Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

ONCOLOGIA, EMATOLOGIA E PATOLOGIA

Ciclo XXX

**Settore Concorsuale: 06-A2**

**Settore Scientifico Disciplinare: MED/04**

# An Integrated Genetic and Epigenetic Analysis of Type 2 Diabetes in the Italian population

**Presentata da:** Elena Marasco

**Supervisore**

**Coordinatore Dottorato** **Prof. Stefano Salvioli**

**Prof. Per Luigi Lollini** **Co-supervisore**

**Dott. Paolo Garagnani**

**Esame finale anno 2018**

# An Integrated Genetic and Epigenetic Analysis of Type 2 Diabetes in the Italian population

# Contents

## Abstract

Type 2 diabetes (T2D) is a common complex metabolic disease and represents one of the main health-care problems worldwide. Only 10-12% of T2D heritability is explained by genetic variability, while many environmental factors have a critical role in the onset of the disease; for these reasons, an important aspect to take into account in this field is the interaction among the different risk factors.

The aim of this thesis is to apply an integrative analysis of genetic and epigenetic data in order to identify new molecular mechanisms and interactions at the basis of T2D considering an Italian population.

First we performed a T2D case-control study on 1352 individuals at genome-wide level using Illumina technology. T2D patients with at least one complication were compared first with the healthy control group and then with centenarians (who never developed the disease), applying the "extreme phenotypes" approach. In the meantime, we conducted a methylation analysis at gene candidate level on 229 diabetic patients and 219 controls. For this analysis we considered three different regions of *TCF7L2* gene. Finally, we applied two different approaches (one based on some phenotypic traits, the genotype of rs7903146 and methylation levels of 5 CpG sites of TCF7L2 gene and the second on DNA methylation of few genes called "epigenetic little clock") in order to develop a disease predictor model able to differentiate T2D patients group from the control group.

In the genetic analyses we identified five genes with multiple associated signals in the comparison between diabetes individuals and controls: *MS4A14* and *THSD4* genes, that have never been related to T2D, and we confirmed association variants in *DNHD1*, *ELMOD1 DLC1* genes. Then, considering the "extreme phenotypes", we compared diabetic patients with centenarians and we found four genes with more than one associated SNP: *FAM13A*, *TCF7L2*, *APBB2* and *MGLL* genes, the last two have never been associated to T2D. Finally, we identified the genes with multiple associated variants that were present in both comparisons. We identified three key genes *TMEM108*, *UPP2* and *TCF7L2* for T2D .

In methylation analysis performed on *TCF7L2* gene, we reported an association between methylation levels of two CpG sites in intron 3 with T2D in our population and we demonstrated that methylation levels of four CpG sites (two in promoter region and two in intron 3) are strongly influenced by the genotype of the most T2D related variant, rs7903146.

We developed a disease prediction model based on a linear regression model of the strongest diabetes related variant (rs7903146 of *TCF7L2* gene), phenotypic trait (BMI, sex, age) and methylation levels of 5 CpG sites in intron 3 region of TCF7L2 gene that would be able to correctly classify 63% of subjects. Finally, we applied an epigenetic little clock to diabetic patients that, surprisingly, did not show an acceleration of epigenetic age.

The overall results highlighted: 1) the genetic variants associated to an increase susceptibility to T2D thanks to the exploitation of a group of Italian centenarians who never developed the disease, 2) the relation between DNA methylation and the genotype of TCF7L2-rs7903146; 3) a first example of a mathematical model for the prediction of the disease 4) that external/environmental factors (such as pharmacological treatment) may influence the epigenetic age of diabetic patients considered in our study.

# 1 Introduction

The increase of people affected by type 2 diabetes (T2D) represents one of the most important health-care problems worldwide and this phenomenon will increase in the next decades in industrialized and developing countries (International Diabetes Federation, 8th edition 2017). This raise is due to the aging of the populations and to the increasing prevalence of obesity (Danaei et al. 2011). T2D is a common complex metabolic disease characterized by an impaired glucose metabolism due to insulin resistance of peripheral tissues. Several risk factors are involved in T2D onset, such as age, obesity and low physical activity (World Health Organization 2004). Among the risk factors the genetic heritability plays a key role in the onset of the disease. In fact, 40% of diabetic patients present a family history of diabetes; nevertheless, the alimentary habits, the physical activity and "healthy" life style, can delay or avoid the onset of the disease. Actually, an important aspect to take into account is the interaction among the different risk factors.

## 1.1 Genetics and T2D

The first linkage analysis studies were performed in the '90s and identified two genes, calpain 10 (CAPN10) and transcription factor 7-like 2 (TCF7L2) associated with T2D (Hanis et al. 1996) (Duggirala et al. 1999). Subsequently, several candidate gene studies have been conducted, focusing on genes potentially involved in pathways affected by T2D, such glucose metabolism, insulin secretion and signalling, and lipid metabolism. By means of this strategy a number of genes were identified: peroxisome proliferator-activated receptor gamma (PPARG), insulin receptor substrate 1 (IRS1) and IRS-2, potassium inwardly-rectifying channel, subfamily J, member 11 (KCNJ11), Wolfram syndrome 1 (wolframin) (WFS1), HNF1 homeobox A (HNF1A), HNF1 homeobox B (HNF1B) and HNF4A (Gaulton et al. 2008).

Using microarray technology, Genome-wide association studies (GWAS) have been conducted and confirmed several T2D-associated loci and identified new ones e.g. HHEX/IDE and SLC30A8 genes (Sladek et al. 2007) (Scott et al. 2007) (Zeggini 2007) (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al. 2007), CDKN2A/2B, and the intronic region of CDKAL1 and IGF2BP2 (Pal and McCarthy 2013).

Finally, in order to increase sample size for the identification of novel associated loci, two meta-analysis studies have been performed. In the first one, the DIAbetes Genetics Replication And

Meta-analysis (DIAGRAM) study, including 10,000 subjects in the "scan" step, and 53,000 individuals in the replication phase, six novel loci have been identified (JAZF1, CDC123-CAMK1D, TSPAN8-LGR5, THADA, ADAMTS9 and NOTCH2) (Zeggini et al. 2008). In the second meta-analysis study (DIAGRAM+), performed on more than 101,000 individuals, 12 additional loci have been associated to T2D (Voight et al. 2010).

## 1.2 Epigenetics and T2D

The term "epigenetics" refers to heritable modifications in gene expression that do not modify nucleotide sequence. Epigenetic mechanisms include DNA methylation, histone acetylation and non-coding RNAs, and they are used to modulate gene expression in response to environmental *stimuli*. The epigenetic profile can persist lifelong and can be transmitted to descendants for up to 2-3 generations (Skinner 2011). Recently, evidences of possible epigenetic involvement in the pathogenesis of diabetes have emerged. Several authors have hypothesized the existence of a metabolic program influenced by environmental conditions and genetic predisposition that contributes to the development of diabetes. It is known that the intrauterine and early infancy environment can modulate the risk of chronic, late-onset diseases like T2D in humans and animal models (Seki et al. 2012). In fact, several evidences suggest that environmental and nutritional exposure in prenatal/neonatal phases modulate the fetal epigenetic profile and can contribute to risk of developing metabolic disorders including type 2 diabetes (Vaiserman 2017). Epigenetic mechanisms such as DNA methylation and histone acetylation alter the expression of genes. There are numerous studies that have shown association between gene expression and methylation with diabetic pathology at the level of the various organs.

The first epigenome-wide study on T2D was performed on DNA from peripheral blood, and differentially methylated regions (DMRs) have been associated to T2D in particular in GWAS susceptibility loci (Toperoff et al. 2012). More recently, three independent studies have found a differential methylation at a CpG site in TXNIP (Chambers et al. 2015) (Florath et al. 2016) (Kulkarni et al. 2015). Moreover, a study conducted on pancreatic islets identified more than 800 genes showing DMRs between diabetics and controls, including genes already known to be associated with T2D like *TCF7L2*, *FTO* and *KCNQ1*.

All these results demonstrate how close is the linkage between genetics and epigenetics in the study of a complex and multifactorial diseases like T2D.

Although in the last decade there has been a dramatic explosion of genetic studies, thanks to technological advancement and innovative analytical methods, only 10-15% of heritability has been explained so far. Numerous efforts have been made in order to explain the "missing heritability". This main obstacle is the small effect of every single variant in the onset of the disease, together with the heterogeneous phenotype of T2D and the coexistence of several risk factors.

## 1.3 Extreme phenotypes

Complicated T2D represents the most severe form of the disease. All the complications share a vascular component, indicating that endothelial damage is a critical point for the progression of diabetes. T2D complications are also influenced by genetic components as demonstrated in three studies that have examined diabetes complications and genetics. Jin and colleagues have showed association of polymorphisms in TMEM217 (transmembrane protein), MRPL14 (mitochondrial ribosomial protein) and GRIK2 (glutamate receptor) genes with retinopathy (Jin and Liu 2008). Another complication of diabetes is nephropathy. In this case an association among rs2268388, mapping in ACAC (Acetyl-CoA Carboxylase Alpha) gene, and rs5186, mapping in AGTR1 (receptor type 1 of angiotensin II) gene, have been described with this complication. The study has been performed on 1,158 patients with T2D from the North and the South of India (Shah et al. 2013). Finally, Achilli and collaborators have performed an association study between T2D and mitochondrial haplogroups. Hypervariable sequences (HVS-I and HVS-II) of mitochondrial DNAs were sequenced in 904 healthy and diabetic subjects from the centre of Italy. No significant association was found with the disease; however, some haplogroups resulted associated with an increased risk of complications. In particular, U3 haplogroup is four times more represented in the patients with nephropathy, haplogroup V resulted associated with renal impairment, while haplogroups H and H3 were associated to increased risk of neuropathy (Achilli et al. 2011).

In the last two centuries, life expectancy is enormously increased, especially in the western countries (Abbott 2004). In fact at the beginning of XIX century life expectancy at birth was around 40 years (Abbott 2004), while in 2000-2005 has reached 71,9 years for men and 79,3 years for women, with exceptions that overcome the 100 (Candore et al. 2006). This phenomenon is due not only to an improvement of socio-economic conditions, but it also has a genetic contribution that explain about 25% of the trait (Herskind et al. 1996) (vB Hjelmborg et al. 2006). In fact, different studies have showed that longevity is the result of a complex interaction among genetic, environmental and stochastics factors (Franceschi et al. 2005). The purpose of the studies performed on centenarians is to detect the genetic components that contribute to longevity and healthy aging.

Centenarians, in fact, constitute the best model of successful aging (Franceschi et al. 2000), as they live much longer than the other members of their birth cohort, largely avoided or postponed the main age-related ailments, and maintained good physical and cognitive performances until very old age. The successful aging could be defined through the following characteristics:

- Absence of disability and diseases
- High-level of cognitive and physical abilities
- Maintenance of a productive or social activity (Franceschi et al. 2008).

Centenarians are characterized not only by an extreme longevity, but also through the ability to escape or delay the most common age-related diseases, such as cardiovascular diseases and diabetes (Franceschi and Bonafè 2003). For this reason, they are also studied for better understanding the genetic component involved in the pathologies. Data obtained from a group of Italian centenarians suggest that the major age-related modifications occur in metabolic pathways and in endocrine functions (Carrieri et al. 2001). It has been also reported that the prevalence of insulin resistance is very low in centenarians, moreover despite glucose and insulin concentrations are the same, between elderly people and centenarians, the latter show highest glucose uptake, comparable to those of middle-age people (Paolisso et al. 1996). Insulin resistance increases during aging up to 80-90 years. Between 90 and 100 years it is observed a significant decrease in insulin resistance (Cevenini et al. 2008). All these data suggest that the energy metabolism has an important role in healthy aging (Roth et al. 2004) (Fontana and Klein 2007) (Bonafè and Olivieri 2009). Moreover, it has been hypothesized that centenarians, in comparison to elderly general population, have a characteristic genetic pattern that preferentially activate an anti-inflammatory response (Salvioli et al. 2006) (Bonafè and Olivieri 2009) (Candore, Caruso, and Colonna-Romano 2010). Therefore, centenarians can be considered a fruitful experimental model to be used for genetic association studies in order to pinpoint genetic determinants of health and longevity.

## 1.4 Disease prediction models

Prediction model is an important tool in both medical practice and research and it may contribute to set up preventive interventions for specific subjects at relatively high risk of developing a disease. In fact, to identify subjects at high risk for diabetes is important in term of targeted prevention strategies, pharmacological focused treatment and follow up.

The initial attempt to create a T2D prediction model was performed in the ‘90s. The first study dates back to 1993, and took into account several factors like demographic, anthropometric, metabolic and hemodynamic parameters. The sensitivity of this multiple logistic regression model was higher (67.7 to 83.3%) respect to impaired glucose tolerance (56.5 to 62.1%); demonstrating that models based on different parameters could improve the efficacy of the test (Stern et al. 1993). Later on, several models have been developed that considered genetic susceptibility loci. A study conducted in Japanese population applied regression methods that included a genetic risk factor in combination with their interactions. The authors identified a model based on a Bayes factor approach and the LASSO regression method in which nine associated SNPs and clinical factors were considered. The authors concluded that the efficacy of their test with an area under the curve (AUC) of 0.8085 was accounted mainly by clinical factors rather than genetic variability (Shigemizu et al. 2014). Very recently, the same analytical method was applied to metabolomics parameters in a European population. The prediction of T2D by using this methods was high (AUC=0.81) and when the fasting glucose value was included, it rose up to AUC=0.89 (Liu et al. 2017). The last reported prediction model for T2D was develop using an analytical platform, the Reverse Engineering and Forward Simulation, and took into account several associated biochemical parameters in a retrospectivity study. The model accurately predicted progression to T2D (AUC = 0.76) in training group, and AUC = 0.78 in validation group samples.

Although several diabetes-related prediction models have been reported, and despite each accuracy seems to be elevate, other efforts will be do in order to found a validated and largely accepted prediction model of diabetes.

## 2 Aim of the study

The aim of this thesis is to apply an integrative analysis of **genetic and epigenetic data in order to identify new risk factors that can underlie yet unknown or underestimated molecular mechanisms at the basis of Type 2 Diabetes** considering an Italian population.

To do this, we decided to evaluate genetic risk factor performing a genome wide association on more than 550,000 SNPs using an approach that considers "extreme phenotypes", *i.e.* persons who never got diabetes in their life, like centenarians, and on the other side, patients with severe forms of T2D. We include in our study T2D patients with at least one complication, and centenarians as well as healthy controls.

In order to answer the first aspect of the question (**genetic** risk factors), we performed association analyses of genetic data for both comparisons (T2D VS controls and T2D VS centenarians) in order to identify the variants that potentially exert biological functions.

In order to answer the second aspect of the question (**epigenetic** risk factors), we considered the methylation status of *TCF7L2* gene, one of the strongest locus related to T2D, in a subset of diabetic patients and healthy controls. We performed the analysis in promoter region and two CpG Islands in intron 3 and 3'UTR of the gene, using a MassARRAY technology (Agena Bioscience).

Finally, we performed an **integrated analysis** of methylation and genetic results of *TCF7L2* gene in order to investigate a possible interaction between these factors. Based on this hypothesis we analysed the methylation status in patients and controls grouped based on the genotype of rs7903146 variant of *TCF7L2*.

# 3 Experimental Procedures

## 3.1 Cohort description

1,395 subjects from Northern/Central Italy were recruited by the Unit of Diabetology of the "National Institute on Health and Science on Aging" (INRCA) of Ancona, Italy; written informed consent was obtained from all subjects and T2D diagnosis was assessed according to the American Diabetes Association criteria (http://www.diabetes.org/). Subsequently, 614 patients (mean age 65.5 years) and 781 unrelated controls (mean age 48.4 years; age range 19-85 years) were studied. Moreover, 351 centenarians (mean age: 100.4 ± 1.4) have been considered in the genetic study (Table 1).

| | **T2D (n=614)** | **Controls (n=781)** | **Centenarians (n=351)** |
|---|---|---|---|
| **Sex (Males/Females)** | 330/284 | 419/362 | 79/272 |
| **Mean Age ± SD (years)** | 65.5±8.5 | 48.4±15.2 | 100.4±1.4 |

**Table 1 Subjects number, sex, age and BMI are reported. T2D: Type 2 Diabetes, BMI: Body Mass Index, SD: Standard Deviation**

A large number of biochemical and endocrinological parameters were collected for all individuals. A detailed clinical history was recorded for controls subjects, in order to exclude the presence of T2D and of any other overt disease.

***Genome Wide Association Study (GWAS) samples***

282 T2D patients with at least one complication or high level of glycated haemoglobin, 737 controls and 333 centenarians were included in GWAS. These subjects were genotyped at whole genome level according to the procedures described in the following section.

***Gene candidate methylation analysis samples***

Among of cohort previously described, 448 subjects, 229 T2D patients and 219 sex-, age- and BMI-matched controls, were selected for methylation analysis of *TCF7L2* gene, details are reported in table 2.

| | T2D (n=229) | Non diabetics (n=219) |
|---|---|---|
| **Gender (Males/Females)** | 115/114 | 92/127 |
| **Mean Age (years)** | 63.8±8.9 | 56.6±12.3 |
| **Mean BMI (kg/m$^2$)** | 28.0±3.9 | 26.9±4.3 |

**Table 2 Subjects number, sex, age and BMI are reported. BMI: Body Mass Index, SD: Standard Deviation**

The analysis was performed at gene region level and the procedure is reported in the following section.

***Disease Prediction Models samples***

*Logistic regression model based on genotypes, phenotypes and methylation value:* the samples encompassed in this analysis are the same including in gene candidate methylation analysis.

*Epigenetic Age Prediction "Little clock" samples:* 298 healthy subjects from 0 to 100 years was included in the training step, in order to test the model. Then, the trained model was applied to 24 centenarians, 24 centenarians' offspring, 16 subjects with Down Syndrome and 181 diabetic patients.

## 3.2 DNA extraction

Genomic DNA was extracted from 1,721 whole blood samples using the QIAmp DNA Blood Kit (Quiagen, Hilden, Germany) following the manufacture procedure except for the elution step, we utilized 100 µl for the first elution and 50 µl for the second one.

## 3.3 DNA quantification, normalization and bisulfite treatment

DNA quantification was obtained with spectrophotometric technology using NanoDrop 1000 (Thermo Fisher Scientific, Inc. MA, USA). For Illumina array 500 ng of DNA were normalized in 10 µl; for gene candidate methylation analysis 1000 ng of DNA were normalized in 10 µl.

1000 ng of genomic DNA was used to bisulfite conversion using the EZ DNA Methylation Kit (Zymo Research, Orange, CA, USA). Procedure was conducted according to the manufacturer's protocol, except for the conversion step; the conversion cycle was the following: 21 cycles at 95°C for 30 seconds and 50°C for 15 minutes.

## 3.4 Genome Wide Association Study

The genetic characterization at genomic level was performed using Infinium CoreExome-24 BeadChip (Illumina, Inc, CA, USA), this technology allows to assess the genotyping of more than 550,000 single nucleotide polymorphisms (SNPs). About 280,000 variants enclose into intronic or gene desert region and more than 265,000 loci encompass onto exonic region. This technology is based silicon-based array device, BeadChip. This peculiar chip manufacturing allows the simultaneous analysis up to 24 samples, which are physically separated within the chip sections. The BeadChip substrate contains microwells in which are held the beads attaching to oligonucleotides probe sequences. The robustness and measurement precision of this technology is allowed by multiple copies of each bead type present in the array; moreover, reproducibility and high-quality of genotyping data is obtained with hybridization-based quality controls of each array. The assay workflow is outlined in figure 1.

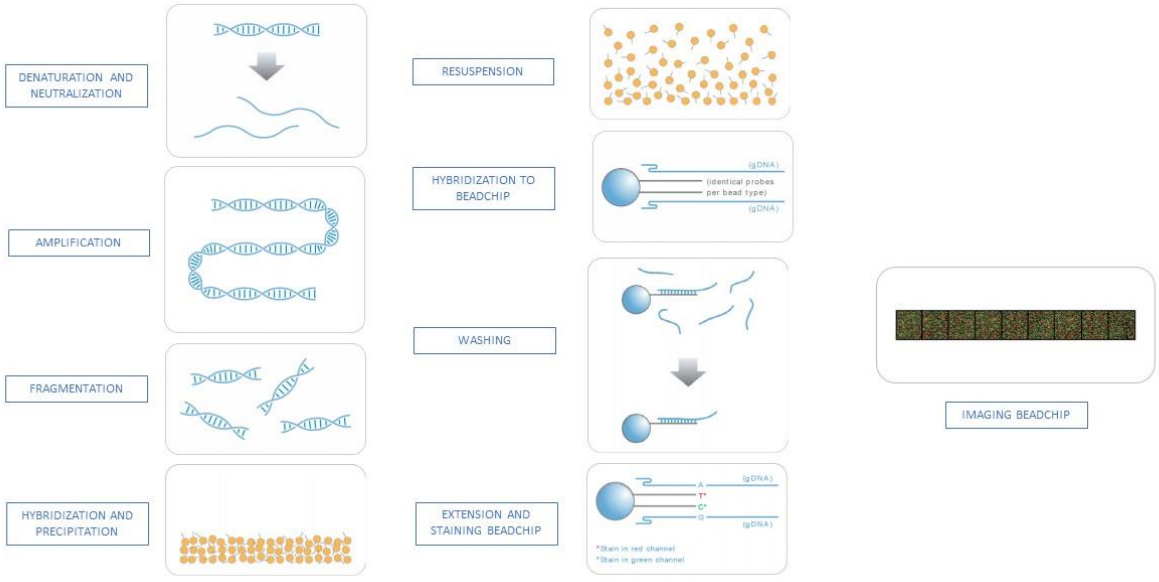


**Figure 1 Workflow of Infinium HTS Assay**

## 3.5 Gene candidate methylation analysis and Epigenetic Little Clock

Methylation levels of gene specific regions were assessed using mass spectrometry technology (Agena Bio, CA, USA). The EpiTYPER protocol, here utilized, is based on amplification

of bisulfite-converted DNA, SAP treatment, following by *in vitro* retro-transcription and T-specific cleavage. The cleaved fragments obtained are than analysed in mass spectrometry. Specific software is able to convert detected quantitative mass spectrometry in DNA methylation levels. This technology allows to interrogate the methylation status of several CpGs in amplicons of up to 700 bps length up to 384 samples in the same experiment. In figure 2 are reported the steps of EpiTYPER protocol.

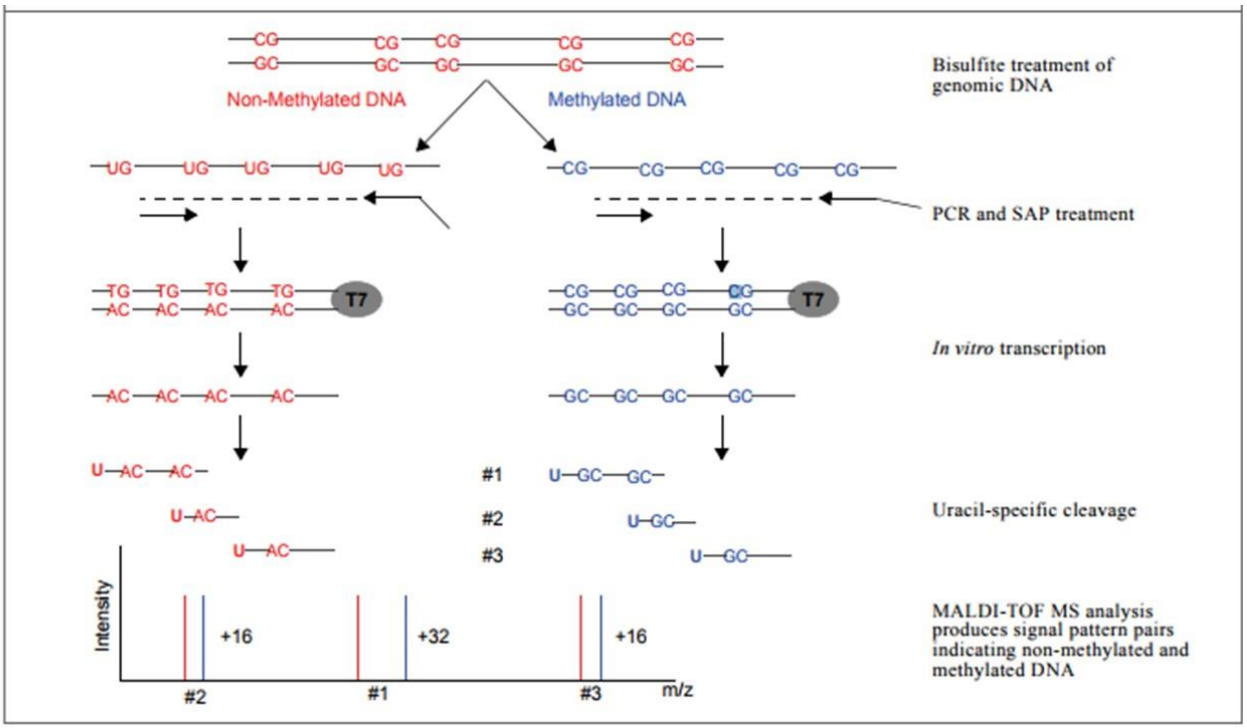


**Figure 2 Overview on EpiTYPER assay**

Moreover, this methodology is high reproducible and sensible, permitting to detect down to 5% differences in methylation. In tables 3 and 4 are reported informations of amplified regions for methylation analyses of *TCF7L2* gene and epigenetic little clock respectively. The design was performed using EpiDesigner software (Agena Bio, CA, USA) (https://www.epidesigner.com/).

| | Chromosome Position | Gene Region | Number of Amplicon | Number of CpG |
|---|---|---|---|---|
| **Promoter Region** | chr10:114709919-114710801 | 5'UTR | 2 | 30 |
| **CpG Island 43** | chr10:114711916-114712744 | Body (Intron 3) | 2 | 54 |
| **CpG Island 16** | chr10:114925323-114925951 | 3'UTR | 1 | 22 |

**Table 3 Amplified regions using for methylation analyses of TCF7L2 gene. Chromosome positions and CpG Island numbers are reported according to GRCh37/hg19 UCSC database.**

| | Chromosome Position | Gene Region | Number of Amplicon | Number of CpG |
|---|---|---|---|---|
| **ELOVL2** | Chr6:11044580-11045153 | Promoter | 1 | 16 |
| **CA3** | Chr8:86350116-86350532 | Promoter | 1 | 6 |
| **LYPD5** | Chr19:44324827-44325168 | Upstream | 1 | 10 |

**Table 4 Amplified regions using for methylation analyses of epigenetic little clock genes. Chromosome positions and number of selected amplicons are reported according to GRCh37/hg19 UCSC database.**

## 3.6 Statistical Analysis

*Genome wide association study:* Association analysis was performed using the PLINK package v.1.06 by means of a logistic model and adding sex as a covariate. The Manhattan plots were calculated using R (package: qqman). Quality controls (QC) were performed on the generated data in attempt to avoid the identification of false positive results when searching for loci potentially involved in longevity and according to protocols and pipelines described in Anderson et al. (Anderson et al. 2010).

*Univariate Logistic Regression, Multivariate Logistic Regression, Partial Least Square (*PLS) *regression*: regression and both Univariate and Multivariate Logistic Regression have been performed in python 2.7 using the module scikit-learn v.0.17.1.

*Epigenetic age (EpiAge) estimation*: estimation model has been computed using the Linear Regression function implemented in python 2.7 within the module scikit-learn v.0.17.1 and the regplot function implemented in seaborn v.0.6.0.

# 4 RESULTS AND DISCUSSION

## 4.1 Genome Wide Association Study

Although many efforts have been made to understand the genetic basis of T2D, only a small percentage of heritability has been explained so far. The difficulty to detect a strong genetic association is due, on one side, to the small effect that each single gene variant has on the phenotype and, on the other side, to the complexity of the disease. Both issues can be overcome by adopting a new strategy in genetic studies of T2D; the innovative approach here applied is the use of "extreme phenotypes". The basic hypothesis is that most serious clinical manifestations should be associated with enrichment in genetic risk variants, and conversely, people who have reached and passed the century of life by avoiding or postponing age-associated diseases should be enriched in protective gene variants. With this strategy, our group demonstrated by a gene-candidate study that the frequency of risk genotypes of rs7903146-T of TCF7L2 is minimal in centenarians and increases progressively according to the severity of the disease, showing that centenarians represent a powerful an informative control group in association studies on clinical heterogeneity disease like T2D (Garagnani et al. 2013).

GWAS was conducted following the hypothesis of the "extreme phenotypes" in order to increase the sensitivity of our study and to identify the variants with a biological function (Garagnani et al. 2013) (Giuliani et al. 2017). First of all, 282 diabetic patients with at least one complication and 773 healthy controls were compared. Then, the same group of diabetic patients (N=282) was compared with 334 centenarians. The genetic characterization was obtained by using Illumina array (Infinium CoreExome-24 BeadChip Kit) that includes 534,333 SNPs. In order to remove false positive associations, Quality controls (QC) were performed according to protocols and pipeline described in literature (Anderson et al. 2010) and described in Materials and Methods section. Figure 3 reports the Manhattan plot performed considering the nominal p-values of the logistic regression between diabetic patients and centenarians (Figure 3A) and diabetic patients and controls (Figure 3B). In the analysis sex has been considered as a covariate.

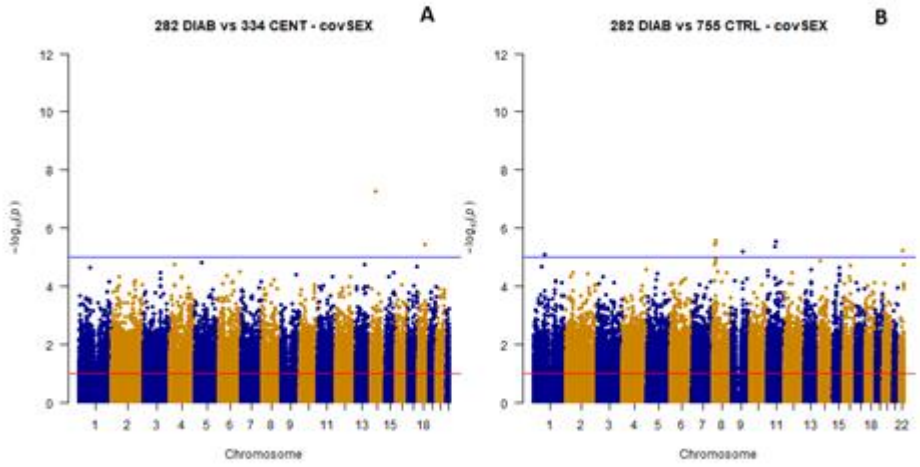


**Figure 2 Manhattan plot of all SNPs for the association analysis of two comparisons; A) associated SNPs considering diabetic patients and centenarians; B) associated SNPs considering diabetic patients and controls. The x axis shows SNPs according to their chromosomal positions. The blue line represents the threshold of $p<10^{-5}$.**

In the comparison of diabetic patients and controls, 245,918 variants have been removed due to minor allele thresholds, and then 270,512 variants and 1,382 individuals have passed QC filters. Subsequently, 73 variants with p value $<10^{-5}$ have been detected (table 5).

| CHR | SNP | BP | A1 | Freq Diab | Freq Cent | A2 | P | OR |
|---|---|---|---|---|---|---|---|---|
| 8 | rs4732756 | 27664686 | A | 0.1685 | 0.09272 | G | 2.82E-06 | 1.974 |
| 11 | rs7925649 | 72970776 | A | 0.1135 | 0.2046 | G | 2.96E-06 | 0.5023 |
| 8 | rs4732748 | 27634064 | A | 0.1667 | 0.09205 | G | 3.41E-06 | 1.962 |
| 8 | rs1729109 | 13194328 | A | 0.1119 | 0.05166 | G | 3.85E-06 | 2.283 |
| 11 | rs3825020 | 60184191 | G | 0.4273 | 0.3185 | A | 4.51E-06 | 1.613 |
| 22 | rs5995572 | 38753311 | G | 0.03929 | 0.006311 | A | 6.04E-06 | 6.159 |
| 9 | rs954124 | 88563064 | A | 0.1578 | 0.251 | G | 6.54E-06 | 0.5522 |
| 1 | rs11161732 | 86557967 | A | 0.531 | 0.4184 | G | 8.63E-06 | 1.564 |
| 8 | rs6987111 | 27600450 | A | 0.1631 | 0.09272 | C | 1.17E-05 | 1.885 |
| 14 | rs4640079 | 32438841 | A | 0.3777 | 0.2742 | G | 1.34E-05 | 1.555 |
| 8 | rs3779620 | 27679981 | G | 0.1684 | 0.09735 | A | 1.44E-05 | 1.855 |
| 22 | rs137878 | 50456791 | A | 0.3564 | 0.2623 | G | 1.90E-05 | 1.599 |
| 8 | rs1671389 | 13181518 | A | 0.1046 | 0.05033 | C | 1.92E-05 | 2.165 |
| 16 | rs7197624 | 55110037 | A | 0.0656 | 0.02517 | G | 1.99E-05 | 2.794 |
| 1 | rs12408292 | 62212822 | A | 0.3404 | 0.4417 | G | 2.18E-05 | 0.636 |
| 15 | rs8032461 | 71956136 | A | 0.2465 | 0.1656 | G | 2.38E-05 | 1.675 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | rs4602536 | 184298419 | G | 0.4592 | 0.357 | A | 2.76E-05 | 1.533 |
| 13 | rs1757889 | 76923323 | G | 0.3333 | 0.4364 | A | 2.83E-05 | 0.647 |
| 11 | rs11605196 | 6524072 | C | 0.2571 | 0.1722 | A | 3.00E-05 | 1.621 |
| 8 | rs7845862 | 38471714 | G | 0.3227 | 0.2311 | A | 3.31E-05 | 1.571 |
| 2 | rs10490411 | 56772337 | G | 0.4982 | 0.394 | A | 3.49E-05 | 1.498 |
| 12 | rs167769 | 57503775 | A | 0.3546 | 0.2616 | G | 3.55E-05 | 1.557 |
| 8 | rs17652171 | 113662583 | C | 0.1613 | 0.2437 | A | 3.60E-05 | 0.5751 |
| 15 | rs1995334 | 71949248 | A | 0.2429 | 0.1642 | G | 3.68E-05 | 1.657 |
| 5 | rs2052481 | 80282815 | C | 0.4078 | 0.3093 | A | 3.69E-05 | 1.526 |
| 2 | rs16862115 | 174680319 | A | 0.266 | 0.1841 | G | 3.89E-05 | 1.624 |
| 11 | rs11603869 | 6541225 | G | 0.2766 | 0.1914 | C | 4.39E-05 | 1.589 |
| 15 | rs338364 | 68173921 | G | 0.3652 | 0.2695 | A | 4.57E-05 | 1.523 |
| 16 | rs12934314 | 49822940 | G | 0.3812 | 0.2834 | A | 4.68E-05 | 1.501 |
| 2 | rs6742210 | 40615520 | A | 0.3865 | 0.4901 | G | 4.89E-05 | 0.6678 |
| 3 | rs13326165 | 52532118 | A | 0.1472 | 0.2305 | G | 4.94E-05 | 0.5795 |
| 11 | rs2233253 | 60157166 | G | 0.3617 | 0.2695 | A | 5.24E-05 | 1.541 |
| 9 | rs17086 | 124125520 | G | 0.539 | 0.4397 | A | 5.59E-05 | 1.498 |
| 6 | rs9388238 | 123731285 | A | 0.2394 | 0.1598 | G | 5.63E-05 | 1.618 |
| 11 | rs4938941 | 60173360 | A | 0.3617 | 0.2702 | G | 5.70E-05 | 1.537 |
| 5 | rs7728693 | 77142481 | G | 0.3227 | 0.4205 | A | 6.34E-05 | 0.6574 |
| 16 | rs1870038 | 55930518 | A | 0.1897 | 0.2762 | G | 6.43E-05 | 0.6126 |
| 1 | rs2902103 | 177341802 | C | 0.1543 | 0.09139 | A | 6.64E-05 | 1.791 |
| 6 | rs3823036 | 99284532 | G | 0.1809 | 0.2656 | A | 6.71E-05 | 0.6059 |
| 15 | rs4777035 | 68605169 | G | 0.5071 | 0.4079 | A | 6.76E-05 | 1.489 |
| 8 | rs1671382 | 13185325 | A | 0.1578 | 0.0947 | G | 6.79E-05 | 1.792 |
| 9 | rs7029374 | 23358448 | A | 0.3227 | 0.2364 | G | 6.81E-05 | 1.554 |
| 1 | rs12091015 | 172019718 | A | 0.5195 | 0.4232 | G | 6.93E-05 | 1.497 |
| 13 | rs1164519 | 76899888 | G | 0.4663 | 0.3695 | A | 6.97E-05 | 1.494 |
| 8 | rs590695 | 70025018 | G | 0.5727 | 0.4741 | A | 6.99E-05 | 1.492 |
| 11 | rs10431058 | 107481518 | A | 0.328 | 0.4258 | G | 7.03E-05 | 0.6611 |
| 7 | rs534126 | 142921234 | A | 0.4911 | 0.3927 | G | 7.20E-05 | 1.491 |
| 11 | rs10890742 | 107474795 | G | 0.3121 | 0.4079 | A | 7.42E-05 | 0.6573 |
| 17 | rs8075721 | 9956321 | G | 0.1259 | 0.2036 | A | 7.59E-05 | 0.5696 |
| 1 | rs17005517 | 219604453 | A | 0.195 | 0.2828 | C | 7.72E-05 | 0.624 |
| 9 | rs10868366 | 88700060 | A | 0.07092 | 0.1318 | C | 7.86E-05 | 0.4816 |
| 18 | rs4401135 | 69242763 | A | 0.3865 | 0.2965 | G | 7.88E-05 | 1.518 |
| 3 | rs1197312 | 133125713 | C | 0.4131 | 0.5093 | A | 7.99E-05 | 0.6663 |
| 16 | rs2903692 | 11238783 | A | 0.3457 | 0.4411 | G | 8.06E-05 | 0.6631 |
| 11 | rs10792493 | 80297206 | G | 0.06383 | 0.1278 | A | 8.13E-05 | 0.4784 |
| 6 | rs17584185 | 80381609 | G | 0.08865 | 0.153 | A | 8.17E-05 | 0.5123 |
| 12 | rs12298170 | 57515363 | G | 0.3918 | 0.2993 | A | 8.31E-05 | 1.497 |
| 11 | rs621942 | 85783738 | A | 0.3227 | 0.2377 | C | 8.40E-05 | 1.545 |
| 12 | rs17096085 | 38472919 | G | 0.0922 | 0.0457 | A | 8.46E-05 | 2.144 |
| 6 | rs214544 | 18284716 | A | 0.05319 | 0.1113 | G | 8.51E-05 | 0.4442 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 20 | rs6074990 | 15980838 | G | 0.2651 | 0.1877 | A | 8.54E-05 | 1.618 |
| 22 | rs137879 | 50457041 | G | 0.3989 | 0.4954 | A | 8.56E-05 | 0.668 |
| 3 | rs2114775 | 16727170 | A | 0.2961 | 0.3901 | C | 8.66E-05 | 0.6541 |
| 13 | rs7330047 | 76925888 | G | 0.3777 | 0.4742 | A | 8.71E-05 | 0.6682 |
| 10 | rs1000280 | 100096148 | A | 0.2961 | 0.2142 | G | 8.77E-05 | 1.562 |
| 2 | rs4675790 | 241859255 | A | 0.4167 | 0.3232 | C | 8.79E-05 | 1.497 |
| 7 | rs17172446 | 55202598 | A | 0.2553 | 0.1781 | G | 8.85E-05 | 1.604 |
| 6 | rs160699 | 2776303 | G | 0.3848 | 0.4861 | A | 9.13E-05 | 0.6832 |
| 11 | rs12790863 | 72979231 | G | 0.1099 | 0.1828 | A | 9.14E-05 | 0.5555 |
| 2 | rs7570320 | 234075691 | A | 0.2926 | 0.3828 | C | 9.37E-05 | 0.647 |
| 22 | rs5771069 | 50435480 | A | 0.3617 | 0.4576 | G | 9.57E-05 | 0.6697 |
| 4 | rs2176304 | 154017805 | A | 0.3387 | 0.2523 | G | 9.85E-05 | 1.522 |
| 18 | rs141913084 | 14490964 | G | 0.1093 | 0.06042 | A | 9.92E-05 | 2.054 |

**Table 5. List of SNPs with a pvalue < $10^{-5}$ in the comparison between diabetic individuals and controls. Chromosome positions are reported according to GRCh37/hg19 UCSC database.**

28 associated polymorphisms are in non-genic regions, while 45 lie in genic loci. Among the genic variants, one is in 3'UTR, one downstream region, 35 are intronic and 9 stands in coding region, all of them are non-synonymous, suggesting a possible modification in the aminoacidic sequence. In this comparison, 5 genes showing more than one associated locus. The *MS4A14* (Membrane Spanning 4-Domains A14) gene shows two associated variants, rs3825020 (p= $4.51^{E-06}$) and rs4938941 (p= $5.70^{E-05}$). The role of the encoding protein is unknown; probably it is part of multimeric receptor complex involved in signal transduction. The second gene with two associated loci is *THSD4* (Thrombospondin Type 1 Domain Containing 4), with rs8032461 (p= $2.38^{E-05}$) and rs1995334 (p= $3.68^{E-05}$). The function of the translated protein is related to peptidase and metallo-endopeptidase activities. No association with T2D of *MS4A14* and *THSD4* has been reported. Another gene showing more than one associated signal (rs11605196, p= $3.00^{E-05}$ and rs11603869, p= $4.39^{E-05}$) is *DNHD1* (DND MicroRNA-Mediated Repression Inhibitor 1); protein encodes from this gene inhibits microRNA-mediated repression, binding to microRNA-targeting sequences of mRNAs. This gene has been associated with several forms of cancer, such as tongue squamous cell and embryonal testis carcinomas, but, recently *DNHD1* has been identified in mixed ethnicity cases of severe proliferative diabetic retinopathy (Ung et al. 2017). Then rs10431058 (p=$7.03^{E-05}$) and rs10890742 (p=$7.42^{E-05}$) mapped in *ELMOD1* (ELMO Domain Containing 1) gene, that encodes for a protein with GTPase activator activity and interestingly, mice with mutation in their paralogue gene Elmod1, show lower blood glucose levels and lower body weight (de Angelis et al. 2015). Finally, three SNPs - rs1729109 (p=$3.85^{E-06}$), rs1671389 (p=$1.92^{E-05}$) and rs1671382 (p=$6.79^{E-05}$)

mapped in *DLC1* (Deleted in liver cancer 1) gene, another gene whose function is involved in GTPase activity regulation, Recently, Shih et al. have demonstrated that the protein encodes by *DLC1* gene regulates the activity of an inhibitors of matrix metalloproteinases, the plasminogen activator inhibitor-1 (*PAI-1*) (Shih, Takada, and Lo 2012). Noteworthy, there are some evidence that *PAI-1* is involved in several pathologies such as renal injury, inflammation, obesity and also diabetes (Ghosh and Vaughan 2012), suggesting a possibly implication of *DLC1* gene in these diseases.

In the comparison considering diabetic patients and centenarians (figure 3A), 270512 variants and 1107 people pass filters and quality controls (QC) check. The first comparison between diabetic patients and centenarians showed no significant SNPs with a nominal p-value < 10-8 but we identified 38 with p value $<10^{-5}$ that are listed in Table 6.

| CHR | SNP | BP | A1 | MAF_A | MAF_U | A2 | P | OR |
|---|---|---|---|---|---|---|---|---|
| 14 | exm1104579 | 58953746 | G | 0.2 | 0.001497 | C | 5.38E-08 | 247.8 |
| 18 | rs1480923 | 49480607 | A | 0.4486 | 0.3174 | G | 3.77E-06 | 1.817 |
| 5 | rs2113083 | 55946967 | G | 0.4681 | 0.3503 | A | 1.54E-05 | 1.791 |
| 13 | rs4307816 | 87730252 | G | 0.1578 | 0.08383 | A | 1.80E-05 | 2.265 |
| 4 | rs9306973 | 39627570 | A | 0.4943 | 0.3859 | G | 1.92E-05 | 1.72 |
| 17 | rs1699597 | 72566351 | G | 0.5585 | 0.4521 | A | 2.18E-05 | 1.721 |
| 1 | exm73271 | 89524657 | C | 0.01241 | 0.0524 | G | 2.29E-05 | 0.1535 |
| 6 | rs911028 | 166070424 | G | 0.2128 | 0.3039 | A | 3.12E-05 | 0.5368 |
| 3 | rs874546 | 127538412 | G | 0.227 | 0.1467 | A | 3.48E-05 | 2.004 |
| 15 | rs12148251 | 91698922 | A | 0.2801 | 0.3802 | G | 3.55E-05 | 0.5662 |
| 9 | rs891720 | 116897213 | A | 0.2642 | 0.1826 | C | 4.20E-05 | 1.853 |
| 6 | rs4710283 | 65506077 | A | 0.5018 | 0.3952 | G | 4.37E-05 | 1.697 |
| 4 | exm2269794 | 89742764 | A | 0.3594 | 0.4641 | G | 4.85E-05 | 0.5918 |
| 11 | rs573264 | 114092582 | A | 0.3954 | 0.2829 | G | 4.87E-05 | 1.717 |
| 2 | rs4671379 | 60495038 | A | 0.4787 | 0.3548 | G | 4.91E-05 | 1.692 |
| 15 | rs11852270 | 46606919 | A | 0.2748 | 0.3817 | G | 4.94E-05 | 0.5748 |
| 12 | rs634264 | 118133956 | A | 0.1738 | 0.268 | G | 5.00E-05 | 0.5316 |
| 4 | rs2869950 | 89742764 | A | 0.3599 | 0.4641 | G | 5.24E-05 | 0.5936 |
| 3 | rs11715363 | 127515984 | G | 0.2199 | 0.1482 | A | 5.85E-05 | 1.958 |
| 8 | rs6471376 | 94405276 | A | 0.3936 | 0.4985 | G | 5.88E-05 | 0.6061 |
| 2 | rs2887234 | 184028531 | G | 0.3434 | 0.4476 | A | 6.43E-05 | 0.5951 |
| 7 | rs414315 | 154812516 | G | 0.266 | 0.3608 | A | 7.06E-05 | 0.5761 |
| 6 | exm570114 | 108677977 | A | 0.06738 | 0.02844 | G | 7.25E-05 | 3.432 |
| 2 | rs6716389 | 176802850 | A | 0.01434 | 0.06287 | G | 7.36E-05 | 0.1936 |
| 1 | rs2067797 | 234264555 | C | 0.5391 | 0.4237 | A | 7.38E-05 | 1.669 |
| 6 | rs3823036 | 99284532 | G | 0.1809 | 0.2979 | A | 7.45E-05 | 0.5649 |
| 6 | rs8192591 | 32185796 | A | 0.0656 | 0.02246 | G | 8.48E-05 | 3.67 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | rs7901695 | 114754088 | G | 0.4557 | 0.3443 | A | 8.81E-05 | 1.663 |
| 4 | rs7697609 | 40956574 | A | 0.2518 | 0.1662 | G | 8.83E-05 | 1.853 |
| 2 | rs6750398 | 236018682 | G | 0.2057 | 0.2934 | A | 9.10E-05 | 0.5522 |
| 2 | rs6714145 | 174454044 | G | 0.2535 | 0.3533 | A | 9.25E-05 | 0.585 |
| 16 | rs4785345 | 49807722 | A | 0.2287 | 0.1467 | G | 9.37E-05 | 1.913 |
| 8 | rs590695 | 70025018 | G | 0.5727 | 0.4611 | A | 9.44E-05 | 1.635 |
| 2 | rs4671386 | 60514993 | C | 0.4876 | 0.3638 | A | 9.45E-05 | 1.65 |
| 11 | rs752849 | 47175327 | A | 0.3191 | 0.2365 | G | 9.51E-05 | 1.749 |
| 4 | rs11737809 | 40895018 | A | 0.2181 | 0.1377 | G | 9.71E-05 | 1.923 |
| 3 | rs1197312 | 133125713 | C | 0.4131 | 0.524 | A | 9.88E-05 | 0.6103 |
| 17 | rs4968282 | 44958937 | G | 0.2447 | 0.3443 | A | 9.94E-05 | 0.5823 |

**Table 6 . List of SNPs with a pvalue < $10^{-5}$ in the comparison between diabetic individuals (N=282) and centenarians (N=334). Chromosome positions are reported according to GRCh37/hg19 UCSC database.**

In addition, the associated variants were computed using snp-nexus tool (http://snp-nexus.org/) in order to obtain more informations about genomic mapping, gene/protein consequences and effect on protein functions. Within these variants, 14 polymorphisms are non-genic and 24 lie in genic region; among these 19 variants are intronic and only 1 in 3'UTR. The last 4 polymorphisms lie in codifying regions and are non-synonymous, implying an aminoacid modification in protein sequence. Further we focus the analysis on those genes that carry more significant SNPs. In particular, we identified two variants -rs7697609 ($p=8.83^{E-05}$) and rs11737809 ($p= 9.71^{E-05}$) - located in *APBB2* gene. The protein encodes by this gene is an amyloid beta precursor binding protein (Family B Member 2) which are involved in signal transduction. Two additional significant loci, rs874546 ($p= 3.48^{E-05}$) and rs11715363 ($p= 5.85^{E-05}$) are located in *MGLL* gene. The hydrolytic protein codifies by this gene is involved in monoacylglycerides and endocannabinoids metabolism. The last gene that carries more than one significant loci is *FAM13A* gene (family with sequence similarity 13, member A) with exm2269794 ($p=4.85^{E-05}$) and rs2869950 ($p=5.24^{E-05}$). The function of the protein codified by this gene is still unknown. While *APBB2* and *MGLL* have never been related to T2D trait, the FAM13A gene has been previously associated with fasting insulin (Scott et al. 2012,) (Yaghootkar et al. 2014). Noteworthy, the rs7901695 ($p=8.81^{E-05}$) of *TCF7L2* gene has confirmed its strong association with T2D also in our cohort.

We recently hypothesized that variants identified in the comparison between centenarians and diabetic patients may add a biological validation of the variants identified in the "classical" case-controls study. Accordingly, we identified those variants that have showed associations in both comparisons (diabetic patients *VS* controls and diabetic patients *VS* centenarians), considering all the

SNPs with p value<$10^{-4}$ in the two independent comparisons described above. Table 7 reported the shared variants.

| rs_ID | Allele | F Diab | F Ctrl | F Cent | P value Db/Ctrl | P value Db/Cent | Gene | Chr | Position |
|---|---|---|---|---|---|---|---|---|---|
| rs2056279 | G | 0.5833 | 0.4848 | 0.4566 | 0.00015 | 0.000119 | ANAPC10 | 4 | 145917557 |
| rs25653 | G | 0.4096 | 0.4934 | 0.5329 | 0.00076 | 0.000152 | ANPEP | 15 | 90349558 |
| rs1197312 | C | 0.4131 | 0.5093 | 0.524 | 8E-05 | 9.88E-05 | BFSP2 | 3 | 133125713 |
| rs17652171 | C | 0.1613 | 0.2437 | 0.2575 | 3.6E-05 | 0.000125 | CSMD3 | 8 | 113662583 |
| rs170020 | G | 0.5372 | 0.4477 | 0.4401 | 0.00032 | 0.000634 | CTB_12O2.1 | 5 | 151344558 |
| rs10889061 | A | 0.3511 | 0.2715 | 0.2635 | 0.0004 | 0.000718 | DAB1 | 1 | 58121100 |
| rs2351921 | A | 0.511 | 0.432 | 0.3952 | 0.00082 | 0.000239 | DENND5B | 12 | 31667973 |
| rs12091015 | A | 0.5195 | 0.4232 | 0.4162 | 6.9E-05 | 0.000677 | DNM3 | 1 | 172019718 |
| rs7854418 | G | 0.4217 | 0.3404 | 0.3338 | 0.00055 | 0.00043 | DOCK8 | 9 | 335545 |
| rs10829661 | A | 0.4291 | 0.3457 | 0.3383 | 0.00049 | 0.000382 | EBF3 | 10 | 131726740 |
| rs4710283 | A | 0.5018 | 0.4172 | 0.3952 | 0.00049 | 4.37E-05 | EYS | 6 | 65506077 |
| rs17086156 | A | 0.1223 | 0.07351 | 0.06437 | 0.00045 | 0.000423 | FRMD3 | 9 | 85945797 |
| rs17130717 | C | 0.01241 | 0.04702 | 0.0524 | 0.00051 | 2.29E-05 | GBP1 | 1 | 89524657 |
| rs45598235 | A | 0.08511 | 0.04371 | 0.04341 | 0.00028 | 0.000321 | GPR110 | 6 | 46977783 |
| rs4777035 | G | 0.5071 | 0.4079 | 0.3859 | 6.8E-05 | 0.00037 | ITGA11 | 15 | 68605169 |
| rs4952336 | C | 0.1489 | 0.2212 | 0.2275 | 0.00028 | 0.000958 | LTBP1 | 2 | 33250571 |
| rs6074990 | G | 0.2651 | 0.1877 | 0.1814 | 8.5E-05 | 0.000509 | MACROD2 | 20 | 15980838 |
| rs954124 | A | 0.1578 | 0.251 | 0.244 | 6.5E-06 | 0.000724 | NAA35 | 9 | 88563064 |
| rs8192591 | A | 0.4947 | 0.4111 | 0.4386 | 0.00065 | 8.48E-05 | NOTCH4 | 6 | 32215796 |
| rs10768450 | A | 0.2101 | 0.2907 | 0.2994 | 0.00026 | 0.00058 | OR51L1 | 11 | 5020832 |
| rs11734405 | G | 0.5727 | 0.4781 | 0.4491 | 0.00027 | 0.000216 | OTUD4 | 4 | 146081096 |
| rs3823036 | G | 0.1809 | 0.2656 | 0.2979 | 6.7E-05 | 7.45E-05 | POU3F2 | 6 | 99284532 |
| rs2272399 | A | 0.4344 | 0.3497 | 0.3428 | 0.00053 | 0.000817 | SLC6A11 | 3 | 10975745 |
| rs2905964 | G | 0.3387 | 0.2556 | 0.247 | 0.00019 | 0.000134 | SPOCK1 | 5 | 136452856 |
| rs167769 | A | 0.3546 | 0.2616 | 0.2545 | 3.5E-05 | 0.000116 | STAT6 | 12 | 57503775 |
| rs7901695 | G | 0.4557 | 0.3742 | 0.3443 | 0.00079 | 8.81E-05 | TCF7L2 | 10 | 114754088 |
| rs7903146 | A | 0.4415 | 0.3623 | 0.3338 | 0.00099 | 0.000122 | TCF7L2 | 10 | 114758349 |
| rs2670179 | A | 0.2943 | 0.3788 | 0.3907 | 0.00034 | 0.000235 | TMEM108 | 3 | 133110636 |
| rs4854583 | G | 0.4089 | 0.496 | 0.506 | 0.0004 | 0.000784 | TMEM108 | 3 | 133112473 |
| rs3732572 | G | 0.411 | 0.4974 | 0.506 | 0.00044 | 0.000897 | TMEM108 | 3 | 133109469 |
| rs12461075 | A | 0.344 | 0.2636 | 0.259 | 0.00033 | 0.000363 | TULP2 | 19 | 49399821 |
| rs6437126 | A | 0.1348 | 0.2013 | 0.2081 | 0.00037 | 0.00041 | UPP2 | 2 | 158865831 |
| rs1860163 | A | 0.2855 | 0.3632 | 0.3772 | 0.00085 | 0.000605 | UPP2 | 2 | 158870618 |
| rs160699 | G | 0.3848 | 0.4861 | 0.4701 | 9.1E-05 | 0.00032 | WRNIP1 | 6 | 2776303 |
| rs4785345 | A | 0.2287 | 0.1623 | 0.1467 | 0.00051 | 9.37E-05 | ZNF423 | 16 | 49807722 |

**Table 7 . List of SNPs with a pvalue < $10^{-4}$ showing association in the two comparisons. Chromosome positions are reported according to GRCh37/hg19 UCSC database.**

Among the 35 shared variants, we found that only one SNP, rs170020, is located in 3'UTR, 28 SNPs lie in intronic regions and 6 SNPs are placed in coding regions, these latter implying an aminoacid substitution in protein sequence (Table 8).

| Gene | rs_ID | P value Db/Ctrl | P value Db/Cent | Gene region | Protein Effect | Trait (GWAS catalogue) |
|---|---|---|---|---|---|---|
| **ANAPC10** | rs2056279 | 0.00015 | 0.000119 | intronic | -- | -- |
| **ANPEP** | rs25653 | 0.00076 | 0.000152 | coding | nonsyn | -- |
| **BFSP2** | rs1197312 | 8E-05 | 9.88E-05 | intronic | -- | -- |
| **CSMD3** | rs17652171 | 3.6E-05 | 0.000125 | intronic | -- | -- |
| **CTB_12O2.1** | rs170020 | 0.00032 | 0.000634 | non-coding intronic | -- | Interstitial lung disease |
| **DAB1** | rs10889061 | 0.0004 | 0.000718 | intronic | -- | -- |
| **DENND5B** | rs2351921 | 0.00082 | 0.000239 | intronic | -- | -- |
| **DNM3** | rs12091015 | 6.9E-05 | 0.000677 | intronic | -- | -- |
| **DOCK8** | rs7854418 | 0.00055 | 0.00043 | intronic | -- | -- |
| **EBF3** | rs10829661 | 0.00049 | 0.000382 | intronic | -- | -- |
| **EYS** | rs4710283 | 0.00049 | 4.37E-05 | intronic | -- | -- |
| **FRMD3** | rs17086156 | 0.00045 | 0.000423 | intronic | -- | -- |
| **GBP1** | rs17130717 | 0.00051 | 2.29E-05 | coding | nonsyn | -- |
| **GPR110** | rs45598235 | 0.00028 | 0.000321 | coding | nonsyn | -- |
| **ITGA11** | rs4777035 | 6.8E-05 | 0.00037 | coding | nonsyn | -- |
| **LTBP1** | rs4952336 | 0.00028 | 0.000958 | intronic | -- | -- |
| **MACROD2** | rs6074990 | 8.5E-05 | 0.000509 | intronic | -- | -- |
| **NAA35** | rs954124 | 6.5E-06 | 0.000724 | intronic | -- | -- |
| **NOTCH4** | rs8192591 | 0.00065 | 8.48E-05 | coding | nonsyn | -- |
| **OR51L1** | rs10768450 | 0.00026 | 0.00058 | coding | nonsyn | -- |
| **OTUD4** | rs11734405 | 0.00027 | 0.000216 | intronic | -- | -- |
| **POU3F2** | rs3823036 | 6.7E-05 | 7.45E-05 | 3utr | -- | -- |
| **SLC6A11** | rs2272399 | 0.00053 | 0.000817 | intronic | -- | -- |
| **SPOCK1** | rs2905964 | 0.00019 | 0.000134 | intronic | -- | -- |
| **STAT6** | rs167769 | 3.5E-05 | 0.000116 | intronic | -- | Eosinophilic esophagitis (pediatric) |
| **TCF7L2** | rs7901695 | 0.00079 | 8.81E-05 | intronic | -- | Coronary heart disease |
| **TCF7L2** | rs7903146 | 0.00099 | 0.000122 | intronic | -- | Type 2 diabetes |
| **TMEM108** | rs2670179 | 0.00034 | 0.000235 | intronic | -- | -- |
| **TMEM108** | rs4854583 | 0.0004 | 0.000784 | intronic | -- | -- |
| **TMEM108** | rs3732572 | 0.00044 | 0.000897 | intronic | -- | -- |
| **TULP2** | rs12461075 | 0.00033 | 0.000363 | intronic | -- | -- |
| **UPP2** | rs6437126 | 0.00037 | 0.00041 | intronic | -- | -- |
| **UPP2** | rs1860163 | 0.00085 | 0.000605 | intronic | -- | -- |
| **WRNIP1** | rs160699 | 9.1E-05 | 0.00032 | intronic | -- | -- |
| **ZNF423** | rs4785345 | 0.00051 | 9.37E-05 | intronic | -- | -- |

**Table 8 List of SNPs with a pvalue < $10^{-4}$ showing association in the two comparisons with more than one variant. Gene region, protein functions and trait are reported.**

Also in this comparison we focus on those genes with multiple associated loci. One of them is *TMEM108* gene, that codes for a transmembrane protein required for the development of neuron circuitry and many other functions related to axonal signal transport and neurons connections. Two GWASs have identified *TMEM108* as possible risk genes for alcohol consumption in European populations (Grant et al. 2009) (Heath et al. 2011), and it is also known as a risk locus for psychotic disorders (Beveridge and Cairns 2012). Remarkably, several GWAS (Speliotes et al. 2010) (Akiyama et al. 2017) and meta-analyses studies (Locke et al. 2015) (Graff et al. 2017)(Justice et al. 2017) have been reported *TMEM18* associated with BMI (Rowlands et al. 2014). *UPP2* (Uridine Phosphorylase 2) gene is a protein involved in purine metabolism and catalysed the reversible phosphorylytic cleavage of uridine and deoxyuridine to uracil and ribose- or deoxyribose-1-phosphate. Noteworthy, a decrease in expression levels of *UPP2* gene have been reported in obese mice (Mollah and Ishikawa 2010) and in animal feeding with high fat diet (Eckel-Mahan et al. 2013). Interestingly and not surprisingly also transcription factor 7-like 2 (*TCF7L2*) gene has shown a multiple association, with rs7901695 and rs7903146. TCF7L2 gene is a Wnt signalling pathway effector, involved in pancreatic beta-cell function, the differentiation of adipocytes and regulation of adipokines (Schinner 2009). rs7903146 has been identified as one of the strongest associated variant in several studies, among different populations. One copy of the minor allele of rs7903146 (T) increased risk for T2D up to 1.37 (Prokopenko, McCarthy, and Lindgren 2008) and carriers of TT genotypes show a great increased of *TCF7L2* expression in pancreatic islet with impairing insulin secretion (Lyssenko et al. 2007).

In the comparison of diabetic patients and controls, several genes with multiple associated variants were identified. In particular, different genes are involved in important cellular mechanism; we found in association gene related to transduction signal (*MS4A14*), to gene expression regulation at post-transcriptional levels (*DNHD1*), involved in metalloprotease activity (*THSD4* and *DLC1*) and finally, genes with GTPase activity (*ELMOD1* and *DLC1*). In addition, also in the comparison between diabetic patients and centenarians, we found two novel genes that show multiple association loci to T2D, *APBB2* and *MGLL*, involved in transduction signal and monoacylglycerides and endocannabinoids metabolism, respectively. In addition, we confirmed *FAM13A*, previously reported related to fasting insulin, and *TCF7L2* genes as implicated in this form of diabetes. Finally, the analysis of shared associated signals, between the "extreme phenotype" approach and the classical case control study, confirmed the crucial role played by TCF7L2 and identified two new genes UPP2 and TMEM108 as potential genetic determinants of type 2 diabetes. While the susceptibility of

*TCF7L2* have been largely reported, also by our group (Garagnani et al. 2013), very scarce findings about the *UPP2* and *TMEM18* gene in T2D have been reported. It is interesting to underline the association of both genes with obesity; in fact it is commonly known that obesity and T2D share numerous genetic associated polymorphisms (Zeggini et al. 2008) (Meyre et al. 2009) (Locke et al. 2015).

**Summary**: in this study I performed a T2D case control study at genome level in the Italian population. T2D patients with at least one complication were compared first with the healthy control group and then with centenarians, applying the "extreme phenotypes" approach. In table 9 are reported those genes that displayed multiple associated variants in the two comparisons.

| Comparison | Associated Genes | Gene functions |
|---|---|---|
| **Diabetics *VS* Controls** | *DNHD1* | Gene expression regulation |
| | *DLC1* | GTPase activity |
| | *ELMOD1* | GTPase activity |
| | ***M4A14*** | Trasduction signals |
| | ***TSHD4*S** | Metallo-protease activity |
| **Diabetics *VS* Centenarians** | ***APBB2*** | Trasduction signals |
| | *FAM13A* | Unkowmn function |
| | ***MGLL*** | Monoacylglycerides and endocannabinoids metabolisms |
| | *TCF7L2* | Trascrition factor involved in wnt-signal pathway |

**Table 9 Genes showing multiple association with p value<$10^{-4}$ in the two comparisons and their functions. Novel associated genes are in bold.**

As summarized in the table, as a whole the present study identified 9 genes with more than one associated variant, and 4 of them have never been reported before. Except for *FAM13A* whose function is still unknown, the other genes are involved in crucial cellular processes. Noteworthy, some genes are involved in the same functions, i.e. intracellular transduction signals, gene expression and GTPase activity regulation.

Finally, we selected those variants that showed multiple associations and were present in both comparisons. The results are reported in table 10.

| Comparison | Shared associated genes | Gene functions |
|---|---|---|
| Diabetics *VS* Controls | *TCF7L2* | Trascrition factor involved in wnt-signal |
| And | | pathway |
| **Diabetic *VS* Centenarians** | *TMEM108* | Transmembrane protein involved in neurons activities |
| | *UPP2* | Purine metabolism |

**Table 10 Genes with multiple association (p value<$10^{-4}$) shared in the two comparisons and their functions.**

Through this analysis 3 genes were identified: *TCF7L2*, already known as one of the most important genes involved in T2D, *UPP2*, which is likely involved in glucose metabolism, and *TMEM108*, another important risk factor for T2D.

While confirming the association of some genes with T2D, this study identified novel associations with previously unreported loci that might be involved in the most severe forms of diabetes. Further investigations to evaluate their role in T2D need to be performed.

## 4.2 Gene-candidate Methylation Analysis

It is known that only 10-12% of T2D heritability is explained by genetic variability (Ali 2013), however many environmental factors also have a critical role in the onset of the disease. In fact, numerous external influences module gene expression via epigenetic mechanisms; the most known and studied is DNA methylation. Studies conduct on models of intrauterine malnutrition and intrauterine environmental exposures have reported that epigenetic changes contribute to pathogenesis of T2D (Hales and Barker 1992) (Dabelea et al. 2000) (Lee et al. 2005) (Hall et al. 2014). For this reason, we decided to investigate the role of the methylation status in the onset of T2D considering one of the strongest susceptibility locus identified in the previous analysis and by recent studies (Mayans et al. 2007) (Sladek et al. 2007) (van Vliet-Ostaptchouk et al. 2007) (Stitzel et al. 2010), i.e. *TCF7L2* gene.

Methylation analysis at gene candidate level was performed on 229 diabetic patients and 219 controls matched for age, sex and BMI. For this analysis we considered three different regions, the first located in the promoter region - previously described as differentially methylated in T2D patients (Canivell et al. 2014), the other regions investigated were the CpG Islands (CpGIs) presented in the entire gene region, one in intron 3 and one in 3'UTR. In Table 11 all informations are reported.

| | Chromosome Position | Gene Region | Number of Amplicons | Number of CpG Sites |
|---|---|---|---|---|
| **Promoter Region** | chr10:114709919-114710801 | 5'UTR | 2 | 30 |
| **CpG Island 43** | chr10:114711916-114712744 | Body (Intron 3) | 2 | 54 |
| **Cpg Island 16** | chr10:114925323-114925951 | 3'UTR | 1 | 22 |

**Table 11 Chromosome positions and CpG Island numbers are reported according to GRCh37 UCSC database.**

Firstly we performed the analysis of variance to assess the differences between the two groups, we analysed methylations of 84 CpG sites in three regions of *TCF7L2*. In table 11 are listed the means of methylation levels of CpG sites resulted differentially methylated in diabetic and controls comparison.

| Gene region | CpG sites | Groups | Methylation levels ± SD | p* (ANOVA) |
|---|---|---|---|---|
| **Intron** | 19_CpG_5 | T2D | 0.6065 ± 0.0796 | 0.0106 |
| | | Ctrl | 0.6265 ± 0.0811 | |
| | 19_CpG_8.9 | T2D | 0.1466 ± 0.0747 | 0.0135 |
| | | Ctrl | 0.1678 ± 0.0760 | |

**Table 12 CpG sites showed a significant different methylation levels in T2D individuals and controls are reported. *p values are corrected for sex, age and BMI**

As reported, only two CpG sites in intronic CpGI of the TCF7L2 gene showed an association signal (CpG_5 p=0.0106; CpG_8.9 p=0.013), while, in the promoter region and in CpGI at 3'UTR no CpG associated sites have been identified. Our results do not confirm those of Canivell et al. and this might be due to the heterogeneity of the diabetic patients. In their study, in fact, the cohort considered including only drug-naive T2D patients, whereas our cohort includes diabetic patients diagnosed since several years and pharmacologically treated. Therefore, we can speculate that the different results might be due to the heterogeneity of patients considered and to the pharmacological treatments that characterized our cohort.

In the light of how reported by Lyssenko et al., which demonstrated that subjects carriers of TT genotypes show a great increased of *TCF7L2* expression in pancreatic islet with impairing insulin secretion (Lyssenko et al. 2007) we investigated if this connection was due to a methylation regulation. Then, we decided to perform an integrated analysis combining genotypes of TCF7L2 rs7903146 and DNA methylation of the surrounding regions. This is because genetic and epigenetic factors may have a combined role in the pathological mechanisms at the basis of the disease. For this reason, we decide to evaluate methylation status in the genetic contest of the *TCF7L2* gene.

Then, DNA methylation profiles were grouped based on the genotype of rs7903146 TCF7L2 of each individual (N=404 including both diabetic patients and controls) and the association between the two factors was assessed using the ANalysis Of VAriance (ANOVA), while correcting for sex, BMI and age. The analysis was done considering 54 CpG sites. The results are reported in table 12.

| Region | CpG Sites | Genotypes (n) | Methylation levels ± DS | p* (ANOVA) | Pairwase Comparison (*p*) |
|---|---|---|---|---|---|
| **Promoter** | 3_CpG_12 | CC (107) | 0.0121 ± 0.0099 | 0.0027 | TT vs TC (0.0355) |
| | | TC (205) | 0.0144 ± 0.0091 | | TT vs CC (0.0019) |
| | | TT (92) | 0.0179 ± 0.0175 | | |
| | 3_CpG_15 | CC (107) | 0.1880 ± 0.0626 | 8.741e-06 | CC vs TC (1.5e-05) |
| | | TC (205) | 0.2235 ± 0.0673 | | CC vs TT (0.00025) |
| | | TT (92) | 0.2235 ± 0.0586 | | |
| **Intron** | 13_CpG_28 | CC (90) | 0.0323 ± 0.0618 | 0.01615 | TT vs TC (0.015) |
| | | TC (178) | 0.0282 ± 0.0870 | | |
| | | TT (80) | 0.0648 ± 0.1414 | | |
| | 19_CpG_1 | CC (104) | 0.3546 ± 0.1371 | 0.01298 | CC vs TC (0.019) |
| | | TC (202) | 0.4006 ± 0.1393 | | CC vs TT (0.030) |
| | | TT (91) | 0.4032 ± 0.1385 | | |

**Table 13 CpG sites showed a significant different methylation levels in three genotypes groups are reported. *p values are corrected for sex, age and BMI**

4 CpG sites, 2 in promoter and 2 intronic regions have shown a significant different methylation values in three groups. The allele specific methylation values here reported show the straight interaction between two risk factors.

Few studies have attempted to integrate genetic and epigenetic data in the study of T2D so far (Bell et al. 2010); Bell et al., have demonstrated a haplotype-specific methylation in the FTO but they did not identify any differentially methylated regions (DMRs) in T2D and controls comparison. More recently, in one study conducted at epigenome-wide level the authors identified DNA methylation variations at level of known GWAS T2D-related loci (Toperoff et al. 2012). In addition, an integrated genetic and epigenenetic study in monozygotic twins, has confirmed an enrichment for DMR in T2D-susceptibility loci (Yuan et al. 2014). A better understanding of genetic and epigenetic interplaying may contribute to improve a disease risk evaluation.

**Summary**: This is the first DNA methylation study of *TCF7L2* where the entire gene region has been deeply investigated. We performed a quantitative methylation analysis of 84 CpG sites spanning in three regions: the promoter and two CpG Islands in intron 3 and in 3'UTR. We reported an association between methylation levels of two CpG sites in intron 3 of *TCF7L2* gene with T2D in

our population. Moreover, we demonstrated that methylation levels of four CpG sites (two in promoter region and two in intron 3) are strongly influenced by the genotype of the most T2D related variant, rs7903146. These results showing a high interaction between two important risk factors indicate the need for integration of genetic and epigenetic data in order to better understand the basis of a complex disease such as T2D.

## 4.3 Disease Prediction Models

At this point it is evident that the search for predisposition to diabetic pathology must take into account not only the individual risk factors but also the interaction between them.

We applied two different approaches in order to develop a disease predictor model able to differentiate T2D patients group from the control group.

Pearson's correlations analysis among all selected variables showed that the variable that has the strongest correlation with diabetes is age (figure 4).

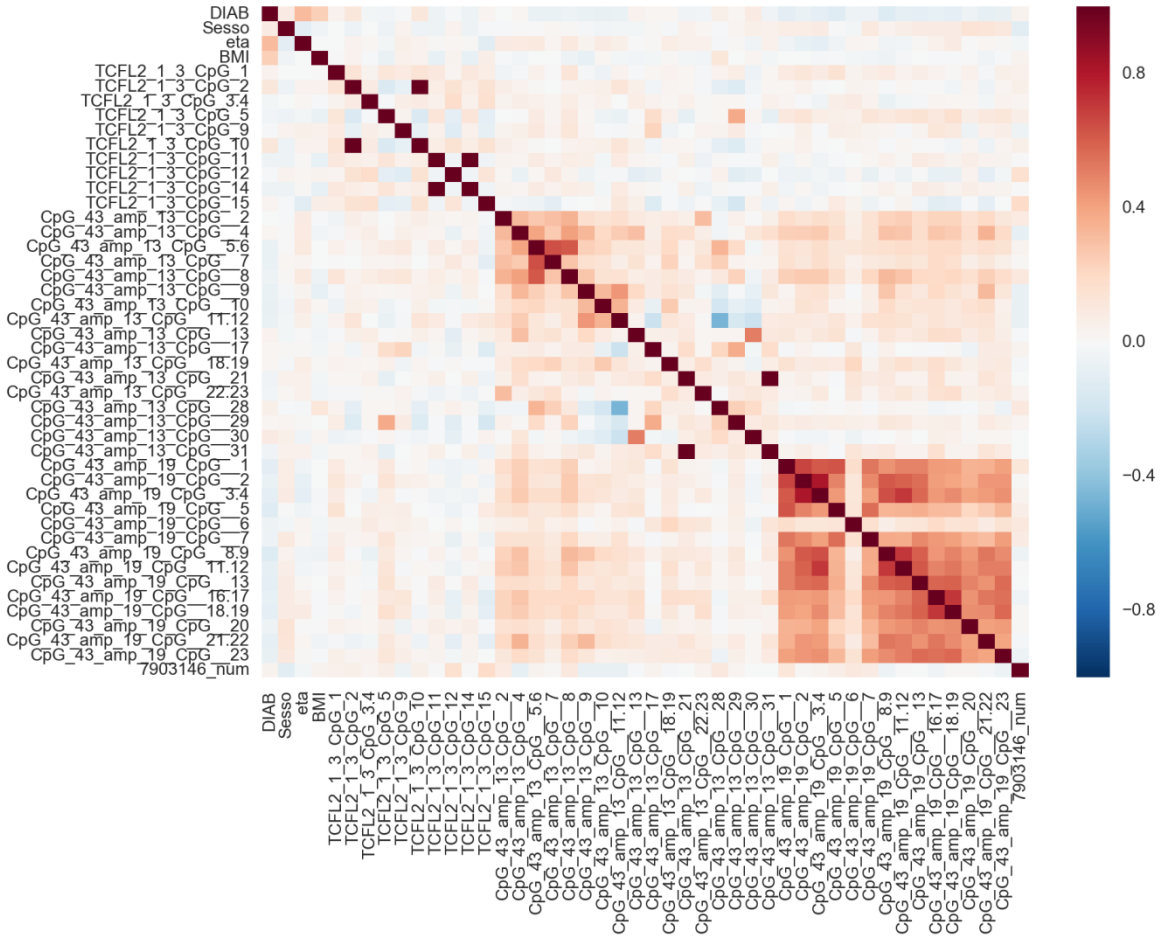


**Figure 3 Heatmap of Pearsons' correlation among selected variables. Colours show positive or negative correlation according to the scale reported on the right of the figure.**

In order to develop a disease predictor model able to differentiate T2D individuals from the control group, we considered two different approaches.

***Disease Prediction Model 1***

The first approach was based on the computation of a logistic regression model that takes into account the following parameters: disease diagnosis, sex, age, BMI, genotypes of rs7903146 and methylation levels of the previously investigated *TCF7L2* gene (54 CpG sites) in 229 diabetic patients and 219 healthy controls.

In order to investigate the association between each epigenetic and genetic covariate (CpG sites and rs7903146) and diabetes, we first performed a 10-fold cross-validated univariate logistic regression in which age, sex and BMI were also added as confounding variables. In order to estimate the regression coefficients variability and to randomize over the choice, the analysis was repeated 50 times. Table 13 reports the mean and standard deviation of the coefficients obtained from the 50 iterations. The first two columns refer to the model coefficient relative to the covariate under investigation (row index). Rows were sorted in decreasing order according to the absolute value of the covariate-specific regression coefficient (column "Coef mean"), i.e. according to the covariates importance. Columns "Age mean", "Age SD", "BMI mean", "BMI SD" and "Sex mean", "Sex SD" refer to the mean and standard deviation of the regression coefficients relative to Age, BMI and Sex respectively.

| | Coef mean | Coef SD | Age mean | Age SD | BMI mean | BMI SD | Sex mean | Sex SD | Accuracy Score mean | Accuracy Score SD |
|---|---|---|---|---|---|---|---|---|---|---|
| **CpG_43_19_CpG_18.19** | -0.31154 | 0.04452 | 0.85176 | 0.03977 | 0.61072 | 0.03329 | 0.35106 | 0.02906 | 0.66786 | 0.06637 |
| **CpG_43_19_CpG_13** | -0.26459 | 0.03806 | 0.83622 | 0.03561 | 0.61900 | 0.03525 | 0.33167 | 0.03543 | 0.66987 | 0.06586 |
| **CpG_43_19_CpG_5** | -0.25858 | 0.03397 | 0.82594 | 0.04605 | 0.61784 | 0.03839 | 0.33127 | 0.03628 | 0.68262 | 0.06514 |
| **CpG_43_19_CpG_8.9** | -0.25569 | 0.03080 | 0.81301 | 0.03938 | 0.59908 | 0.04146 | 0.33513 | 0.03495 | 0.66880 | 0.06931 |
| **CpG_43_19_CpG_21.22** | -0.24744 | 0.04516 | 0.83035 | 0.03405 | 0.59678 | 0.03413 | 0.35155 | 0.03158 | 0.68133 | 0.06635 |
| **CpG_43_19_CpG_16.17** | -0.24484 | 0.02904 | 0.83043 | 0.04234 | 0.60842 | 0.04284 | 0.33287 | 0.03866 | 0.66417 | 0.06677 |
| **CpG_43_19_CpG_1** | -0.23991 | 0.04067 | 0.82809 | 0.04158 | 0.62219 | 0.03365 | 0.33480 | 0.04183 | 0.67053 | 0.06645 |
| **CpG_43_13_CpG_5.6** | -0.23978 | 0.04125 | 0.84970 | 0.03292 | 0.60832 | 0.03729 | 0.32685 | 0.03895 | 0.66987 | 0.07125 |
| **CpG_43_13_CpG_10** | -0.23556 | 0.03232 | 0.84838 | 0.04589 | 0.60854 | 0.04201 | 0.31399 | 0.03427 | 0.66978 | 0.06836 |
| **TCFL2_1_3_CpG_2** | 0.22138 | 0.03281 | 0.81686 | 0.03998 | 0.60686 | 0.04734 | 0.33997 | 0.03515 | 0.67078 | 0.06230 |
| **TCFL2_1_3_CpG_10** | 0.22050 | 0.03180 | 0.80677 | 0.04273 | 0.59743 | 0.03701 | 0.33186 | 0.03663 | 0.67040 | 0.06422 |
| **CpG_43_13_CpG_30** | 0.20596 | 0.03708 | 0.83193 | 0.04022 | 0.60218 | 0.04159 | 0.29620 | 0.03808 | 0.67770 | 0.06502 |
| **CpG_43_19_CpG_23** | -0.20169 | 0.03003 | 0.81607 | 0.04186 | 0.60742 | 0.02806 | 0.35236 | 0.03141 | 0.67988 | 0.06154 |
| **CpG_43_13_CpG_11.12** | -0.19895 | 0.03381 | 0.82987 | 0.04078 | 0.60241 | 0.03808 | 0.31781 | 0.02563 | 0.68765 | 0.06749 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **CpG_43_19_CpG_11.12** | -0.18810 | 0.02980 | 0.81927 | 0.04241 | 0.60226 | 0.03698 | 0.32600 | 0.02716 | 0.67765 | 0.06351 |
| **CpG_43_19_CpG_3.4** | -0.18077 | 0.03054 | 0.80834 | 0.04187 | 0.60568 | 0.03610 | 0.33689 | 0.03330 | 0.67222 | 0.06559 |
| **CpG_43_13_CpG_8** | -0.17668 | 0.03052 | 0.84234 | 0.04162 | 0.60043 | 0.03523 | 0.32022 | 0.03837 | 0.66386 | 0.06706 |
| **CpG_43_19_CpG_20** | -0.17015 | 0.03268 | 0.83970 | 0.03774 | 0.61375 | 0.03094 | 0.34883 | 0.02932 | 0.65545 | 0.06737 |
| **CpG_43_13_CpG_7** | -0.16922 | 0.03469 | 0.82340 | 0.03144 | 0.61012 | 0.03167 | 0.31668 | 0.02992 | 0.67554 | 0.06287 |
| **CpG_43_19_CpG_7** | -0.15616 | 0.03389 | 0.83195 | 0.04356 | 0.60539 | 0.04358 | 0.33907 | 0.04103 | 0.65902 | 0.06817 |
| **CpG_43_13_CpG_31** | -0.14690 | 0.03139 | 0.83253 | 0.03873 | 0.60358 | 0.04183 | 0.32999 | 0.03957 | 0.66807 | 0.06658 |
| **CpG_43_13_CpG_21** | -0.14555 | 0.02944 | 0.83552 | 0.04510 | 0.60331 | 0.03158 | 0.33107 | 0.03494 | 0.66856 | 0.06792 |
| **CpG_43_13_CpG_13** | 0.12611 | 0.04219 | 0.82937 | 0.03432 | 0.61772 | 0.03073 | 0.32764 | 0.03029 | 0.67540 | 0.06804 |
| **CpG_43_13_CpG_22.23** | -0.12603 | 0.04302 | 0.82489 | 0.03408 | 0.59060 | 0.03946 | 0.32296 | 0.03411 | 0.67021 | 0.06131 |
| **CpG_43_19_CpG_2** | -0.11296 | 0.04138 | 0.81233 | 0.04242 | 0.60693 | 0.02615 | 0.32401 | 0.03929 | 0.66248 | 0.06488 |
| **TCFL2_1_3_CpG_15** | 0.11002 | 0.03425 | 0.82840 | 0.03601 | 0.59032 | 0.03428 | 0.32364 | 0.03279 | 0.66646 | 0.06847 |
| **CpG_43_13_CpG_9** | -0.10988 | 0.04218 | 0.83561 | 0.04585 | 0.60837 | 0.04468 | 0.31667 | 0.03951 | 0.66708 | 0.06850 |
| **7903146_num** | 0.10653 | 0.02831 | 0.82604 | 0.04518 | 0.59966 | 0.03551 | 0.33278 | 0.03633 | 0.66465 | 0.06781 |
| **CpG_43_19_CpG_6** | -0.10473 | 0.03373 | 0.83081 | 0.04195 | 0.60075 | 0.03883 | 0.31151 | 0.03278 | 0.66458 | 0.06997 |
| **TCFL2_1_3_CpG_1** | 0.10343 | 0.03424 | 0.82765 | 0.04649 | 0.60495 | 0.03494 | 0.31609 | 0.03153 | 0.67322 | 0.06791 |
| **TCFL2_1_3_CpG_9** | -0.10021 | 0.03943 | 0.82525 | 0.03717 | 0.60931 | 0.03551 | 0.31384 | 0.03361 | 0.65887 | 0.06337 |
| **CpG_43_13_CpG_28** | 0.07721 | 0.04120 | 0.82623 | 0.04300 | 0.59951 | 0.04006 | 0.31827 | 0.03636 | 0.67354 | 0.07063 |
| **TCFL2_1_3_CpG_3.4** | -0.07619 | 0.02818 | 0.83253 | 0.03961 | 0.60132 | 0.03751 | 0.31550 | 0.03303 | 0.66476 | 0.06752 |
| **TCFL2_1_3_CpG_14** | -0.07124 | 0.03483 | 0.82644 | 0.04137 | 0.61132 | 0.04058 | 0.31953 | 0.03751 | 0.66800 | 0.06429 |
| **TCFL2_1_3_CpG_11** | -0.07043 | 0.02980 | 0.82356 | 0.04626 | 0.59808 | 0.04114 | 0.31192 | 0.03500 | 0.66877 | 0.06880 |
| **TCFL2_1_3_CpG_12** | -0.06850 | 0.03638 | 0.82573 | 0.03803 | 0.60678 | 0.04015 | 0.31797 | 0.03843 | 0.66709 | 0.06640 |
| **CpG_43_13_CpG_2** | -0.05141 | 0.02659 | 0.82158 | 0.03851 | 0.59728 | 0.03742 | 0.31433 | 0.03559 | 0.66318 | 0.06471 |
| **CpG_43_13_CpG_17** | 0.04904 | 0.03367 | 0.80994 | 0.03094 | 0.60251 | 0.03730 | 0.31412 | 0.03564 | 0.66796 | 0.06581 |
| **CpG_43_13_CpG_29** | 0.02459 | 0.03123 | 0.82793 | 0.03912 | 0.60780 | 0.03970 | 0.31177 | 0.03409 | 0.66629 | 0.06657 |
| **CpG_43_13_CpG_18.19** | -0.02002 | 0.04604 | 0.82443 | 0.03501 | 0.60714 | 0.03911 | 0.31779 | 0.03485 | 0.66981 | 0.06502 |
| **CpG_43_13_CpG_4** | -0.01899 | 0.03123 | 0.82115 | 0.04480 | 0.60548 | 0.04453 | 0.31835 | 0.04465 | 0.66649 | 0.06779 |
| **TCFL2_1_3_CpG_5** | -0.01743 | 0.03373 | 0.83462 | 0.04114 | 0.61421 | 0.03402 | 0.31878 | 0.03272 | 0.66652 | 0.06545 |

**Table 14 Means and standard deviations (SD) of the coefficients (coef) obtained from the 50 iterations for the first prediction model; mean and standard deviation of the regression coefficients relative to age, BMI and sex are also reported.**

It is worthy of note that age is always the most important covariate in explaining the diabetes group, followed by BMI and sex, in agreement with the Pearson's correlation results in Figure 2. Finally, the last two columns report the accuracy scores (again, mean and standard deviation of the 50 repetitions) of each logistic model. As explained in statistical analysis section, the accuracy score is equal to the number of samples that are correctly classified by the model divided by the total number of samples, i.e. the fraction of samples that are correctly classified. As shown in Table 13, the percentage of samples for which the model is able to predict the correct classification ranges from 66.7% to 67.0% according to the covariate taken into account.

The results obtained from the univariate logistic regressions were exploited to select a subset of covariates that were subsequently used to train a multivariate model. In particular, we sorted the CpG sites according to their univariate average coefficient (absolute value). Then, starting from the top of the list, we considered an increasing number of CpG sites to be included in the model and, for each subset of CpGs, we performed a double 10-fold cross-validated logistic regression with ridge penalization and we estimated the accuracy of the model in predicting T2D.

Each model also included sex, age, BMI and the genotype of rs7903146 and was re-computed 50 times to randomize over the *cross validation* folds choice. Then, for each model, the percentage of correctly classified samples over the 50 iterations was computed. Here, a sample was considered correctly classified if the model predicts its correct class at least 10% of the times (5 times over 50). The results are reported in figure 5.

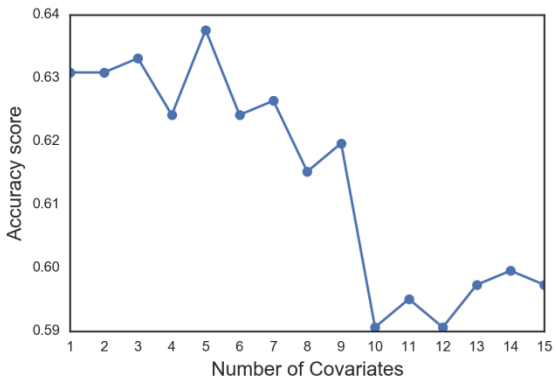


**Figure 4 Accuracy score based on number of covariates included**

As shown, the model with the highest accuracy score turned out to be the one with 5 covariates, in particular 5 CpG sites in the CpG Island in the intron 3 (CpG43, amp19 CpG sites: 13, 21.22, 18.19, 5 and 8.9).

Over the 50 iterations, 63% of samples were correctly classified at least 10% of the times, as shown in figure 6.

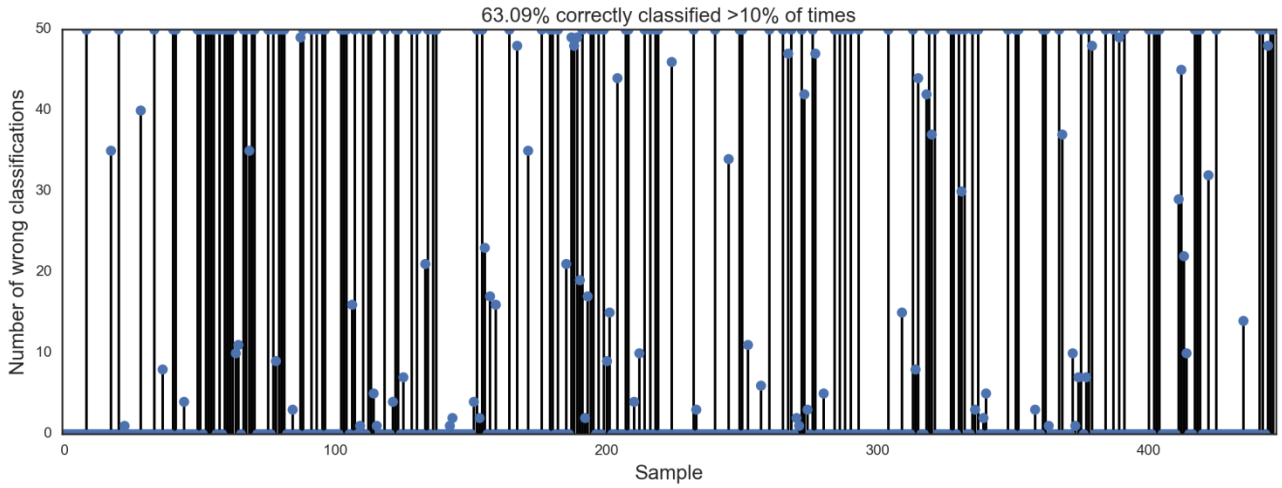


**Figure 5 Graphical view of prediction model results. In the figure are reported each sample (x-axis) and the number of each sample are wrongly classified (y-axis).**

Notice that some samples are always correctly classified, while others are never assigned to the right class. This means that the classification is constant but not precise, as the model is able to classify correctly only 63% of subjects.

Finally, we performed a Partial Least Squares (PLS) analysis, that is similar to Principal Component Analysis (PCA) but instead of finding the combinations of covariates that maximize the variance, it maximizes the covariance with the outcome variable (in our case T2D). Here, we used the following covariates: age, sex, BMI, genotype of rs7903146 and the 5 CpG sites that were selected according to our method. Results are shown in figure 7.

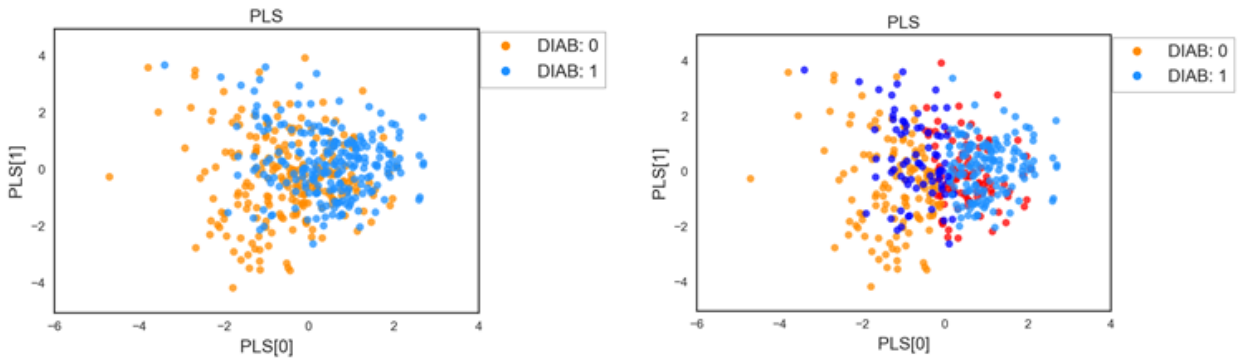


**Figure 6 PLS analysis of the results of the first prediction model. On the left graphic subjects are plotted based on two principle components, on the right one samples that are wrongly classified in Liner regression model are further identified by red (for controls) and blue (for T2D patients) colour.**

The figure on the left shows the patients (points) plotted in the space defined by the first 2 PLS components. As it can be appreciated, the two groups are largely overlapping. If we identify with darker colours (red for healthy controls, blue for T2D patients) the samples that were wrongly classified by the Logistic regression model, we see that PLS confirms the result: healthy controls that were wrongly classified by the logistic model overlap with T2D patients and *vice versa*.

***Disease Prediction Model 2***

The second approach that we applied was based on the development of a predictive model for the biological age of an individual based on DNA methylation measurements.

At variance with the first model, the regions that have been selected resulted to be associated with biological age in previous studies. In particular, we took into account at least 4 sets of CpG sites whose level of methylation is age-dependent: a) the CpG sites included in Horvath's clock (Horvath 2013); b) the CpG sites included in Hannum's clock (Hannum et al. 2013); c) the CpG sites included in a recent meta-analysis that we performed to identify age-associated DNA methylation changes in blood (Bacalini et al. 2015); d) the CpG sites which show an increase in inter-individual variability in DNA methylation with age, according to a recent study (Slieker et al. 2016). According to the correlation with chronological age and to the association with age-related conditions results obtained from this investigation, three target loci have been chosen: ELOVL2, CA3 and LYPD5. We performed a design specific EpiTYPER assays onto the surrounding CpG sites not assessed by the Infinium platform, in order to evaluate the DNA methylation status of the above identified genes that allows investigation of 16 CpG sites on promoter region of ELOVL2 gene, 6 CpG sites on the promoter region of CA3 gene and 10 on the 3' untranslated region of LYPD5 gene.

This analysis was performed on 298 healthy controls from 0 to 100 years, 24 centenarians, 24 centenarians' offspring, 16 subjects with Down Syndrome and 181 T2D patients. DNA methylation

levels of 32 CpG sites were assessed using EpiTYPER technology (Agena Bio, San Diego, CA, USA). The regions have been selected as described in experimental procedures.

Epigenetic Age (EpiAge) was estimated performing a Linear Regression on the control group based on the 32 selected CpGs. EpiAge of the remaining groups was predicted using the coefficients of the model trained on controls. Results are shown in figure 8.

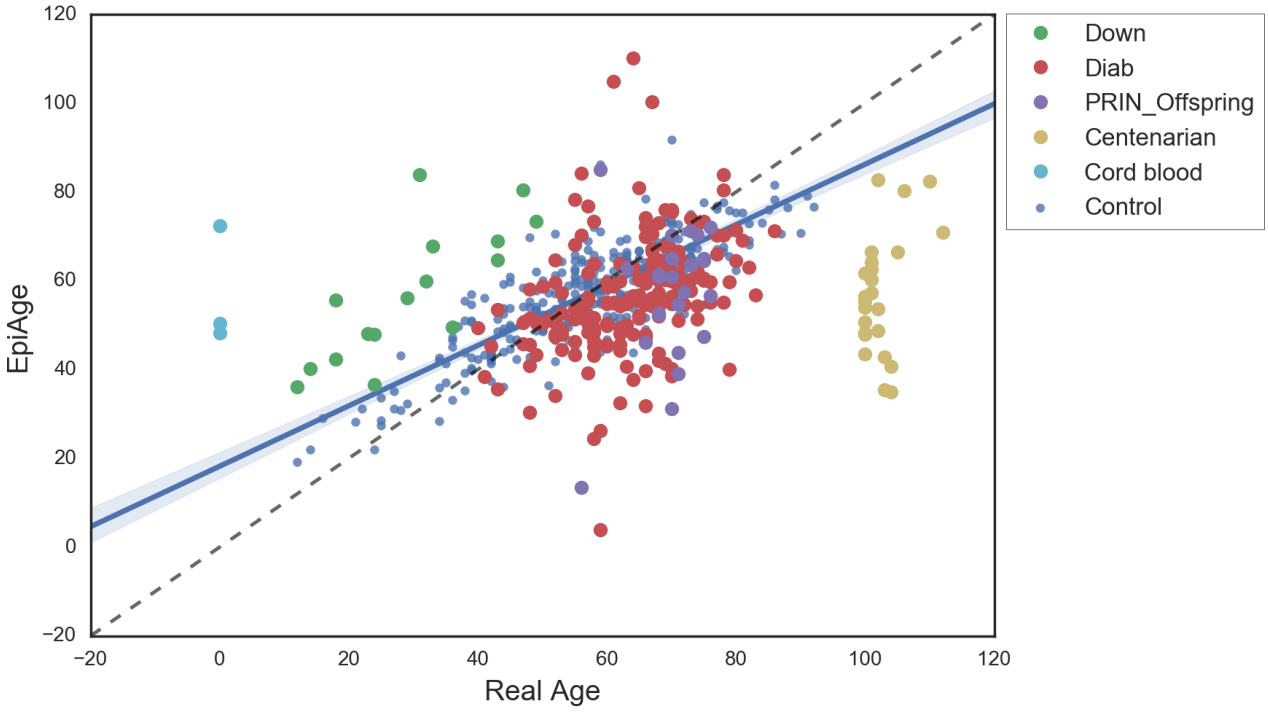


**Figure 7 Scatter plot of the EpiAge of subjects as a function of their Real Age. Colours refer to different groups according to the figure legend. The dashed line corresponds to the bisector (EpiAge = Real Age). The blue line represent the regression line obtained for the healthy control group and the shadow around it indicates the 95% confidence interval.**

The slope and $r^2$ value for each analysed group are reported in table 14.

| Group | Slope value | $r^2$ value |
|---|---|---|
| Control | 0.68 | 0.83 |
| Centenarians | 1.73 | 0.41 |
| Centenarians' Offspring | 0.86 | 0.29 |
| Cord blood | na | 0.0 |
| Down Syndrome | 1.01 | 0.78 |
| T2D | 0.54 | 0.35 |

**Table 15 Slope and $r^2$ values resulted from linear regression model are reported for each sample group considered.**

As reported in the scatter plot, the healthy control group spreads around the central line showing a high correlation between chronological age (real age) and epigenetic predicted age

(EpiAge). The centenarians group and the majority of their offspring showed a deceleration of their EpiAge. In contrast, subjects with Down Syndrome have shown an acceleration of EpiAge. Quite surprisingly, the group of T2D patients did not display an acceleration of EpiAge, on the contrary some T2D patients showed a deceleration of EpiAge.

**Summary**: The first prediction model based on linear regression analysis was developed using one of the strongest diabetes related variant (rs7903146 of *TCF7L2* gene), phenotypic trait (BMI, sex, age) and methylation levels of 5 CpG sites in intron 3 region of TCF7L2 gene. Although 63% of samples were correctly classified as healthy controls or T2D patients, the accuracy of our model is not satisfactory and needs to be improved. As reported in PLS analysis, in fact, the two groups are largely overlapped. These results suggesting that other variables (phenotypic trait, biochemistry parameters and clinical informations) are requested in order to improve the accuracy of the model. Another important aspect with likely impact in the model prediction performance is the high phenotypic heterogeneity involved in T2D. To this regard, Wen and Liu demonstrated that the use of more refined subphenotypes facilitates the identification of new predictors and leads to improved risk prediction models (Wen and Lu 2013).

The second prediction model was based on the estimation of age derived from epigenetic levels of particular CpG sites on the genome. Taking a lesson from Epigenetic Clock of Horvath's, the most popular epigenetic age (EpiAge) estimator reported so far (Horvath 2013), we developed a "epigenetic little clock" based on methylation values of only 32 CpG sites belonging to three genes (ELOVL2, CA3 and LYPD5). As many studies have been reported a modification of EpiAge in presence of age-related diseases, cancer or Down Syndrome (Levine et al. 2015) (Horvath and Ritz 2015) we decided to evaluate epigenetic clock modification in our T2D samples. Here we applied our "epigenetic little clock" for biological age estimation in diabetes patients, in order to evaluate a possible acceleration of epigenetic age in T2D patients due to the disease. Surprisingly, T2D patients did not show an acceleration of epigenetic age, conversely, some of them seem to have a deceleration of EpiAge. A possible explanation for these results might be the interaction with environmental factors or pharmaceutical therapies that might affect the methylation status.

# 5 Conclusions

Unravelling the etiopathology of Type 2 diabetes (T2D) is a major challenge due to the extreme heterogeneity of the disease. Actually, T2D is a common complex and multifactorial disease caused by the interaction of genetics, epigenetics and environmental factors; studies that evaluate the combined effect of these factors represent a novel field of research for this complex disorder.

In the present thesis we present an integrated genetic and epigenetic analysis of a cohort of Italian subjects. First, we performed a GWAS of 1352 individuals (cases=282 and controls=1070), following the approach of the "extreme phenotypes" that we recently suggested in the study of age-related diseases, in order to increase sensitivity and to identify the variants with a possible biological function. In the first comparison between 282 T2D patients with at least one vascular complication and 737 healthy controls we found 73 associated variants with nominal p values $<10^{-5}$ in the logistic regression analysis. Then, we focus on those genes that carry at least two significant SNPs ($p<10^{-4}$); by this analysis we identified 5 genes with multiple associated signals: *MS4A14* and *THSD4* genes, involved in transduction signal and post transcriptional gene expression, respectively, that had never been associated to T2D. Conversely, *DNHD1* gene, that regulates gene expression by binding to mRNA at level of microRNA target sequences, has been related of severe proliferative diabetic retinopathy (Ung, 2017). Finally, two genes that encode for GTPase activity proteins have been found in association with multiple signals in our study: *ELMOD1* gene, previously associated with low blood glucose levels and low body weight in mice (de Angelis et al. 2015) and *DLC1* gene, recently identifies as a modulator of the plasminogen activator inhibitor-1 (*PAI-1*) involved in several pathologies such as renal injury, inflammation, obesity and also diabetes (Shih, Takada, and Lo 2012).

When we considered 282 T2D patients with at least one vascular complication with 334 centenarians, we found 38 associated variants with nominal p value $<10^{-5}$. Moreover, we selected those genes with more than one associated SNP with $p<10^{-4}$,. We identified two novel signals, *APBB2* gene, that encodes for a protein involved in transduction signal, and *MGLL*, which participates in monoacylglycerides and endocannabinoids metabolism. Moreover, we identified two genes previously reported associated with the disease: *FAM13A*, found in association with fasting insulin in two studies (Scott et al. 2012;) (Yaghootkar et al. 2014) and *TCF7L2*, the strongest locus related to T2D to date and previously described by our group (Garagnani et al. 2013).

Then, we identified the number of associated variants that were present in the two comparisons described above (T2D vs ctrl and T2D vs centenarians). This analysis revealed a total of 35 common variants that we hypothesized to be the variants with a high biological value.

As illustrated above, we focused on variants showing multiple associations among those having p value $<10^{-4}$. Then we identified three genes shared between the two comparisons: *TMEM108*, that codes for a transmembrane protein involved in neuronal activities, *UPP2* gene, involved in purine metabolism and once again *TCF7L2* gene. Numerous evidences have been reported for *TMEM108* gene as related with BMI (Locke et al. 2015) (Graff et al. 2017) (Justice et al. 2017) (Rowlands et al. 2014); obese mice (Mollah and Ishikawa 2010) and animal fed with high fat diet (Eckel-Mahan et al. 2013) have showed a decrease in the expression levels of *UPP2* gene. Finally, *TCF7L2* gene was confirmed as one of the most important predictors for T2D genetic risk. In conclusion the analysis of extreme phenotypes allows us to describe significant associations even if the number of our samples is not very high. This suggests that the cohorts identified and the use of centenarians as group of "supercontrols" maximize the differences observed from a biological point of view.

Since TCF7L2 is considered the master gene for T2D, we conducted a gene-candidate methylation analysis of this gene, (Mayans et al. 2007 Sladek et al. 2007; van Vliet-Ostaptchouk et al. 2007; Stitzel et al. 2010). The analysis was performed considering 229 T2D patients and 219 age and sex matched controls, on the promoter region, previously associated with T2D in a Spanish population (Canivell et al. 2014), as well as in two CpG Islands (CpGI) in intron 3 and in 3' UTR of *TCF7L2* gene. The analysis of variance (ANOVA) performed on 70 CpG sites has shown two CpG sites in intronic CpGI having different methylation values between T2D patients and healthy controls. Subsequently, we performed an integrated analysis combining genotypes of *TCF7L2* rs7903146 and DNA methylation of the surrounding regions. The ANOVA analysis conducted on 404 samples grouped based on the genotypes of rs7903146 has shown a different methylation level in three groups. In conclusion we suggested that a DNA methylation variability exists that separates T2D patients from controls, and that *TCF7L2* rs7903146 influences methylation profile of the gene independently from phenotype.

Finally, we developed and tested two different disease predictor models considering some of these genetic and epigenetic variables, in order to check the possibility to differentiate T2D patients group from the controls group, and eventually to identify persons at risk of developing T2D before the onset of the disease.

The first method was based on the computation of a logistic regression model on the following parameters: disease diagnosis, sex, age, BMI, genotypes of rs7903146 and methylation levels of 5 CpG sites in *TCF7L2* gene in 229 diabetic subjects and 219 controls.

This model was able to correctly classify 63% of subjects, resulting mildly accurate and for this reason it will be integrated with more parameters in order to improve its accuracy.

The second approach was based on a predictive model for the biological age of an individual based on DNA methylation measurements. We developed a "epigenetic little clock" trough the methylation status of 32 CpG sites of three genes (ELOVL2, CA3 and LYPD5). The analysis was performed on 298 healthy subjects (age range=0-99 years), 24 centenarians, 24 centenarians' offspring, 16 subjects with Down Syndrome and 181 T2D patients. The centenarians group and the majority of their offspring showed a deceleration of their epigenetic age, conversely, subjects with Down Syndrome showed an acceleration of epigenetic age. Unexpectedly, this analysis has shown that diabetic patients did not show an acceleration of epigenetic age. These results might be influenced by environmental factors, such as physical activity or diet, or drugs therapies that might modify methylation status. Recently evidence shown that metformin, one of the most used anti-diabetic drug, have an anti-aging effect (Pryor and Cabreiro 2015; Podhorecka, Ibanez, and Dmoszyńska 2017).

In conclusion, in here described GWAS analysis, including 1389 Italians individuals, we identified several loci previously described in association with T2D and some novel genes possibly involved in the disease identified thanks to the extreme phenotype approach. Remarkably, the "extreme phenotypes" approach gives more value of our study and it could be a valid tool in the study of complex age-related disease allowing the identification of variants with biological significance. The gene candidate methylation study indicates an association of methylation for *TCF7L2* gene and T2D and moreover it demonstrates a straight interaction between genotypes and methylation, underling the need to perform more comprehensive studies for better improving a disease risk evaluation. Linear regression model for prediction of the disease has shown a weak accuracy and it will be implemented with other parameters. Finally, T2D patients did not show an acceleration of epigenetic age in our samples, further investigations are needed in order to evaluate how the environment, drugs or lifestyle can influence them.

# 6 Bibliography


Abbott, Alison
2004 Ageing: Growing Old Gracefully. Nature 428(6979): 116–118.

Achilli, Alessandro, Anna Olivieri, Maria Pala, et al.
2011 Mitochondrial DNA Backgrounds Might Modulate Diabetes Complications Rather than T2DM as a Whole. PloS One 6(6): e21029.

Akiyama, Masato, Yukinori Okada, Masahiro Kanai, et al.
2017 Genome-Wide Association Study Identifies 112 New Loci for Body Mass Index in the Japanese Population. Nature Genetics 49(10): 1458–1467.

Ali, Omar
2013 Genetics of Type 2 Diabetes. World Journal of Diabetes 4(4): 114–123.

Anderson, Carl A., Fredrik H Pettersson, Geraldine M Clarke, et al.
2010 Data Quality Control in Genetic Case-Control Association Studies. Nature Protocols 5(9): 1564–1573.

de Angelis, Martin Hrabě, George Nicholson, Mohammed Selloum, et al.
2015 Analysis of Mammalian Gene Function through Broad-Based Phenotypic Screens across a Consortium of Mouse Clinics. Nature Genetics 47(9): 969–978.

Bacalini, Maria Giulia, Alessio Boattini, Davide Gentilini, et al.
2015 A Meta-Analysis on Age-Associated Changes in Blood DNA Methylation: Results from an Original Analysis Pipeline for Infinium 450k Data. Aging 7(2): 97–109.

Bell, Christopher G., Sarah Finer, Cecilia M. Lindgren, et al.
2010 Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus. PloS One 5(11): e14040.

Beveridge, Natalie J., and Murray J. Cairns
2012 MicroRNA Dysregulation in Schizophrenia. Neurobiology of Disease 46(2): 263–271.

Bonafè, Massimiliano, and Fabiola Olivieri
2009 Genetic Polymorphism in Long-Lived People: Cues for the Presence of an Insulin/IGF-

Pathway-Dependent Network Affecting Human Longevity. Molecular and Cellular Endocrinology 299(1): 118–123.

Candore, Giuseppina, Carmela R. Balistreri, Florinda Listì, et al.
2006 Immunogenetics, Gender, and Longevity. Annals of the New York Academy of Sciences 1089: 516–537.

Candore, Giuseppina, Calogero Caruso, and Giuseppina Colonna-Romano
2010 Inflammation, Genetic Background and Longevity. Biogerontology 11(5): 565–573.

Canivell, Silvia, Elena G. Ruano, Antoni Sisó-Almirall, et al.
2014 Differential Methylation of TCF7L2 Promoter in Peripheral Blood DNA in Newly Diagnosed, Drug-Naïve Patients with Type 2 Diabetes. PLoS ONE 9(6). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4051650/, accessed November 21, 2017.

Carrieri, G., M. Bonafè, M. De Luca, et al.
2001 Mitochondrial DNA Haplogroups and APOE4 Allele Are Non-Independent Variables in Sporadic Alzheimer's Disease. Human Genetics 108(3): 194–198.

Cevenini, E., L. Invidia, F. Lescai, et al.
2008 Human Models of Aging and Longevity. Expert Opinion on Biological Therapy 8(9): 1393–1405.

Chambers, John C., Marie Loh, Benjamin Lehne, et al.
2015 Epigenome-Wide Association of DNA Methylation Markers in Peripheral Blood from Indian Asians and Europeans with Incident Type 2 Diabetes: A Nested Case-Control Study. The Lancet. Diabetes & Endocrinology 3(7): 526–534.

Dabelea, D., R. L. Hanson, R. S. Lindsay, et al.
2000 Intrauterine Exposure to Diabetes Conveys Risks for Type 2 Diabetes and Obesity: A Study of Discordant Sibships. Diabetes 49(12): 2208–2211.

Danaei, Goodarz, Mariel M. Finucane, Yuan Lu, et al.
2011 National, Regional, and Global Trends in Fasting Plasma Glucose and Diabetes Prevalence since 1980: Systematic Analysis of Health Examination Surveys and Epidemiological Studies with 370 Country-Years and 2·7 Million Participants. Lancet (London, England) 378(9785): 31–40.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Richa Saxena, Benjamin F. Voight, et al.
2007 Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science (New York, N.Y.) 316(5829): 1331–1336.

Duggirala, R., J. Blangero, L. Almasy, et al.

1999 Linkage of Type 2 Diabetes Mellitus and of Age at Onset to a Genetic Location on Chromosome 10q in Mexican Americans. American Journal of Human Genetics 64(4): 1127–1140.

Eckel-Mahan, Kristin L., Vishal R. Patel, Sara de Mateo, et al.
2013 Reprogramming of the Circadian Clock by Nutritional Challenge. Cell 155(7): 1464–1478.

Florath, Ines, Katja Butterbach, Jonathan Heiss, et al.
2016 Type 2 Diabetes and Leucocyte DNA Methylation: An Epigenome-Wide Association Study in over 1,500 Older Adults. Diabetologia 59(1): 130–138.

Fontana, Luigi, and Samuel Klein
2007 Aging, Adiposity, and Calorie Restriction. JAMA 297(9): 986–994.

Franceschi, C., and M. Bonafè
2003 Centenarians as a Model for Healthy Aging. Biochemical Society Transactions 31(2): 457–461.

Franceschi, C., L. Motta, S. Valensin, et al.
2000 Do Men and Women Follow Different Trajectories to Reach Extreme Longevity? Italian Multicenter Study on Centenarians (IMUSCE). Aging (Milan, Italy) 12(2): 77–84.

Franceschi, Claudio, Luciano Motta, Massimo Motta, et al.
2008 The Extreme Longevity: The State of the Art in Italy. Experimental Gerontology 43(2): 45–52.

Franceschi, Claudio, Fabiola Olivieri, Francesca Marchegiani, et al.
2005 Genes Involved in Immune Response/Inflammation, IGF1/Insulin Pathway and Response to Oxidative Stress Play a Major Role in the Genetics of Human Longevity: The Lesson of Centenarians. Mechanisms of Ageing and Development 126(2): 351–361.

Garagnani, Paolo, Cristina Giuliani, Chiara Pirazzini, et al.
2013 Centenarians as Super-Controls to Assess the Biological Relevance of Genetic Risk Factors for Common Age-Related Diseases: A Proof of Principle on Type 2 Diabetes. Aging 5(5): 373–385.

Gaulton, Kyle J., Cristen J. Willer, Yun Li, et al.
2008 Comprehensive Association Study of Type 2 Diabetes and Related Quantitative Traits with 222 Candidate Genes. Diabetes 57(11): 3136–3144.

Ghosh, Asish K., and Douglas E. Vaughan
2012 PAI-1 in Tissue Fibrosis. Journal of Cellular Physiology 227(2): 493–507.

Giuliani, Cristina, Chiara Pirazzini, Massimo Delledonne, et al.

2017 Centenarians as Extreme Phenotypes: An Ecological Perspective to Get Insight into the Relationship between the Genetics of Longevity and Age-Associated Diseases. Mechanisms of Ageing and Development 165(Pt B): 195–201.

Graff, Mariaelisa, Robert A. Scott, Anne E. Justice, et al.
2017 Genome-Wide Physical Activity Interactions in Adiposity - A Meta-Analysis of 200,452 Adults. PLoS Genetics 13(4): e1006528.

Grant, Julia D., Arpana Agrawal, Kathleen K. Bucholz, et al.
2009 Alcohol Consumption Indices of Genetic Risk for Alcohol Dependence. Biological Psychiatry 66(8): 795–800.

Hales, C. N., and D. J. Barker
1992 Type 2 (Non-Insulin-Dependent) Diabetes Mellitus: The Thrifty Phenotype Hypothesis. Diabetologia 35(7): 595–601.

Hall, Elin, Petr Volkov, Tasnim Dayeh, et al.
2014 Effects of Palmitate on Genome-Wide MRNA Expression and DNA Methylation Patterns in Human Pancreatic Islets. BMC Medicine 12: 103.

Hanis, C. L., E. Boerwinkle, R. Chakraborty, et al.
1996 A Genome-Wide Search for Human Non-Insulin-Dependent (Type 2) Diabetes Genes Reveals a Major Susceptibility Locus on Chromosome 2. Nature Genetics 13(2): 161–166.

Hannum, Gregory, Justin Guinney, Ling Zhao, et al.
2013 Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell 49(2): 359–367.

Heath, Andrew C., John B. Whitfield, Nicholas G. Martin, et al.
2011 A Quantitative-Trait Genome-Wide Association Study of Alcoholism Risk in the Community: Findings and Implications. Biological Psychiatry 70(6): 513–518.

Herskind, A. M., M. McGue, N. V. Holm, et al.
1996 The Heritability of Human Longevity: A Population-Based Study of 2872 Danish Twin Pairs Born 1870-1900. Human Genetics 97(3): 319–323.

vB Hjelmborg, Jacob, Ivan Iachine, Axel Skytthe, et al.
2006 Genetic Influence on Human Lifespan and Longevity. Human Genetics 119(3): 312–321.

Horvath, Steve
2013 DNA Methylation Age of Human Tissues and Cell Types. Genome Biology 14(10): R115.

Horvath, Steve, and Beate R. Ritz
2015 Increased Epigenetic Age and Granulocyte Counts in the Blood of Parkinson's Disease Patients. Aging 7(12): 1130–1142.

Jin, Tianru, and Ling Liu
2008 The Wnt Signaling Pathway Effector TCF7L2 and Type 2 Diabetes Mellitus. Molecular Endocrinology (Baltimore, Md.) 22(11): 2383–2392.

Justice, Anne E., Thomas W. Winkler, Mary F. Feitosa, et al.
2017 Genome-Wide Meta-Analysis of 241,258 Adults Accounting for Smoking Behaviour Identifies Novel Loci for Obesity Traits. Nature Communications 8: 14977.

Kulkarni, Hemant, Mark Z. Kos, Jennifer Neary, et al.
2015 Novel Epigenetic Determinants of Type 2 Diabetes in Mexican-American Families. Human Molecular Genetics 24(18): 5330–5344.

Lee, Yun Yong, Kyong Soo Park, Youngmi Kim Pak, and Hong Kyu Lee
2005 The Role of Mitochondrial DNA in the Development of Type 2 Diabetes Caused by Fetal Malnutrition. The Journal of Nutritional Biochemistry 16(4): 195–204.

Levine, Morgan E., Ake T. Lu, David A. Bennett, and Steve Horvath
2015 Epigenetic Age of the Pre-Frontal Cortex Is Associated with Neuritic Plaques, Amyloid Load, and Alzheimer's Disease Related Cognitive Functioning. Aging 7(12): 1198–1211.

Liu, Jun, Sabina Semiz, Sven J. van der Lee, et al.
2017 Metabolomics Based Markers Predict Type 2 Diabetes in a 14-Year Follow-up Study. Metabolomics: Official Journal of the Metabolomic Society 13(9): 104.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, et al.
2015 Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology. Nature 518(7538): 197–206.

Lyssenko, Valeriya, Roberto Lupi, Piero Marchetti, et al.
2007 Mechanisms by Which Common Variants in the TCF7L2 Gene Increase Risk of Type 2 Diabetes. Journal of Clinical Investigation 117(8): 2155–2163.

Mayans, Sofia, Kurt Lackovic, Petter Lindgren, et al.
2007 TCF7L2 Polymorphisms Are Associated with Type 2 Diabetes in Northern Sweden. European Journal of Human Genetics: EJHG 15(3): 342–346.

Meyre, David, Jérôme Delplanque, Jean-Claude Chèvre, et al.

2009 Genome-Wide Association Study for Early-Onset and Morbid Adult Obesity Identifies Three New Risk Loci in European Populations. Nature Genetics 41(2): 157–159.

Mollah, Md Bazlur R, and Akira Ishikawa
2010 A Wild Derived Quantitative Trait Locus on Mouse Chromosome 2 Prevents Obesity. BMC Genetics 11: 84.

Pal, A., and M. I. McCarthy
2013 The Genetics of Type 2 Diabetes and Its Clinical Relevance. Clinical Genetics 83(4): 297–306.

Paolisso, G., A. Gambardella, S. Ammendola, et al.
1996 Glucose Tolerance and Insulin Action in Healty Centenarians. The American Journal of Physiology 270(5 Pt 1): E890-894.

Podhorecka, Monika, Blanca Ibanez, and Anna Dmoszyńska
2017 Metformin - Its Potential Anti-Cancer and Anti-Aging Effects. Postepy Higieny I Medycyny Doswiadczalnej (Online) 71(0): 170–175.

Prokopenko, Inga, Mark I. McCarthy, and Cecilia M. Lindgren
2008 Type 2 Diabetes: New Genes, New Understanding. Trends in Genetics: TIG 24(12): 613–621.

Pryor, Rosina, and Filipe Cabreiro
2015 Repurposing Metformin: An Old Drug with New Tricks in Its Binding Pockets. The Biochemical Journal 471(3): 307–322.

Roth, George S., Julie A. Mattison, Mary Ann Ottinger, et al.
2004 Aging in Rhesus Monkeys: Relevance to Human Health Interventions. Science (New York, N.Y.) 305(5689): 1423–1426.

Rowlands, David S., Rachel A. Page, William R. Sukala, et al.
2014 Multi-Omic Integrated Networks Connect DNA Methylation and MiRNA with Skeletal Muscle Plasticity to Chronic Exercise in Type 2 Diabetic Obesity. Physiological Genomics 46(20): 747–765.

Salvioli, S., F. Olivieri, F. Marchegiani, et al.
2006 Genes, Ageing and Longevity in Humans: Problems, Advantages and Perspectives. Free Radical Research 40(12): 1303–1323.

Schinner, S.
2009 Wnt-Signalling and the Metabolic Syndrome. Hormone and Metabolic Research = Hormon-Und Stoffwechselforschung = Hormones Et Metabolisme 41(2): 159–163.

Scott, Laura J., Karen L. Mohlke, Lori L. Bonnycastle, et al.
2007 A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. Science (New York, N.Y.) 316(5829): 1341–1345.

Scott, Robert A., Vasiliki Lagou, Ryan P. Welch, et al.
2012 Large-Scale Association Analyses Identify New Loci Influencing Glycemic Traits and Provide Insight into the Underlying Biological Pathways. Nature Genetics 44(9): 991–1005.

Seki, Yoshinori, Lyda Williams, Patricia M. Vuguin, and Maureen J. Charron
2012 Minireview: Epigenetic Programming of Diabetes and Obesity: Animal Models. Endocrinology 153(3): 1031–1038.

Shah, Viral N., Balneek Singh Cheema, Rajni Sharma, et al.
2013 ACACβ Gene (Rs2268388) and AGTR1 Gene (Rs5186) Polymorphism and the Risk of Nephropathy in Asian Indian Patients with Type 2 Diabetes. Molecular and Cellular Biochemistry 372(1–2): 191–198.

Shigemizu, Daichi, Testuo Abe, Takashi Morizono, et al.
2014 The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort. PloS One 9(3): e92549.

Shih, Yi-Ping, Yoshikazu Takada, and Su Hao Lo
2012 Silencing of DLC1 Upregulates PAI-1 Expression and Reduces Migration in Normal Prostate Cells. Molecular Cancer Research: MCR 10(1): 34–39.

Skinner, Michael K.
2011 Environmental Epigenetic Transgenerational Inheritance and Somatic Epigenetic Mitotic Stability. Epigenetics 6(7): 838–842.

Sladek, Robert, Ghislain Rocheleau, Johan Rung, et al.
2007 A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes. Nature 445(7130): 881–885.

Slieker, Roderick C., Maarten van Iterson, René Luijk, et al.
2016 Age-Related Accrual of Methylomic Variability Is Linked to Fundamental Ageing Mechanisms. Genome Biology 17(1): 191.

Speliotes, Elizabeth K., Cristen J. Willer, Sonja I. Berndt, et al.
2010 Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index. Nature Genetics 42(11): 937–948.

Stern, M. P., P. A. Morales, R. A. Valdez, et al.
1993 Predicting Diabetes. Moving beyond Impaired Glucose Tolerance. Diabetes 42(5): 706–714.

Stitzel, Michael L., Praveen Sethupathy, Daniel S. Pearson, et al.
2010 Global Epigenomic Analysis of Primary Human Pancreatic Islets Provides Insights into Type 2 Diabetes Susceptibility Loci. Cell Metabolism 12(5): 443–455.

Toperoff, Gidon, Dvir Aran, Jeremy D. Kark, et al.
2012 Genome-Wide Survey Reveals Predisposing Diabetes Type 2-Related DNA Methylation Variations in Human Peripheral Blood. Human Molecular Genetics 21(2): 371–383.

Ung, Cindy, Angie V. Sanchez, Lishuang Shen, et al.
2017 Whole Exome Sequencing Identification of Novel Candidate Genes in Patients with Proliferative Diabetic Retinopathy. Vision Research.

Vaiserman, Alexander M.
2017 Early-Life Nutritional Programming of Type 2 Diabetes: Experimental and Quasi-Experimental Evidence. Nutrients 9(3). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5372899/, accessed December 5, 2017.

van Vliet-Ostaptchouk, J. V., R. Shiri-Sverdlov, A. Zhernakova, et al.
2007 Association of Variants of Transcription Factor 7-like 2 (TCF7L2) with Susceptibility to Type 2 Diabetes in the Dutch Breda Cohort. Diabetologia 50(1): 59–62.

Voight, Benjamin F., Laura J. Scott, Valgerdur Steinthorsdottir, et al.
2010 Twelve Type 2 Diabetes Susceptibility Loci Identified through Large-Scale Association Analysis. Nature Genetics 42(7): 579–589.

Wen, Yalu, and Qing Lu
2013 A Multiclass Likelihood Ratio Approach for Genetic Risk Prediction Allowing for Phenotypic Heterogeneity. Genetic Epidemiology 37(7): 715–725.

Yaghootkar, Hanieh, Robert A. Scott, Charles C. White, et al.
2014 Genetic Evidence for a Normal-Weight “Metabolically Obese” Phenotype Linking Insulin Resistance, Hypertension, Coronary Artery Disease, and Type 2 Diabetes. Diabetes 63(12): 4369–4377.

Yuan, Wei, Yudong Xia, Christopher G. Bell, et al.
2014 An Integrated Epigenomic Analysis for Type 2 Diabetes Susceptibility Loci in Monozygotic Twins. Nature Communications 5: 5719.

Zeggini, E.
2007 A New Era for Type 2 Diabetes Genetics. Diabetic Medicine 24(11): 1181–1186.

Zeggini, Eleftheria, Laura J. Scott, Richa Saxena, et al.
2008 Meta-Analysis of Genome-Wide Association Data and Large-Scale Replication Identifies Additional Susceptibility Loci for Type 2 Diabetes. Nature Genetics 40(5): 638–645.

# 7 Figures Legend



# 8 Tables Legend