UNIVERSITY OF VERONA

DEPARTMENT OF NEUROSCIENCES, BIOMEDICINE AND MOVEMENT SCIENCES

Graduate School of Health and Life Sciences

PhD program in Applied Health and Life Sciences

Cycle XXXIII

GENOME-WIDE DNA METHYLATION PROFILING OF OBESE INSULIN RESISTANT CHILDREN

S.S.D. MED/03

Coordinator:	Prof. Giovanni Malerba
Tutor:	Prof. Giovanni Malerba

PhD candidate: Lucas Moron Dalla Tor

SOMMARIO (ITA)

Introduzione: L'insulino resistenza si presenta quando la risposta delle cellule all'insulina è diminuita causando un drammatico innalzamento dei livelli di zucchero nel sangue. I diversi fattori di rischio per l'insulino resistenza includono uno stile di vita sedentario, obesità, storia familiare di diabete e invecchiamento. Negli ultimi anni, il diabete di tipo 2, l'insulino resistenza e l'obesità sono considerevolmente aumentate nella popolazione contribuendo all'incremento in morbidità e mortalità nel mondo. I molti meccanismi proposti per spiegare il funzionamento dell'insulino resistenza includono varianti genetiche, deregolazioni trascrittomiche e modificazioni epigenetiche, come per esempio la metilazione del DNA.

Scopo: L'obiettivo di questa tesi comprende l'utilizzo di metodologie bioinformatiche applicate allo studio della metilazione del DNA lungo tutto il genoma usando l'Infinium Human Methylation EPIC array (~850k CpGs) per studiare la componente epigenetica dell'insulino resistenza in una coorte di 186 soggetti pediatrici obesi, uniformemente divisi in due gruppi (insulino resistenti/insulino sensibili).

Risultati: L'analisi bioinformatica della metilazione a livello genomico, suggerisce una forte modulazione della composizione in termini di tipi cellulari nei soggetti insulino resistenti, suggerendo un possibile ruolo dell'infiammazione nella malattia. Inoltre, l'analisi della metilazione differenziale su singoli CpG o regioni, accompagnata da un'analisi di "gene set enrichment", mette in evidenza diverse vie collegate al metabolismo di carboidrati e grassi. In aggiunta, associando i probes differenzialmente metilati con risultati di studi riportati in letteratura, emergono ulteriori fattori che si potrebbero considerare durante lo studio di questa condizione.

Conclusioni: In conclusione, abbiamo utilizzato diversi approcci bioinformatici applicandoli ad una numerosa coorte di individui per studiare la metilazione del DNA a livello genomico nel contesto dell'insulino resistenza, con risultati che supportano l'ipotesi che la metilazione sia più legata a cambiamenti globali piuttosto che cambiamenti localizzati in pochi loci.

ABSTRACT (ENG)

Background: Insulin resistance occurs when the response of cells to insulin is decreased hence causing blood sugar levels to rise dramatically. Among others, the most common risk factors for insulin resistance include sedentary lifestyle, obesity, family history of diabetes and advanced age. In the last few decades, type 2 diabetes, insulin resistance, and obesity have increased dramatically in the general population contributing to an increase in morbidity and mortality around the world. Among others, the main mechanisms proposed for the action of insulin resistance include genetic variants, transcriptomic dysregulations, and epigenetic changes such as DNA methylation.

Aim: The aim of this thesis is to employ bioinformatic methods to genome-wide DNA methylation using the Infinium Human Methylation EPIC array (~850k CpGs) to study the epigenetic component of insulin resistance in a cohort of 186 obese pediatric individuals equally divided in two groups (insulin resistant/insulin sensitive).

Results: Bioinformatic analysis of the genome-wide methylation data, suggests a strong modulation of cell type composition in insulin resistant subjects proposing a role of inflammation in this disease. Furthermore, differential methylation of single CpG or regions, coupled with gene set enrichment analysis highlighted several pathways involved with carbohydrates and fat metabolism.

Additionally, associating differentially methylated probes with previously reported studies highlights additional factors that may be useful to consider when studying this condition.

Conclusions: In conclusion, we employed different bioinformatics strategies applied to a large cohort of individuals to study genome-wide DNA methylation in IR, with results supporting the hypothesis that methylation is more related to general methylation landscape changes rather than methylation variations in a few loci.

INTRODUCTION

Insulin resistance and type 2 diabetes

Insulin is a hormone produced by beta cells in the pancreas and, in healthy subjects is released in response to rising levels of sugar in the bloodstream due to food intake, acting like a key and allowing glucose to enter cells in various tissues such as muscles and fat.

Insulin resistance (IR) occurs when the response of cells to insulin is decreased. This has several adverse metabolic consequences, like impaired inhibition of gluconeogenesis and impaired glucose uptake by insulin-dependent tissues, leading to progressive elevation of fasting and post-load blood glucose up to type 2 diabetes (T2D); impaired lipolysis by muscle and adipose tissue, which causes increased synthesis of lipoproteins and increased fatty acids deposition in the liver, leading to increased VLDL (Very Low Density Lipoprotein) triglycerides and nonalcoholic fatty liver disease (NAFLD) respectively, elevation of insulin concentration, which causes suppression of nitric oxide synthase (NOS) and hyperactivation of the renin-angiotensin-aldosterone (RAA) system, leading to hypertension, as well as chronic stimulation of the ovary theca cells, leading to the polycystic ovary syndrome spectrum. Major risk factors for IR include sedentary lifestyle, family history of diabetes and advanced age. It is possible to reverse the effects of insulin resistance before it develops into full T2D and other major complications with change in lifestyle, diet, and pharmacological treatment. In the last few decades, T2D and obesity have increased dramatically in the general population contributing to an increase in morbidity and mortality around the world[1,2].

Recent studies have also connected conditions in the intrauterine and postnatal environment to deleterious metabolic outcomes in children, these phenomena fall into the definition of *Developmental Origin of Health and Disease* (DOHaD) which states that exposure to certain environmental conditions during critical steps of development (before and/or after birth) may have significant effects on an individual short- and long- term health status. Some of the most studied environmental conditions include poor nutrition (micro- or macro- nutrient deficiency), exposure to chemical agents, hormonal and metabolic perturbations[1,3].

Contrarily to T1D which is caused mainly by an autoimmune response of the body leading to pancreatic beta cells destruction, T2D is a multigenic complex disease with heritability risk from 25% to 80%, involving some known genetic variants accounting only for 10% of the total risk. In the recent years, the focus has shifted from genome studies to epigenome studies, since epigenetic modifications are now considered to be a strong plausible mechanism of action for the disease because, they act like a fast environmental adaptation[2,4].

In the past decade, many studies tried to link insulin resistance and type 2 diabetes to epigenetic modifications, gene expression dysregulations and single nucleotide polymorphisms (SNPs) in the adult, providing a list of potential markers for diagnosis and drug testing although not many studies focused on IR among obese children.

Epigenetics

Diversity is a natural consequence of both evolution and inheritance and, for decades, scientists have attempted to understand the underlying mechanisms driving life both at macro- and microscopic levels. Darwin's theory of evolution in 1859 together with Watson, Crick and Franklin's discovery of the DNA structure in 1953 started a revolution in life sciences leading up to genomics, transcriptomics and epigenomics studies as they are known today. At first the variation seen within a population was attributed to nucleotide changes in the DNA double helix molecule leading to aminoacidic changes in the final protein. Epigenetics added another layer of complexity, where previously heritable changes were thought to be propagated exclusively through DNA polymorphisms, a set of newly discovered mechanisms would ensure fine regulation of gene expression and constitute a strategy for a quick but soft adaptation to the environment[5].

The epigenome consists of a set of chemical changes in DNA and supporting structures (like histones) acting at the cellular level for the maintenance and regulation of DNA-related processes. Therefore, changes to the epigenome usually result in modification of the chromatin structure and regulation system of the genome without changes in the nucleotide sequence itself. These changes can be passed down to the offspring via transgenerational epigenetic inheritance. Although virtually all cells within an organism contain the same genetic information, not all cells express the same genes at the same time. In a broad sense epigenetic modification mediate the diversified gene expression profiles in a variety of cells and tissues in a multicellular organism in fact the roles of the epigenome are to regulate gene expression, development, tissue differentiation and suppression of transposable elements. Unlike the genome which generally remains unchanged within an individual, the epigenome can also be dynamically altered by environmental conditions.

In general Epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence itself. Specific epigenetic processes include paramutation, bookmarking, imprinting, gene silencing, X chromosome inactivation, position effect, DNA methylation reprogramming and regulation of histone modification. In the past decade, epigenetics has become increasingly important in life sciences since various epigenetic changes are associated with development and disease making them a potential target for new pharmacological treatments. Moreover, high-throughput technologies offer an opportunity to study epigenetic alterations in an “epigenetic wide” fashion by considering a very high number of loci (depending on the specific technology) and trying to associate their epigenetic status to phenotypic traits including diseases. With the rise of genome/epigenome wide approaches, bioinformatics and biostatistics have become key disciplines in order to obtain strong, reproducible and accurate results in every phase of an epigenomics project allowing us to better understand complex phenomena.

The epigenetic machinery

In eukaryotes, the DNA is contained in the nucleus and it’s tightly packed, allowing for a better organization/storage of the genetic material. DNA organized in chromatin restrict its accessibility to some parts of the genome while giving easier access to other specific areas. There are two forms of chromatin, euchromatin which is loosely condensed, allowing active transcription, and heterochromatin which is tightly condensed, hence preventing transcription. Chromatin regulation is dynamic and can change according to cell cycle stage, cell type, environment, and many other factors, actively influencing what can be produced and, at which efficiency level. The systematic arrangement of chromatin allows for the creation of a structure called nucleosome consisting of DNA packed around octamers, which are made of five histone proteins H2A, H2B, H3, H4 and H1 (for an increased stability - **Figure 1**).

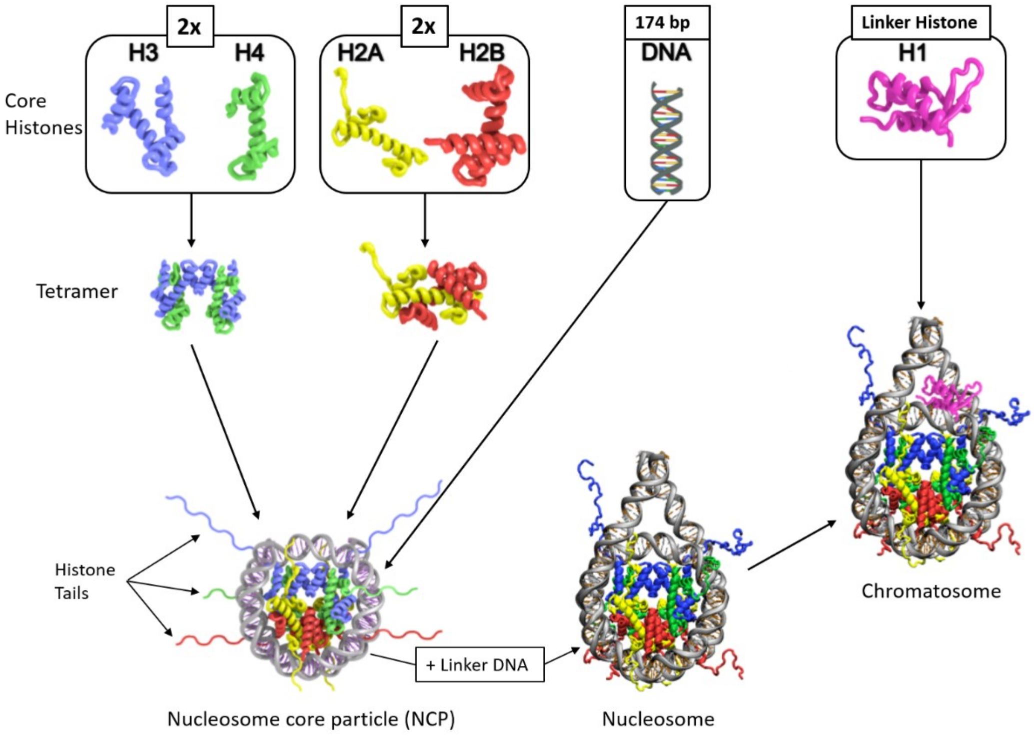


Figure 1: Nucleosome and chromatosome formation*. Schematic representation of histone proteins and how they are assembled to form complete nucleosomes and chromatosome*[6].

The combination of histones and nucleosome have an N-terminal tail which is prone to modifications such as acetylation and methylation of lysine residues or phosphorylation of serine residues. Histones are involved in epigenetic changes through the remodeling process which alters the compaction status of the chromatin, rendering DNA more, or less accessible to a wide variety of cellular processes and components like, small non-coding RNA mediated regulation or DNA methylation[7].

DNA methylation

DNA methylation (DNAm) occurs when a methyl- group (CH_3-) is covalently attached to the carbon 5 within a cytosine ring forming 5-methylcytosine (5mC – **Figure 2**) by the action of DNA methyltransferase enzymes (DNMT) and S-adenosylmethionine (SAM) as a methyl- group donor[8]. Moreover, is one of the best studied and most mechanistically understood epigenetic modifications, it is also well conserved among animals, plants, and fungi[9,10]. DNA methylation is seen as a key player in epigenetic silencing of transcription, it may coordinately regulate the chromatin status via the interaction of DNMT with other modifications and with components of the machinery mediating those marks. DNA methylation status is heritable, and it can be found in cytosines both in a CpG dinucleotide context and in a CHG or CHH trinucleotide contexts, however the vast majority of methylated cytosines in humans is found in a CpG context (around 98%) in adult human somatic cells while almost 25% of methylation in embryonic stem cells is found in other contexts[10].

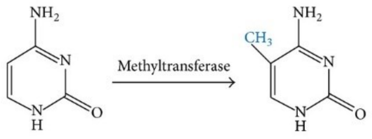


Figure 2: From cytosine to 5-methylcytosine. *Cytosine methylation to form 5-methylcytosine[11].*

There are three DNMT enzymes, DNMT3A, DNMT3B and DNMT1 which alter the methylation status in different ways, DNMT3A and DNMT3B can transfer a methyl group to previously unmodified cytosines hence creating new methylated regions or loci. DNMT1 works during DNA replication and can only copy the methylation patterns from the template strand to the newly synthetized strand (**Figure 3**). All three DNMT enzymes are heavily involved during embryo development but, by the time cells reach a terminal differentiation status, their expression is highly reduced[5].

During zygote formation, most of DNA methylation is removed, only to be re-established in the embryo in a short period of time after implantation. DNA methylation is essential for normal development since it plays a paramount role in several processes like genomic imprinting, X chromosome inactivation, suppression of repetitive elements and many others.

Human development from zygote into a complex adult organism requires a set of highly specific cellular processes. Gene expression regulation is primarily encoded in cis elements directed by transcription factors. Moreover, heritability of covalent DNA modifications and chromatin rearrangements often contribute to respond to developmental pressure.

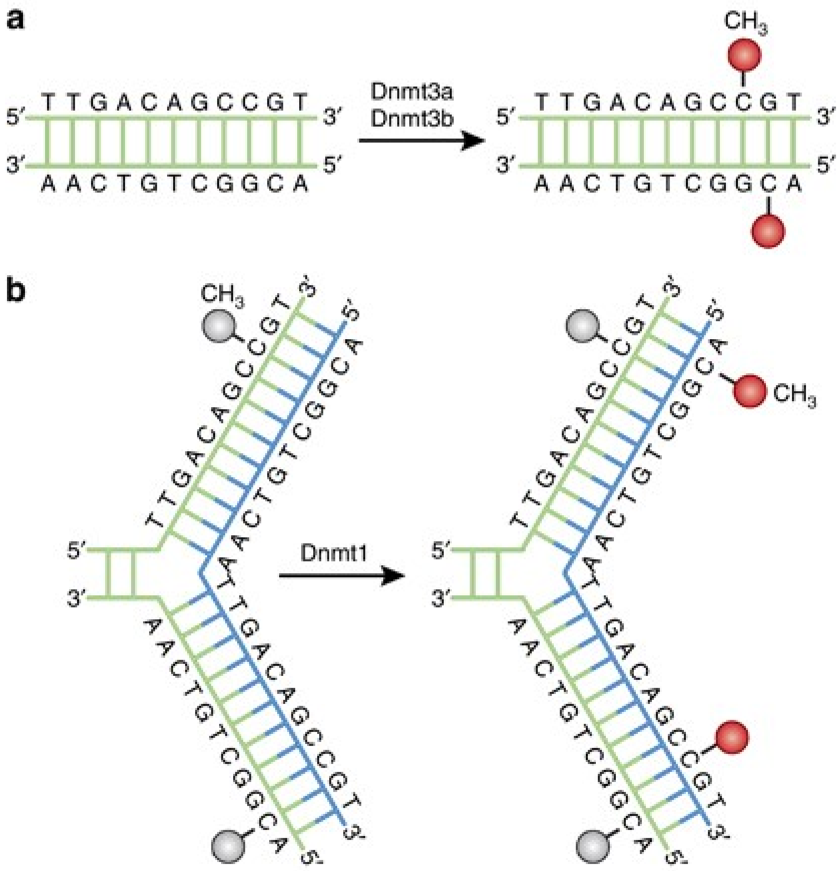


Figure 3: DNA methyltransferase enzymes in action. *Graphical representation of DNMT enzymes de novo methylation (a), and methylation copying process during DNA replication (b)* [5].

Genomic elements and DNA methylation

As previously mentioned, the majority of DNA methylation occurs on cytosines that precede a guanine nucleotide also known as CpG sites. Overall, mammalian genomes are depleted of CpG sites this may result from the mutagenic potential of 5mC that can deaminate to thymine. The remaining CpG sites are spread out across the genome where they are heavily methylated except for CpG islands. DNA methylation is essential for silencing retroviral elements, regulating tissue-specific gene expression, genomic imprinting, and X chromosome inactivation. Importantly, DNA methylation in different genomic regions may exert different influences on gene activities based on the underlying genetic sequence[5].

CpG islands

CpG islands (CGIs) are stretches of DNA roughly 1000 base pairs long, with a content of CpG dinucleotides around 50%, they're generally unmethylated and can be found in more than half (~70%) of human gene promoters[7] or distributed in other functionally relevant regions like transcription start sites (TSS). In particular, promoters for housekeeping genes are often embedded in CpG islands, and generally highly conserved between mice and humans[12]. Therefore, to maintain a hypomethylated status, the activity of DNMT enzymes has to be continuously blocked in active promoters to prevent unwanted gene silencing. Moreover, promoter methylation (or unmethylation) status is cell type or tissue dependent, and it is established during development but even if a large part of the mechanisms behind CGI regulated gene expression are known, there are still many molecular aspects being uncovered especially during developmental age. It appears that CpG islands have been evolutionarily conserved to promote gene expression by regulating the chromatin structure and transcription factor binding, but when methylated, CGIs can impair these mechanisms by also recruiting repressor (in the form of methyl-binding) proteins for a stable silencing of gene expression[5]. A good example highlighting the importance of CGI-related mechanisms can be seen with the disruption of their methylation patterns taking place in neoplastic cells[13].

Furthermore, DNA methylation can be found at enhancers and is also highly dynamic varying according to physiological and pathological conditions, especially in cancer where enhancer methylation plays an important role as promoter methylation does[7].

Many other genomic elements carrying methylation have gained popularity and attention over time, like CpG shores which are regions distant at most 2kb from a CGI. Shores usually present a lower CpG density compared to CGIs, but their methylation status is much more variable, and they are in fact considered among the most variable regions in the genome. Furthermore, methylation in CpG shores is also associated to transcriptional inactivation.

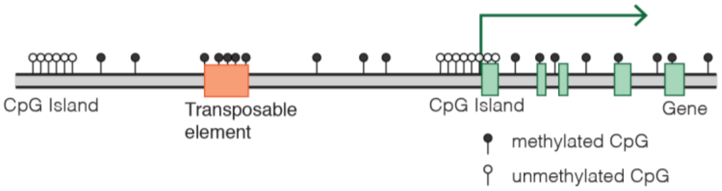


Figure 4: DNA methylation schematic landscape. *Showing different possible situations: CpG islands are unmethylated when allowing transcription, while intergenic transposable elements are methylated to prevent their activation.*

Intergenic regions

In the human genome, intergenic regions often contain endogenous transposable elements (nearly 40% of mammalian genomes), they are organized in three classes, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs). LINEs and LTRs encode for elements with strong promoters. Moreover, transposable elements can be potentially harmful since their expression would cause them to replicate and insert themselves in another genomic region (trough duplication), permanently altering or disrupting the functionality of the targeted area (**Figure 4**). To block these elements and prevent their activation, over a long period of time, DNA methylation is used causing constitutive hypermethylation[9]. Moreover, spontaneous mutations induced by 5-methylcytosine deamination to thymine permanently fix this block[5,14].

Gene body

As previously mentioned, DNA methylation often affects enhancers, promoters and TSS causing gene silencing, but gene bodies can also be methylated especially the first exon which also acts as a silencer when methylated. While methylation in other exons is associated to a higher gene expression during cell division while in non-dividing cells like neurons, gene body methylation (beyond the first exon) has not been associated with expression alterations or fluctuations[5] (**Figure 4**). Furthermore, methylation in intronic or intragenic regions still has unclear functions.

DNA methylation using microarrays

Fast and accurate methylation analysis methods are a key element for discovering the role that methylation plays in several contexts from life sciences to medicine. The main principle on which DNA methylation detection is founded, relies on the chemical action of sodium bisulfite. When DNA is treated with sodium bisulfite, unmethylated cytosines are converted by deamination to uracil, while 5mC remain unchanged[15]. Among other methods to assess DNA methylation there are tiling hand-made microarrays and sequencing of bisulfite-treated DNA (whole genome or target regions based). Although effective, sequencing methods require a moderately high amount of DNA and labor, and they make bioinformatic analysis challenging for large-case studies. A very good and cost-effective way to limit the resources needed for analysis of a large cohort is to employ microarrays such as the *Infinium Human Methylation EPIC* (EPIC) by *Illumina* which can measure methylation levels of approximately 850000 single CpG dinucleotides using a small amount of genomic DNA (~250ng) while analyzing 8 samples per array, with 16 being the base configuration (2 arrays) reaching up to 96 samples (12 arrays). The EPIC array was designed to incorporate ~99% of RefSeq genes, ~95% of known CGIs and ~80% of FANTOM5 enhancers also including open chromatin and enhancers from ENCODE and genes, promoters, and UTRs from GENCODE, for a detailed description of these features[16], see the *List of Abbreviations* section.

The EPIC array uses beads linked to long target-specific probes designed to query single CpGs within a sample, methylation is measured in a quantitative way by "genotyping" bisulfite converted DNA. Each array contains two types of assays (or chemistries) called *Infinium I* and *Infinium II* which complement each other strengths benefitting the array overall sample coverage. The *Infinium I* assay uses two types of probes per CpG locus, the first one for methylation and the second for unmethylation matching the status of the analyzed site and they are designed under the assumption that methylation is regionally correlated within a 50 base pair span[17,18]. The *Infinium II* assay uses only one probe complementing the base immediately upstream the considered nucleotide then, during array preparation a single base extension (SBE) allows the addition of a labeled guanine (G) or adenine (A) which are complementary to the locus of interest that can contain either a cytosine (methylated locus) or thymine (unmethylated locus). Therefore, the *Infinium II* assay enables methylation status detection regardless of the previously mentioned assumptions about neighboring CpGs while maintaining a high correlation with *Infinium I* probes detected methylation[16] (>90% - **Figure 5**).

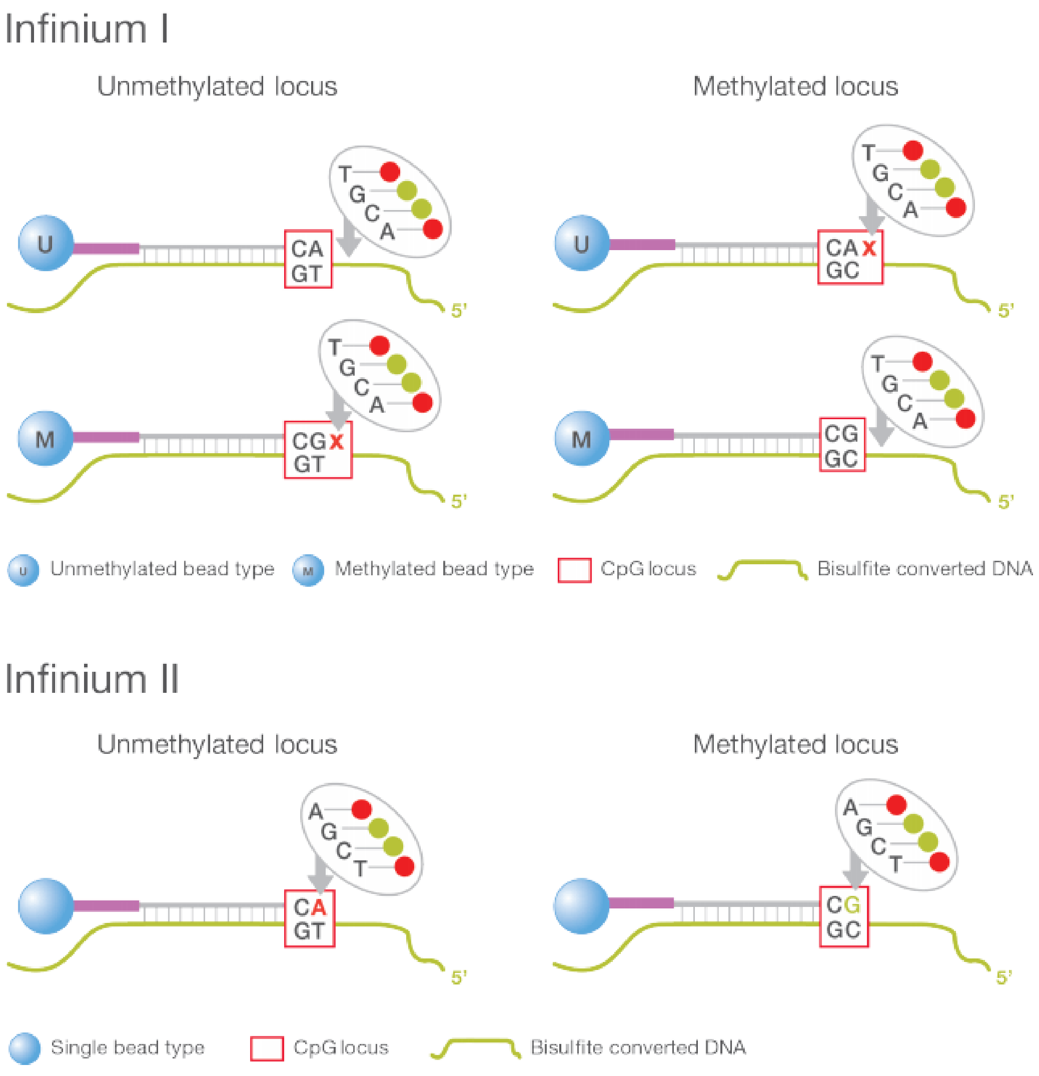


Figure 5: Graphical representation of Infinium I and Infinium II chemistries. *Two probes per locus are used to evaluate methylation or unmethylation status at a CpG site assuming correlation on a 50bp range (Infinium I) while a single probe allowing for SBE and detection of methylation without assuming its genomic distribution (Infinium II).*

Since looking at single probe luminescence would be challenging to analyze, two metrics are used to associate each locus to its methylation status, the β-value and the M-value which are computed locus-wise using methylation (M) and unmethylation (U) values directly derived from the array or the sequencing (after some pre-processing steps).

$$\beta\text{-}value = \frac{M}{(M+U)} \qquad with\text{: } 0 \leq \beta\text{-}value \leq 1$$

$$M\text{-}value = \log_2\left(\frac{M}{U}\right) \qquad with\text{: } -5 \leq M\text{-}value \leq 5$$

Methylation ratios computed using the β-value or the M-value express the same concept, while the β-value ranges from 0 to 1, the M-value ranges from -5 to 5. Moreover, for both β-value and M-value the lower bound corresponds to full unmethylation and the upper bound to full methylation with values in between corresponding to hemi-methylation. Furthermore, the use of one value or the other doesn't have a severe impact on any downstream analysis although it has been reported that using the M-value for differential analysis may produce slightly more statistically significant results while ensuring a better reproducibility[19,20].

AIM OF THE PROJECT

The main goal of this project is to investigate the association of genome-wide DNA methylation profiles with insulin resistance in pediatric obese subjects. Since insulin resistance is often present alongside obesity, a cohort of obese children was chosen to remove from the analyses the effect of obesity on insulin levels. A secondary aim is to define a reliable bioinformatic workflow to efficiently analyze DNA methylation array data, allowing for integration with biological knowledge (**Figure 6**).

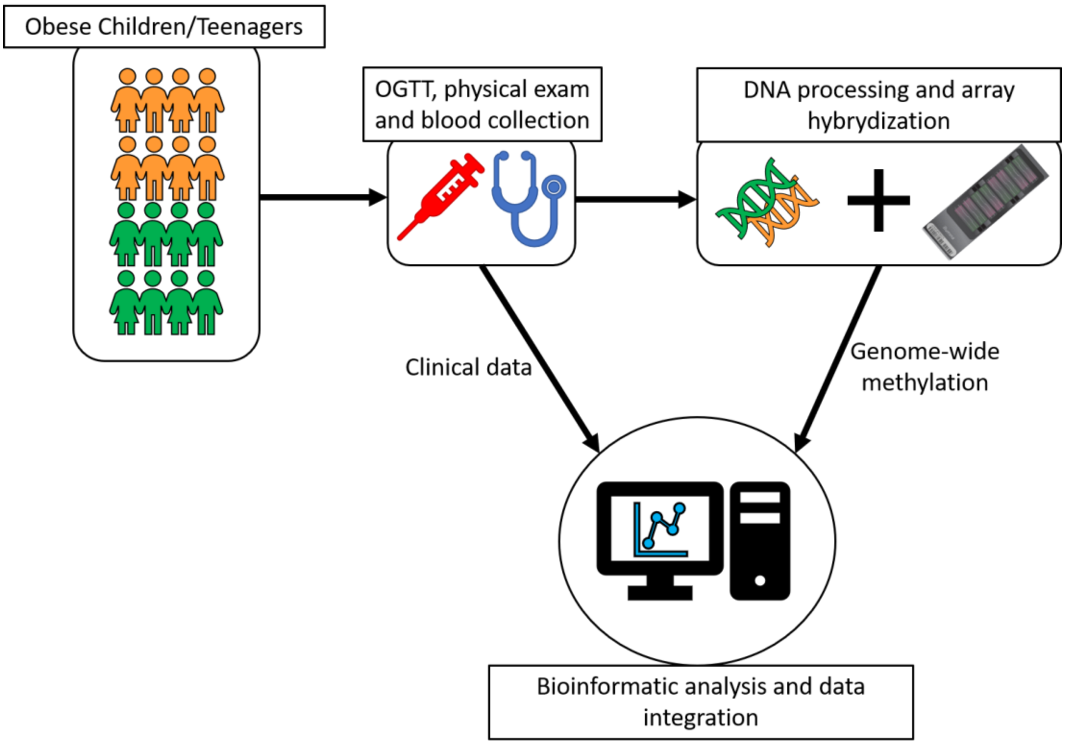

Obese Children/Teenagers
OGTT, physical exam and blood collection
DNA processing and array hybrydization
Clinical data
Genome-wide methylation
Bioinformatic analysis and data integration

Figure 6: Schematic representation of the study design.

MATERIALS & METHODS

Dataset composition

The dataset composed of obese pediatric subjects enrolled by the team of professor *Miraglia del Giudice (University of Campania Luigi Vanvitelli)* was divided in two groups, insulin resistant (96 samples; “R” prefix), and insulin sensitive (96 samples; “N” prefix). Subjects of the 2 groups were matched for BMI, sex, and age values. Furthermore, the insulin resistance status was assessed using the HOMA-IR and WBISI indexes as described below. The insulin resistance status (*Sample_Group* feature; IR = insulin resistant, IS = insulin sensitive) investigated in the present study was defined according to WBISI values only.

Clinical information

For each subject, information regarding several clinical features were collected to use as covariates during the bioinformatic analysis (**Table 1**).

			Sample_Group	
Feature Name	**Description**	**Units**	**IR mean (±sd)**	**IS mean (±sd)**
Weight	Subject's weight	Kg	76.11 (±17.98)	76.12 (±18.94)
BMI	Body Mass Index	-	32.43 (±5.1)	32.67 (±5.3)
Height	Subject's height	m	1.52 (±0.1)	1.52 (±0.1)
Waist	Waist circumference	cm	93.98 (±13.86)	92.96 (±11.72)
PAD	Diastolic arterial pressure	mmHg	65.85 (±10.68)	67.38 (±8.21)
PAS	Systolic arterial pressure	mmHg	117.1 (±12.27)	110.8 (±11.84)
Age	Subject's age	years	11.86 (±1.99)	11.78 (±1.9)
HOMA-IR	Homeostatic Model Assessment for Insulin Resistance	-	8.28 (±3.17)	1.77 (±0.68)
Average Glucose	Mean of G_t values	mmol/L	111.6 (±12.9)	101.4 (±11.94)
Average Insulin	Mean of I_t values	µU/L	158.21 (±56.6)	38.35 (±14.23)
WBISI	Whole Body Insulin Sensitivity Index	-	1.48 (±0.81)	6.58 (±1.55)
IGI	Insulinogenic Index (ratio between I_0-I_{30}/G_0-G_{30})	µU/mmol	3.83 (±2.54)	2.1 (±4.42)
Sex	Subject's sex (M: male; F: female)	-	M: 52 - F: 44	M: 53 - F: 43
G_t	Subject's glycemia at timepoint 0'≤ t ≤120'	mmol/L	-	-
I_t	Subject's insulin level at timepoint 0'≤ t ≤120'	µU/L	-	-

Table 1: Main clinical features collected for the cohort*. Features descriptions with measurement units and average value (± standard deviation, when applicable) for Sample_Group (IR = insulin resistant, IS = insulin sensitive).*

Insulin resistance assessment

A standardized method to measure glucose and insulin fluctuations is called *Oral Glucose Tolerance Test* (OGTT), which is performed by orally administering a standard amount of glucose to a subject, collecting blood samples at regular time intervals (usually every 30 minutes for 2 hours) to measure insulin and glucose amounts with an additional sample collected before the test to measure fasting glucose and insulin levels. The several samples collected are used to determine how long it takes for blood glucose to go back to a baseline level. The OGTT is mainly used for insulin resistance, diabetes, and beta cell functionality testing, but it is sometimes used for reactive hypoglycemia, acromegaly, and disorders in carbohydrates metabolism.

As mentioned above the *Oral Glucose Tolerance Test* is used to quantify glycemic fluctuations over time but the OGTT itself cannot discriminate between IR, T2D or beta cell disfunctions, therefore two OGTT-derived estimates were used to assess fasting and post-load insulin sensitivity, the *Homeostatic Model Assessment for Insulin Resistance* (HOMA-IR) and the *Whole-Body Insulin Sensitivity Index* (WBISI), by using data collected during the test such as fasting glucose (G_0 [mmol/L]), fasting insulin (I_0 [μU/L]), mean glucose (G_μ [mmol/L]), mean insulin (I_μ [μU/L]) and they are computed as follows[21–23]:

$$HOMA\text{-}IR = \frac{G_0 \cdot I_0}{22.5}$$

$$WBISI = \frac{10000}{\sqrt{(G_0 I_0) \cdot (G_\mu I_\mu)}}$$

It should be noted that HOMA-IR optimal range is between 0.5 and 1.4, with values like 1.9 being a sign of early IR, while values around 2.9 being significant IR, on the other hand WBISI optimal value is above 3, lower values progressively tend to IR and, in some studies, IR subjects were found with WBISI threshold around 2.5[24].

Sample collection & DNA processing

DNA processing and genome-wide methylation profiling were conducted by the team of Professor *Claudio Maffeis (University of Verona),* starting from peripheral venous blood, genomic DNA for each sample was extracted using *DNeasy Blood and Tissue kit* (Qiagen) according to manufacturer's protocol then DNA concentration was assessed using a Nanodrop spectrophotometer and samples were adjusted through serial dilution to 48-52 ng/µL.

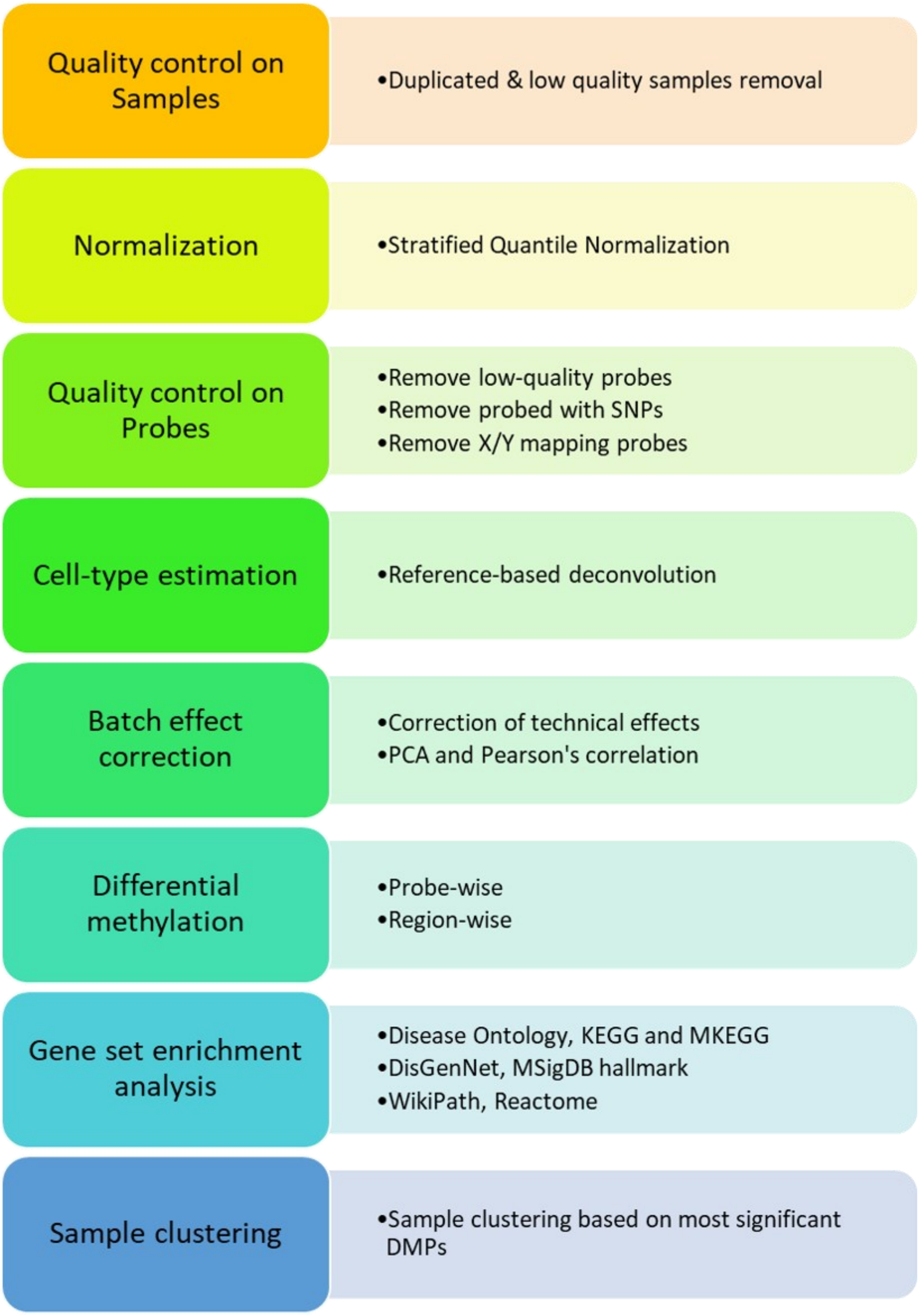
Bisulfite conversion of genomic DNA was performed using *EZ DNA methylation Kit* (Zymo Research), using 600ng of DNA per sample, followed by DNA amplification, fragmentation, and precipitation, were performed following manufacturer's protocol.

Genome-wide methylation profiling

Genome-wide methylation profiling was carried out on bisulfite-converted DNA using the *Illumina Human Methylation EPIC array* (Illumina Inc.) following Illumina standard protocol, using 250 ng of bisulfite converted DNA per sample to perform array hybridization. DNA of four samples were split in two forming duplicates and randomly placed in the arrays, to serve as a microarray quality control during the bioinformatic analysis. Once the arrays were ready, a single base extension and staining were performed, and data were collected using an *iScan* (Illumina Inc.) system producing standard "**.idat"* files, containing probes fluorescence raw per-CpG methylation levels.

Bioinformatic analysis

The bioinformatic workflow employed in this thesis is described as summarized in **Figure 7**.



Quality control on Samples	•Duplicated & low quality samples removal
Normalization	•Stratified Quantile Normalization
Quality control on Probes	•Remove low-quality probes •Remove probed with SNPs •Remove X/Y mapping probes
Cell-type estimation	•Reference-based deconvolution
Batch effect correction	•Correction of technical effects •PCA and Pearson's correlation
Differential methylation	•Probe-wise •Region-wise
Gene set enrichment analysis	•Disease Ontology, KEGG and MKEGG •DisGenNet, MSigDB hallmark •WikiPath, Reactome
Sample clustering	•Sample clustering based on most significant DMPs

Figure 7: Summarized steps of the bioinformatic analysis. *The image reports the main steps of the workflow with a brief description of the analytical strategies employed.*

Quality Control

Microarray quality control was performed in three phases:

- Quality control on samples
    - Evaluation of microarray quality through duplicated samples
    - Detection and removal of low-quality samples
- Normalization
    - Inter-array normalization using quantiles
    - Intra-array normalization filtering outliers
- Quality control on probes to remove low quality probes and SNPs related probes

Quality control on samples, was performed by computing probe-wise detection p-value (detP) which is a statistical indicator of probe reliability, it derives from the sum of raw methylation and unmethylation signals compared to the background signal, estimated using negative control positions, assuming they're normally distributed. Then, a pairwise comparison was performed on duplicated samples to assess the signal quality on different arrays and positions, using methylation values (β-values), detP and absolute methylation difference ($|\Delta\beta|$) per CpG. Average per-sample detection p-value was calculated to get a sample-wise quality estimation allowing the removal of duplicated samples by picking, for each couple, the one with the lower detP value. Furthermore, all samples with mean detP ≥ 0.05 were also removed since their low quality may affect downstream analyses.

Samples were then normalized to correct for systematic measurement errors in the data. Errors may be introduced by several factors such as the data acquisition method/platform (i.e., signal noise), differences in probe labeling, subtle variations in target DNA concentration, efficiency of hybridization. For the purpose of this study, normalization was performed using *stratified quantile normalization* (SQN) approach. This method starts by normalizing the signal of type II probes across samples, interpolating a reference distribution used to normalize type I probes. Given the different nature of type I and II probes, this process is stratified by genomic region before applying the interpolation (inter-array normalization). Moreover, sex probes (on chromosomes X and Y) are normalized separately for female and male samples using the sex information provided in the sample sheet. This method does not perform background correction, but it filters out methylation outliers by thresholding intensities close to zero (similar to intra-array normalization). SQN was chosen because it relies on the assumption that samples have similar distribution regardless of their class and it is more appropriate where the methylation difference (by class) does not involve global changes such as in cancer-normal comparison[25,26].

Subsequently, a probe filtering step was done to remove from the dataset probes with detP ≥ 0.01 in at least one sample then, probes containing SNPs at the CpG site or single base extension site (SBE – type II Infinium probes) were also removed. As last step in quality control, probes mapping on X and Y chromosomes were removed to minimize the impact of sex that may constitute an unwanted source of variation leading to biased results in the following steps of the analysis.

These steps were performed starting from **.idat* files produced by the array scanner, as previously mentioned. They were loaded into *R* using a sample sheet, which is a file containing a complete list of samples with array-related information, to which clinical features were also added. Another annotation file called microarray manifest was loaded since it contains information regarding mappings and overlapping genes as well as CpG position within functional elements such as CGIs, CpG shores, CpG shelves, open sea, DNAse hypersensitivity sites, TSS, enhancers, promoters and more, as reported on the Illumina support page[27]. All quality control sub-steps described above were performed using the *minfi*[26] package under *R*.

Per-sample cell type composition

Peripheral venous blood, by nature, is a mixture of different cell types that can be present in different amounts in each sample. This variability may be due to several biologically and clinically important conditions or reflect changes that are a consequence of the disease state, potentially causing the analysis to be biased. Therefore, it would be useful to estimate the proportion of each blood cell type. This can be achieved with a statistical process usually called cell type deconvolution or estimation. Among the many ways to perform deconvolution, a reference-based approach was chosen since the main constituent cell types within blood are known. With this method, a reference file or database containing molecular markers representative of each possible cell type in the tissue of interest, is used along with a multivariate regression to estimate individual cell type ratios. In this case a specific reference for blood cells built using sorted cells and the EPIC array was used to identify CD4+ T lymphocytes, CD8+ T lymphocytes, B cells, Natural Killer cells, Monocytes and Neutrophils/Granulocytes according to the low methylation of known gene markers such as RPTOR (cg04162316), CD8A (cg25939861), BLK (cg03860768), CLASP1 (cg14047092), SLFN5 (cg02647842) and NFIA (cg22451300) respectively[28,29]. A two-sided t-test was performed on cell type proportions versus the *Sample_Group* (disease state) and *sex* variables, not only to see if these factors can influence the sample composition and therefore the downstream analysis, but also to see if some cell types tend to be overabundant in IR subjects.

Batch effects correction and exploratory analysis

Batch effects are defined as a technical source of variation which is not embedded in the samples themselves, but it is unwillingly introduced during the several steps where the samples are manually handled for example different processing times, different operators, different machines or, in this case, even different arrays. Applying a batch effect correction method can be useful to clean the data by estimating and removing these unwanted sources of variation. To correct for batch effect, an empirical Bayes procedure was used to directly remove known technical sources of variation, returning corrected methylation data. *Sample_Plate* was used as main technical factor while *Sample_Group* (disease state), *sex*, *BMI*, *age,* and *cell type proportions* were used as outcome and other covariates. The correction was performed by using the *ComBat* method from the *sva R* package[30,31].

The correction effect was then assessed by computing Pearson's correlation between a *Principal Component Analysis* (PCA) performed on M-values and clinical variables using the *prcomp_irlba* method from the *irlba* package[32,33].

Differential methylation analysis

After removing batch effect and reducing the effect of possible covariates, it was possible to proceed with probe-wise and region-wise differential methylation analysis. It consists in fitting a linear model to represent the phenotypical difference of interest, in this case comparing insulin resistant subjects with insulin sensitive ones. The fitted model (one per probe set) was then corrected for multiple testing and microarray annotation was merged to the result obtaining a comprehensive list of differentially methylated probes (DMPs) ranked by Benjamini & Hochberg (BH) adjusted p-value. Probe-wise differential analysis was performed using *limma* package in *R*[34,35].

Sometimes it may be interesting to assess the methylation status of larger genomic regions rather than single loci, and how they are modulated in the two groups. This step relies on previously discovered DMPs and, for each chromosome computes two smoothed estimates: one weighted with CpG-wise t-statistics, and one not (for a null comparison). Then, the two estimates are compared using the *Satterthwaite*[36] approximation which accounts for different sample variances in the two estimates. The entire process, detects regions in which methylation between the two groups of interest is statistically different, this step was performed using the *dmrcate* method from the homonymous *R* package[37,38].

Gene set enrichment analysis

When dealing with high throughput technologies it is very easy to obtain a huge amount of result (i.e., many different loci across the genome), so data interpretation becomes a key aspect with this type of project which ultimately leads to a better understanding of the biological process associated with the phenotype studied. Since probe-wise differential methylation analysis produces a list of loci and gives a logarithmic methylation fold change (logFC) between the groups compared, it is possible to apply a *gene set enrichment analysis* (GSEA) to highlight which metabolic pathways, according to an annotated list of gene-sets, are likely to be affected by the variability of the trait examined. This step was performed by extracting from the DMP results, gene names (when present) linked to CpGs (with adjusted p-value < 0.05 together with their respective logFC values) then, since in microarrays multiple probes can map on the same gene, to avoid overrepresentation bias, the mean logFC for each group of redundant genes was used. Therefore, using multiple specific gene sets, it was possible to retrieve the category to which each gene belongs from a pathway standpoint. Furthermore, GSEA was performed limiting the minimum number of genes per category to 50, meaning that only pathways containing at least 50 genes from the differential analysis are reported. The gene sets used for this analysis are from the *Disease Ontologies*[39] (DO), *DisGenNet*[40–42] (DGN), *Kyoto encyclopedia of Genes and Genomes*[43–45] (KEGG), *Hallmark curated gene signatures*[46–48], *WikiPathways*[49] (WP) and *Reactome pathways*[50–52] (RP) which are all handled by the *clusterProfiler*[53,54] and *DOSE*[55,56] packages.

DMP based sample clustering

Since in literature, the adjusted p-value cut-off for DMPs filtering is often set to 1e^{-7}, loci meeting this criterion were selected for further analysis, and their associated β-values were used to perform a hierarchical complete clustering (based on Euclidean distance) on samples to see if they would group into two different clusters according to their IR/IS status. The clustering results were then studied to understand the underlying factors behind cluster formation, dividing the dendrogram to form two and three cluster. Moreover, the impact of this DMP-driven separation on the insulin resistance status was evaluated by computing (per sample) t-tests (Welch's two sided) and Fisher's (exact test for count data) tests against numerical and categorical clinical features available. Furthermore, these *"super-significant"* loci were cross-referenced with the *EWAS catalogue*[57,58] database to understand if they were found in other epigenomics studies, which phenotype/trait these studies were focusing on.

RESULTS

Quality control

A total of 192 samples were analyzed with clinical data and microarray annotation, each sample (corresponding to a single individual) contains 866091 raw CpG-mapping probes. As previously reported in the *Bioinformatic analysis* section, microarray quality control was performed in three phases:

- Quality control on samples
    - Evaluation of microarray quality through duplicated samples
    - Detection and removal of low-quality samples
- Normalization
    - Inter-array normalization using quantiles
    - Intra-array normalization filtering outliers
- Quality control on probes to remove low quality probes and SNPs related probes

As mentioned above, the dataset included four paired-duplicated samples (N10, N10_2, N46, N46_2, N80, N80_2, R45, R45_2) that were used to measure the detection reliability of methylation profiles between different arrays and positions, using methylation values (β-values), absolute difference in beta value (|Δβ|) and detP as shown in **Figure 8**. For each duplicate-pair, the sample with the higher mean detP was removed (N10_2, N46, N80, R45_2) from the following analyses. Moreover, two individual samples (N25 and N44) were then removed because of their poor quality (mean detP ≥ 0.05 – **Figure 9**).

Stratified quantile normalization (with genomic region stratification, outlier filtering) was applied and 865859 were retained.

Probes were then filtered out according to signal/background-noise ratio (detP ≥ 0.01 - in at least one sample), the presence of CpG and SBE mapping SNPs (40801 + 26638), their localization on sex chromosomes (17587) retaining a total of 780842. Methylation distribution post normalization can be seen in **Figure 10**.

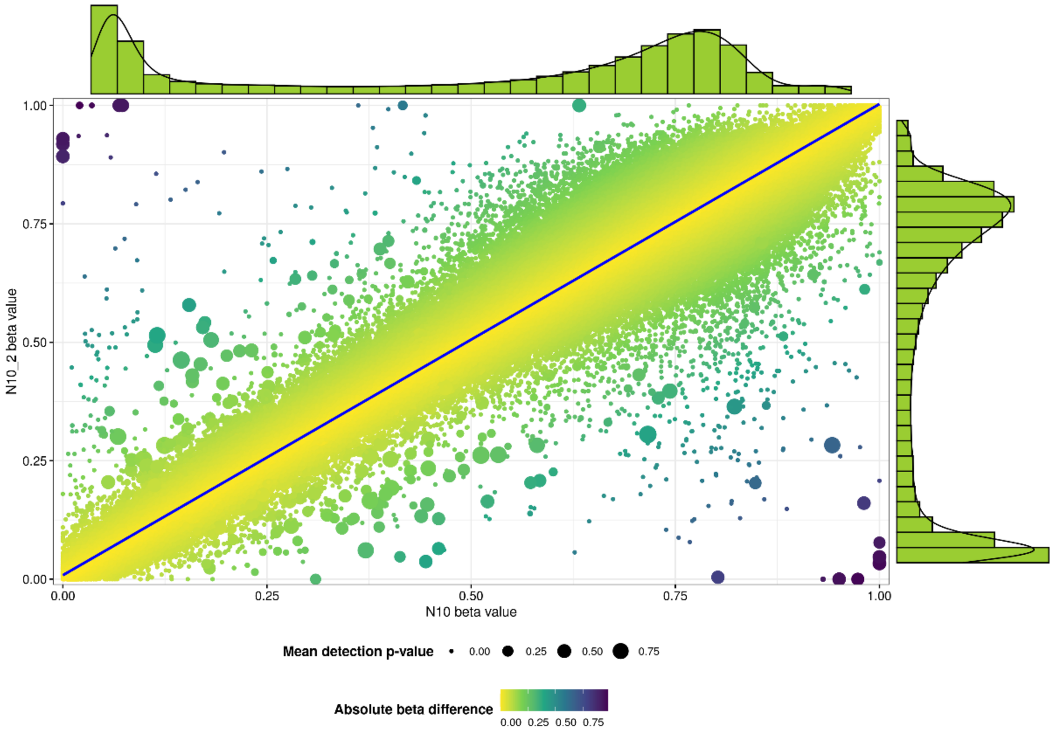


Figure 8: DNA methylation profiles of a duplicate sample. *The figure shows the DNAm value of a duplicated sample N10 (x-axis N10; y-axis N10_2). Marginal distributions show the overall methylation profiles for the pair. Smaller points represent lower detP; Darker points represent higher mean absolute difference between β-values (of the couple).*

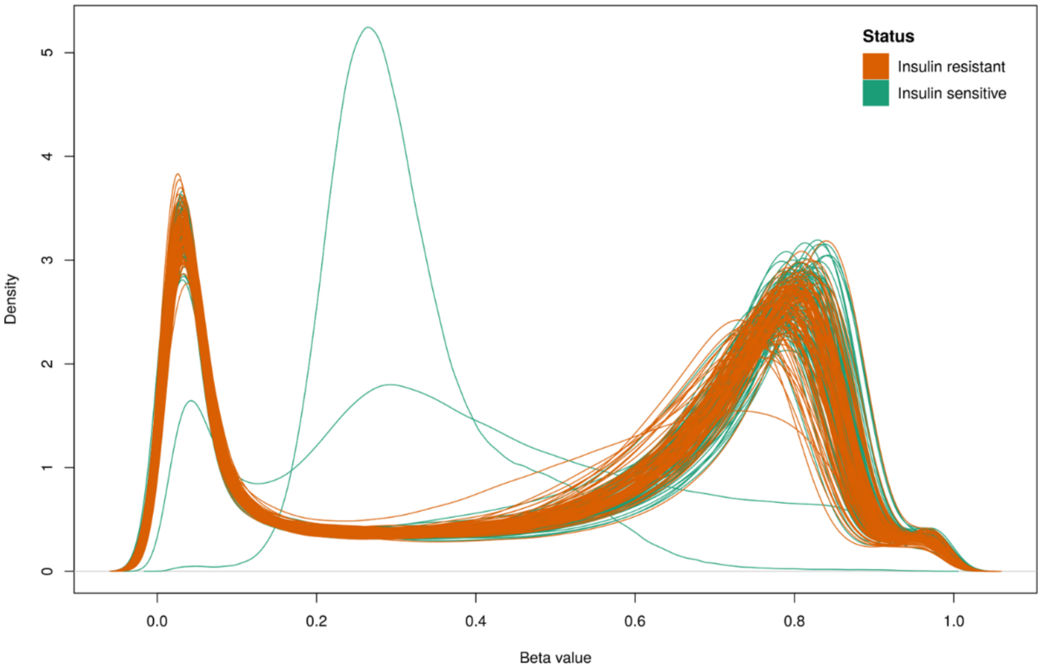


Figure 9: DNA methylation profiles on raw data. *Figure shows probe density (y-axis) according to their methylation value (x-axis). IR and IS samples are colored in orange and green, respectively. Two IS samples (green), clearly show an irregular methylation profile therefore, they were discarded.*

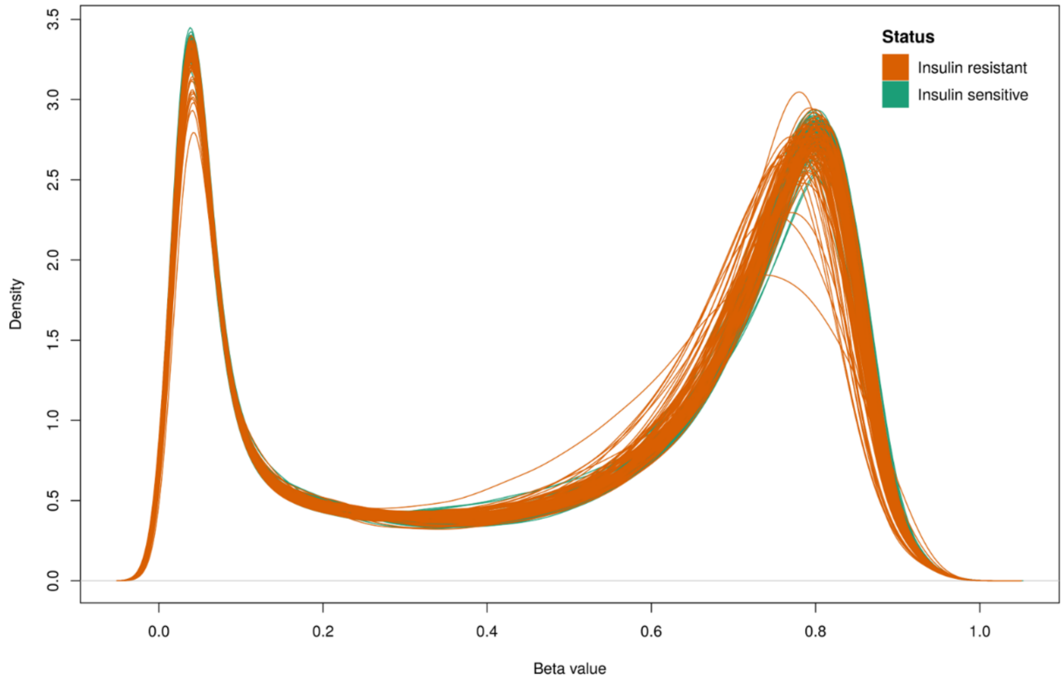


Figure 10: DNA methylation profile at the end of QC. *Figure shows probe density (y-axis) according to their methylation value (x-axis). IR and IS samples are colored in orange and green, respectively. It is possible to appreciate the differences caused by QC when compared with Fig 8.*

Cell type composition

Cell type proportions for each sample were estimated with a reference-based approach (6 reference cell types) as previously described. Cell type amounts of each sample were tested to evaluate their association with phenotypical status and sex (IR/IS – M/F; Welch's two-sided t-test). Per group (IR/IS) cell type testing revealed that B cells, CD8+ T cells, Monocytes and Neutrophils proportions are statistically significant between IR and IS samples (p-value<0.05). Furthermore, B cells, CD8T+ cells and Monocytes have a higher estimated mean in the IR group. On the contrary, Neutrophils have an estimated mean higher for IS samples (**Table 2** - **Figure 11**). Per sex composition testing led to a very different situation returning a significant result (p-value<0.05) only for NK cells with estimated mean value for female subjects being lower than the males as showed in **Table 3** and **Figure 12**.

Cell type	Proportion in IR	Proportion in IS	p-value	
B cells	0.083	0.064	1.53E-05	*
CD4+ T cells	0.106	0.111	0.3781	
CD8+ T cells	0.153	0.132	0.0077	*
Monocytes	0.068	0.057	0.0042	*
Neutrophils	0.514	0.552	0.0042	*
NK cells	0.047	0.045	0.6643	

Table 2: T-test results on disease status*. Results of statistical testing on insulin resistance/sensitivity status using mean cell type proportions. The red "*" symbol marks rows associated with statistically significant p-value (<0.05).*

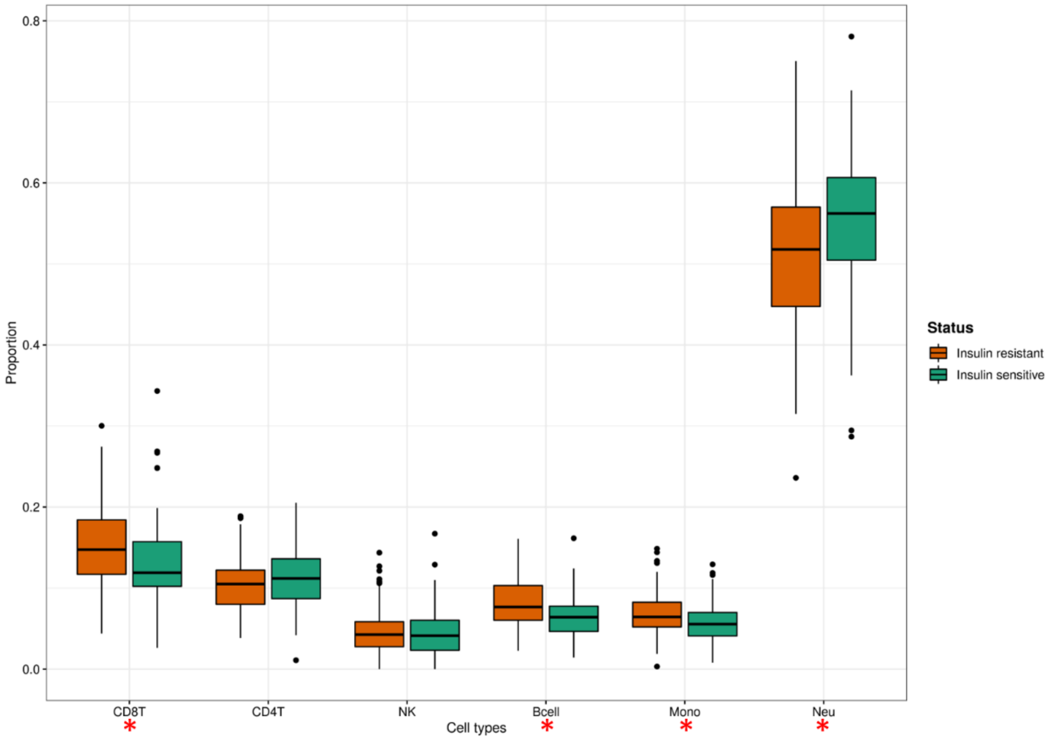


Figure 11: Cell type composition. *Boxplot showing the amount (y-axis) of cell types (x-axis) for the 2 groups. Insulin resistant in orange and insulin sensitive in green. The red "*" symbol marks rows associated with statistically significant p-value (<0.05).*

Cell type	Proportion in F	Proportion in M	p-value	
B cells	0.07143308	0.075744943	0.345952	
CD4+ T cells	0.11328711	0.10448452	0.095739	
CD8+ T cells	0.14086403	0.144368256	0.658005	
Monocytes	0.05961453	0.064647134	0.204548	
Neutrophils	0.5419071	0.525565823	0.221046	
NK cells	0.04093377	0.050008325	0.03383	*

Table 3: T-test results on sex. *Results of statistical testing on sex (Male/Female) using mean cell type proportions. The red "*" symbol marks rows associated with statistically significant p-value (<0.05).*

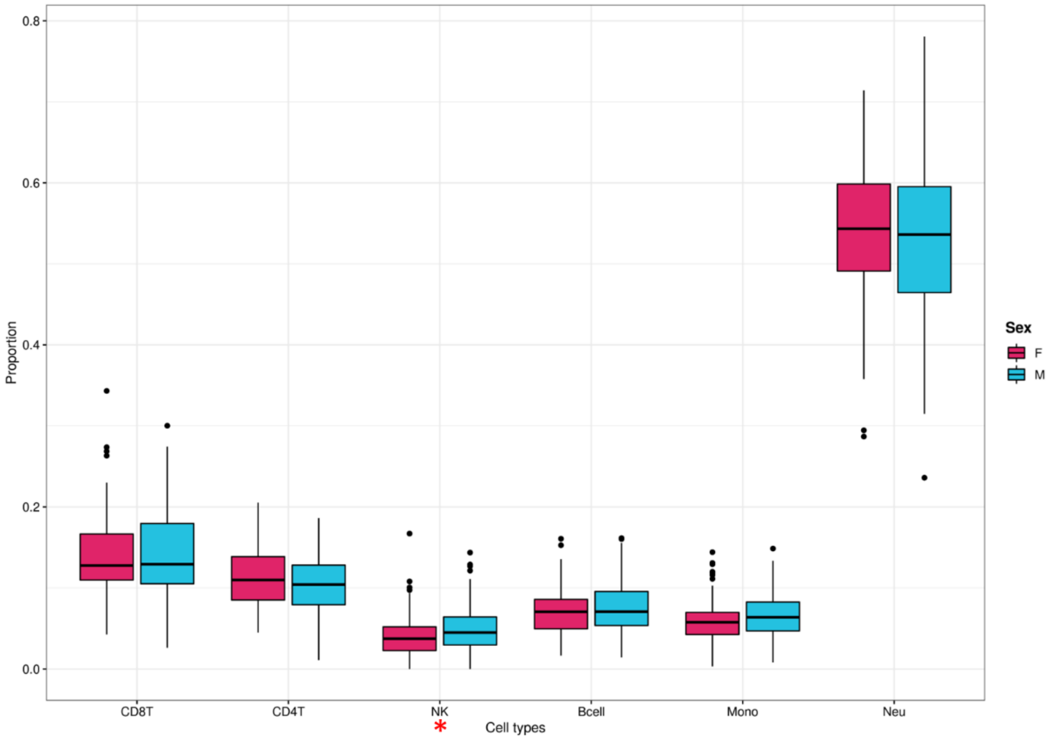


Figure 12: Cell type composition. *Boxplot showing the amount (y-axis) of cell types (x-axis) for the 2 groups. Males (M) in turquoise and Females (F) in magenta. The red "*" symbol marks rows associated with statistically significant p-value (<0.05).*

Batch effect correction and exploratory analysis

Technical effects were corrected using *comBat* as previously described. Then, PCA was performed to visualize sample data distribution along principal components 1, 2 and 3 (PC1, PC2, PC3). Coloring the data according to IR/IS status (**Figure 13**), shows no clear separation between IR samples (orange) and IS samples (green) although in panel A and C, insulin sensitive samples seem to be more concentrated around values lower than 0 for PC1 whether insulin resistant samples appear more scattered towards positive values of the principal component. Coloring points according to sex (**Figure 14**) shows an almost complete overlapping of point clouds representing males (turquoise) and females (magenta) thus confirming that samples variance associated with the first three principal components is not affected by sex. To better understand the relationship between clinical data and methylation, PCs were used to compute Pearson's correlation with clinical data, results are shown in **Figure 15**.

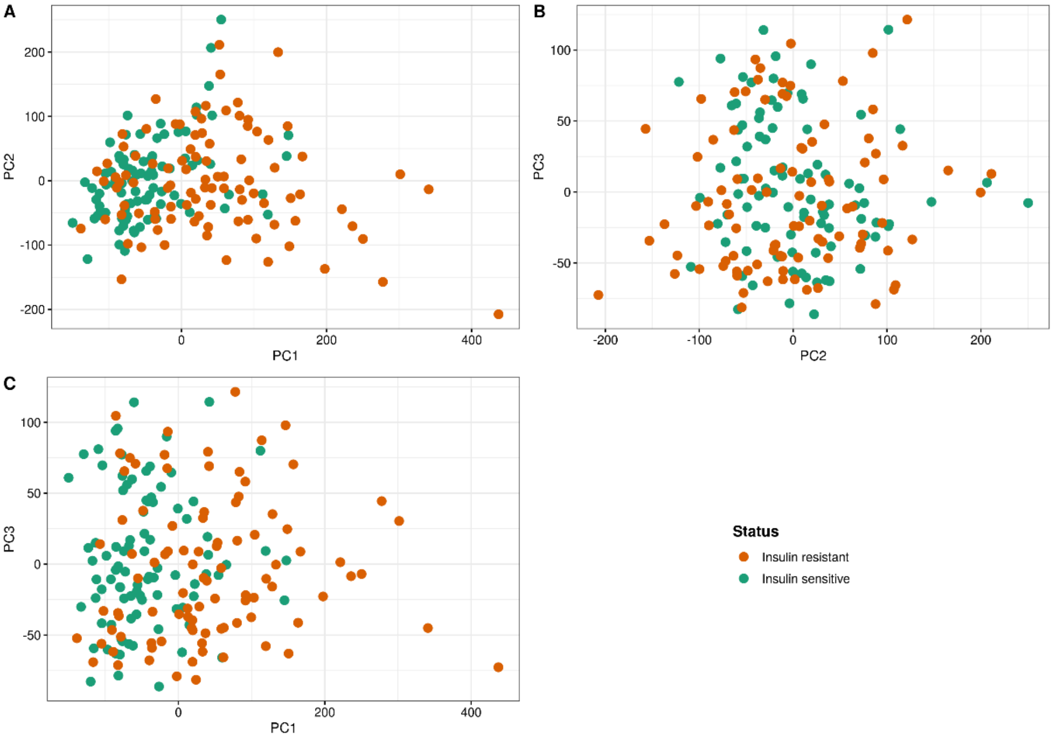


Figure 13: Principal component analysis colored by disease status. *PCA plot of IR samples (orange) vs IS samples (green). Visualizations of PC1 vs PC2 (A), PC2 vs PC3 (B), and PC1 vs PC3 (C).*

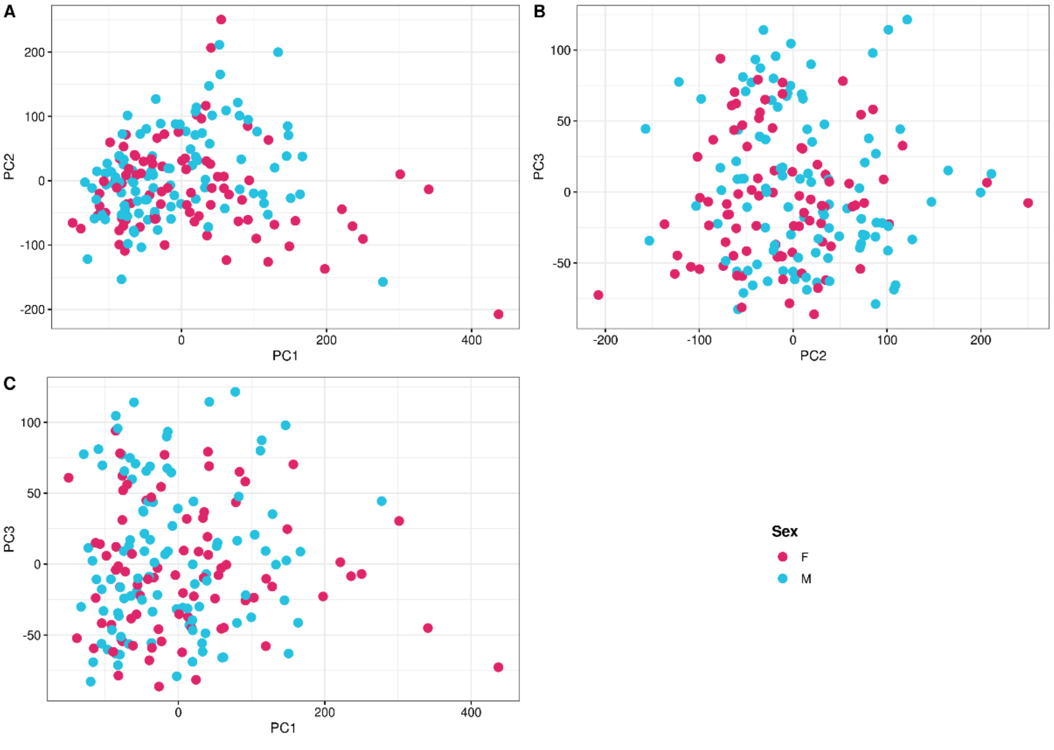


Figure 14: Principal component analysis of samples grouped by sex. *PCA plot of Males (turquoise) vs Females (magenta). Visualization of PC1 vs PC2 (A), PC2 vs PC3 (B), PC1 vs PC3 (C).*

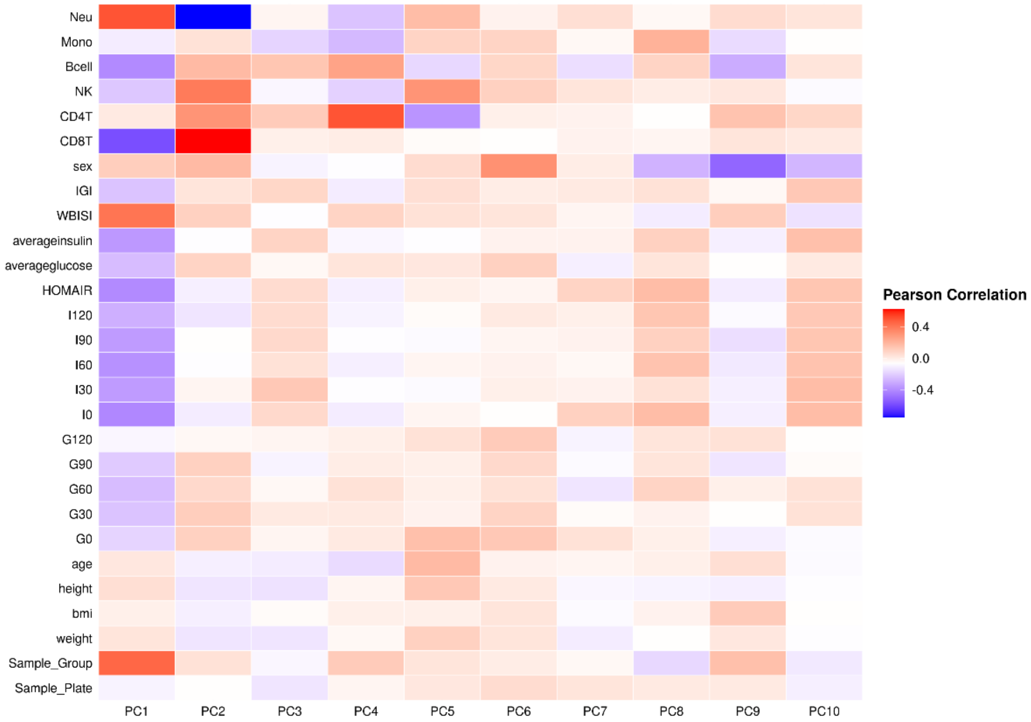


Figure 15: Heatmap of Pearson's correlation between clinical features and Principal Components. *The heatmap shows Pearson's correlation computed between clinical features and PCs derived from genome-wide methylation profiles in color scale from blue (inverse correlation) to red (direct correlation) going through white (no correlation).*

Differential methylation analysis

Probe-wise differential methylation was performed on 780842 single loci, producing a total of 280095 statistically significant sites (BH adjusted p-value < 0.05; 29946 both protein and non-protein coding) and 500747 non-significant sites (**Table 4** - **Table 5**). In this analysis the comparison performed (IR-IS) can be divided in 132367 hypermethylated DMPs and 147728 DMPs hypomethylated in IR samples. **Table 6** and **Figure 16** shows the top 10 DMPs ranked by BH-adjusted p-value, a complete list of differentially methylated probes with complete annotation can be found in the *Supplementary results* section.

Region-wise differential analysis produced 41305 regions of variable length ranging from 3 bp to 22659 bp containing a minimum of 2 CpGs and a maximum of 193 CpGs per region. The number of statistically significant regions according to Fisher's multiple comparison statistics (which is the default method) is 41246 (overlapping 20770 genes both protein and non-protein coding). A graphical representation of the top four DMRs can be seen in **Figure 17**, while a comprehensive table of the top 10 DMRs is displayed in **Table 7**. Some of the genes overlapping the top four DMRs are related to small nucleolar RNAs, MHC, inhibition of transcriptional activity, repressive histone methylation, RNA 28S sub-unit methylation and cell defense against toxic, carcinogenic, and pharmacologically active electrophilic compounds.

	IR - IS
Up	132367
Non-significant	500747
Down	147728

Table 4: Up/Down differential methylation. *The table summarizes the outcome of the differential statistical testing on probes for the comparison between IR and IS (baseline) samples. Non-significant refers to the number of probes with p-value≥0.05. Up and Down refer to the number of probes Hyper- or Hypo- methylated (p<0.05).*

p-value adjusted ranking	Number of DMPs
$< 1e^{-7}$	647
$< 1e^{-5}$	17219
$< 1e^{-3}$	102197
$< 1e^{-2}$	187696
$< 5e^{-2}$	280095

Table 5: Number of differentially methylated proves under a specific p-value threshold. *The table shows the number of DMPs below the specific p-value thresholds reported in the first column.*

Chr	Pos	Strand	CpG name	Log-FC	AvMeth	p-value	Adj p-val	Gene
chr19	42133492	-	cg19187564	-0.231	-2.675	1.06E-15	8.24E-10	CECAM4
chr1	158789404	+	cg27082765	-0.377	-4.063	2.13E-15	8.31E-10	-
chr22	31949147	+	cg02350510	-0.285	-2.017	6.71E-15	1.75E-09	-
chr2	218808843	-	cg13677779	-0.152	-1.217	2.69E-14	4.21E-09	TNS1
chr8	128746896	+	cg17160660	-0.309	-3.698	2.70E-14	4.21E-09	MYC
chr1	46268207	-	cg04984052	0.207	-1.111	3.25E-14	4.23E-09	MAST2
chr11	124647347	-	cg02682092	-0.370	-3.494	4.05E-14	4.52E-09	MSANTD2
chr13	52738164	-	cg01886593	-0.257	-0.229	5.14E-14	5.02E-09	-
chr22	16228563	+	cg03531177	0.190	0.282	6.71E-14	5.25E-09	LA16c-89F12.6
chr20	62588672	-	cg20593868	-0.205	-2.338	7.80E-14	5.25E-09	UCKL1

Table 6: Top 10 differentially methylated probes ranked by BH adjusted p-value (Adj p-val). *Each row describes a probe with Chromosome (Chr), mapping position (Pos), mapping strand (Strand), probe name (CpG name), logarithmic fold change (Log-FC), average methylation across all samples tested (AvMeth), raw p-value (p-value), BH adjusted p-value and gene mapping ("-" = not present).*

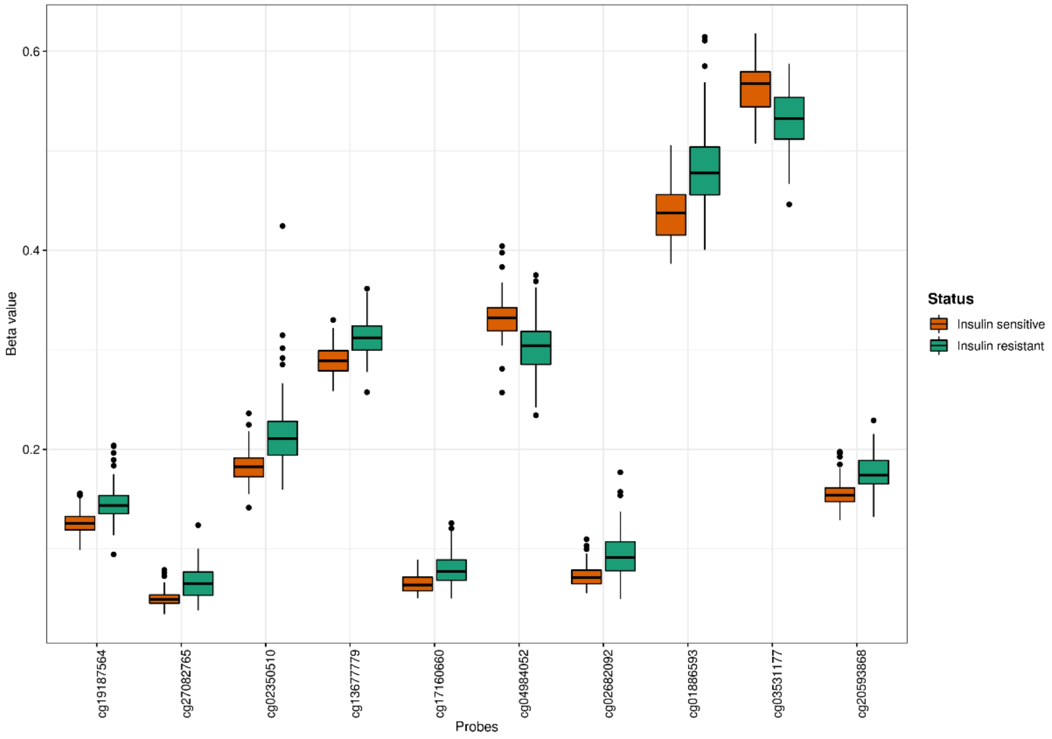


Figure 16: Variability of top 10 DMPs. *Differentially methylated probes reported in* ***Table 6*** *(x-axis) and their methylation value (β-value; y-axis). Orange boxes represent IR samples while green boxes represent IS samples.*

Chr	start	end	Width (bp)	# CpGs	Fisher p-value	IS vs IR	# Genes
chr6	32035188	32057846	22659	193	1.24E-127	+	2
chr6	32153575	32172871	19297	146	3.48E-89	+	4
chr12	16757954	16764406	6453	53	3.88E-89	-	2
chr6	31845110	31857100	11991	107	1.36E-77	+	3
chr6	31623842	31640160	16319	151	7.84E-68	+	8
chr2	66662218	66668012	5795	35	3.62E-64	-	3
chr6	30649909	30659692	9784	100	1.33E-62	-	2
chr6	31743769	31750174	6406	59	1.87E-62	+	3
chr12	7259717	7263232	3516	27	2.03E-62	-	2
chr6	33167187	33181870	14684	173	3.14E-61	+	6

Table 7: Top 10 differentially methylated regions. *DMRs grouping together several CpGs; each row contains chromosome (Chr), start position (hg19), end position, width of the region (in bp), number of CpGs within the region (# CpGs) and Fisher multiple comparison statistics (Fisher p-value). The column "IS vs IR" marks the result of the comparison between the two sample groups, "+" means that the region is Hypermethylated in IS samples while "-" means that the region is Hypermethylated in IR samples. The last column (# Genes) contains the number of genes overlapping the region (regardless if they are protein coding or not).*

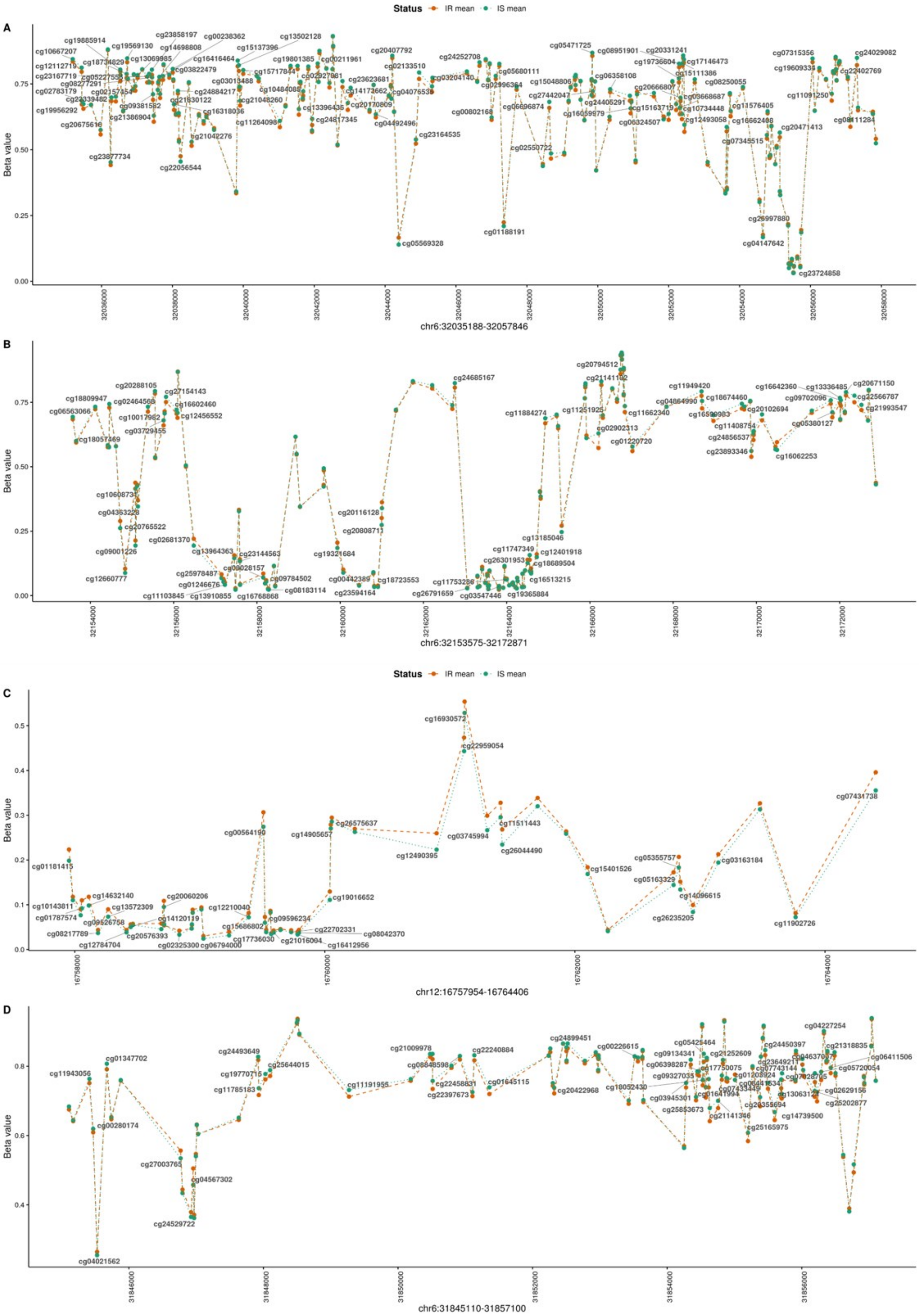


Figure 17: Top 4 DMRs. *All four panels (A; B; C; D) show the methylation landscape of the top 4 differentially methylated regions using genomic coordinates (x-axis) and β-values (y-axis). Orange dots correspond to single probes in the region, for IR samples while green dots display it for IS samples. Dashed lines connect the points for a better understanding of the methylation fluctuations within each region. Grey labels are present only for CpGs that present a statistically significant difference in methylation between the two groups.*

Gene set enrichment analysis

Gene set enrichment analysis was performed on genes covering all statistically significant differentially methylated probes (BH adjusted p-value < 0.05). As previously mentioned, the gene sets used are the following *Disease Ontologies*[39] (DO), *DisGenNet*[40–42] (DGN), *Kyoto encyclopedia of Genes and Genomes*[43–45] (KEGG), *Hallmark curated gene signatures*[46–48], *WikiPathways*[49] and *Reactome pathways*[50–52] (RP). Moreover, the minimum number of genes per gene set was limited to 50, therefore preventing the analysis from returning pathways with a small number of DMP-associated genes. In this section are highlighted only results for the *Disease Ontology* (**Figure 18**, **Figure 19** and **Table 8**) set where categories were selected among those relevant for the phenotype studied. Furthermore, only diseases/pathways sharing at least two genes are displayed to highlight their interactive nature. All complete and unfiltered lists of GSEA results (for all gene sets), as also all other figures are reported in the *Supplementary results* section.

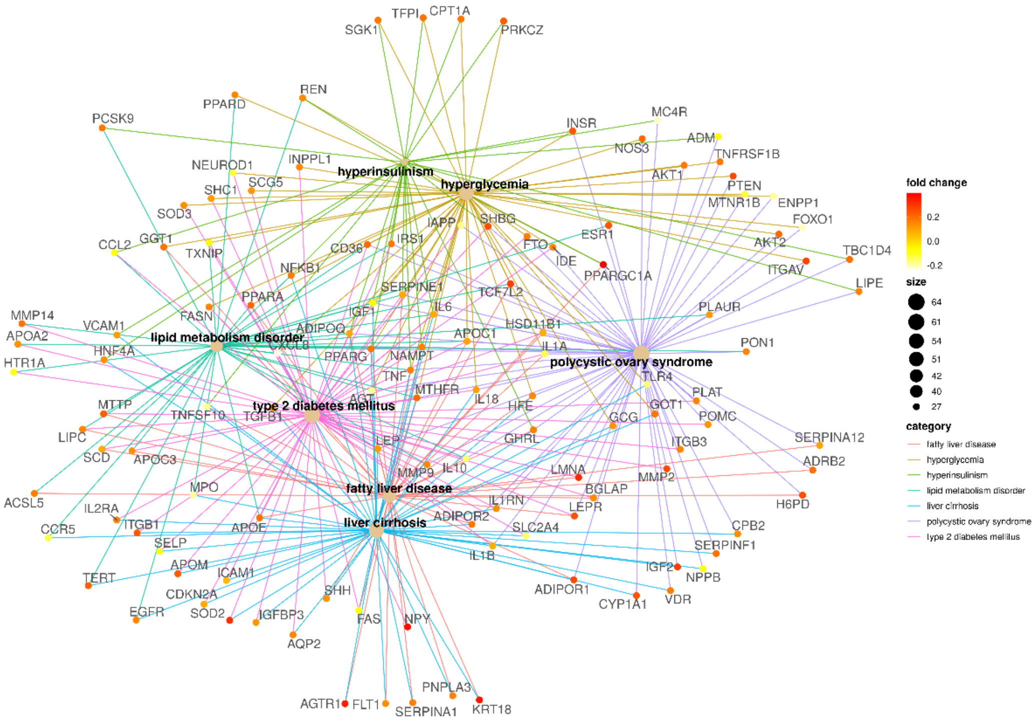


***Figure 18: Network visualization from Disease Ontologies.** The graph shows nodes corresponding to Disease Ontology (DO) categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the DO category, while gene nodes are colored according to the estimated fold change. The size of DO categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*

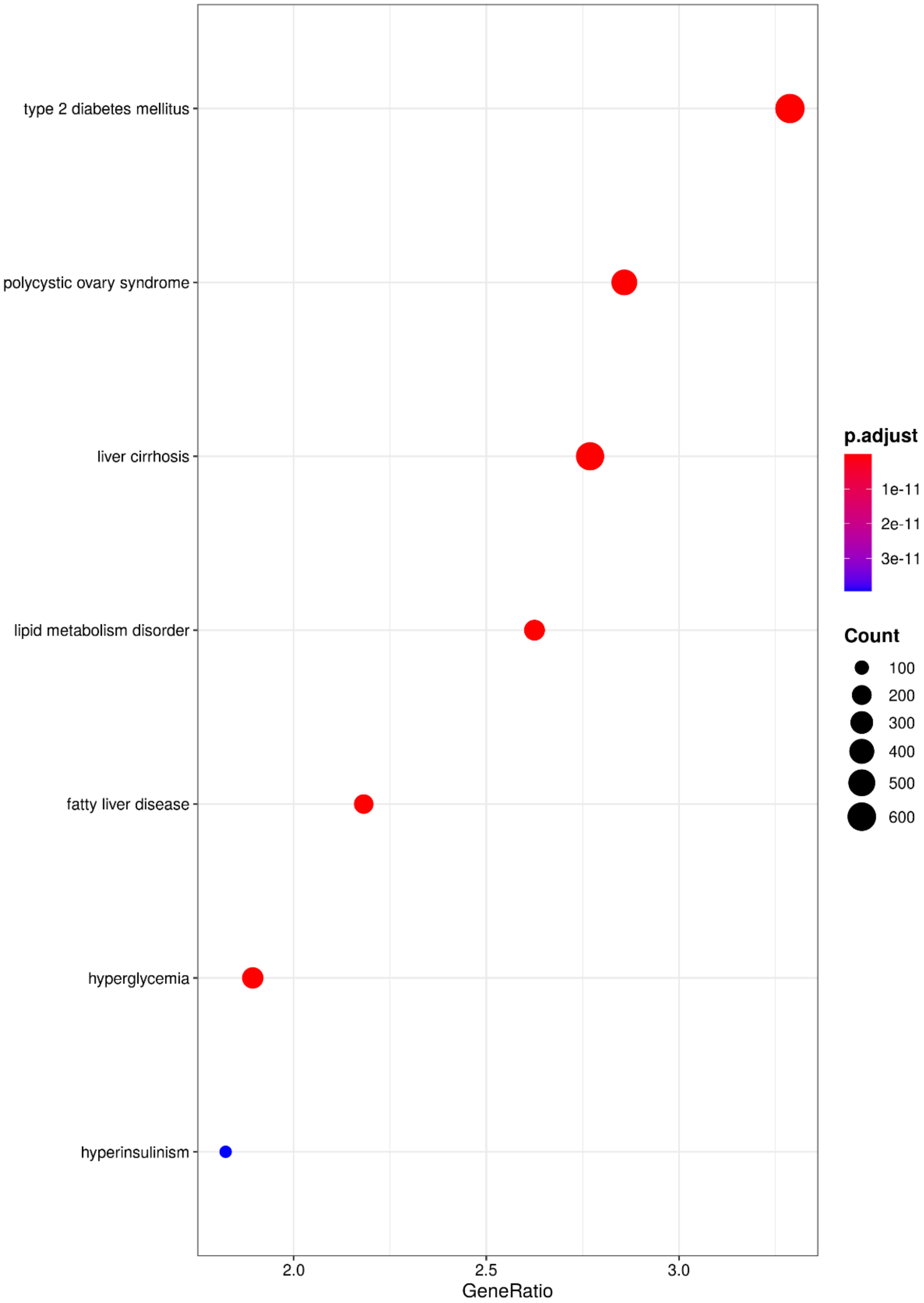


Figure 19: Disease Ontology gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis; GeneRatio) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value do not reflect the data shown but the total data from which this graph was derived.*

Ontology ID	Description	Set size	p-value	adj p-value
DOID:5082	liver cirrhosis	208	5.84E-52	6.52E-51
DOID:9352	type 2 diabetes mellitus	198	1.76E-51	1.81E-50
DOID:11612	polycystic ovary syndrome	155	8.27E-43	4.25E-42
DOID:4195	hyperglycemia	132	1.25E-27	3.02E-27
DOID:9452	fatty liver disease	88	1.85E-25	4.13E-25
DOID:3146	lipid metabolism disorder	88	1.02E-17	1.46E-17
DOID:2018	hyperinsulinism	51	7.26E-12	7.71E-12

Table 8: Disease Ontology GSEA results of IR related pathways. *Table showing results displayed in **Figure 18** and **Figure 19**. Each row shows the Ontology ID (DOID), a brief description of the pathway function, the total number of genes from the analysis within the pathway and, both raw p-value and adjusted p-value.*

DMP based samples clustering

As previously reported in **Table 5**, filtering DMP results by BH adjusted p-value < 1e^{-7} returns 647 CpGs. To investigate the association between these loci ("*super-significant")* and the disease status, their methylation values were used to cluster samples, (hierarchical complete clustering with Euclidean distance) as showed in **Figure 20** and **Figure 21**. By cutting the tree it was possible to generate two and three clusters (k=2, k=3). Generating two clusters (**Table 9** - **Figure 20**), eight IR samples formed a small community (red samples R45, R90, R94, R76, R36, R58, R41, R34 – Cluster1) while all the other samples formed a bigger one (black – Cluster2). Generating three clusters, the smaller community of IR samples (Cluster1) remained unchanged while the bigger cluster divided in two sub-clusters as shown in **Table 10** and **Figure 21**.

Cluster	# IR samples (R)	# IS samples (N)
1	8	0
2	87	91

Table 9: Samples per cluster for k = 2. *Summary of clustering results for k=2, with number of IR/IS samples per cluster **Figure 20**..*

Cluster	# IR samples (R)	# IS samples (N)
1	8	0
2A	18	66
2B	69	25

Table 10: Samples per cluster k = 3. *Summary of clustering results for k=3, with number of IR/IS samples per cluster **Figure 21**.*

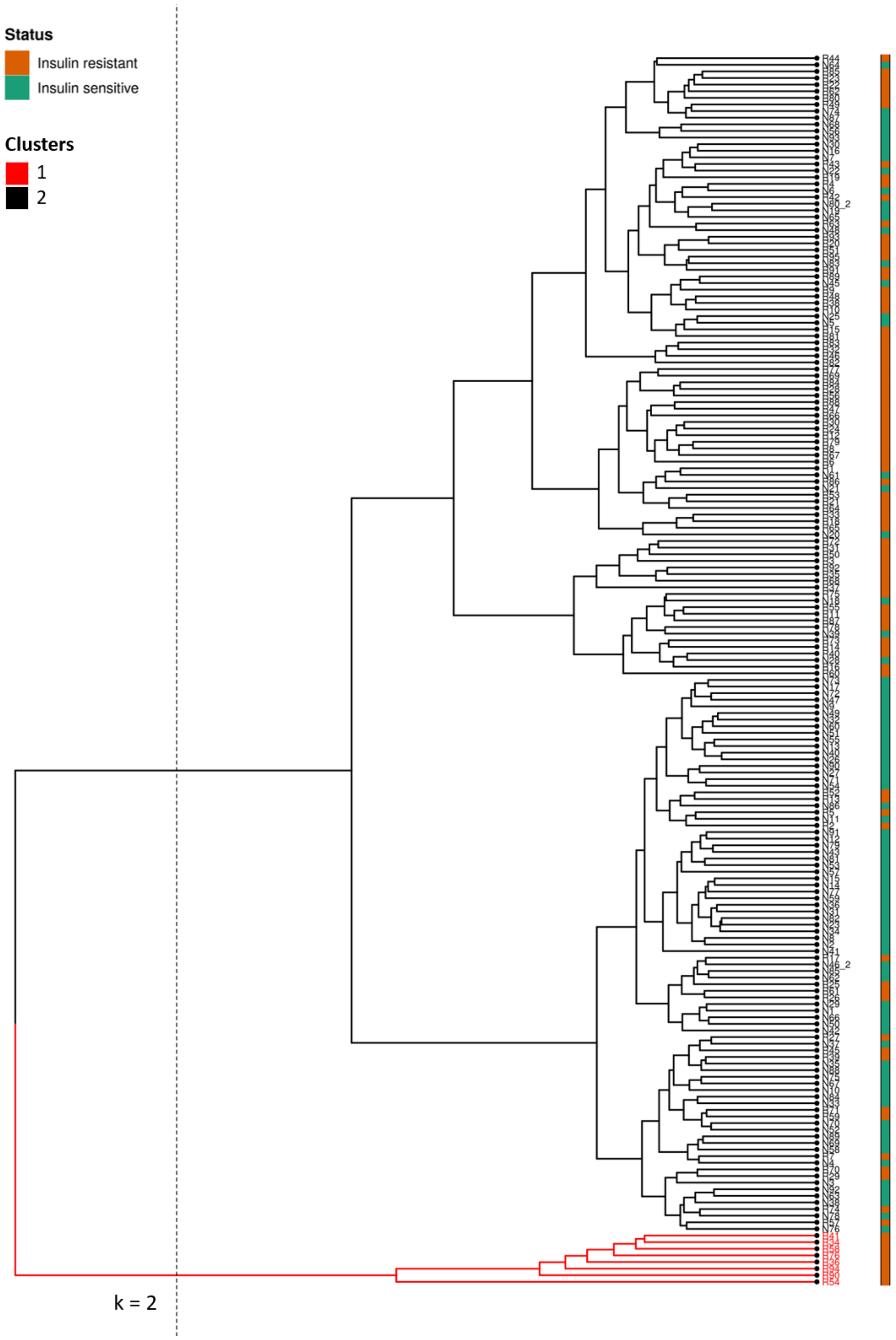


Figure 20: Hierarchical clustering on samples. *Dendrogram for hierarchical clustering showing the 2 clusters (red; black). The dashed line marks the cut performed to obtain 2 clusters (k=2). The orange/green bar on the right marks the distribution of samples by phenotype (IR in orange; IS in green).*

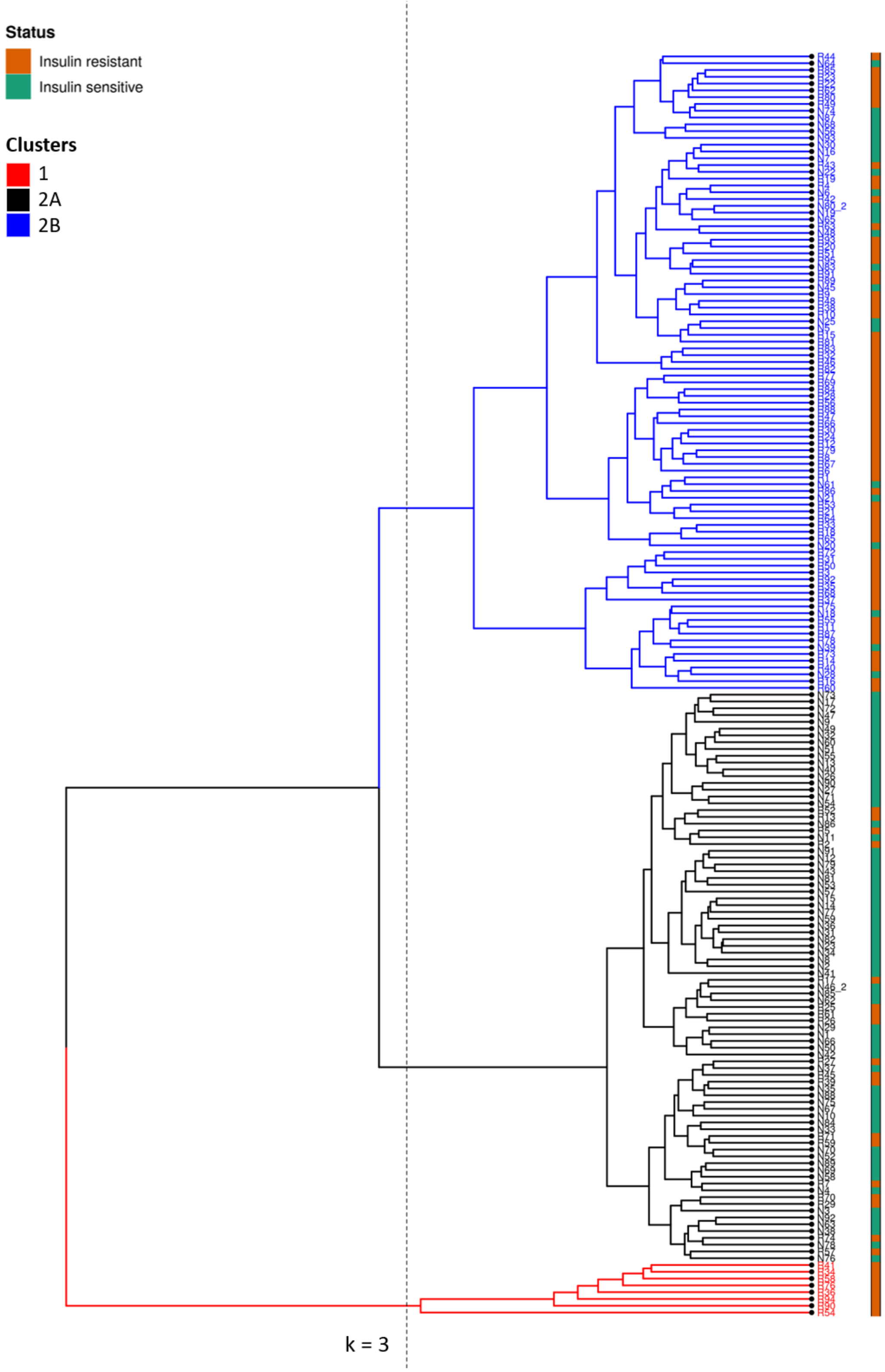


***Figure 21: Hierarchical clustering on samples**. Dendrogram for hierarchical clustering showing the 3 clusters (red; blue; black). The dashed line marks the cut performed to obtain the clusters (k=3). The orange/green bar on the right marks the distribution of samples by phenotype (IR in orange; IS in green).*

Welch's t-test and Fisher's exact test were employed to evaluate which clinical features are involved in the separation process. All statistical testing p-values were grouped together to allow for a better exploration of the data for both tests (**Table 12 - Table 13**). A table including all clustering-based comparisons performed is shown below (**Table 11**). An additional comparison between Cluster 1 and all other IR samples was performed to study their differences.

Clustering	Comparisons
k =2	Cluster 1 vs Cluster 2
k = 3	Cluster 1 vs Cluster 2A
	Cluster 2A vs Cluster 2B
	Cluster 1 vs Cluster 2B
-	Cluster 1 vs all other IR

Table 11: Clustering-derived comparisons performed. *Small table summarizing the comparisons performed starting from the two clustering (k =2; k=3), and a final comparison between Cluster 1 (8 IR samples) and all other IR samples.*

	p-value				
Features tested	**Cluster1 vs Cluster2**	**Cluster1 vs Cluster2A**	**Cluster2A vs Cluster2B**	**Cluster1 vs Cluster2B**	**Cluster1 vs all other IR**
Weight	0.4872	0.1733	0.6022	0.2309	0.8204
BMI	0.6006	0.1648	0.7783	0.3957	0.8925
Height	0.3915	0.2194	0.7116	0.3257	0.9556
Waist	0.6368	0.0669	0.5964	0.4438	0.5138
PAD	0.7079	0.5994	0.8373	0.6880	0.2807
PAS	0.3166	0.3919	0.3585	0.4223	0.5210
Age	0.9112	0.0681	0.5043	0.9878	0.8710
G_0	0.5708	0.4549	0.1356	0.5664	0.6065
G_{30}	0.0446	0.3987	0.0433	0.0261	0.1571
G_{60}	0.4126	0.9293	0.2006	0.1572	0.0317
G_{90}	0.7083	0.9947	0.0281	0.2599	0.9244
G_{120}	0.6723	0.6380	0.0407	0.4014	0.9924
I_0	4.58E-05	4.38E-05	8.19E-19	4.76E-05	0.3949
I_{30}	0.0011	0.0010	1.37E-06	9.94E-04	0.1575
I_{60}	0.0004	0.0004	3.87E-12	3.71E-04	0.5512
I_{90}	0.0016	0.0015	9.46E-08	1.49E-03	0.5698
I_{120}	0.0054	0.0025	0.0015	0.0046	0.6564
HOMA-IR	0.0002	0.0002	1.35E-19	1.72E-04	0.5839
Avg Glucose	0.2377	0.9265	0.0394	0.0631	0.2368
Avg Insulin	0.0003	0.0003	1.17E-10	2.38E-04	0.9346

WBISI	1.01E-06	7.10E-10	3.45E-06	7.49E-07	0.5043
IGI	0.0070	0.0022	0.0002	0.0064	0.0005
CD8+ T cells	0.0059	0.4527	0.0022	0.0021	0.0533
CD4+ T cells	0.1849	0.1665	0.8984	0.2522	0.7353
NK cells	0.0967	0.8409	0.1286	0.2638	0.2932
B cells	0.0560	0.0729	0.0351	0.0128	0.1333
Monocytes	0.4821	0.7371	0.0021	0.3972	0.4654
Neutrophils	0.0786	0.6402	0.0029	0.0256	0.3424

Table 12: T-test comparing clusters for all clinical features. *T-test p-values for each comparison are reported with red highlighting for statistically significant results (p-value<0.05).*

	p-value				
	Cluster1 vs Cluster2	**Cluster1 vs Cluster2A**	**Cluster2A vs Cluster2B**	**Cluster1 vs Cluster2B**	**Cluster1 vs all other IR**
Sample_Group	0.0068	0.1943	2.23E-12	1.68E-05	-
Gender	0.0238	0.0228	0.8802	0.0257	0.0228

Table 13: Fisher's test comparing clusters for categorical features. *Fisher's test p-values for each comparison are reported with red highlighting for statistically significant results (p-value<0.05). In the last column the comparison with Sample_Group was skipped since all samples tested belong to the same group (IR).*

Cross-referencing the *"super-significant"* CpGs with *EWAS catalogue*[57,58], 922 study-CpG association were found, covering only 290 of the 647 loci isolated from the differential analysis. Furthermore, the vast majority of these CpG are associated with different traits hence different phenotypes. Although the *EWAS catalogue* does not contain information regarding the specific pathways involved in their studies, it was still possible to look at the macroscopic function of these CpGs through the phenotypic trait reported. Moreover, **Table 14** reports phenotypes studied in the *EWAS catalogue* involving the associations found. For a complete list of CpG-study associations with phenotypic traits and other information, see the *Supplementary results* section.

Trait studied	# of Matches
Tissue	255
Gestational age	156
Smoking	65
HIV infection	49
Rheumatoid arthritis	43
Fetal vs adult liver	41
Clear cell renal carcinoma	29
Sex	29
Maternal smoking in pregnancy	28
Age	26
Primary Sjogrens syndrome	20
Schizophrenia	19
Age 4 vs age 0	15
Pancreatic ductal adenocarcinoma	11
Frontotemporal dementia	9
Body mass index	8
Alcohol consumption per day	7
Attention deficit hyperactivity disorder	5
Child abuse	4
Juice consumption	4
Progressive supranuclear palsy	4
Air pollution exposure	3
Ulcerative colitis	3
Ageing	2
Arm fat mass	2
C-reactive protein	2
Cholesterol esters in large VLDL	2
Hypertensive disorders of pregnancy	2
Inflammatory bowel disease	2
Lung function decline	2
Mean diameter for HDL particles	2
Substance use	2
Waist circumference	2

Table 14: Matching traits in the Epigenome-wide association study catalogue[57,58]**.** *Traits/phenotypes studied in the EWAS catalogue matching with most significant differentially methylated probes, with number of probes (matches) per trait.*

DISCUSSIONS

The main objective of this project is to investigate whether epigenetic modifications, and in particular DNA methylation, play a role on insulin resistance in clinically obese pediatric individuals.

It is known that the insulin resistance is associated with several factors including BMI, age, sex, obesity, and others. The subjects studied are children or pre-teenager (age range: 8 – 15 years old), with or without insulin resistance (IR vs IS), assessed by standard tests (OGTT and blood samples collection).

Since obesity is one of the main risk factors for insulin resistance in the general population, we focused the study on obese individuals to control the influence of obesity on IR. Genome-wide DNA methylation profiling was studied because epigenetic modifications are supposed to be important players in strict connection with individual genomic profiles, transcriptomic regulations, and the environmental conditions in many complex phenotypes, including IR.

The analysis investigated the methylation of 850k CpGs, using the Infinium Human Methylation EPIC microarray (by Illumina), in 186 obese children (95 IR and 91 IS).

The proportion of cell types in each individual was estimated using reference methylation profiles from blood cell types. A clear modulation of different cell types was observed between IR and IS. The different abundance of several cell types (i.e., B-cells, CD8, monocytes and neutrophils; see **Table 2**) suggested a possible role of general inflammation in insulin resistance. It could be hypothesized that common factors affecting IR contribute to cell type dysregulation, or that some cell populations contribute to disease or its severity. The present study cannot distinguish the true relationship between the two hypotheses. Furthermore, it would be interesting to study more in detail the impact of methylation in insulin resistance stratifying by sex although in the present study, the methylation component of sex was balanced out by design. It is noteworthy to mention, that the estimate of the cell type distributions derives from a mathematical model (deconvolution) and not from an analytical molecular assay (e.g., cytofluorimetry) and therefore these results should be interpreted carefully.

Interestingly, the analysis conducted using PCA method, showed a strong correlation of PC1 (main source of variability) with several IR-related parameters and distributions of 3 cell types (negative correlation: average glucose and insulin, CD8+ T lymphocytes, B cells, HOMA-IR; positive correlation: neutrophils, WBISI – see **Figure 15**). This confirms the previously observed association between IR and immune cell types and, suggests that the overall genome methylation (i.e., many CpG sites across the genomes) is associated with IR or related phenotypes.

Of note, PC1 and PC2 correlated with cell type amounts, but following an opposite strength. The contrasting but still present association shows that the 2 main source of PCA variability (PC1 and PC2) are both linked to cell populations. Specifically, the correlation involved CD8+ T cell, neutrophil, and NK cells (both PC1 and PC2). Although the PCA was conducted using autosome probes only, sex showed a correlation with PC6 and PC9 (opposite strength) suggesting that sex influences methylation status on many loci at genome level.

Differential methylation analysis was performed both on single probes (DMPs) or on CpG regions (DMRs), to study the association of IR with methylation of either single CpGs or chromosomal regions (of variable length).

The strongest 10 associations of IR with methylation of chromosomal regions (41305 regions in total overlapping 20770 genes both protein and non-protein coding; see *Supplementary results*) were observed to mainly map on 3 chromosomes (chromosome 6, 12 and 2; see **Table 7**). The genes mapping in the top 10 DMRs (35 genes including TNXB, PBX, NOTCH4, MGST1, EHMT2, APOM, MEIS1, PPP1R18, VWA, RING1, etc.), to the author's knowledge, have never been reported to be associated with insulin resistance. Additionally, DMR results are difficult to interpret since methylation affects genes in different locations (regulatory, coding, and inter-genic regions), making the connection of gene functions with the herein described methylation difficult. In the future it might be interesting to apply additional filtering (e.g., absolute Log-FC>0.6) to both DMPs and DMRs results retaining only probes/regions with higher difference between the two groups under the hypothesis that those probes/regions would be biologically more relevant in IR.

Since methylation is generally thought to play a role when multiple close CpGs are affected, we arbitrarily decided not to directly investigate the possible role of single CpGs and therefore we moved to study probe-associated genes under the hypothesis that they may belong to common functional pathways associated with IR.

Gene set enrichment analysis was performed on statistically significant DMPs (~280K), employing multiple gene sets (as described above in **Gene set enrichment analysis** section), showing a high number of pathways/diseases, with several of them connected to insulin resistance or related metabolic conditions (**Figure 18**, **Supplementary figure 4** - **Supplementary figure 8**).

With the aim to investigate the individual's methylation profile using the 647 most important DMPs (adjusted p-value<1e-7), we performed a hierarchical clustering (as described in the **DMP based sample clustering** section) of the 186 individuals regardless of IR status. The analysis showed 3 main clusters including 8 (cluster 1), 84 (cluster 2A), and 94 (cluster 2B) individuals, respectively (**Figure 21**). Cluster 1 was entirely made up of IR subjects, whereas Cluster 2A was mainly composed by IS individuals (IS frequency: 78.6%; 66 IS and 18 IR) and Cluster 2B by IR individuals (IR frequency: 73.4%; 25 IS and 69 IR). The 8 samples of cluster 1 may identify a subgroup of IR individuals presenting shared features associated to a common condition as disease severity that would justify their separation from the others, which may also be due to other unknown factors that were not taken into consideration during the study design such as ethnicity, macro- or micro- nutrient deficiencies, development related conditions (e.g. polycystic ovary syndrome), maternal gestational diabetes or others[59–61]. Epigenetics factors are likely to play in important role in IR since we observed the over-representation of IS and IR individuals in cluster 2A and 2B, respectively. Other important factors, such as gene variants and environmental/lifestyle factors, not investigated in this thesis should be taken into consideration together with methylation in the future studies. In addition, the 647 probes used for clustering, were also associated to traits reported in other studies, taken from the epigenome-wide association catalogue (EWAS catalogue[57,58]) supporting the hypothesis that other factors may need to be investigated when studying insulin resistance such as, gestational age, maternal smoking during pregnancy, and others (**Table 14**). Intriguingly, preliminary results not reported in this thesis, show that the overall genome-wide methylation of insulin resistant subjects is different from the insulin sensitive ones, suggesting the involvement of DNA methylation in IR individuals at genome level.

Finally, we point out that the approach employed, although accurate and reliable, can only detect DNA methylation and not all possible epigenetic modifications that might be involved in insulin resistance, which will be investigated in future studies.

CONCLUSIONS

In conclusion, we employed different bioinformatics strategies applied to a large cohort of individuals to study genome wide DNA methylation in IR. The results support the hypothesis that methylation plays a role in IR and that this condition is more related to general methylation landscape changes rather than methylation variations in a few loci. In the future, it might be interesting to replicate the results presented here on a different cohort as well as exploit the connection between epigenetics, genomics, transcriptomics, and IR to dissect the underlying mechanisms of this condition.

List of abbreviations

BH: Benjamini, Hochberg (method for p-value adjustment)

BMI: Body Mass Index

CGI: CpG island

detP: Detection p-value

DGN: DisGenNet (gene set)

DMP: Differentially methylated probe (sometimes L instead of P for Loci)

DMR: Differentially methylated region

DNA: Deoxyribonucleic acid

DNAm: DNA methylation

DNMT: DNA methyltransferase

DO: Disease ontology (gene set)

DOHaD: Developmental origin of health and disease

EPIC: Refers to: Illumina Infinium Methylation array EPIC

EWAS: Epigenome-wide association study

F: (in this work refers to) Female (sex)

FANTOM5: Functional Annotation of the Mammalian Genome (collaborative project for the identification of functional elements within the genome)

GENCODE: Scientific consortium for the annotation of genomic elements, part of the ENCODE effort (Encyclopedia of DNA elements).

GSEA: Gene set enrichment analysis

G_t: Glucose at timepoint “t” (during OGTT)

HOMA-IR: Homeostatic model assessment for insulin resistance

IGI: Insulinogenic index

IR: Insulin resistant

IS: Insulin sensitive

I_t: Insulin at timepoint “t” (during OGTT)

KEGG: Kyoto encyclopedia of genes and genomes

LINE: Long interspersed nuclear element

LogFC: Logarithmic fold change

LTR: Long terminal repeat

M: (in this work refers to) Male (sex)

NAFLD: Non-alcoholic fatty liver disease

NOS: Nitric oxidase synthase

OGTT: Oral glucose tolerance test

PAD: Diastolic arterial pressure

PAS: Systolic arterial pressure

PC: Principal component

PCA: Principal component analysis

RAA: Renin-angiotensin-aldosterone

RefSeq: Reference sequences database of the National Center for Biotechnology Information

RNA: Ribonucleic acid

RP: Reactome pathways (gene set)

SAM: S-adenosyl methionine

SBE: Single base extension

SINE: Short interspersed nuclear element

SNP: Single nucleotide polymorphism

SQN: Stratified quantile normalization

T1D: Type 1 diabetes

T2D: Type 2 diabetes

TSS: Transcription start site

VLDL: Very low-density lipoprotein

WBISI: Whole body insulin sensitivity index

WP: WikiPathways (gene set)

UTR: Untranslated region

List of figures

List of tables

Software information

All the analyses presented in this manuscript were carried out using *R-4.0.3* software for statistical computing. Many different packages were used to handle, process, and visualize data. The most important packages include:

minfi:
https://bioconductor.org/packages/release/bioc/html/minfi.html
limma:
https://bioconductor.org/packages/release/bioc/html/limma.html
RColorBrewer:
https://cran.r-project.org/web/packages/RColorBrewer/index.html
missMethyl:
http://bioconductor.org/packages/release/bioc/html/missMethyl.html
DMRcate:
https://bioconductor.org/packages/release/bioc/html/DMRcate.html
stringr:
https://www.rdocumentation.org/packages/stringr/versions/1.4.0
FlowSorted.BloodEPIC:
https://bioconductor.org/packages/release/data/experiment/html/FlowSorted.Blood.EPIC.html
IlluminaHumanMethylationEPICanno.ilm10b4.hg19:
https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylationEPICanno.ilm10b4.hg19.html
IlluminaHumanMethylationEPICmanifest:
https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylationEPICmanifest.html
ENmix:
https://bioconductor.org/packages/release/bioc/html/ENmix.html
sva:
https://bioconductor.org/packages/release/bioc/html/sva.html
ggplot2:
https://cran.r-project.org/web/packages/ggplot2/index.html

purrr:

https://cran.r-project.org/web/packages/purrr/index.html

broom:

https://cran.r-project.org/web/packages/broom/index.html

reshape2:

https://cran.r-project.org/web/packages/reshape2/index.html

pheatmap:

https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12

dedextend:

https://cran.r-project.org/web/packages/dendextend/index.html

irlba:

https://cran.r-project.org/web/packages/irlba/index.html

dplyr:

https://cran.r-project.org/web/packages/dplyr/index.html

ggpubr:

https://cran.r-project.org/web/packages/ggpubr/index.html

enrichplot:

http://bioconductor.org/packages/release/bioc/html/enrichplot.html

clusterProfiler:

https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

DOSE:

https://www.bioconductor.org/packages/release/bioc/html/DOSE.html

org.Hs.eg.db:

https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

Supplementary results

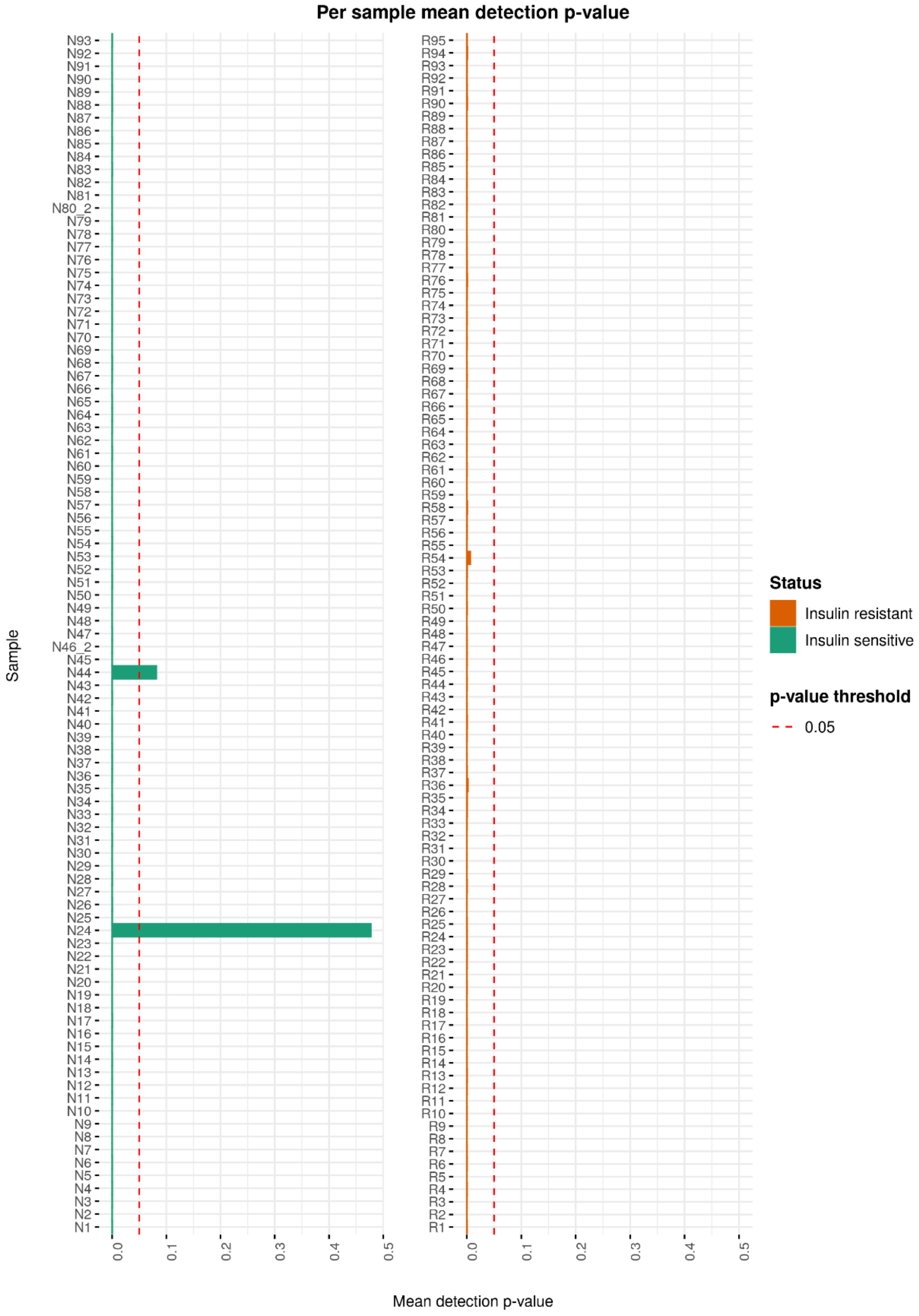
NOTE: Results for the following analyses:

Probe-wide differential methylation analysis, region-wise differential methylation analysis, Gene set enrichment analysis for all gene sets previously cited, EWAS catalogue cross-referencing (from the DMP based sample clustering section); **could not** be added to this manuscript due to size limits, they will be available at the following OneDrive link in pdf format (file size ~114MB):

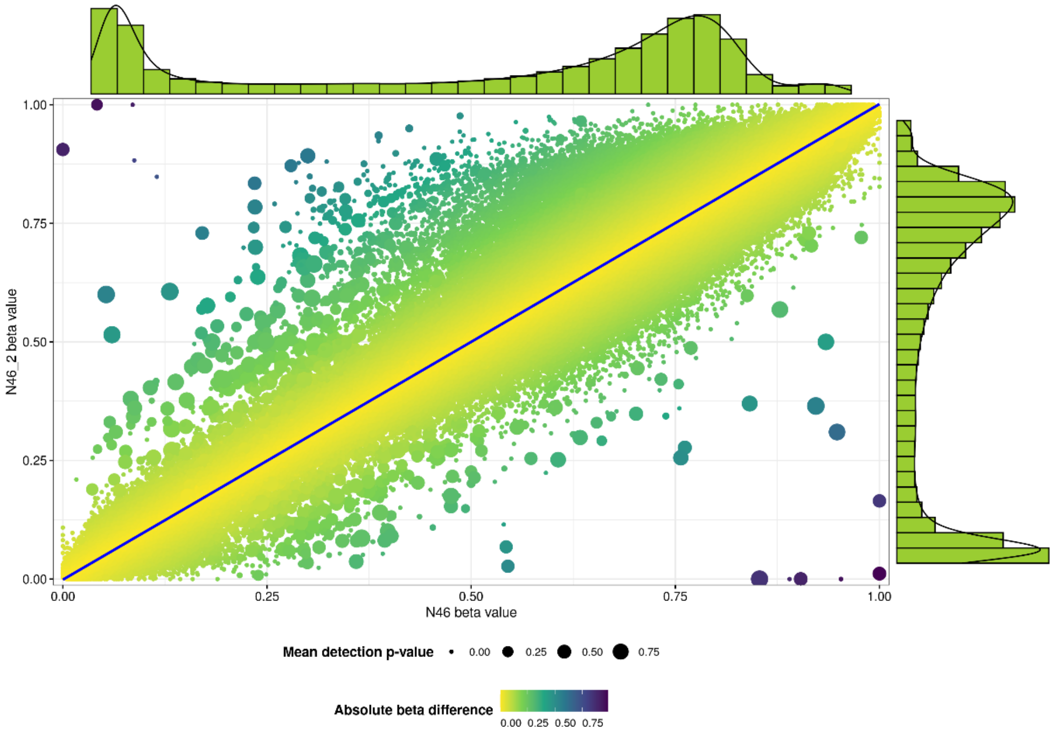
https://univr-my.sharepoint.com/:b:/g/personal/lucas_morondallator_univr_it/EXMB50I-3ytNgIDbzEROLyYBq8PbConmP3DWgVVt30FB7A?e=HuDP2Y

A complete list of the supplementary tables present in the external file can be found at page 59 of this document, accompanied by a brief description and page number.

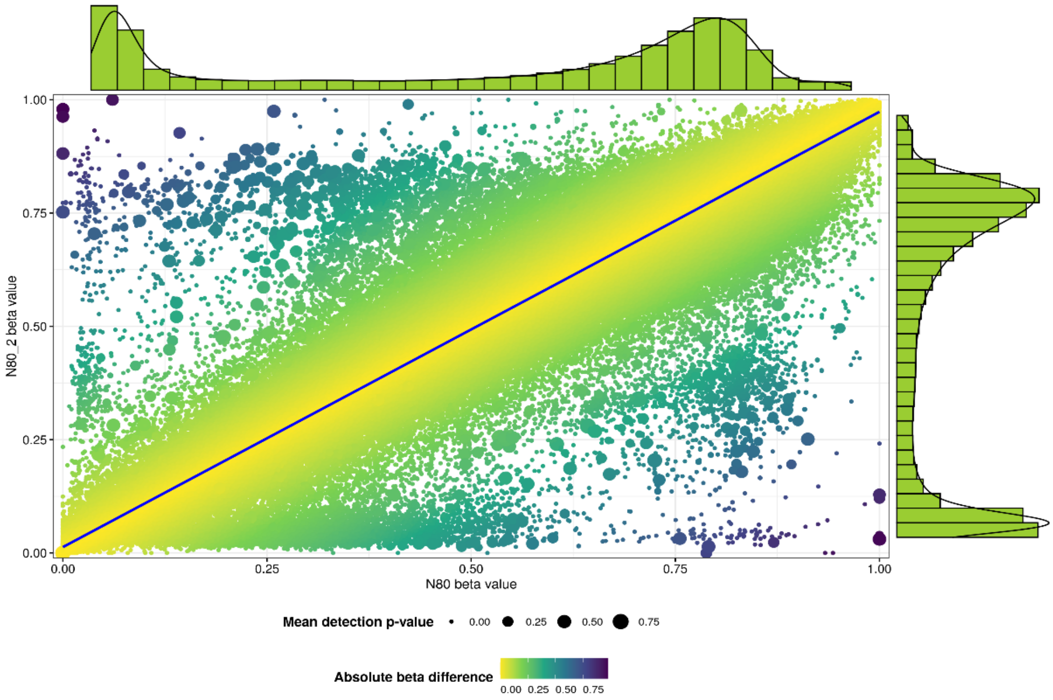
If the link is broken or non-functioning, please contact mdt.lucas@live.it and request a new link or a copy of the pdf document.



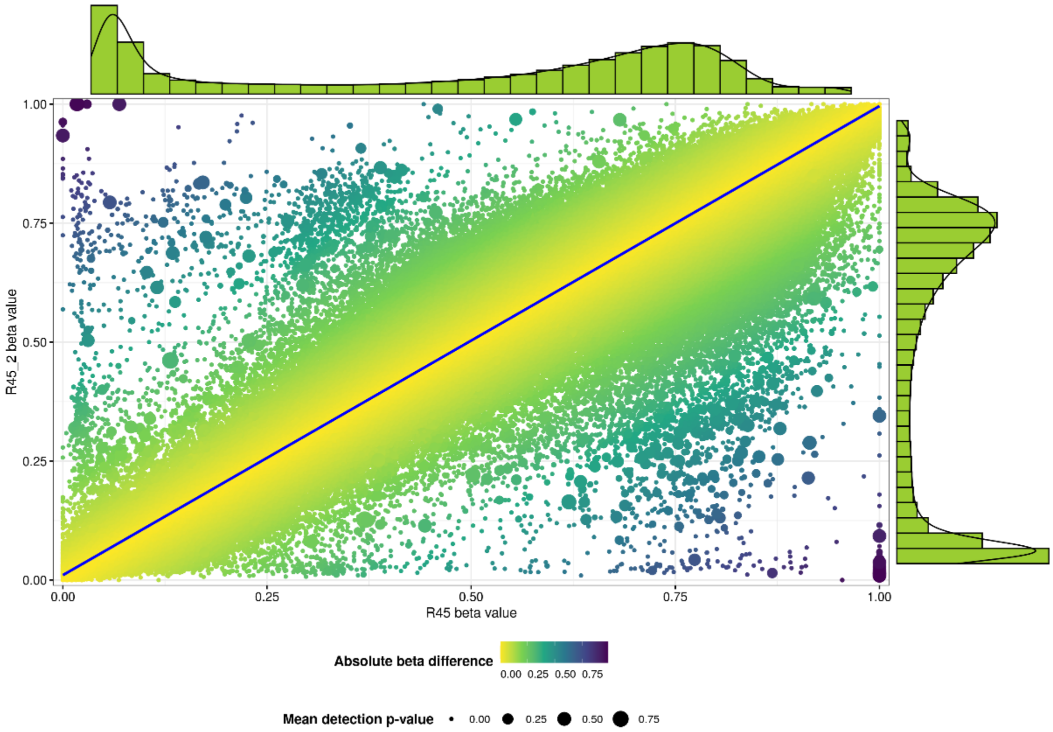
Supplementary table 1: Samples quality. *Quality control plot showing the mean detection p-value for each sample. The red line marks the threshold above which samples are considered low quality (p-value≥0.05). Samples are divided in insulin resistant (orange) and insulin sensitive (green).*



Supplementary figure 1: DNA methylation profiles of a duplicate sample. *The figure shows the DNAm value of a duplicated sample N46 (x-axis N46; y-axis N46_2). Marginal distributions show the overall methylation profiles for the pair. Smaller points represent lower detP; Darker points represent higher mean absolute difference between β-values (of the couple).*



Supplementary figure 2: DNA methylation profiles of a duplicate sample. *The figure shows the DNAm value of a duplicated sample N80 (x-axis N80; y-axis N80_2). Marginal distributions show the overall methylation profiles for the pair. Smaller points represent lower detP; Darker points represent higher mean absolute difference between β-values (of the couple).*



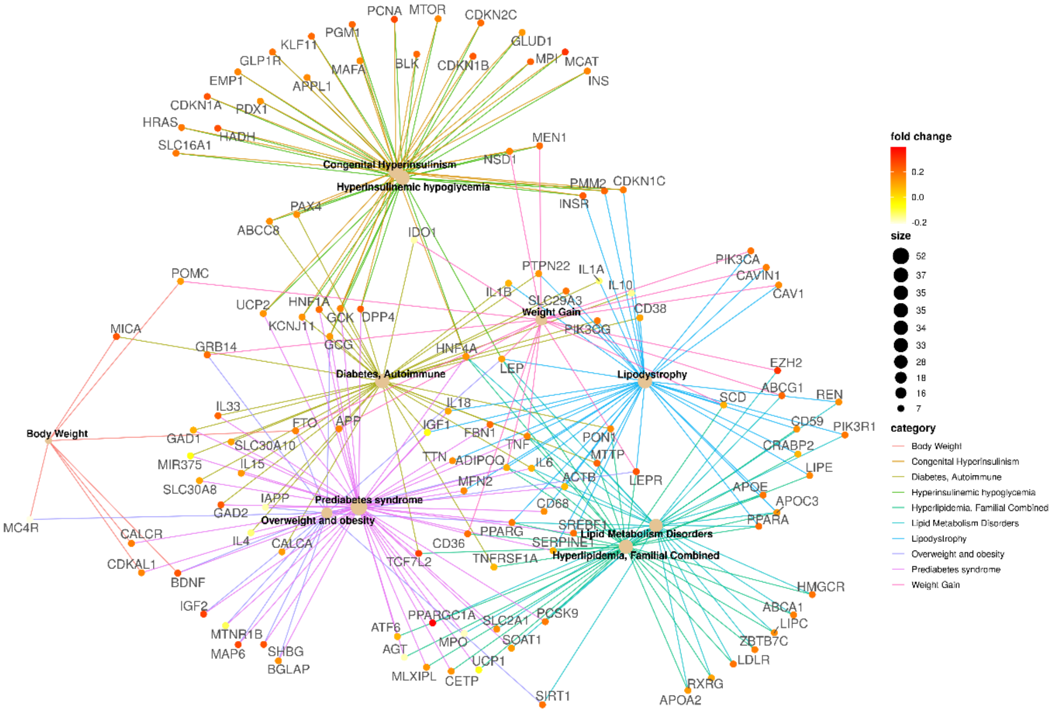
Supplementary figure 3: DNA methylation profiles of a duplicate sample. *The figure shows the DNAm value of a duplicated sample R45(x-axis R45; y-axis R45_2). Marginal distributions show the overall methylation profiles for the pair. Smaller points represent lower detP; Darker points represent higher mean absolute difference between β-values (of the couple).*

Cell type	Mean IR	Mean IS	t-statistics	p-value	degrees of freedom	CI - low	CI - high
B cells	0.0832	0.0640	4.4459	1.53E-05	178.7916	0.0107	0.0278
CD4+ T cells	0.1062	0.1108	-0.8836	0.3781	183.6445	-0.0150	0.0057
CD8+ T cells	0.1531	0.1321	2.6964	0.0077	183.5945	0.0056	0.0363
Monocytes	0.0678	0.0567	2.8976	0.0042	183.9883	0.0036	0.0187
Neutrophils	0.5143	0.5524	-2.8967	0.0042	183.5700	-0.0641	-0.0122
NK cells	0.0468	0.0449	0.4347	0.6643	183.6433	-0.0067	0.0104

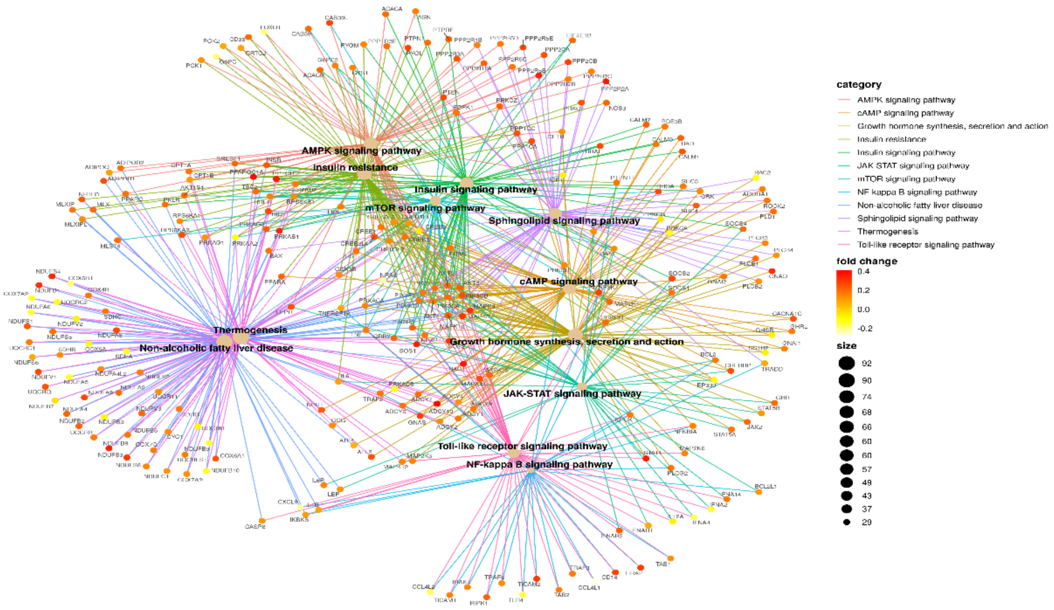
Supplementary table 2: Insulin resistance impact on cell composition. *Results of t-test comparing insulin resistance status to Cell type proportions estimated during deconvolution. Each row contains the cell type tested, mean values for both groups (IR/IS), t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

Cell type	Mean F	Mean M	t-statistics	p-value	degrees of freedom	CI - low	CI - high
B cells	0.0714	0.0757	-0.9449	0.3460	180.3480	-0.0133	0.0047
CD4+ T cells	0.1133	0.1045	1.6749	0.0957	175.6552	-0.0016	0.0192
CD8+ T cells	0.1409	0.1444	-0.4434	0.6580	181.5896	-0.0191	0.0121
Monocytes	0.0596	0.0646	-1.2735	0.2045	174.0514	-0.0128	0.0028
Neutrophils	0.5419	0.5256	1.2279	0.2210	183.6243	-0.0099	0.0426
NK cells	0.0409	0.0500	-2.1382	0.0338	183.0779	-0.0174	-0.0007

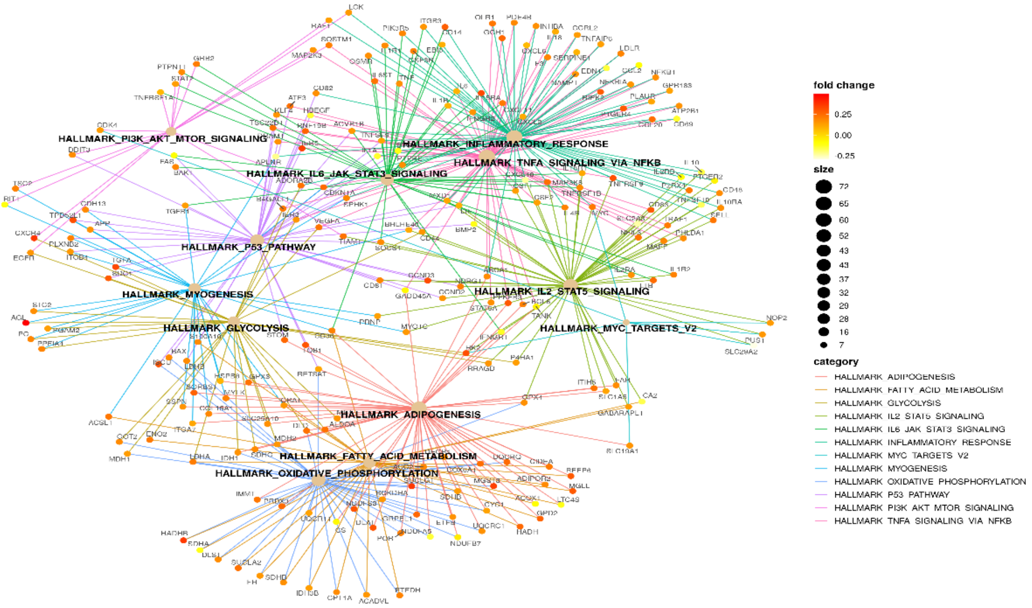
Supplementary table 3: Sex impact on cell composition. *Results of t-test comparing sex to Cell type proportions estimated during deconvolution. Each row contains the cell type tested, mean values for both groups (M/F), t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*



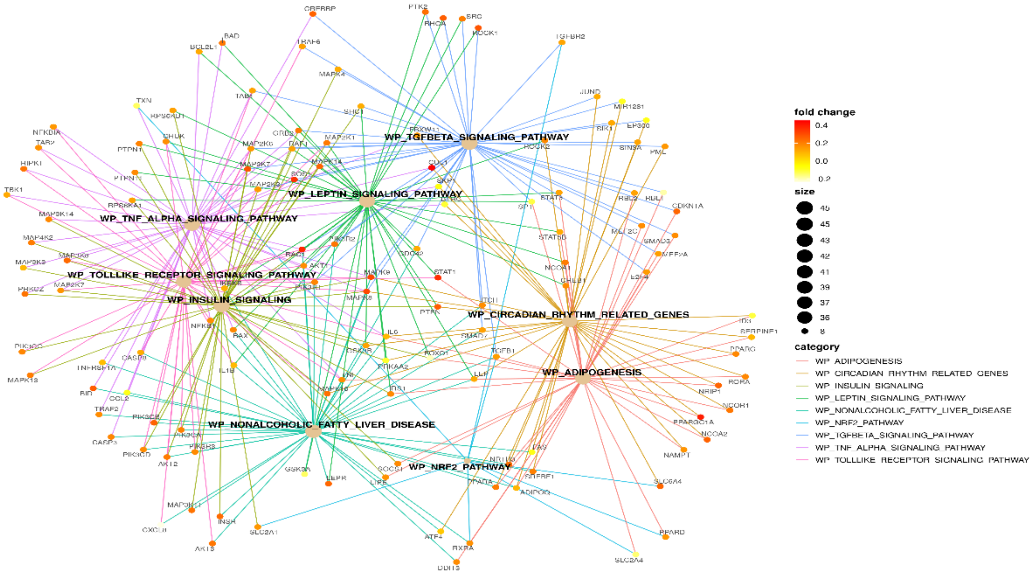
Supplementary figure 4: Network visualization from DisGenNet Ontologies. *The graph shows nodes corresponding to DisGenNet Ontology (DGN) categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the DGN category, while gene nodes are colored according to the estimated fold change. The size of DGN categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*


AMPK signaling pathway
Insulin resistance
Insulin signaling pathway
mTOR signaling pathway
Sphingolipid signaling pathway
cAMP signaling pathway
Thermogenesis
Non-alcoholic fatty liver disease
Growth hormone synthesis, secretion and action
JAK-STAT signaling pathway
Toll-like receptor signaling pathway
NF-kappa B signaling pathway
category
AMPK signaling pathway
cAMP signaling pathway
Growth hormone synthesis, secretion and action
Insulin resistance
Insulin signaling pathway
JAK STAT signaling pathway
mTOR signaling pathway
NF kappa B signaling pathway
Non-alcoholic fatty liver disease
Sphingolipid signaling pathway
Thermogenesis
Toll-like receptor signaling pathway
fold change
0.4
0.2
0.0
-0.2
size
92
90
74
68
66
60
60
57
49
43
37
29

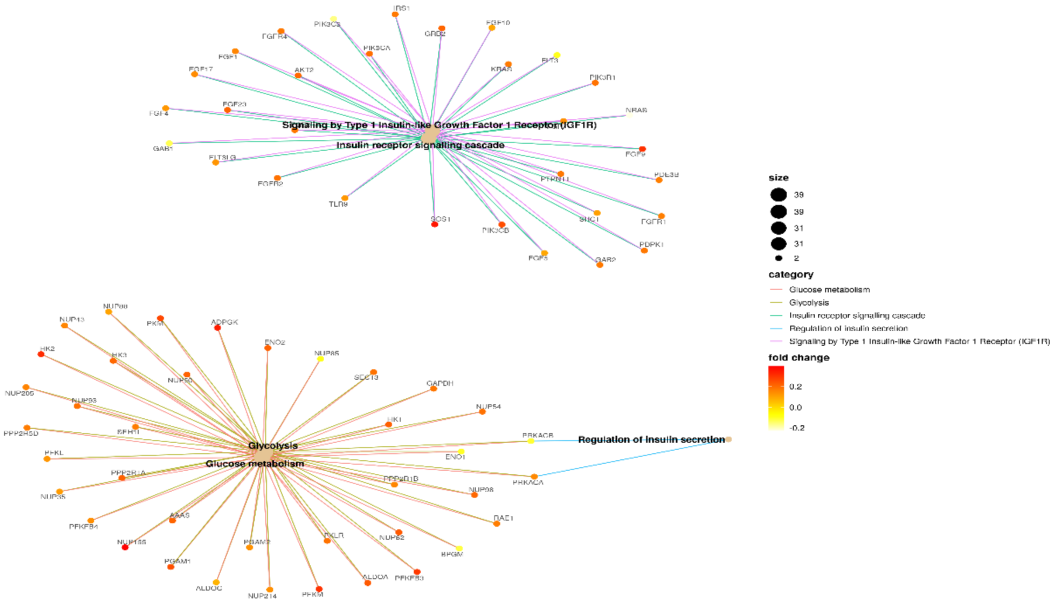
Supplementary figure 5: Network visualization from KEGG. *The graph shows nodes corresponding to KEGG Ontology categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the KEGG category, while gene nodes are colored according to the estimated fold change. The size of KEGG categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*


HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_P53_PATHWAY
HALLMARK_MYOGENESIS
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_GLYCOLYSIS
HALLMARK_MYC_TARGETS_V2
HALLMARK_ADIPOGENESIS
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_OXIDATIVE_PHOSPHORYLATION
fold change
0.25
0.00
-0.25
size
72
65
60
52
43
43
37
32
29
28
16
7
category
HALLMARK ADIPOGENESIS
HALLMARK FATTY ACID METABOLISM
HALLMARK GLYCOLYSIS
HALLMARK IL2 STAT5 SIGNALING
HALLMARK IL6 JAK STAT3 SIGNALING
HALLMARK INFLAMMATORY RESPONSE
HALLMARK MYC TARGETS V2
HALLMARK MYOGENESIS
HALLMARK OXIDATIVE PHOSPHORYLATION
HALLMARK P53 PATHWAY
HALLMARK PI3K AKT MTOR SIGNALING
HALLMARK TNFA SIGNALING VIA NFKB

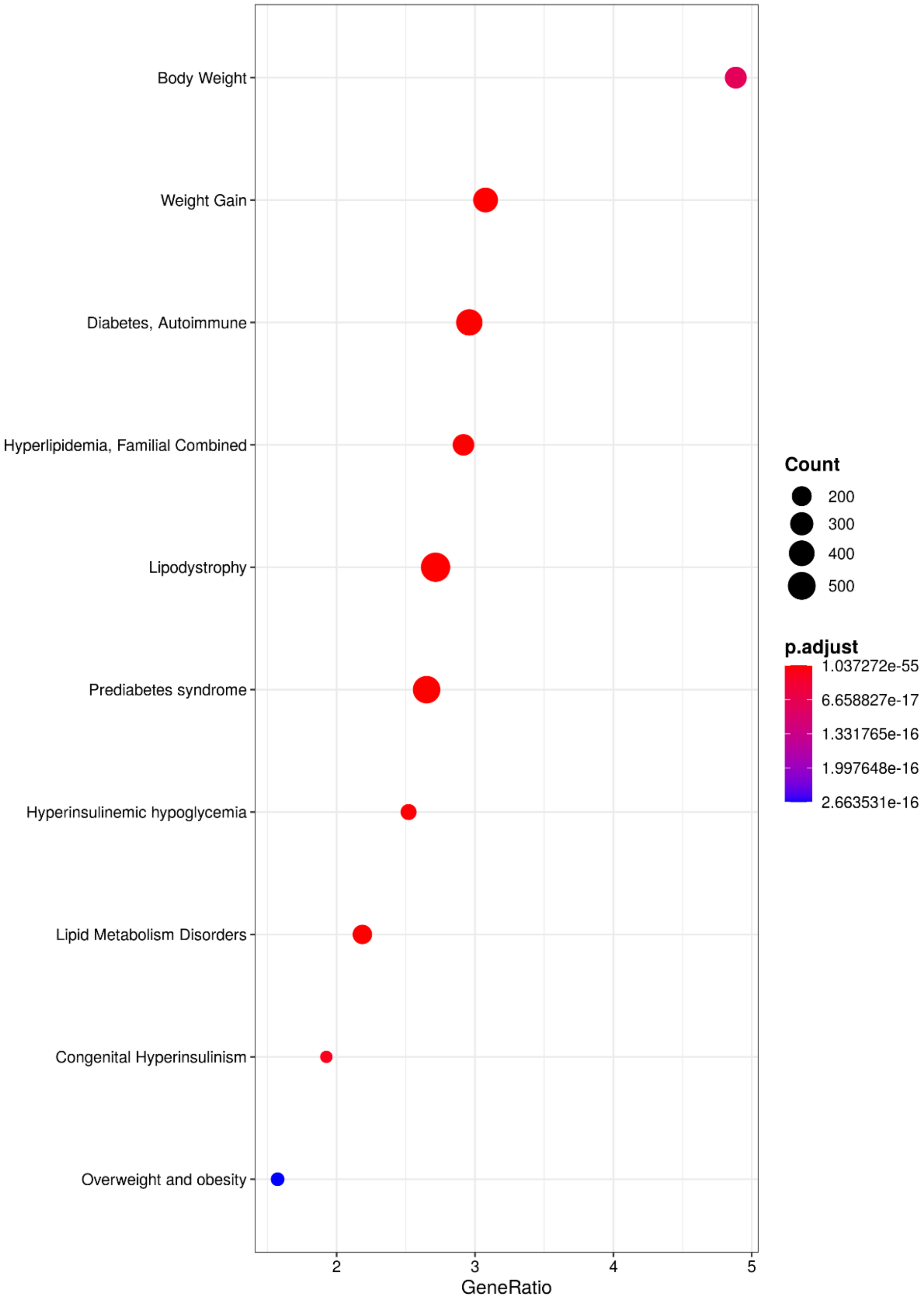
Supplementary figure 6: Network visualization from Hallmark gene Ontologies. *The graph shows nodes corresponding to mSigDB Hallmark gene Ontology categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the category, while gene nodes are colored according to the estimated fold change. The size of categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*


WP_TGFBETA_SIGNALING_PATHWAY
WP_LEPTIN_SIGNALING_PATHWAY
WP_TNF_ALPHA_SIGNALING_PATHWAY
WP_TOLLLIKE_RECEPTOR_SIGNALING_PATHWAY
WP_INSULIN_SIGNALING
WP_CIRCADIAN_RHYTHM_RELATED_GENES
WP_ADIPOGENESIS
WP_NONALCOHOLIC_FATTY_LIVER_DISEASE
WP_NRF2_PATHWAY
fold change
0.4
0.2
0.0
0.2
size
45
45
43
42
41
39
37
36
8
category
WP_ADIPOGENESIS
WP_CIRCADIAN_RHYTHM_RELATED_GENES
WP_INSULIN_SIGNALING
WP_LEPTIN_SIGNALING_PATHWAY
WP_NONALCOHOLIC_FATTY_LIVER_DISEASE
WP_NRF2_PATHWAY
WP_TGFBETA_SIGNALING_PATHWAY
WP_TNF_ALPHA_SIGNALING_PATHWAY
WP_TOLLLIKE_RECEPTOR_SIGNALING_PATHWAY

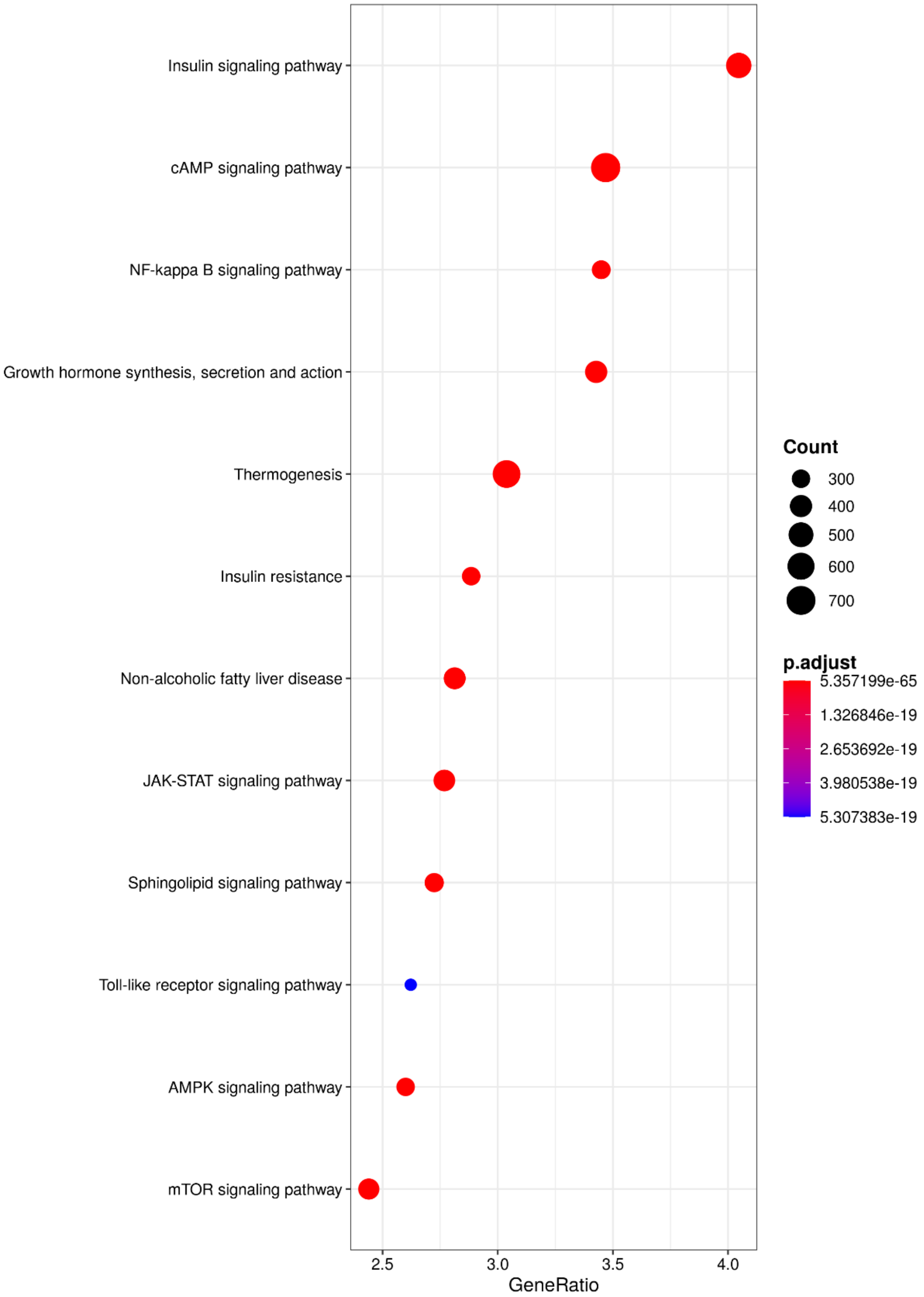
Supplementary figure 7: Network visualization from WikiPathways. *The graph shows nodes corresponding to WikiPathways (WP) categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the WP category, while gene nodes are colored according to the estimated fold change. The size of WP categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*


Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
Insulin receptor signalling cascade
Glycolysis
Glucose metabolism
Regulation of insulin secretion
size
39
39
31
31
2
category
Glucose metabolism
Glycolysis
Insulin receptor signalling cascade
Regulation of insulin secretion
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)
fold change
0.2
0.0
-0.2

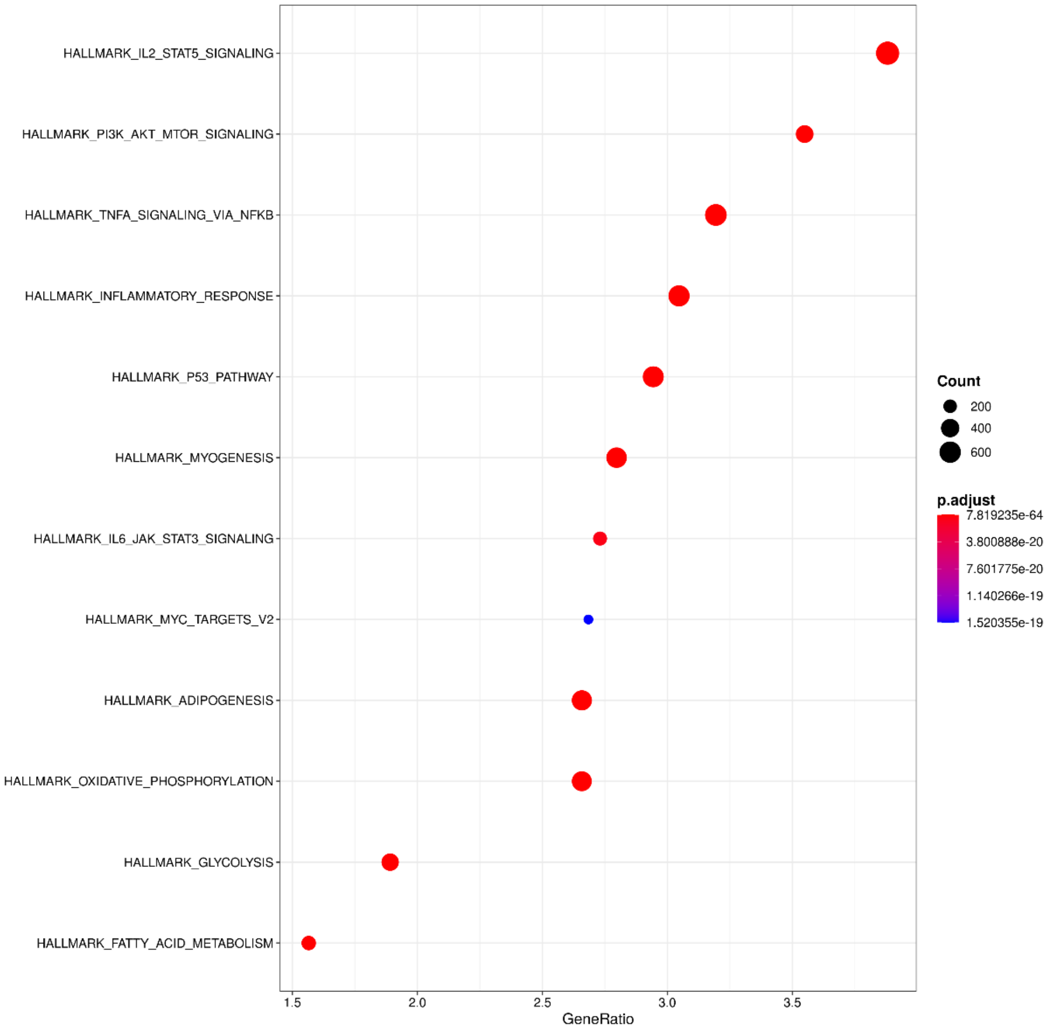
Supplementary figure 8: Network visualization from Reactome pathways. *The graph shows nodes corresponding to Reactome pathways (RP) categories (Bold; black) and their shared associated genes (grey). Each edge is colored according to the RP category, while gene nodes are colored according to the estimated fold change. The size of RP categories dots is proportional to the number of genes within that category. For visualization purposes only genes shared among at least two categories are displayed.*



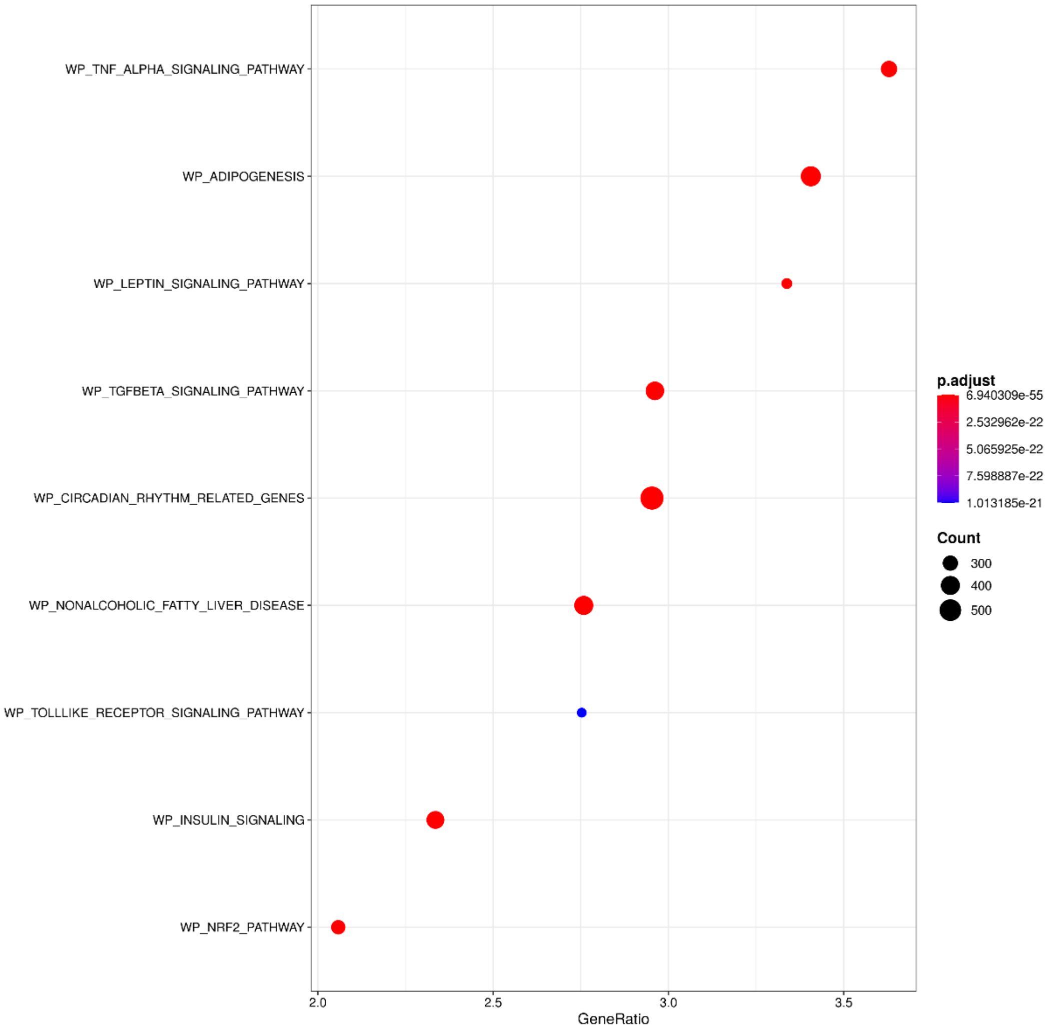
Supplementary figure 9: DisGenNet gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value don't reflect the data shown but the total data from which this graph was derived.*



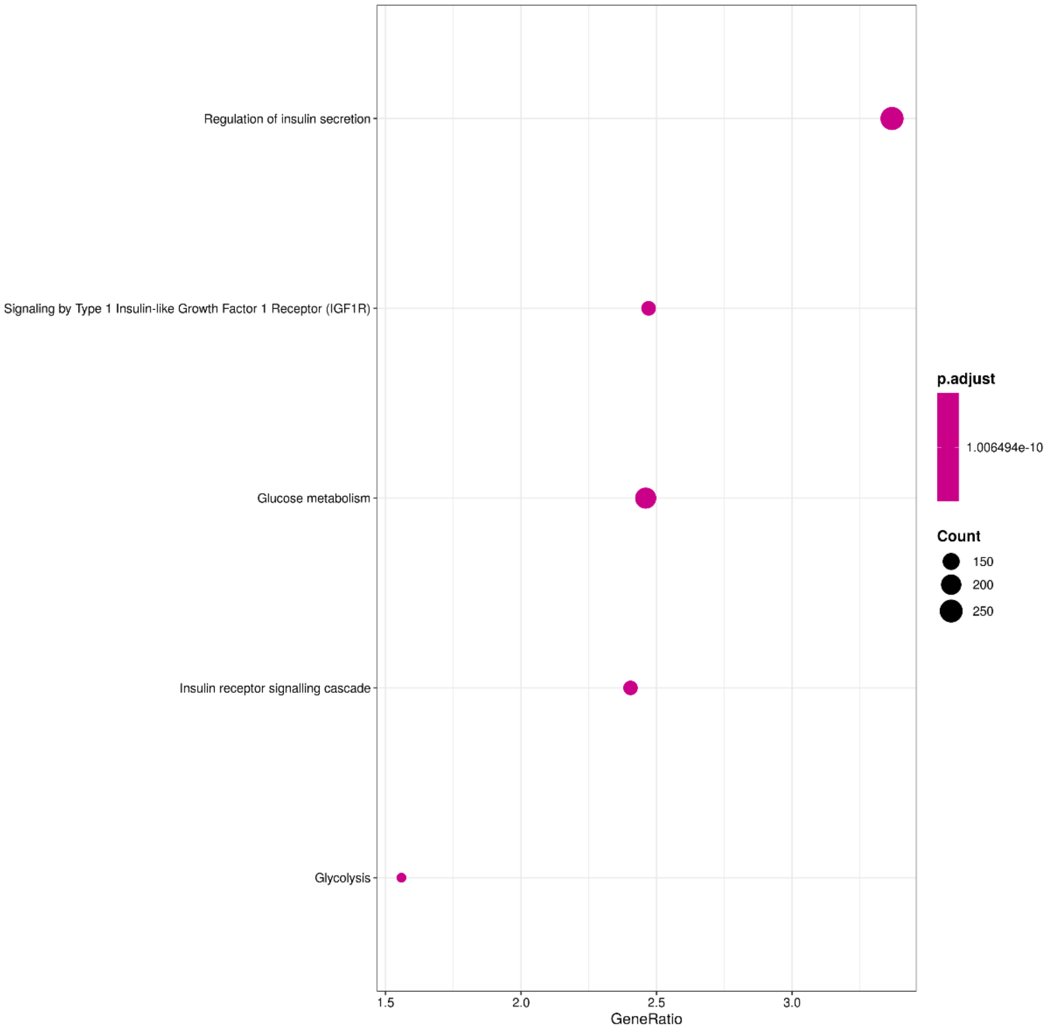
Supplementary figure 10: KEGG gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value don't reflect the data shown but the total data from which this graph was derived.*



Supplementary figure 11: mSigDB Hallmark Ontology gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value don't reflect the data shown but the total data from which this graph was derived.*



Supplementary figure 12: WikiPathways gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value don't reflect the data shown but the total data from which this graph was derived.*



Supplementary figure 13: Reactome Pathways gene-set enrichment analysis. *For each category (y-axis) the relative number of genes is showed (x-axis) while the absolute number is encoded in the dot size. Dots are colored according to the GSEA estimated adjusted p-value. Due to software limitations, legend scales for Count and adjusted p-value don't reflect the data shown but the total data from which this graph was derived.*

Feature tested	Mean Cluster1	Mean Cluster2	t-statistics	p-value	degrees of freedom	CI - low	CI - high
Weight	79.0625	73.6600	0.7144	0.4872	13.4426	-10.8798	21.6848
BMI	33.3798	32.1330	0.5365	0.6006	13.0865	-3.7706	6.2641
Height	1.5373	1.5044	0.8814	0.3915	15.5573	-0.0463	0.1120
Waist	94.0000	91.8333	0.4834	0.6368	13.1108	-7.5087	11.8421
PAD	70.1250	67.5714	0.3830	0.7079	12.9496	-11.8554	16.9625
PAS	115.3750	108.7143	1.0496	0.3166	10.8931	-7.3232	20.6446
Age	12.0595	12.1493	-0.1133	0.9112	15.6643	-1.7724	1.5928
G0	78.0000	75.0000	0.5790	0.5708	15.6811	-8.0014	14.0014
G30	137.0000	115.6000	2.1961	0.0446	14.6658	0.5889	42.2111
G60	122.8750	114.0000	0.8415	0.4126	15.8150	-13.5037	31.2537
G90	112.6250	108.4000	0.3819	0.7083	14.0058	-19.5028	27.9528
G120	108.0000	103.5000	0.4337	0.6723	11.7998	-18.1497	27.1497
I0	43.3625	8.4500	8.2259	4.58E-05	7.6629	25.0499	44.7751
I30	240.0500	64.7000	4.7845	0.0011	8.6805	91.9760	258.7240
I60	227.9375	43.4400	6.0381	0.0004	7.7576	113.6514	255.3436
I90	194.9625	46.2600	4.6738	0.0016	7.9137	75.1951	222.2099
I120	166.1250	57.7800	3.5864	0.0054	9.4859	40.5353	176.1547
HOMA-IR	8.4142	1.5511	6.9079	0.0002	7.4886	4.5445	9.1816
Avg glucose	111.7000	103.3000	1.2270	0.2377	15.9232	-6.1190	22.9190
Avg insulin	174.4875	44.1260	6.2003	0.0003	7.9345	81.8078	178.9152
WBISI	1.3170	6.4897	-10.7286	1.01E-06	9.7741	-6.2503	-4.0950
IGI	3.3982	1.4582	3.2575	0.0070	11.8033	0.6400	3.2401
CD8+ T cells	0.1896	0.1286	3.3170	0.0059	12.3436	0.0210	0.1009
CD4+ T cells	0.0933	0.1165	-1.3866	0.1849	15.7659	-0.0589	0.0124
NK cells	0.0517	0.0338	1.8214	0.0967	10.6479	-0.0038	0.0395
B cells	0.1047	0.0703	2.0781	0.0560	14.4641	-0.0010	0.0698
Monocytes	0.0534	0.0477	0.7196	0.4821	16.0000	-0.0111	0.0226
Neutrophils	0.4987	0.5671	-1.8842	0.0786	15.3637	-0.1457	0.0088

Supplementary table 4: Features impact on clustering. *Comparing Cluster 1 and 2 (k=2) with all clinical features and cell type proportions using t-test. Each row contains the feature tested, mean value for Cluster 1 and 2, t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

Feature tested	estimate	p-value	CI - low	CI - high
Sample_Plate	7.4239	0.0645	0.9232	340.5305
Sample_Group	0.0000	0.0068	0.0000	0.5829
Sex	0.1100	0.0238	0.0024	0.8849

Supplementary table 5: Features impact on clustering. *Comparing Cluster 1 and 2 (k=2) with all clinical features and cell type proportions using Fisher's test. Each row contains the feature tested, mean value estimated, p-value and both bounds of the confidence interval (CI).*

Feature tested	Mean Cluster1	Mean Cluster2A	t-statistics	p-value	degrees of freedom	CI - low	CI - high
Weight	79.0625	68.9600	1.4343	0.1733	14.0902	-4.9958	25.2008
BMI	33.3798	30.6799	1.4597	0.1648	15.1661	-1.2387	6.6384
Height	1.5373	1.4868	1.2812	0.2194	15.1086	-0.0334	0.1343
Waist	94.0000	86.5556	1.9790	0.0669	14.6860	-0.5885	15.4774
PAD	70.1250	67.0000	0.5400	0.5994	11.6117	-9.5298	15.7798
PAS	115.3750	110.3333	0.8927	0.3919	10.5837	-7.4493	17.5326
Age	12.0595	10.6519	1.9567	0.0681	15.9983	-0.1174	2.9326
G0	78.0000	81.2000	-0.7763	0.4549	10.3743	-12.3405	5.9405
G30	137.0000	126.9000	0.8671	0.3987	15.9542	-14.5980	34.7980
G60	122.8750	121.9000	0.0901	0.9293	15.6635	-21.9991	23.9491
G90	112.6250	112.7000	-0.0068	0.9947	13.9991	-23.7971	23.6471
G120	108.0000	113.0000	-0.4829	0.6380	11.7322	-27.6147	17.6147
I0	43.3625	7.9900	8.3591	4.38E-05	7.5756	25.5184	45.2266
I30	240.0500	55.6300	5.2143	0.0010	7.6021	102.1118	266.7282
I60	227.9375	44.6800	5.9724	0.0004	7.8826	112.3162	254.1988
I90	194.9625	42.7100	4.8668	0.0015	7.4168	79.1119	225.3931
I120	166.1250	39.8700	4.4520	0.0025	7.5401	60.1578	192.3522
HOMA-IR	8.4142	1.5840	6.9060	0.0002	7.3577	4.5144	9.1460
Avg glucose	111.7000	111.1400	0.0938	0.9265	15.4641	-12.1339	13.2539
Avg insulin	174.4875	38.1760	6.6661	0.0003	7.1233	88.1275	184.4955
WBISI	1.3170	6.2966	-18.2423	7.10E-10	11.5527	-5.5769	-4.3823
IGI	3.3982	1.0771	4.3511	0.0022	8.3212	1.0992	3.5430
CD8+ T cells	0.1896	0.1736	0.7709	0.4527	15.1260	-0.0282	0.0601
CD4+ T cells	0.0933	0.1193	-1.4528	0.1665	15.2800	-0.0643	0.0121
NK cells	0.0517	0.0543	-0.2040	0.8409	15.9942	-0.0302	0.0249
B cells	0.1047	0.0767	2.0188	0.0729	9.4024	-0.0032	0.0592
Monocytes	0.0534	0.0562	-0.3417	0.7371	15.9103	-0.0204	0.0148
Neutrophils	0.4987	0.4834	0.4788	0.6402	12.6470	-0.0540	0.0847

Supplementary table 6: Features impact on clustering. *Comparing Cluster 1 and 2A (k=3) with all clinical features and cell type proportions using t-test. Each row contains the feature tested, mean value for Cluster 1 and 2A, t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

Feature tested	estimate	p-value	CI - low	CI - high
Sample_Plate	7.1900	0.0613	0.8690	335.3593
Sample_Group	0.0000	0.1943	0.0000	1.7595
Sex	0.1078	0.0228	0.0023	0.8929

Supplementary table 7: Features impact on clustering. *Comparing Cluster 1 and 2A (k=3) with all clinical features and cell type proportions using Fisher's test. Each row contains the feature tested, mean value estimated, p-value and both bounds of the confidence interval (CI).*

Feature tested	Mean Cluster2A	Mean Cluster2B	t-statistics	p-value	degrees of freedom	CI - low	CI - high
Weight	74.2378	71.3300	0.5356	0.6022	11.7507	-8.9479	14.7634
BMI	32.1537	31.6436	0.2885	0.7783	10.9139	-3.3848	4.4050
Height	1.5102	1.4971	0.3792	0.7116	11.1802	-0.0628	0.0890
Waist	92.8500	91.0000	0.5411	0.5964	14.9880	-5.4372	9.1372
PAD	66.5571	67.5000	-0.2121	0.8373	8.0492	-11.1844	9.2987
PAS	115.7429	109.0000	0.9767	0.3585	7.6788	-9.2937	22.7794
Age	11.6016	12.0720	-0.6903	0.5043	11.0298	-1.9696	1.0289
G0	80.8936	75.0000	1.6260	0.1356	9.8183	-2.2030	13.9902
G30	127.2447	113.9000	2.2450	0.0433	12.6583	0.4675	26.2218
G60	119.0426	107.0000	1.3643	0.2006	10.6637	-7.4609	31.5460
G90	113.2128	100.9000	2.4736	0.0281	12.8755	1.5484	23.0771
G120	110.7872	99.2000	2.3189	0.0407	10.9780	0.5865	22.5879
I0	33.5335	8.3300	10.9446	8.19E-19	100.2953	20.6349	29.7721
I30	163.6517	62.2500	6.2324	1.37E-06	25.9209	67.9529	134.8505
I60	162.8007	46.3400	8.5732	3.87E-12	62.3336	89.3089	143.6125
I90	129.1704	44.5200	6.5521	9.46E-08	38.4551	58.5064	110.7945
I120	110.7851	54.5400	3.8336	0.0015	16.0342	25.1483	87.3420
HOMA-IR	6.6984	1.5266	11.3019	1.35E-19	100.3879	4.2640	6.0797
Avg glucose	110.2362	99.2000	2.3428	0.0394	10.8061	0.6452	21.4271
Avg insulin	119.9883	43.1960	8.4534	1.17E-10	42.6231	58.4677	95.1169
WBISI	2.8598	6.7982	-7.2003	3.45E-06	14.6984	-5.1064	-2.7705
IGI	3.2153	1.4249	4.2210	0.0002	27.9883	0.9215	2.6592
CD8+ T cells	0.1671	0.1154	3.8133	0.0022	12.8155	0.0224	0.0810
CD4+ T cells	0.1106	0.1123	-0.1307	0.8984	10.9601	-0.0300	0.0266
NK cells	0.0516	0.0384	1.6227	0.1286	13.0084	-0.0044	0.0307
B cells	0.0827	0.0593	2.3998	0.0351	11.0811	0.0020	0.0449
Monocytes	0.0681	0.0471	3.6447	0.0021	16.2809	0.0088	0.0332
Neutrophils	0.4901	0.5883	-3.7501	0.0029	11.6164	-0.1554	-0.0409

Supplementary table 8: Features impact on clustering. *Comparing Cluster 2A and 2B (k=3) with all clinical features and cell type proportions using t-test. Each row contains the feature tested, mean value for Cluster 2A and 2B, t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

Feature tested	estimate	p-value	CI - low	CI - high
Sample_Plate	1.0538	0.8816	0.5610	1.9818
Sample_Group	0.1004	2.23E-12	0.0464	0.2081
Sex	1.0624	0.8802	0.5620	2.0084

Supplementary table 9: Features impact on clustering. *Comparing Cluster 2A and 2B (k=2) with all clinical features and cell type proportions using Fisher's test. Each row contains the feature tested, mean value estimated, p-value and both bounds of the confidence interval (CI).*

Feature tested	Mean Cluster1	Mean Cluster2B	t-statistics	p-value	digrees of freedom	CI - low	CI - high
Weight	79.0625	71.3300	1.2477	0.2309	15.2723	-5.4567	20.9217
BMI	33.3798	31.6436	0.8752	0.3957	14.5210	-2.5044	5.9767
Height	1.5373	1.4971	1.0159	0.3257	15.0762	-0.0440	0.1243
Waist	94.0000	91.0000	0.7855	0.4438	15.8016	-5.1051	11.1051
PAD	70.1250	67.5000	0.4101	0.6880	13.8597	-11.1180	16.3680
PAS	115.3750	109.0000	0.8345	0.4223	10.6630	-10.5035	23.2535
Age	12.0595	12.0720	-0.0155	0.9878	15.5526	-1.7233	1.6983
G0	78.0000	75.0000	0.5859	0.5664	15.5662	-7.8798	13.8798
G30	137.0000	113.9000	2.4983	0.0261	13.4779	3.1964	43.0036
G60	122.8750	107.0000	1.4852	0.1572	15.7396	-6.8147	38.5647
G90	112.6250	100.9000	1.1896	0.2599	10.6897	-10.0447	33.4947
G120	108.0000	99.2000	0.8732	0.4014	10.8860	-13.4092	31.0092
I0	43.3625	8.3300	8.2882	4.76E-05	7.5421	25.1815	44.8835
I30	240.0500	62.2500	4.8364	0.0010	8.7768	94.3138	261.2862
I60	227.9375	46.3400	5.9108	0.0004	7.9208	110.6262	252.5688
I90	194.9625	44.5200	4.6966	0.0015	8.1177	76.7623	224.1227
I120	166.1250	54.5400	3.6496	0.0046	9.8699	43.3382	179.8318
HOMA-IR	8.4142	1.5266	6.9580	0.0002	7.3826	4.5713	9.2039
Avg glucose	111.7000	99.2000	1.9982	0.0631	15.8681	-0.7705	25.7705
Avg insulin	174.4875	43.1960	6.1932	0.0002	8.1835	82.5964	179.9866
WBISI	1.3170	6.7982	-11.1218	7.49E-07	9.7399	-6.5833	-4.3791
IGI	3.3982	1.4249	3.2936	0.0064	11.9854	0.6677	3.2789
CD8+ T cells	0.1896	0.1154	3.7441	0.0021	14.2984	0.0318	0.1166
CD4+ T cells	0.0933	0.1123	-1.1880	0.2522	15.9688	-0.0529	0.0149
NK cells	0.0517	0.0384	1.1617	0.2638	14.7629	-0.0111	0.0376
B cells	0.1047	0.0593	2.8716	0.0128	13.3857	0.0113	0.0795
Monocytes	0.0534	0.0471	0.8707	0.3972	15.4654	-0.0091	0.0216
Neutrophils	0.4987	0.5883	-2.4727	0.0256	15.3320	-0.1667	-0.0125

Supplementary table 10: Features impact on clustering. *Comparing Cluster 1 and 2B (k=3) with all clinical features and cell type proportions using t-test. Each row contains the feature tested, mean value for Cluster 1 and 2B, t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

Feature tested	estimate	p-value	CI - low	CI - high
Sample_Plate	7.5610	0.0590	0.9074	354.1202
Sample_Group	0.0000	1.68E-05	0.0000	0.1829
Sex	0.1147	0.0257	0.0024	0.9565

Supplementary table 11: Features impact on clustering. *Comparing Cluster 1 and 2B (k=3) with all clinical features and cell type proportions using Fisher's test. Each row contains the feature tested, mean value estimated, p-value and both bounds of the confidence interval (CI).*

Feature tested	Mean Cluster1	Mean all other IR	t-statistics	p-value	digrees of freedom	CI - low	CI - high
Weight	79.0625	80.8700	-0.2316	0.8204	13.1698	-18.6489	15.0339
BMI	33.3798	33.7112	-0.1378	0.8925	12.8091	-5.5358	4.8729
Height	1.5373	1.5392	-0.0566	0.9556	15.9654	-0.0750	0.0711
Waist	94.0000	96.3500	-0.6678	0.5138	15.9900	-9.8102	5.1102
PAD	70.1250	63.7500	1.1300	0.2807	11.9256	-5.9256	18.6756
PAS	115.3750	119.7500	-0.6613	0.5210	11.9627	-18.7955	10.0455
Age	12.0595	11.9548	0.1652	0.8710	14.9256	-1.2464	1.4558
G0	78.0000	80.2000	-0.5309	0.6065	10.5411	-11.3693	6.9693
G30	137.0000	124.0000	1.5158	0.1571	11.2885	-5.8172	31.8172
G60	122.8750	102.4000	2.3711	0.0317	14.8957	2.0584	38.8916
G90	112.6250	111.6000	0.0968	0.9244	12.8781	-21.8761	23.9261
G120	108.0000	108.1000	-0.0097	0.9924	11.4841	-22.5897	22.3897
I0	43.3625	38.4400	0.8754	0.3949	15.2943	-7.0424	16.8874
I30	240.0500	318.9860	-1.4831	0.1575	15.9871	-191.7730	33.9010
I60	227.9375	201.4970	0.6088	0.5512	15.9330	-65.6586	118.5396
I90	194.9625	173.2720	0.5835	0.5698	12.7546	-58.7800	102.1610
I120	166.1250	150.9500	0.4557	0.6564	12.5341	-57.0403	87.3903
HOMA-IR	8.4142	7.6791	0.5596	0.5839	15.1803	-2.0619	3.5321
Avg glucose	111.7000	105.2600	1.2431	0.2368	12.4181	-4.8053	17.6853
Avg insulin	174.4875	176.6290	-0.0836	0.9346	13.8878	-57.1553	52.8723
WBISI	1.3170	1.4137	-0.6832	0.5043	15.8667	-0.3967	0.2034
IGI	3.3982	6.2694	-4.7189	0.0005	12.4462	-4.1916	-1.5508
CD8+ T cells	0.1896	0.1475	2.1032	0.0533	14.5096	-0.0007	0.0849
CD4+ T cells	0.0933	0.0992	-0.3442	0.7353	15.6536	-0.0422	0.0304
NK cells	0.0517	0.0395	1.0913	0.2932	14.2647	-0.0117	0.0360
B cells	0.1047	0.0788	1.5933	0.1333	14.0560	-0.0089	0.0607
Monocytes	0.0534	0.0622	-0.7519	0.4654	13.0464	-0.0341	0.0165
Neutrophils	0.4987	0.5359	-0.9788	0.3424	15.8070	-0.1178	0.0434

Supplementary table 12: Features impact on clustering. *Comparing Cluster 1 and other IR samples with all clinical features and cell type proportions using t-test. Each row contains the feature tested, mean value for Cluster 1 and remaining IR samples, t-statistics, p-value, degrees of freedom and both bounds of the confidence interval (CI).*

	estimate	p-value	CI - low	CI - high
Sample_Plate	6.7305	0.0647	0.8095	314.8494
Gender	0.1079	0.0228	0.0023	0.8976

Supplementary table 13: Features impact on clustering. *Comparing Cluster 1 and remaining IR with all clinical features and cell type proportions using Fisher's test. Each row contains the feature tested, mean value estimated, p-value and both bounds of the confidence interval (CI).*

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Giovanni Malerba for believing in my research abilities and for his many valuable advices during my PhD.

I would also like to thank Professor Matteo Pellegrini, his team, and collaborators at the University of California Los Angeles for everything they taught me and for all the funny moments inside and outside the lab.

I also want to thank the teams of Professors Claudio Maffeis, Anita Morandi and Professor Emanuele Miraglia del Giudice for choosing me to carry out this inspiring and positively challenging project.

The completion of my doctorate could not have been successful without the support of my colleagues Dr. Michela Deiana, Dr. Samuele Cheri, Dr. Maria Carelli, Dr. Laura Veschetti, Dr. Marco Castelli, Dr. Elena Locatelli, Dr. Elisa De Tomi, Dr. Mirko Treccani and last but not least, Dr. Cristina Patuzzo.

Finally, I want to thank my mother, grandmother and best friends Giulia and Monica for being the best support system in the world.

References

1. Tomar, A. S. *et al.* Intrauterine Programming of Diabetes and Adiposity. *Curr. Obes. Rep.* **4**, 418–428 (2015).

2. Rosen, E. D. Epigenomic and transcriptional control of insulin resistance. *J. Intern. Med.* **280**, 443–456 (2016).

3. Mandy, M. & Nyirenda, M. Developmental Origins of Health and Disease: the relevance to developing nations. *Int. Health* **10**, 66–70 (2018).

4. Szabó, M., Máté, B., Csép, K. & Benedek, T. Epigenetic Modifications Linked to T2D, the Heritability Gap, and Potential Therapeutic Targets. *Biochem. Genet.* **56**, 553–574 (2018).

5. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2013).

6. Draizen, E. J. *et al.* HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants. *Database J. Biol. Databases Curation* **2016**, (2016).

7. Marabita, F., Tegnér, J. & Gomez-Cabrero, D. Introduction to Data Types in Epigenomics. in *Computational and Statistical Epigenomics* (ed. Teschendorff, A. E.) vol. 7 3–34 (Springer Netherlands, 2015).

8. Niculescu, M. D. & Zeisel, S. H. Diet, Methyl Donors and DNA Methylation: Interactions between Dietary Folate, Methionine and Choline. *J. Nutr.* **132**, 2333S-2335S (2002).

9. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

10. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).

11. Ghavifekr Fakhr, M., Farshdousti Hagh, M., Shanehbandi, D. & Baradaran, B. DNA Methylation Pattern as Important Epigenetic Criterion in Cancer. *Genet. Res. Int.* **2013**, 1–9 (2013).

12. Illingworth, R. S. *et al.* Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLOS Genet.* **6**, e1001134 (2010).

13. Illingworth, R. S. & Bird, A. P. CpG islands - ‘A rough guide’. *FEBS Lett.* **583**, 1713–1720 (2009).

14. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20**, 116–117 (1998).

15. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89**, 1827–1831 (1992).

16. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).

17. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).

18. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20**, 883–889 (2010).

19. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

20. Xie, C. *et al.* Differential methylation values in differential methylation analysis. *Bioinformatics* **35**, 1094–1097 (2019).

21. Matsuda, M. & DeFronzo, R. A. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care* **22**, 1462–1470 (1999).

22. Singh, B. & Saxena, A. Surrogate markers of insulin resistance: A review. *World J. Diabetes* **1**, 36–47 (2010).

23. Gutch, M., Kumar, S., Razi, S. M., Gupta, K. K. & Gupta, A. Assessment of insulin sensitivity/resistance. *Indian J. Endocrinol. Metab.* **19**, 160–164 (2015).

24. Stern, S. E. *et al.* Identification of Individuals With Insulin Resistance Using Routine Clinical Measurements. *Diabetes* **54**, 333–339 (2005).

25. Liu, J. & Siegmund, K. D. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics* **17**, 469 (2016).

26. Andrews, K. D. H., Martin Aryee, Rafael A. Irizarry, Andrew E. Jaffe, Jovana Maksimovic, E. Andres Houseman, Jean-Philippe Fortin, TimTriche, Shan V. *minfi*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.MINFI.

27. Infinium MethylationEPIC BeadChip Product Files. https://emea.support.illumina.com/array/array_kits/infinium-methylationepic-beadchip-kit/downloads.html.

28. Salas, L. A. *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.* **19**, 64 (2018).

29. Lucas A., D. C. *FlowSorted.Blood.EPIC*. (Bioconductor, 2018). doi:10.18129/B9.BIOC.FLOWSORTED.BLOOD.EPIC.

30. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

31. Torres, J. T. L. C., W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, John D. Storey, Yuqing Zhang, Leonardo Collado. *sva*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.SVA.

32. Baglama, J. & Reichel, L. Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. *SIAM J. Sci. Comput.* **27**, 19–42 (2005).

33. Baglama, J. & Reichel, L. Restarted block Lanczos bidiagonalization methods. *Numer. Algorithms* **43**, 251–272 (2007).

34. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

35. Choi, G. S., Yifang Hu, Matthew Ritchie, Jeremy Silver, James Wettenhall, Davis McCarthy, Di Wu, Wei Shi, Belinda Phipson, AaronLun, Natalie Thorne, Alicia Oshlack, Carolynde Graaf, Yunshun Chen, Mette Langaas, EgilFerkingstad, Marcus Davy, Francois Pepin, Dongseok. *limma*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.LIMMA.

36. Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics* **2**, 110–114 (1946).

37. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8**, 6 (2015).

38. Peters, T. *DMRcate*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.DMRCATE.

39. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).

40. Pinero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, bav028–bav028 (2015).

41. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).

42. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* gkz1021 (2019) doi:10.1093/nar/gkz1021.

43. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

44. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

45. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

46. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

47. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

48. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

49. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).

50. Wu, G. & Haw, R. Functional Interaction Network Construction and Analysis for Disease Discovery. *Methods Mol. Biol. Clifton NJ* **1558**, 235–253 (2017).

51. Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).

52. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

53. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).

54. *clusterProfiler*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.CLUSTERPROFILER.

55. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).

56. Dall'Olio, G., Y., Li-Gen Wang, Vladislav Petyuk. *DOSE*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.DOSE.

57. Battram, T. *et al.* The EWAS Catalog: a database of epigenome-wide association studies. (2021) doi:10.31219/osf.io/837wn.

58. MRC-IEU EWAS Catalog. http://www.ewascatalog.org/.

59. Holness, M. J. & Sugden, M. C. Epigenetic regulation of metabolism in children born small for gestational age: *Curr. Opin. Clin. Nutr. Metab. Care* **9**, 482–488 (2006).

60. Arpón, A. *et al.* Epigenome-wide association study in peripheral white blood cells involving insulin resistance. *Sci. Rep.* **9**, 2445 (2019).

61. Ling, C. & Rönn, T. Epigenetics in Human Obesity and Type 2 Diabetes. *Cell Metab.* **29**, 1028–1044 (2019).