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-8(2007).

34. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P.C.W., Leunissen, J.A.M., Brunner, H.G. & Vriend, G. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res. 33, W758-61(2005).

35. Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B. & Hide, W.A. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res. 33, 1544-52(2005).

36. López-Bigas, N. & Ouzounis, C.A. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32, 3108-14(2004).

37. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6, 55(2005).

38. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22, 773-4(2006).

39. Lord, P.W., Stevens, R.D., Brass, A. & Goble, C.A. Investigating semantic similarity measures across the Gene Ontology: the relationship between

sequence and annotation. Bioinformatics 19, 1275-83(2003).

40. Stumvoll, M., Goldstein, B.J. & van Haeften, T.W. Type 2 diabetes: principles of pathogenesis and therapy. Lancet 365, 1333-46(2005).

41. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-402(1997).

42. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M. et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33, D39-45(2005).

43. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. Nat. Genet. 31, 316-9(2002).

44. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M.A. G2D: a tool for mining genes associated with disease. BMC Genet. 6, 45(2005).

45. Turner, F.S., Clutterbuck, D.R. & Semple, C.A.M. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol. 4, R75(2003).

46. Awata, T., Kurihara, S., Takata, N., Neda, T., Iizuka, H., Ohkubo, T. et al. Functional VEGF C-634G polymorphism is associated with development of diabetic macular edema and correlated with macular retinal thickness in type 2 diabetes. Biochem. Biophys. Res. Commun. 333, 679-85(2005).

47. Eisenstein, R.S. Iron regulatory proteins and the molecular control of mammalian iron metabolism. Annu. Rev. Nutr. 20, 627-62(2000).

48. Swaminathan, S., Fonseca, V.A., Alam, M.G. & Shah, S.V. The role of iron in diabetes and its complications. Diabetes Care 30, 1926-33(2007).

Chapter 3: Pro12Ala polymorphism of the PPARγ locus modulates the relationship between energy intake and body weight in type 2 diabetic patients

Introduction

Over the last two decades, the prevalence of overweight and obesity has increased worldwide [1]. Although the epidemic of obesity is largely caused by dietary and other lifestyle-related factors, the genetic background likely plays a role in determining the differences among individuals in gaining weight under the same environmental conditions.

Among the genetic factors potentially involved in the etiology of obesity, the gene encoding for the peroxisome proliferator–activated receptor (PPAR)γ, a nuclear receptor that regulates adipocyte differentiation, lipid storage, fat-specific gene expression, and insulin action, has attracted much attention. This genetic variation has been extensively investigated in relation to obesity with apparently controversial results [2-4]. In cross-sectional studies, the Ala variant has been associated with lower or higher BMI, whereas the few available longitudinal data indicate a tendency for the Ala carriers to gain more weight over time than noncarriers. The relationship between the Pro12Ala polymorphism and environmental, lifestyle-related factors has been little explored. The few available studies have been conducted in nondiabetic individuals, and, although not entirely

consistent, they support the idea that the impact of this polymorphism on weight and metabolic features is modulated by lifestyle-related factors [5-11].

Obesity is a common feature of type 2 diabetes, and dietary treatment plays a key role in the management of these patients; it is therefore particularly relevant to explore the means to identify diabetic patients who are more sensitive to weight gain/loss under given conditions.

The aim of the study was to explore, in a population-based sample of type 2 diabetic patients, the relation among BMI, habitual diet, and the Pro12Ala polymorphism in PPARγ2.

Methods

The study design was cross-sectional and observational. Participants were 343 unrelated type 2 diabetic patients, aged 40–70 years, consecutively seen at the outpatient diabetes clinic of a health district of the province of Naples. The study was approved by the local ethics committee; informed consent was obtained from all participants. Patients with serum creatinine≥2mg/dl or cardiovascular events in the previous 6 months were excluded. All participants were regularly followed up at the clinic by their own doctors, according to current guidelines for good clinical practice. The study investigations were conducted by ad hoc trained observers unaware of the participant's genotype status. No intervention was implemented; the prescribed diet and medications were not modified by the study investigators. Weight, height, and waist circumference were measured with participants wearing light clothing and no shoes. BMI was calculated as body weight in kilograms divided by the square of height in meters. Dietary habits were investigated with the use of a 138-item semiquantitative food frequency questionnaire administered by trained dietitians and designed on the basis of previous validity and reliability studies [12, 13]. Briefly, participants were asked how often, on average, they had consumed a specified portion of a given food during the previous year. Daily nutrient intake was calculated by multiplying the nutrient content of the specified portion of a food item by the frequency of its daily consumption and then summing the results of all the items. Food values for energy and nutrients were taken from the tables of the European Institute of Oncology [14]. Energy intake (kcal/day) and total saturated and polyunsaturated fat (g/day) were calculated; the

polyunsaturated-to-saturated fatty acid ratio (P/S) and the glycemic load of the diet were also computed. Energy expenditure due to physical activity was evaluated by a standardized questionnaire [15]. Participants were asked to fill in a questionnaire on habitual physical activity at work and during leisure time, which consisted of four increasing activity levels.

Genomic DNA was isolated from whole blood using Biorobot EZ1 Qiagen. By PCR, all samples were genotyped for the Pro12Ala single nucleotide polymorphism. All the oligoprimers were tested by gradient PCR to optimize melting temperature. Genotyping was performed by an allele-specific amplification method using SYBR Green detection in a real-time ABI PRISM 7000 apparatus (PE Applied Biosystem).

Statistical analysis

Data are given as means and SDs or percentages. For non–normally distributed variables, log-transformed values were used in the analyses; the original values are given in the text and tables as geometric means and interquartile ranges. Group means were compared by unpaired Student's t test or ANOVA, as appropriate. Proportions were compared by contingency tables and χ^2 analysis. The separate and combined effect of the Pro12Ala polymorphism and diet on BMI was explored across quartiles of caloric intake using two-way ANOVA. Due to the different distribution of energy intake between men and women, sex-specific quartiles were computed. Multivariate analysis was conducted by linear regression, with BMI as the outcome variable; the independent variables included

in the model were the Pro12Ala polymorphism, estimated daily energy intake, total fat, saturated fat, P/S, glycemic load, age, sex, hypoglycemic medications, A1C, and physical activity. The χ^2 goodness-of-fit test was used to assess deviation from Hardy-Weinberg equilibrium of the genotypic frequency by calculating expected frequencies of genotypes. A p value < 0.05 (two tailed) was considered significant. All statistical analyses were conducted using SPSS for Windows version 12.0

RESULTS

The general characteristics of the study participants are shown in Table 1, together with the PPARγ2 genotype. As expected for type 2 diabetic patients, participants were middle aged and generally overweight. As for the PPARγ2 genotype, 301 subjects (88%) were Pro homozygotes, 41 (11.7%) were Pro/Ala heterozygotes, and only 1 was a homozygote for the Ala variant. Therefore, in subsequent analyses, those with the Ala/Ala or Pro/Ala genotype were considered as one group and were referred to as "Ala carriers," whereas individuals with the Pro/Pro genotype were referred to as "non-Ala carriers." The genotype distribution is in Hardy-Weinberg equilibrium. Ala carriers and non-carriers were comparable with respect to age, diabetes duration (Table 3.1).

Table 3.1: characteristics of the sample			
	Non-Ala carriers	**Ala carriere**	
n	301	42	
Diabetes duration (y)	15.8 ± 8.9	14.06 ± 7.73	0.738
Age (y)	57.8 ± 8.6	58.9 ± 6.7	0.463
BMI (kg/m^2)	31.3 ± 5.8	33.6 ± 7.1	0.022
BMI, Body Mass Index			

BMI was significantly higher in carriers than non-carriers (p=0.02), whereas no significant differences were observed for waist circumference and waist-to-hip ratio between the two groups (Table 3.1). Differences in BMI were not explained by differences in dietary habits. Estimated energy intake or the macronutrient composition of the diet (i.e., intake of total fat, saturated fat, polyunsaturated fat,

P/S, and carbohydrates) was not significantly different between the two groups. Medications for diabetes are known to affect body weight; however, we did not observe any significant difference in the proportion of patients using insulin, insulin secretagogues, or insulin sensitizers (namely metformin, as the thiazolidenidiones were not marketed in Italy at the time the study was conducted) between the two genotype groups (Table 3.2). Study participants were generally sedentary; the proportion of physically active participants was low in both groups, with no significant differences between Ala carriers and non-Ala carriers (13.3 vs. 21.4%, respectively, χ^2 p=0.24).

Table 3.2: Nutrient intake and physical activity by genotype

Nutrient Intake	**Non-Ala carriers**	**Ala carriere**	**p**
Total Fat (g)	60.35 ± 20.84	60.2 ± 19.72	0.967
Saturated fat (g)	20.42 ± 9.18	20.39 ± 8.20	0.986
Glycemic load	131.30 ± 65.95	121.26 ± 30.1	0.172
P/S	0.53 ± 0.20	0.56 ± 0.24	0.440
Physically active (%)	40 (13.3)	9 (21.4)	0.239
Data are means (SD) or n (%)			

To explore the separate and combined effect of the Pro12Ala polymorphism and diet on BMI, participants were stratified according to sex-specific quartiles of energy intake and genotype. BMI increased progressively with increasing energy intake in both genotype groups with a significant linear trend (p=0.03 for the effect of energy intake; p=0.016 for the effect of genotype, with no significant interaction).

Figure 3.1 clearly shows that in the first quartile of energy intake BMI was similar

in carriers and non-carriers (30.0 vs. 30.1 kg/m2 , p = 0.1), whereas in the highest quartile of caloric intake the Ala carriers had a significantly greater BMI than Pro/ Pro homozygotes (36.0 vs. 32.1 kg/m^2 , p = 0.016). Interestingly, average daily energy intake and diet composition (i.e., total fat, saturated fat, P/S, and carbohydrates) were comparable within each quartile for carriers or non-carriers of the Ala allele (Table 3.3).

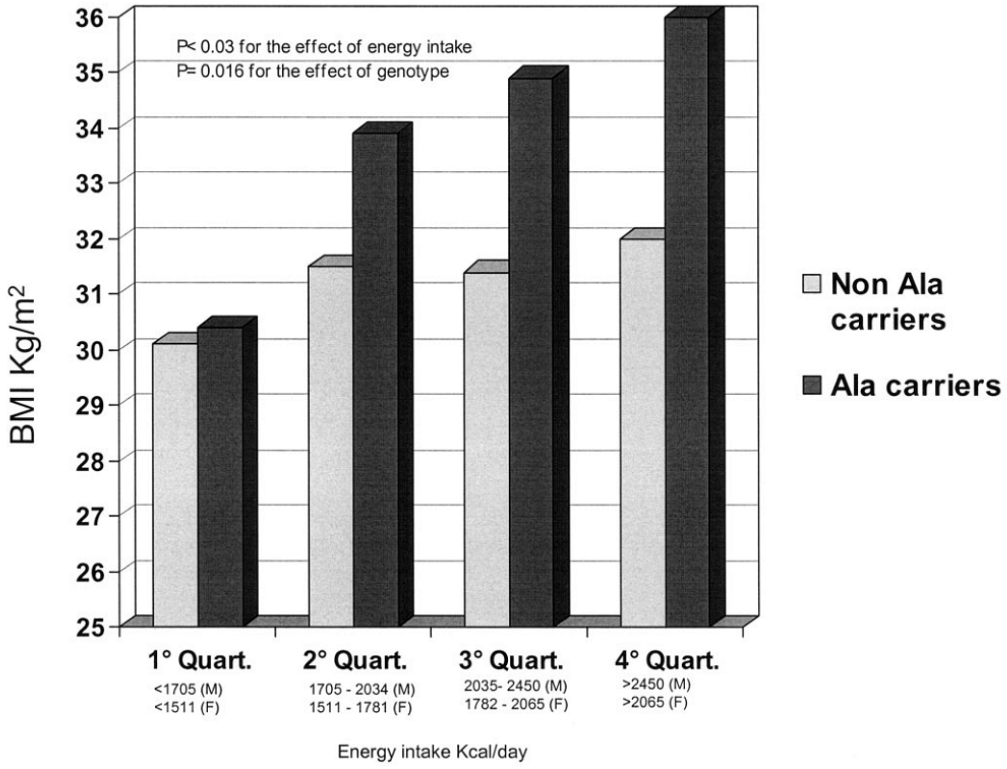


Figure 3.1: BMI in Ala carriers and non-Ala carriers according to sex-specific quartiles of daily energy intake

Table 3.3: Average daily intake of energy and macronutrients by sex-specific quartiles of energy intake and genotype					
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**	**p**
n					
Non-Ala carriers	75	75	76	76	
Ala carriers	11	11	10	10	
Energy (kcal/day)					
Non-Ala carriers	1382 ± 202	1730 ± 132	2023 ± 171	2671 ± 449	Qp=0.001
Ala carriers	1250 ± 276	1761 ± 154	2102 ± 116	2595 ± 517	Gp= 0.787
Energy (kcal/kg body wt)					
Non-Ala carriers	19.2 ± 3.9	22.6 ± 4.2	26.7 ± 5.0	33.6 ± 7.5	Qp=0.001
Ala carriers	17.7 ± 4.0	22.0 ± 4.3	23.9 ± 4.7	30.7 ± 8.2	Gp= 0.04
Total fat (g/day)					
Non-Ala carriers	42 ± 9	53 ± 7	61 ± 10	83 ± 17	Qp=0.001
Ala carriers	41 ± 14	54 ± 9	64 ± 7	79 ± 19	Gp= 909
Saturated fat (g/day)					
Non-Ala carriers	12 ± 3	13 ± 3	20 ± 5	30 ± 8	Qp=0.001
Ala carriers	12 ± 4	18 ± 4	21 ± 4	28 ± 8	Gp= 0.903
P/S					
Non-Ala carriers	0.62 ± 0.24	0.53 ± 0.17	0.51 ± 0.17	0.44 ± 0.16	Qp=0.001
Ala carriers	0.66 ± 0.22	0.54 ± 0.33	0.56 ± 0.16	0.47 ± 0.16	Gp= 0.317
Glycemic load					
Non-Ala carriers	103 ± 37	120 ± 24	138 ± 63	164 ± 99	Qp=0.05
Ala carriers	76 ± 18	117 ± 16	134 ± 36	156 ± 29	Gp= 0.298
Data are means (SD), Qp = p value for the quantile classification, Gp = p value for the Ala/Non-Ala carriers classification					

Relative to the non-carriers, Ala carriers had a significantly lower energy intake

per kilogram of bodyweight, thus suggesting that the Ala allele is associated with a higher food efficiency (i.e., for the same body weight, a lower energy intake is required to maintain a stable body weight). Possible confounders such as glycemic control, physical activity, and proportion of patients on insulin, sulfonylureas, or metformin were comparable between Ala carriers and non-Ala carriers within quartiles.

Multivariate regression analysis (Table 3.4) was performed to explore the independent effect on BMI of energy intake, diet composition, and genotype (presence/absence of the Ala allele); since age and sex are associated with both BMI and energy intake, these two variables were included in the model. Among the variables included in the model, only energy intake and presence of the Pro12Ala polymorphism were significantly and independently associated with BMI. This finding did not change when type of hypoglycemic medications and physical activity were also included in the model.

Table 3.4: Multivariate regression analysis of the association between genotype, diet variables, and BMI

	β	**SE β**	***p***
Energy (kcal)	0.004	0.001	0.004
Sex	4.967	0.667	<0.001
Genotype (Ala/non-Ala)	2.356	0.954	0.014
Age (y)	-0.014	0.039	0.709
Total fat (g)	0.025	0.070	0.721
Saturated fat (g)	-0.194	0.162	0.230
P/S	0.761	2.203	0.730
Glycemic load	-0.002	0.006	0.661

CONCLUSIONS

This study shows that type 2 diabetic patients carrying the Pro12Ala polymorphism of PPAR2 have a significantly higher BMI than non-carriers despite a similar energy intake. As a matter of fact, BMI increases progressively with increasing energy intake in both groups; however, Ala carriers had a significantly lower energy intake per kilogram body weight, thus suggesting that the Ala allele is associated with a higher food efficiency. The confounding effect of hypoglycemic medications, glycemic control, and physical activity was ruled out, thus conferring consistency to the finding. Very few studies have assessed the impact of genetic polymorphisms and diet on weight, and none of these were performed in diabetic individuals. PPARγ2 is one of the most promising candidate genes of common obesity, although so far results of association studies have been somewhat inconsistent. Cross-sectional studies have shown no difference or a lower or modestly greater BMI in Ala carriers compared with non-carriers; the few available prospective studies suggest that the Pro12Ala polymorphism is associated with higher insulin sensitivity and may confer increased susceptibility to weight gain over time, particularly in obese individuals [4]. However, no information on habitual diet was collected in these studies. Results of intervention studies in non diabetic patients indicate that the Pro12Ala polymorphism may modulate physiological responses to dietary fat intake in humans [5-11]. In the Quebec Family Study, the Ala carriers had higher BMI, waist circumference, and fat mass than non carriers but were more resistant to weight gain and metabolic deterioration when exposed to a high fat intake [6]. At least three other studies

have confirmed that the weight response to the amount and type of dietary fat differs according to the PPAR2 genotype [5, 7, 9]. In our study, the Pro12Ala polymorphism did not seem to modulate the impact of the fat content of the diet on BMI. It is relevant to note in this regard that the present study was conducted in a Mediterranean region where on average the habitual dietary fat intake is lower than in Northern European and American countries. Furthermore, the study was conducted in diabetic patients who are usually prescribed, as part of their treatment, a diet reduced in both total and saturated fat. All study participants were regularly attending a diabetes clinic, and, although most patients were not fully compliant with the prescribed diet, the average intake of total fat and saturated fat was substantially lower in this sample than in previous studies (i.e., average total fat intake was 60 g in our study, 90 g in the Canadian study [6], and 72 g in the finnish study [7]). Likewise, P/S was higher in our study than in others. It is possible that the modifying effect of the Pro12Ala variation on the relationship between dietary fat intake and BMI may not be evident for a low total fat intake, predominantly of the unsaturated type.

The results are compatible with the hypothesis that Ala carriers have a higher food efficiency (i.e., for the same body weight, they need a lower energy intake to keep their weight stable). As to mechanisms responsible for the effects of the Ala variant on individual weight regulation, we can only make speculations. The cellular and molecular mechanisms by which PPAR affects adipogenesis are not entirely clear; it has been suggested that the Pro12Ala polymorphism is associated with greater insulin sensitivity, and this could be linked to a greater increase in

bodyweight [16-19].

The role of a combined gene-environment effect in the etiology of complex traits such as obesity and insulin resistance needs to be further explored, as it may provide a basis for identifying at-risk individuals at a young age and enable the selection of more responders to preventive measures based on lifestyle modifications.

References

1. Flegal, K.M., Carroll, M.D., Ogden, C.L. & Johnson, C.L. Prevalence and trends in obesity among US adults, 1999-2000. JAMA 288, 1723-7(2002).
2. Masud, S., Ye, S. Effect of the peroxisome proliferator activated receptor-gamma gene Pro12Ala variant on body mass index: a meta-analysis. J. Med. Genet. 40, 773-80(2003).
3. Stefanski, A., Majkowska, L., Ciechanowicz, A., Frankow, M., Safranow, K., Parczewski, M. et al. Lack of association between the Pro12Ala polymorphism in PPAR-gamma2 gene and body weight changes, insulin resistance and chronic diabetic complications in obese patients with type 2 diabetes. Arch. Med. Res. 37, 736-43(2006).
4. Ek, J., Urhammer, S.A., Sørensen, T.I., Andersen, T., Auwerx, J. & Pedersen, O. Homozygosity of the Pro12Ala variant of the peroxisome proliferation-activated receptor-gamma2 (PPAR-gamma2): divergent modulating effects on body mass index in obese and lean Caucasian men. Diabetologia 42, 892-5(1999).
5. Luan, J., Browne, P.O., Harding, A.H., Halsall, D.J., O'Rahilly, S., Chatterjee, V.K. et al. Evidence for gene-nutrient interaction at the PPARgamma locus. Diabetes 50, 686-9(2001).
6. Robitaille, J., Després, J., Pérusse, L. & Vohl, M. The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Québec Family Study. Clin. Genet. 63, 109-16(2003).
7. Lindi, V., Sivenius, K., Niskanen, L., Laakso, M. & Uusitupa, M.I. Effect of the Pro12Ala polymorphism of the PPAR-gamma2 gene on long-term weight change in Finnish non-diabetic subjects. Diabetologia 44, 925-6(2001).
8. Franks, P.W., Luan, J., Browne, P.O., Harding, A., O'Rahilly, S., Chatterjee, V.K.K. et al. Does peroxisome proliferator-activated receptor gamma genotype

(Pro12ala) modify the association of physical activity and dietary fat with fasting insulin level?. Metab. Clin. Exp. 53, 11-6(2004).

9. Pisabarro, R.E., Sanguinetti, C., Stoll, M. & Prendez, D. High incidence of type 2 diabetes in peroxisome proliferator-activated receptor gamma2 Pro12Ala carriers exposed to a high chronic intake of trans fatty acids and saturated fatty acids. Diabetes Care 27, 2251-2(2004).
10. Memisoglu, A., Hu, F.B., Hankinson, S.E., Manson, J.E., De Vivo, I., Willett, W.C. et al. Interaction between a peroxisome proliferator-activated receptor gamma gene polymorphism and dietary fat intake in relation to body mass. Hum. Mol. Genet. 12, 2923-9(2003).
11. Lindi, V., Schwab, U., Louheranta, A., Laakso, M., Vessby, B., Hermansen, K. et al. Impact of the Pro12Ala polymorphism of the PPAR-gamma2 gene on serum triacylglycerol response to n-3 fatty acid supplementation. Mol. Genet. Metab. 79, 52-60(2003).
12. Panico, S., Dello Iacovo, R., Celentano, E., Galasso, R., Muti, P., Salvatore, M. et al. Progetto ATENA, a study on the etiology of major chronic diseases in women: design, rationale and objectives. Eur. J. Epidemiol. 8, 601-8(1992).
13. Trevisan, M., Krogh, V., Ferro-Luzzi, A., Riccardi, G., Freudenheim, J., Sette, S. et al. [Food questionnaire for epidemiological studies on large cohorts for use in Italy]. Ann. Ist. Super. Sanita 28, 397-401(1992).
14. Salvini, S. A food composition database for epidemiological studies in Italy. Cancer Lett. 114, 299-300(1997).
15. Saltin, B. & Grimby, G. Physiological analysis of middle-aged and old former athletes. Comparison with still active athletes of the same ages. Circulation 38, 1104-15(1968).
16. Lazar, M.A. PPAR gamma, 10 years later. Biochimie 87, 9-13(2005).
17. Stumvoll, M. & Häring, H. Reduced lipolysis as possible cause for greater weight gain in subjects with the Pro12Ala polymorphism in PPARgamma2?.

Diabetologia 45, 152-3(2002).

18. Kao, W.H.L., Coresh, J., Shuldiner, A.R., Boerwinkle, E., Bray, M.S., Brancati, F.L. Pro12Ala of the peroxisome proliferator-activated receptor-gamma2 gene is associated with lower serum insulin levels in nonobese African Americans: the Atherosclerosis Risk in Communities Study. Diabetes 52, 1568-72(2003).

19. Tremblay, A., Boulé, N., Doucet, E. & Woods, S.C. Is the insulin resistance syndrome the price to be paid to achieve body weight stability?. Int J Obes (Lond) 29, 1295-8(2005).

Chapter 4: Statistical method to identify gene-environment interaction

Introduction

Although mapping strategies, as linkage analysis, had allowed identification of many genes implicated in monogenic conditions, they are less efficient in identifying genes that are involved in the complex forms of disease [1]. Moreover, other approaches as candidate-gene and genome-wide association find it hard in identifying genetic predisposition of complex diseases [2, 3]. One of the reasons of this failure could be that studies in this area basically examined the relationship between genetic factors and traits, without jointly considering environmental determinants [4].

The study of Gene-Environment Interactions (hereafter denoted by GxE) could be useful for several reasons. First, if only the separate contributions of genes and environment is estimated, ignoring their interactions, it may be incorrectly estimate the effects of genes also leading to inconsistence of replication. Second, the identification of GxE provides direct evidence that involved biological pathways are relevant to specific traits allowing further focused researches. Third, understanding GxE might focus medical intervention, identifying sub-groups of individuals who are differently sensible to defined environmental exposure [4]. For example, in classical complex traits, like in the metabolic disorders, there have been positive reports that, beside the pharmaceutical therapies, intensive diet

and exercise interventions reduce the incidence of T2DM [5]. Indeed, studies have demonstrated that each lifestyle intervention is not effective for all obese patients [6] then, in order to provide the best treatment, could be of great interest to know which group of individuals will benefit from each intervention. Ultimately, from an epidemiological point of view, not considering the gene-environment interaction could lead to a serious underestimation of the disease risk. In fact, a low relative risk for single genetic marker does not imply its irrelevance, since it could be involved in an interaction with an environmental trigger resulting in a high risk interplay.

Despite a lot of information have been collected about both genetic and environmental risk factors, there are relatively few examples of gene-environment interaction in epidemiological literature. The main reason is that the majority of the studies have been designed to examine the main effect of single factors instead of examining the interactions [4]. This is also due to the limitations of statistical methods. They would require very large case-control data sets to identify gene-gene and gene-environment interactions, considering the loss of statistical power caused by the increasing dimensionality of the data [7, 8].

The Feature Selection Methods (FSMs) are statistical tools aimed to point out among all observed variables the ones more correlated to the status of individuals. Presently, some of the mainly used FSMs are Logistic Regression [9, 10], Linear Discriminant Analysis [11], and MDR [7]. We expect that not all methods are equally sensitive to detect the whole phenomenon. Some of them, in fact, could be more prone to reveal additive behavior though could not detect epistatic or

complex interactions, whereas others could be good to detect complex interaction but fail to point out simple single factor effect. For example, BLR can detect interactions between factors only if they of linear type. For this reason it would be important to determine the power of each specific method in this field of application. Hence, to accomplish this purpose one would need a considerably large number of trial data sets to test the methods response, for example against benchmarks or synthetic populations [12].

A method to overcome the weakness of single FSMs may be to joint them in an ensemble. An ensemble of FSMs is a set of FSMs whose individual decisions are combined to analyze new example. Ensemble are often much more accurate that the individual elements that make them up. One condition to make an ensemble accurate is that its individual members use different strategies [13]. In the case of FSMs the strategies are very different, thus the ensemble should greatly improve the overall ability to identify involved factors.

Therefore, a representative set of FSMs were selected among those more frequently used in biomedical studies. Then a set of simulated populations were created and each single FSM were challenged against them. Further, an ensemble of these FSMs were defined and were challenged against the simulated populations. Finally the ensemble were used to identify genetic and environmental factors involved in a real sample of Type 2 Diabetes Mellitus. The ensemble successfully individuated among a set of four genes and three environmental factors, the gene TCF7L2, age, and BMI as associated with T2DM.

Methods

Feature Selection Methods

FSMs have the aim to point out, among many genetic and environmental variables, the relevant ones in determining the disease occurrence.

The main techniques adopted in epidemiological biomedical studies are Univariate methods (UNI), Backward stepwise Logistic Regression (BLR), Multifactor Dimensionality Reduction (MDR), and Linear Discriminant Analysis (LDA).

Univariate methods

Although univariate methods are not able to detect interactions, they can detect marginal effect of each single factor. Univariate methods can be, nevertheless, considered the reference methods of analyses able to identify with reliability the involvement of a factors in a phenomenon. For this reason, I performed a χ^2 on genetic, discrete, variables and t-test on environmental, continuous, ones. After that, I consider as final answer the subset of variables which have passed the tests at a given confidence level.

Backward stepwise logistic regression

Logistic Regression is a classical statistical method used to select from a set of variables those that predict the dependent, binary, variable. The goal of logistic regression is to correctly provide the category (in this case, affected/not affected) of an individual using the most parsimonious model. To accomplish this goal, a

model is created that includes the predictor variables that are useful in leading the response variable. Logistic regression directly models the probability of one individual to belong to a given class. In particular, for a two-class problem ($Y \in \{0, 1\}$), the conditioned probabilities for a set of N observations $x \equiv (x_1, ..., x_N)$ are

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp\{\beta_0 + \sum_i \beta_i x_i\}}$$
$$P(Y = 0|\mathbf{x}) = 1 - P(Y = 1|\mathbf{x})$$

The unknown parameters β_i are usually estimated by Maximum Likelihood method.

Backward stepwise regression begins with a full or saturated model (i.e. including all the predictor variables) and subsequently variables are eliminated from the model, one by one, in an iterative process. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model without loosing prediction power, the analysis is over.

Multifactor Dimensionality Reduction

Multifactor Dimensionality Reduction (MDR) is an interesting approach for detecting combinations of variables that affect the dependent variable. In particular, MDR was designed specifically to identify interactions among discrete variables that influence a binary outcome. It can be considered as a nonparametric alternative to traditional statistical methods such as logistic regression.

MDR is a constructive induction algorithm that, at each step, converts two or more variables in a single one that, possibly, is a better predictor for the outcome variable. First of all, a set of n variable is selected from the pool of all factors and all the possible combinations of their values are computed. Then, the ratio of the number of cases to the number of controls is estimated within each combination. Next step is labeling each of that ratio as ”high-risk” if the cases/controls ratio meets or exceeds some threshold (e.g. 1.0), or as ”low-risk” if that threshold is not exceeded. In this way, a new attribute is formed by pooling high-risk cells into one group and low-risk cells into another group. This reduces the n-dimensional model to a one-dimensional model. Among all of the n-factor combinations, a single model that maximizes the cases/controls ratio of the high-risk group is selected, namely the one that will have the minimum classification error among the n-locus models.

To decide which subset of n variables better predicts the outcome variable, the prediction error of each model is estimated by a 10-fold cross-validation. Here, the data are randomly partitioned into 10 equipopulated parts. The model is developed for each possible 9/10 of the subjects and then is used to make predictions about the disease status of each possible 1/10 of the subjects excluded. The proportion of subjects for which an incorrect prediction was made is an estimation of the prediction error. To reduce the possibility of poor estimates of the prediction error that are due to chance divisions of the data set, the 10-fold cross-validation is repeated 10 times. To choose the final model, for each of the 10 repetitions and for each of the 10 models (due to cross-validation) of each

repetition, the number of occurrences of each model is counted and it is considered as the ”best” the one satisfying, in this order, the criteria of being more frequent, with the lower classification error and having the minimum number of variables.

Linear Discriminant Analysis

Linear discriminant methods determine linear functions, which divide the domain space into two regions. Learning processes adjust the parameters of linear models to obtain an optimal correspondence between the half-spaces of the feature space and the data categories (classes). Linear discriminants are numerical functions defined by linear combinations of the argument vector components, namely a set of N observations $x \equiv (x_1, ..., x_N)$:

$$f(x) = \mathbf{w}^T \mathbf{x} + b \ , \qquad (1)$$

where w is a weight vector and b stands for a threshold. If x satisfies the condition $f(x) > 0$, then the model assigns the label of the positive category to it, otherwise the label of the negative class is assigned. The instances for which $f(x) = 0$ define the hyperplane, which splits the whole space into the two regions. The linear discriminant function can be determined in different ways. The idea of Fisher’s linear discriminant lies in the choose of w that maximizes the separation between the two classes i.e. the ratio of the variance between the classes to the variance within the classes. This can be interpreted as maximizing the total scatter of the

data while minimizing the within scatter of the classes.

In the cases of association study, it can be used for feature selection combining a backward feature selection with an LDA classifier. At each step, one of the variables is excluded, an LDA classifier is trained and tested using a 10-fold cross validation and an averaged empirical risk is estimated. To reduce the possibility of poor estimation due to chance divisions of the data set, the 10-fold cross-validation is repeated 10 times and the final model chosen is the one more frequent.

Population simulator

Gene-Environment iNteraction Simulator (GENS) is a software generating simulated populations for case-control studies. In these populations a gene-environment interaction between two factors modulates the disease risk. In complex disease involved factors only increase or decrease the disease risk, being very far to be a deterministic process. For this reason in GENS, the involved factors only increase or decrease the disease risk, but do not controls all the disease probability.

GENS allows to evaluate the power and the minimal sample size of a method of analysis. This is performed by generating a set of populations having the same underlying interaction model, but with different strength of interaction and/or different size.

By GENS it is possible to simulate all the way genetic and environmental factors can interact, allowing complex, non-linear and epistatic relationships, though

respecting the biological constraints. In GENS the main effort designing it was to respect the standard epidemiological biomedical parameters, like Relative Risk of a population with respect to another one (hereafter denoted by RR), and the type of genetic dominance (dominant, co-dominant, and recessive).

The main purpose of the present study was to determine the performances of a set of statistical methods in terms of ability to identify interacting factors involved in a complex disease. For this reason we have restricted the analysis to a realistic, relatively simple situation: a one gene-one environment interaction. Furthermore, we have select two type of interactions, an additive model (ADD) and a modulative (MOD). As depicted in the figure 4.1, panel A, in the ADD model, both gene and environmental exposure influence the risk, but in an independent way. In this scenario the resulting risk is the sum of the genetic and environmental risk. While in MOD model (figure 4.1, panel B) the gene modules the response to the environmental exposure. In this scenario, the effect of the environmental depends by the genotype and at the same environmental exposure can lead different levels of disease risk on the basis of the genotype.

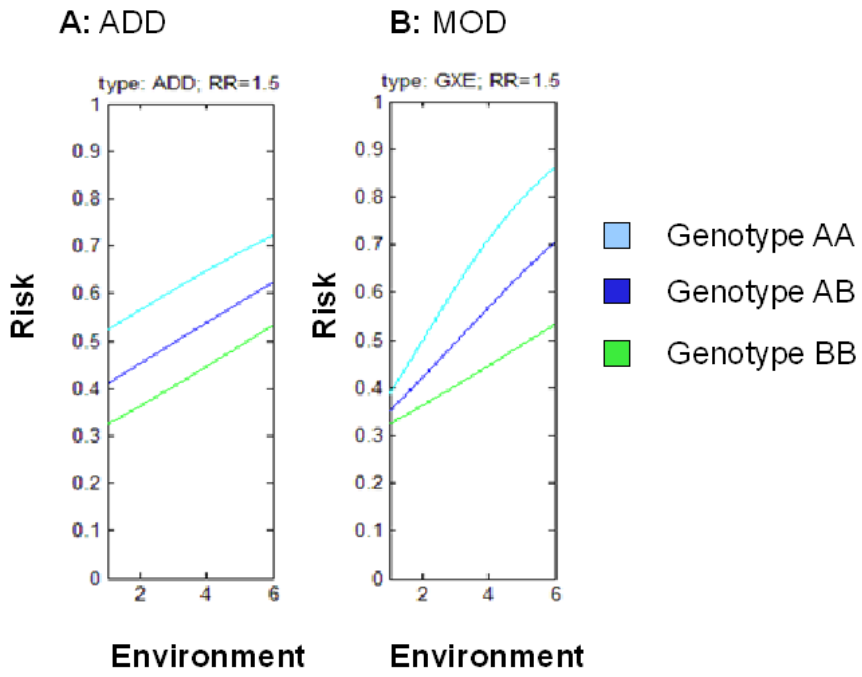


Fig 4.1: disease risk by environmental exposure and genotype. In both graphs, each line represent the relationship environmental exposure-disease risk for individuals with a specific genotype. On the X axis is represented the environmental exposure and on the Y axis the disease risk. In the panel A, is represented the ADD model, when the amount of environmental exposure increases, the risk increases of the same amount for each genotype. In panel B, is represented the MOD model, where the increase of the same amount of environmental factor exposure leads to different variations of risk in different genotypes.

In details, each individual in the simulated population randomly receives its

genotype and levels of environmental exposure. These variables are assigned in order to obey, at the population level, to user defined genetic frequencies and environmental distributions. To define the risk of an individual, see figure 4.1, is necessary to know the type of GxE interaction (MOD or ADD), the genotype (AA, AB or BB), and the level of environmental exposure. These information let identify the risk on the Y axis of the graph. Furthermore, a random number is generated, from 0 to 1, and if the number is higher that the risk calculated for the individual then he is considered affected otherwise is not-affected. Further, at each individual beside those involved in the disease a set of environmental and genetic factors are assigned that behave as noise.

To compute the Disease Risk (R) we imply a multi-logistic method that allows both epidemiologically well accepted definitions of disease risk and modeling of free non-linear interactions. By a mathematical point of view, disease risk in individuals, carrying a specific genotype, is function of a basal constant risk in addition to an environment-based component. The first one is the basal genetic risk (R_g), while the second one is the risk, in individuals carrying that genotype, in response to the environmental exposure level (R_e).

$$R \propto R_g + R_e$$

To allow the non-linear interactions among factors, we designed a relationship for each genotype:

$$Genotype\ AA: R_{AA} \propto R_{g,AA} + R_{e,AA}$$

$$Genotype\ AB: R_{AB} \propto R_{g,AB} + R_{e,AB}$$

$$Genotype\ BB : R_{BB} \propto R_{g,BB} + R_{e,BB}$$

According to this approach, ADD model have same R_e and different R_g in each genotype, whether MOD model have same R_g and different R_e in each genotype. In other terms, in the ADD case individual carrying different genotype have the different basal genetic risk. However among different genotype the contribution of the environment is the same, thus, the two factors act additively. In the MOD case, the basal genetic risk is the same among genotypes, however it is the environmental risk that is different in each genotype.

We based our approach on a logistic expression for the disease risk. In particular, the logit (log-odds) of being affected is defined as a linear function of environmental variable. We selected the logistic function because that the intensity of a biological response is often proportional to the logarithm of the stimulus extent, according to the Weber–Fechner law [14]. Moreover, the logistic function naturally leads to risk values ranging from 0 to 1. Furthermore, from an epidemiological point of view, the coefficients of the logistic function correspond to the logarithm of the relative risks due to a one-unit increase in the covariate value [9].

The main design effort was to allow the user to generate simulate populations according to provided epidemiological parameters, but respecting a set of constraints to make the simulator behavior biologically meaningful.

To produce a simulated population GENS needs a set of basal data:

- **The total number of individuals in the population.**

It is possible to simulate population of any sample size, in order to test, for instance, the ability of a FSM to correctly identify factors in condition of different sample sizes.

- **The fraction of overall affected individuals in the populations.**

It is possible to simulate populations with the desired proportion of affected individuals, to check whether the case/control ratio could affect the ability of the FSMs.

- **The number of genetic markers of each individual.**

By this value is possible to set the total number of genetic factors that were assigned to an individual. Only a fraction of them (one in present case) are involved in the disease, whereas the others act as noisy background.

- **The number of environmental exposures of each individual.**

By this value is possible to set the total number of environmental exposure that were assigned to an individual. Only a fraction of them (one in present case) are involved in the disease, whereas the others act as noisy background.

- **The allelic or genotypic frequency of gene.**

For each genetic factor, either involved in the disease or noisy background, is possible to define the allelic or the genotypic frequency.

- **The distribution of the environmental exposure (i.e. gaussian).**

For each environmental factor, either involved in the disease or noisy background, is possible to define the distribution in the population.

- **The risk associated to the genetic factor**

At the genetic factor involved in the disease is associated a value of risk. In particular, it is always considered as reference an homozygote genotype (i.e. AA) and this values represent the Relative Risk the other homozygote genotype (i.e. BB)

- **The type of genetic dominance (dominant, co-dominant, recessive)**

By this values, it possible to modulate the risk of the heterozygothe individuals. In details is possible to have a dominant, recessive o co-dominant behavior.

- **The risk associated to the environmental factor**

By this value is possible to define the value of Odds Ratio of a one-unit increase of the value of the environmental factor.

- **The type of GxE interaction.**

All the parameter above can be combined in different ways, by this parameter is possible to decide if the the genetic and environmental factors interact in a additive (ADD) or modulative (MOD) way.

Analysis strategy

To analyze the ability of different FSMs to identify factors involved in complex disease we challenged FSMs against populations with different characteristics. For this purpose, we developed a software that recursively performed an analysis of each method on each simulated population. Furthermore, the computational power necessary for the analyses required to adopt a parallel computing strategy.

The first step to allow a comparison among methods was to identify a set of parameters for each methods that resulted in a 90% specificity. The specificity of a test is the reliability of the test. In other terms, when a test identify a factor how is reliable this result? The specificity is calculated as:

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

We considered "True Positive" the factors that were considered involved by FSM and were set as involved during the process of population simulation. While we considered "False Positive" factors that were identified by FSM and were not set as involved during the process of population simulation. On the contrary, "True Negative" were factors not identified by FSM and were not set as involved, while "False Negative" were factors that the FSM did not considered involved but were set as involved.

To this aim we performed all analyses with different parameters and afterward selected those that allowed 90% specificity. The parameters range were:

BLR, single factor remove p-value = {0.1, 0.05, 0.01, 0.001}

Univariate, p-value = {0.1, 0.05, 0.01, 0.001}

LDA, Cross Validation = {2, 3, 4, 5, 6, 7, 8, 9}

MDR Cross Validation = {2, 3, 4, 5, 6, 7, 8, 9}

The second step was to use collect the results of the analyses with the selected parameters and calculate the sensibility. The sensibility is the ability of a test to identify the largest number of involved factors, and is calculated:

$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}$$

In this way we challenged all the FSMs using a common maximum 10% of false positive results.

Ensemble

In order to overcome limits of single FSMs we developed a simple ensemble of these methods. An ensemble of FSMs is a set of FSMs whose individual decisions are combined to classify new example.

In this study we adopted a simple unweighted bayesian ensemble. This method is also called voting method, because each FSM in the ensemble acts as a voters and, in particular, votes the involvement of each factors in the disease. The factors that reach the majority is considered involved in the disease. Shortly, each FSM votes the involvement of a factor in the disease, if three FSMs (of four) vote a factors that factors is considered involved by the ensemble.

Although further ensemble methods have been developed that have greatest performance than unweighted bayesian, the purpose of this study was to prove the principle that an ensemble of FSMs could improve the statistical ability to identify factors involved in complex disease. Beside, it is not possible to define whether a factors was incorrectly associated to a disease, then is also conceptually difficult to define a weight system for FSMs.

Real World sample

To use FSMs and the ensemble in a real world case, we analyzed a dataset of a

case-control study on T2DM. T2DM is a complex disease where the GxE interactions are considered frequent [15].

The sample was randomly selected by those viewed at the Diabetes Outpatient Clinic of the University “Federico II”, Naples. A cohort of unrelated non-diabetic glucose-tolerant control subjects were randomly selected among telephone company employees taking part in a company sponsored health screening. All study participants were Caucasians of Italian origin, unrelated and residents of the same geographical area. The study was approved by the local ethics committee; informed consent was obtained from all participants. No intervention was implemented; the prescribed medications were not modified by the study investigators. Weight and height were measured with participants wearing light clothing and no shoes. BMI was calculated as body weight in kilograms divided by the square of height in meters. DNA were extracted by peripheral blood cells by using standard laboratory techniques, and genotyping were performed by a genotyping service by the Kbioscience laboratory in Hoddesdon Herts, UK.

The genetic factors analyzed were: TCF7L2 (rs7903146), UCP3 (rs1800849), PPARγ (rs1801282), FTO (rs9939609). The environmental factors collected were: age, BMI, systolic blood pressure, gender.

Results

Comparison of different FSM against simulated populations

In order to analyzed the behavior of different FSMs we challenged each method against a set of simulated population where the factors involved were already known. In these populations further characteristics were already known, as the type of genetic dominance, the type GxE, the entity of risk associated to genetic and environmental factors.

To create the simulated populations we adopted the parameters reported in table 4.1, that could be considered reasonably similar to those of real populations.

Table 4.1: parameters used in the creation of simulated populations.

Parameter	**Values**
Number of individuals	500, 1500, 4500, 13500
Frequency of disease	50 %
Number of gene factors	20, 1 involved in disease risk
Number of environmental factors	20, 1 involved in disease risk
Allelic frequencies	For the involved gene factor: 0.12, 0.25, 0.5 For others: random > 0.1
Distribution of environmental exposure	Gaussian, mean = 0, standard deviation = 1
Risk of genetic factor (RR)	1.1, 1.2, 1.5, 2.0, 3.0
Type of genetic dominance	Dominant, recessive, co-dominant
Risk of the environmental exposure (OR)	1.2
Type of GxE interaction	ADD, MOD

In particular, we created 100 populations with each combination of the characteristics reported in table 4.1, obtaining a set 3600 populations. Being the creation of populations based on a random process, each population was different by others also if they shared the same characteristics.

The FSMs that we selected for the analysis were those mostly used in biomedical studies, namely backward stepwise logistic regression (BLR), multifactorial dimensionality reduction (MDR), linear discriminant analysis (LDA), χ^2, and t-test. Although the two latter are univariate methods, they are widely used in association study. The χ^2 designed for discrete factors (like genetic ones) and the t-test for continue factors (like environmental factors).

To analyze the simulated populations with FSMs we developed a software that allowed the comprehensive analysis of all the 3600 populations. Requiring an high computational power, the analysis were performed on the GRID parallel computing facility of the University of Naples “Federico II”.

FSMs are conceptually different among them, and each one may be performed using different parameters. We had to set up starting conditions for each FSM to allow further comparisons. Thus, we firstly identified the parameters that allowed an equal specificity among them. To this aim, we recursively analyzed all the populations with a wide et of parameters of each FSM and, afterward, we selected those parameters that allowed a minimum of 90% of specificity. Results of this preliminary analyses are reported in figure 4.2.

According to these results the following analysis were performed with the

parameters reported in table 4.2. In this way we could compare the sensibility of each methods, allowing a maximum of 10% false positive results.

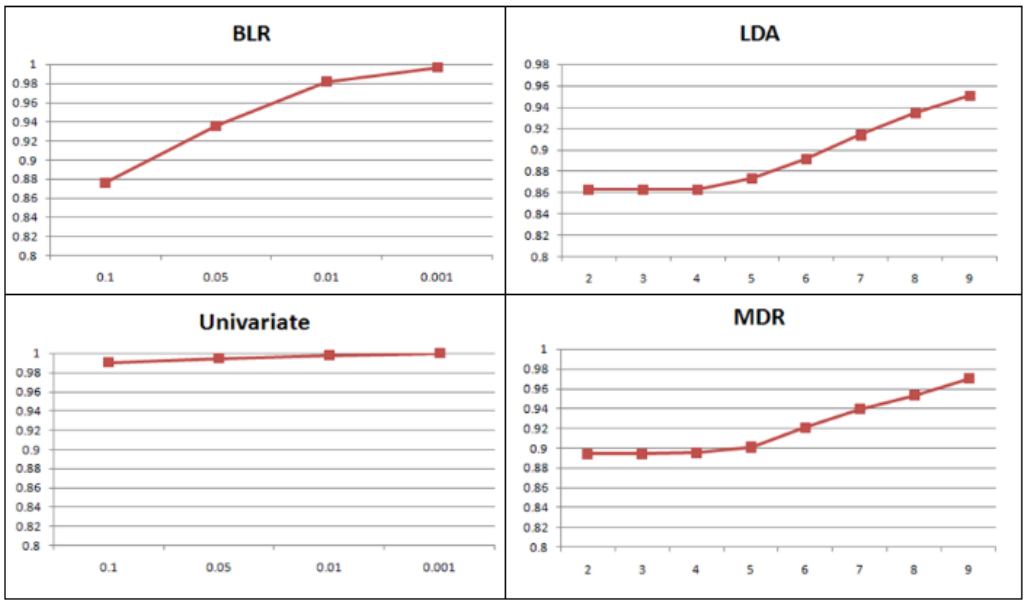


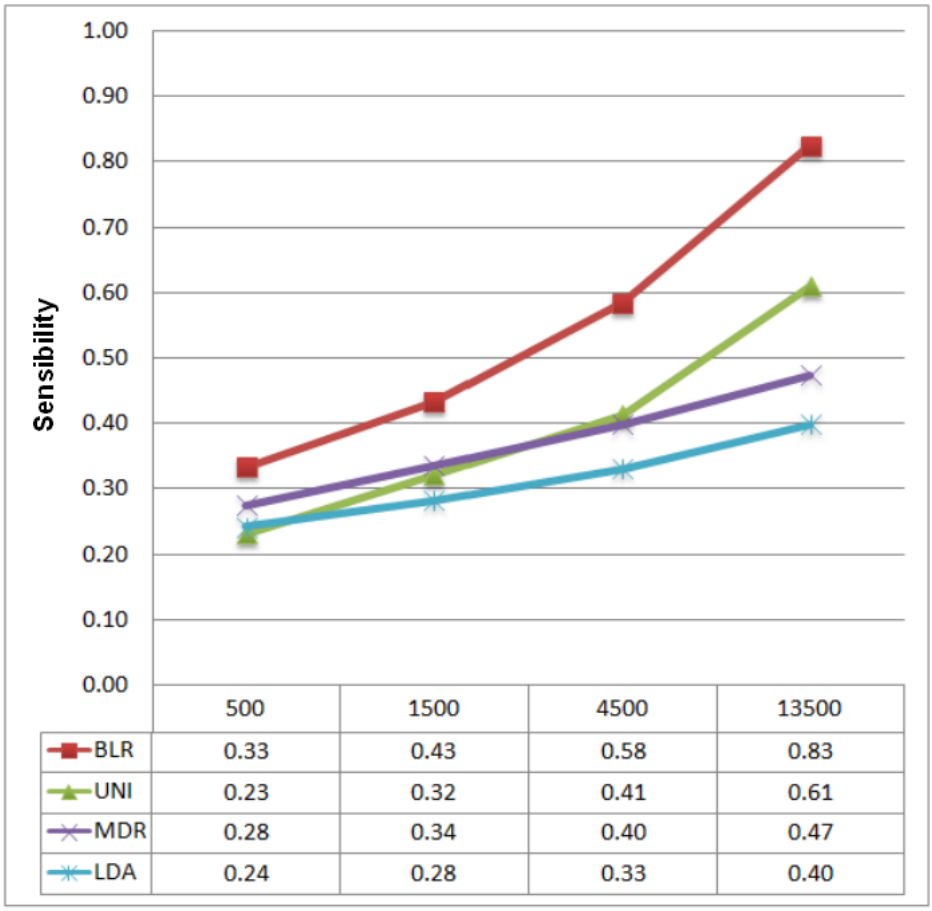
Figure 4.2: In this figure are showed the results of 86400 analyses, performed to set up common conditions for each FSM.. On the Y axis is reported the specificity of FSM, and on the X axis the range of the parameter checked. For each FSM, the minim value of the parameter that allowed a specificity > 90 % were selected.

Table 4.2: parameters of the FSMs used in the analysis of simulated populations and on the real world dataset. These parameters warrantied a specificity higher that 90 %

FSM parameter	**value**
BLR, single factor remove p-value	0.05
Univariate, p-value	0.1
LDA, Cross Validation	7
MDR Cross Validation	5

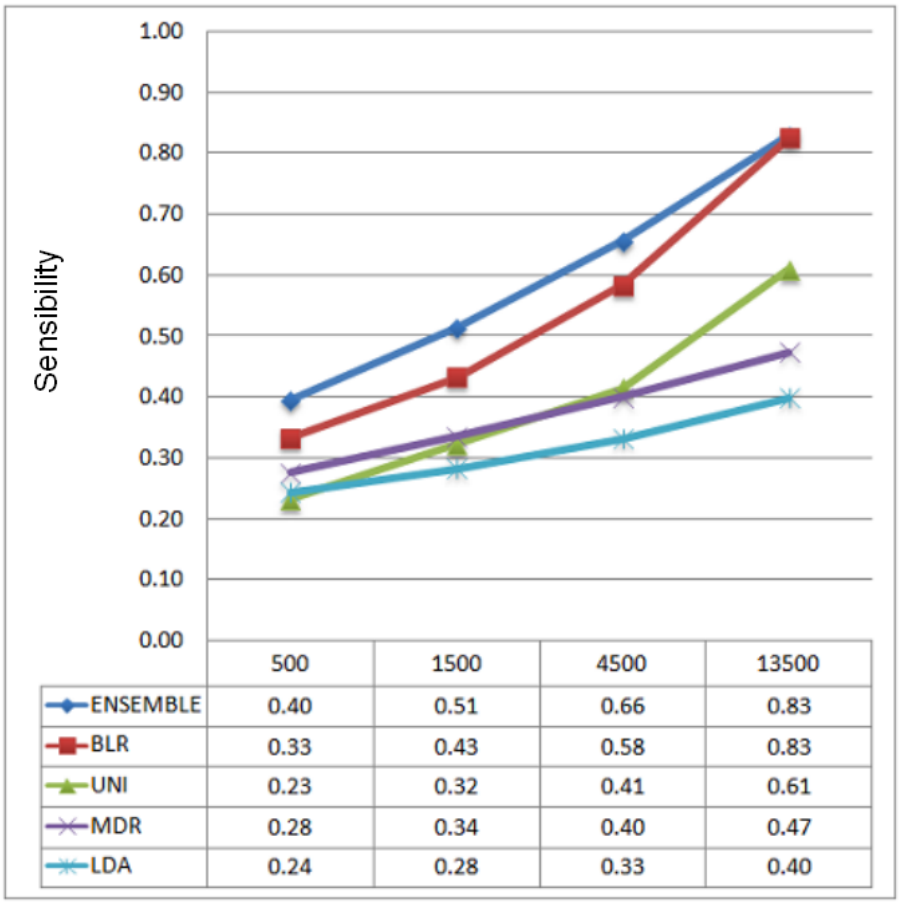
In figure 4.3 are reported sensibility of FSMs, clustered for sample size. As expected all FSMs increased in sensibility when the sample size increased. However the performance of the BLR was better of any others. Moreover, although MDR and LDA had an overall lower sensibility, seemed to be lessen influenced by sample size. Indeed, should be highlighted that BLR is based on logistic functions and this could give some advantage to BLR than other FSMs. However our simulator is based on a multi-logistic model that made relationship among factors quite different from simple logistic ones.

To overcome the limits of single FSMs we developed an ensemble. The ensemble of FSMs is a set of FSMs whose individual decisions are combined to classify new example. We adopted a simple unweighted bayesian ensemble. The results of the ensemble sensibility are reported in the figure 4.5 together with those of the single FSMs. The ensemble had better performance in each sample size, except that in 13500 individuals, where the results were similar to those of RBL. FSMs, different from BLR, had an important role in low sample sizes and in these conditions the ensemble can more benefit of their role. Indeed, MDR have been developed to analyze situations with many factors and low sample size [16]. However the situations largely more frequent in real world are those with 500-1500 samples. Thus, in these conditions the ensemble may effectively aids the identifications of factors involved in complex disease.



	500	1500	4500	13500
BLR	0.33	0.43	0.58	0.83
UNI	0.23	0.32	0.41	0.61
MDR	0.28	0.34	0.40	0.47
LDA	0.24	0.28	0.33	0.40

Figure 4.3: Sensibility of FSMs on the simulated populations. On the Y axis is reported the sensibility and on the X axis the sample size of the analyzed populations. In total are represented data from 3600 population analyzed by each FSM. BLR: backward stepwise logistic regression, UNI: t-test and χ^2, MDR: multifactorial dimensionality reduction, LDA: linear discriminant analysis.



	500	1500	4500	13500
ENSEMBLE	0.40	0.51	0.66	0.83
BLR	0.33	0.43	0.58	0.83
UNI	0.23	0.32	0.41	0.61
MDR	0.28	0.34	0.40	0.47
LDA	0.24	0.28	0.33	0.40

Figure 4.5: Sensibility of FSMs on the simulated populations. On the Y axis is reported the sensibility and on the X axis the sample size of the analyzed populations. In total are represented data from 3600 population analyzed by each FSM. BLR: backward stepwise logistic regression, UNI: t-test and χ^2, MDR: multifactorial dimensionality reduction, LDA: linear discriminant analysis.

Real World Sample

We analyzed a real case-control cohort of Type 2 Diabetes Mellitus by each FSMs and by ensemble. The patients were collected by the Diabetes Outpatient

Clinic of the University “Federico II”, Naples and the non-diabetic glucose-tolerant controlss were randomly selected among telephone company employees taking part in a company sponsored health screening. Principal characteristic are reported in table 4.3. As expected, T2DM patients were generally older and had an higher BMI than controls.

The genetic factors analyzed were: TCF7L2 (rs7903146), UCP3 (rs1800849), PPARγ (rs1801282), FTO (rs9939609). TCF7L2 is the most important T2DM susceptibility gene identified to date, with common intronic variants strongly associated with diabetes in all major racial groups [17]. PPARgamma belong to a subfamily of nuclear receptors that form heterodimers with retinoid X receptors (RXRs) and these heterodimers regulate transcription of various genes. PPAR-gamma is a regulator of adipocyte differentiation. It have been implicated in T2DM in several association study [18, 19] and is the target of the antidiabetic drugs thiazolidinediones [20]. FTO is an unknown function gene that have been associated with T2DM in several genome-wide association studies [18, 19, 21]. With the capacity to participate in thermogenesis and energy balance, UCP3 is an important obesity candidate gene [22, 23] that have been associated with the disease in candidate gene studies [24]. In table 4.4 are reported the genotypic frequencies found in cases and controls.

Table 4.3: characteristics of the sample			
	Cases	**Controls**	**t-test p-value**
N	448	305	
Age (y)	59 ± 7.4	54 ± 6.6	< 0.001
Gender (% F)	39	38	0.31
BMI	29 ± 4.5	27 ± 4.5	< 0.001
N, number of individuals in the category, BMI, body mass index ($m \cdot kg^{-2}$)			

Table 4.4: Genotipic frequencies in cases and controls.					
Gene		**Genotype 1**	**Genotype 2**	**Genotype 3**	**χ^2 p-value**
FTO	**Cases**	90	213	117	0.23
	Controls	47	107	80	
PPARγ	**Cases**	368	47	3	0.52
	Controls	209	21	3	
UCP3	**Cases**	13	97	311	0.069
	Controls	2	67	166	
TCF7L2	**Cases**	110	214	93	<0.001
	Controls	93	112	26	
For TCF7L2 (rs7903146) genotype 1, 2 and 3 are respectively TT, CT and CC; for UCP3 (rs1800849) AA, AG, and GG, for PPARγ (rs1801282) CC, CG, and GG, for FTO (rs9939609) AA, TA, and TT					

The sample was analyzed by the FSMs and by the ensamble. Results are reported in table 4.5. Univariate and BLR both found an association between a genetic factor, the TCF7L2, and two environmental factors, age and gender. MDR found that the three environmental factors were associated with the disease. Whereas LDA found associated TCF7L2 and the three environmental factors (age, gender, and BMI). The ensemble summarized the results in indicating as associated TCF7L2, age, and BMI.

Table 4.5: results of FSMs and ensmble on the T2DM cohort	
FSM	**Factors associated**
Univariate	Age, BMI, TCF7L2
BLR	Age, BMI, TCF7L2
MDR	Age, BMI, Geneder
LDA	Age, BMI, Geneder, TCF7L2
Ensemble	*Age, BMI, TCF7L2*

In order to assess the performances of FSMs and ensemble while decreasing the sample size, we made subsample of the whole dataset by a random picking process. We made four subpopulations of 90%, 50%, 25% and 10% of the whole sample size and, to avoid biases, we made ten subpopulations of each size. In total we obtained 40 populations to analyze with the 4 FSMs and the ensemble. We checked the performances of the FSMs and of the ensemble by comparing each result with the result of the ensemble on whole dataset (Age, BMI, TCF7L2). In this way we could demonstrate that ensemble have better performances of other FSMs when the sample size decreased, see figure 4.6. Te overall performance of the ensemble is similar to to that of the BLR where the sample size was relatively high (90% and 50%). However, when the sample size further decreased the ensemble gave better results, probably thanks to MDR and LDA. It is noteworthy that MDR performance that was bad in higher sample size was poorly affected by the sample reduction, indeed in the smallest sample size (10%) MDR had better performances than other FSMs. Nevertheless MDR were designed with the specific purpose to works in low-sample size situations [7].

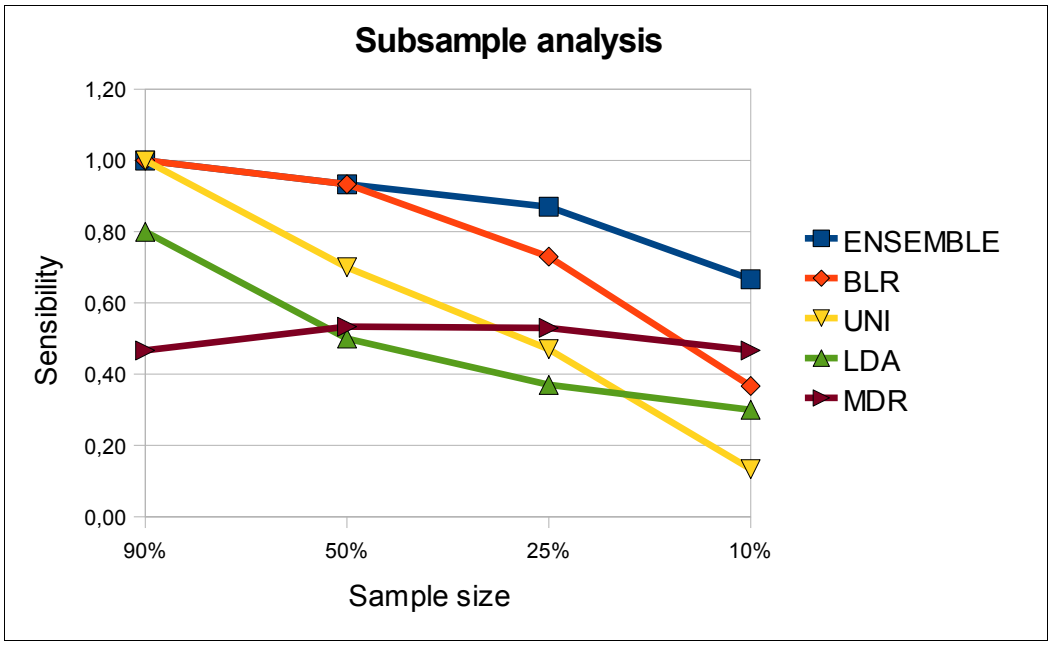


Figure 4.6: Analysis of subsample of the whole dataset. On the Y axis is reported the sensibility calculated as the ability to replicate the same result of the ensemble on the whole dataset. On the X axis are reported decreasing size of the subsample. BLR: backward stepwise logistic regression, UNI: t-test and χ^2, MDR: multifactorial dimensionality reduction, LDA: linear discriminant analysis.

Conclusion

The strategy of simulations can be useful to select feature selection methods, and to check their characteristic. For example, backward stepwire regression had better performance when the sample size was larger, both in simulated populations than in the real sample. On the contrary, MDR and LDA, had lower performances in larger sample size, however their performances increased in small sample size, both in simulated populations than in the real sample. The ensemble had performances higher or equal to the best FSM in each situations, because it collected the strength of each method in each situation.

References

1. Tabor, H.K., Risch, N.J. & Myers, R.M. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. 3, 391-7(2002).
2. Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G. et al. Replicating genotype-phenotype associations. Nature 447, 655-60(2007).
3. Ioannidis, J.P.A. Genetic associations: false or true?. Trends Mol Med 9, 135-8(2003).
4. Hunter, D.J. Gene-environment interactions in human diseases. Nat. Rev. Genet. 6, 287-98(2005).
5. American Diabetes Association Nutrition Recommendations and Interventions for Diabetes: a position statement of the American Diabetes Association. Diabetes Care 30 Suppl 1, S48-65(2007).
6. Berrettini, W., Bierut, L., Crowley, T.J., Cubells, J.F., Frascella, J., Gelernter, J. et al. Setting priorities for genomic research. Science 304, 1445-7; author reply 1445-7(2004).
7. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 138-47(2001).
8. Cocozza, S. Methodological aspects of the assessment of gene-nutrient interactions at the population level. Nutr Metab Cardiovasc Dis 17, 82-8(2007).
9. Hosmer, D., Lemeshow, S., Wiley, J. & InterScience, W. Applied logistic regression (ed.) (Wiley New York, 2000).
10. Bagley, S.C., White, H. & Golomb, B.A. Logistic regression in the medical

literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 54, 979-85(2001).

11. McLachlan, G. Discriminant analysis and statistical pattern recognition (ed.) (Wiley New York, 1992).
12. Schmidt, M., Hauser, E.R., Martin, E.R. & Schmidt, S. Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: environmental covariates, gene-gene and gene-environment interaction. Stat Appl Genet Mol Biol 4, Article15(2005).
13. Hansen, L. & Salamon, P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 993-1001(1990).
14. Bliss, C. The calculation of the dosage-mortality curve. Annals of Applied Biology 22, 134-167(1935).
15. Romao, I. & Roth, J. Genetic and environmental interactions in obesity and type 2 diabetes. J Am Diet Assoc 108, S24-8(2008).
16. Hahn, L.W., Ritchie, M.D. & Moore, J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376-82(2003).
17. Hattersley, A.T. Prime suspect: the TCF7L2 gene and type 2 diabetes risk. J. Clin. Invest. 117, 2077-9(2007).
18. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316, 1341-5(2007).
19. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H. et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316, 1336-41(2007).
20. Lehmann, J.M., Moore, L.B., Smith-Oliver, T.A., Wilkison, W.O., Willson, T.M. & Kliewer, S.A. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma).

J. Biol. Chem. 270, 12953-6(1995).

21. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M. et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316, 889-94(2007).

22. Boss, O., Samec, S., Paoloni-Giacobino, A., Rossier, C., Dulloo, A., Seydoux, J. et al. Uncoupling protein-3: a new member of the mitochondrial carrier family with tissue-specific expression. FEBS Lett. 408, 39-42(1997).

23. Millet, L., Vidal, H., Andreelli, F., Larrouy, D., Riou, J.P., Ricquier, D. et al. Increased uncoupling protein-2 and -3 mRNA expression during fasting in obese and lean humans. J. Clin. Invest. 100, 2665-70(1997).

24. Pinelli, M., Giacchetti, M., Acquaviva, F., Cocozza, S., Donnarumma, G., Lapice, E. et al. Beta2-adrenergic receptor and UCP3 variants modulate the relationship between age and type 2 diabetes mellitus. BMC Med. Genet. 7, 85(2006).