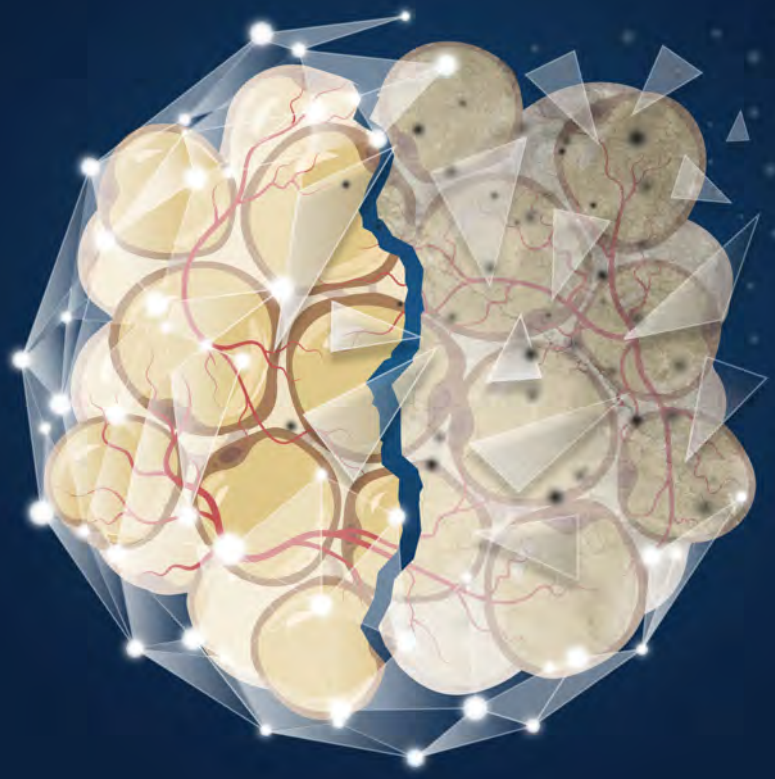
PROGRAMA DE DOCTORADO EN NUTRICION Y CIENCIAS DE LOS ALIMENTOS

Multi-Omics integration and machine learning for the identification of molecular markers of insulin resistance in prepubertal and pubertal children with obesity

AUGUSTO MIGUEL ANGUITA RUIZ



UNIVERSIDAD DE GRANADA

MULTI-OMICS INTEGRATION AND MACHINE LEARNING FOR THE IDENTIFICATION OF MOLECULAR MARKERS OF INSULIN RESISTANCE IN PREPUBERTAL AND PUBERTAL CHILDREN WITH OBESITY

Editor: Universidad de Granada. Tesis doctorales
Cover design & Layout: The Voice of Science, SL and Susana Izquierdo, Gráficas Granada, SL.

PROGRAMA DE DOCTORADO EN NUTRICION Y CIENCIAS DE LOS ALIMENTOS

Multi-Omics integration and machine learning for the identification of molecular markers of insulin resistance in prepubertal and pubertal children with obesity

AUGUSTO MIGUEL ANGUITA RUIZ

MULTI-OMICS INTEGRATION AND MACHINE LEARNING FOR THE IDENTIFICATION OF MOLECULAR MARKERS OF INSULIN RESISTANCE IN PREPUBERTAL AND PUBERTAL CHILDREN WITH OBESITY

Editor: Universidad de Granada. Tesis doctorales
Cover design & Layout: The Voice of Science, SL and Susana Izquierdo, Gráficas Granada, SL.

INTERNATIONAL DOCTORAL THESIS / TESIS DOCTORAL INTERNACIONAL

Multi-Omics integration and machine learning for the identification of molecular markers of insulin resistance in prepubertal and pubertal children with obesity

Integración multi-ómica y aprendizaje automático para la identificación de marcadores moleculares de resistencia a la insulina en niños prepúberes y púberes con obesidad

PROGRAMA DE DOCTORADO EN NUTRICION Y CIENCIAS DE LOS ALIMENTOS

DEPARTAMENTO DE BIOQUIMICA Y BIOLOGIA MOLECULAR II
FACULTAD DE FARMACIA
UNIVERSIDAD DE GRANADA

AUGUSTO MIGUEL ANGUITA RUIZ

2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Augusto Miguel Anguita Ruiz
ISBN: 978-84-1117-028-4
URI: http://hdl.handle.net/10481/70696

El doctorando D. Augusto Miguel Anguita Ruiz ha realizado la presente Tesis Doctoral Internacional como beneficiario de una beca-contrato con cargo al programa i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016 (referencia: IFI17/00048) del Instituto de Salud Carlos III, perteneciente al Ministerio de Ciencia e Innovación del Gobierno de España, por resolución de 05 de diciembre de 2017.

A mis padres, María José y Augusto, por inspirarme día a día, por enseñarme con su ejemplo el valor del esfuerzo, y por creer en mi incondicionalmente.

A Inés, mi compañera de camino, por hacerlo todo tan fácil.

A Chiqui, la mejor directora, maestra y amiga que podría haber deseado durante esta etapa.

"La ciencia es una empresa colectiva que abarca muchas culturas y se extiende a muchas generaciones.

En todas las épocas, y a veces en los lugares menos probables, hay gentes que desean apasionadamente comprender el mundo.

Gracias a la ciencia, hemos descubierto que las moléculas de la vida se forman fácilmente bajo condiciones comunes por todo el cosmos, hemos delineado mapas de máquinas moleculares en el corazón de la vida, hemos descubierto un microcosmos en una gota de agua, nos hemos asomado al caudal sanguíneo y al interior de nuestro planeta para ver la tierra como un solo organismo, hemos encontrado volcanes en otros mundos y explosiones en el sol, hemos escuchado los púlsares, y hemos buscado otras civilizaciones.

Por el momento, hemos caminado mucho… ¿o no?

No hay forma de saber de dónde va a surgir el próximo gran descubrimiento,

¿Cuál será el sueño de la mente que rehaga el mundo?"

"Cosmos: A Personal Voyage". CARL SAGAN

Table of contents

Research projects and funding

THE PRESENT DOCTORAL THESIS WAS FUNDED BY THE FOLLOWING INSTITUTIONS:

- Institute of Health Carlos III - Personal funding: "Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016, Project number: IFI17/00048".

- Institute of Health Carlos III - Health Research Fund (ERDF FUNDS) - Projects numbers PI05/1968, PI11/01425, PI11/02042, PI11/02059, PI16/01301, PI16/01205, PI16/00871 and PI20/00711.

- Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015).

- Mapfre Foundation ("Research grants by Ignacio H. de Larramendi 2017").

- ERDF/Regional Government of Andalusia/Ministry of Economic Transformation, Industry, Knowledge and Universities (grant number P18-RT-2248).

- University of Granada, Plan Propio de Investigación 2016, Excellence actions: Units of Excellence; Unit of Excellence on Exercise and Health (UCEES).

abstract

ABSTRACT

Abstract

CHILDHOOD OBESITY can develop early in life leading to the appearance of metabolic alterations such as insulin resistance (IR). If maintained during adulthood, obesity and IR usually derive into the development of more serious conditions like Type II Diabetes and cardiovascular disease, which considerably increase morbidity and mortality in affected populations. As many other complex disorders, obesity and its associated cardiometabolic comorbidities constitute a complex phenotype arising from the interaction between an at-risk molecular profile (involving genomics, transcriptomics, epigenomics and proteomics disturbances) and environmental exposures. On this sense, one of the most promising fields of research in obesity has involved the identification of early-life predictive molecular biomarkers able to stratify patients according to their risk for developing cardiometabolic complications later in life. Interestingly, the ideal and most robust biomarker discovery approach would involve the simultaneous analysis of multiple omics data layers at a once, allowing tracking a molecular disturbance from all its possible dimensions. Due to the complexity of omics data, nevertheless, new and innovative analytics approaches have been demanded. In the middle of this need, bioinformatics and artificial intelligence (AI) have experienced a remarkable boost due to their ability to automatically obtain descriptive or predictive models from massive amounts of data (Big Data). The present Doctoral Thesis gathers a series of research works in which bioinformatics and AI are conveniently applied to several obesity observational omics research projects for identifying new molecular biomarkers of IR and metabolic alterations in children and adolescents with obesity. Study populations are composed of more than 2000 Spanish children with ages ranging from 2-18 years. In summary, the results presented in the present Doctoral Thesis indicate that; 1) obesity is a complex disorder resulting from the interaction between genetic and environmental factors, 2) the creation of predictive tools based on the combination of small-risk effects genetic variants is an interesting but simple strategy for predicting future obesity, 3) multi-omics research approaches in obesity are necessary to understand the complex molecular mechanisms underlying disease, and 4) the application of eXplainable Artificial Intelligence (XAI) machine learning (ML) models can help us to unravel the complex relationships between omics molecular elements. The application of multi-omics research approaches and the use of complex analytical tools (such as bioinformatics and AI) are the correct way for approaching a true implementation of a personalized care in obesity. Further studies like those presented in the present Doctoral Thesis and as well as larger cohorts projects should be encouraged in order to validate presented findings. This will require a close collaboration between clinicians and basic researchers, and the creation of multidisciplinary teams, in which the presence of mixed bioinformatics profiles will be of great importance.

Keywords

Adolescent; Child; DNA Methylation; Epigenetics; Epigenome-Wide Association Study, EWAS; Gene expression; Genetics; Genome-Wide Association Study; GWAS; Insulin resistance; Multi-omics; Pediatric obesity; Puberty; RNAseq.

Resumen

LA OBESIDAD INFANTIL puede desarrollarse en etapas tempranas de la vida y dar lugar a la aparición de alteraciones metabólicas como la resistencia a la insulina (RI). Si se mantienen durante la edad adulta, la obesidad y la RI suelen derivar en el desarrollo de afecciones más graves como la diabetes tipo II y las enfermedades cardiovasculares, que aumentan considerablemente la morbilidad y la mortalidad en las poblaciones afectadas. Como muchos otros trastornos complejos, la obesidad y sus comorbilidades cardiometabólicas constituyen un fenotipo complejo que surge de la interacción entre un perfil molecular de riesgo (que implica alteraciones genómicas, transcriptómicas, epigenómicas y proteómicas) y las exposiciones ambientales. En este sentido, uno de los campos de investigación más prometedores en materia de obesidad ha consistido en la identificación de biomarcadores moleculares predictivos durante las etapas tempranas de la vida, capaces de estratificar a los pacientes en función de su riesgo de desarrollar complicaciones cardiometabólicas en la edad adulta. Uno de los enfoques más interesantes y robustos para el descubrimiento de biomarcadores implicaría el análisis simultáneo de múltiples capas de datos ómicos a la vez, permitiendo el estudio de una alteración molecular desde todas sus posibles dimensiones. Sin embargo, debido a la complejidad de los datos ómicos, se requieren enfoques analíticos innovadores. En medio de esta necesidad, la bioinformática y la inteligencia artificial (IA) han experimentado un notable impulso debido a su capacidad para obtener automáticamente modelos descriptivos o predictivos a partir de cantidades masivas de datos (Big Data). La presente Tesis Doctoral recoge una serie de trabajos de investigación en los que la bioinformática y la IA se aplican convenientemente a varios estudios ómicos de obesidad para identificar nuevos biomarcadores moleculares de RI y alteraciones metabólicas en niños y adolescentes con obesidad. Las poblaciones de estudio de la presente tesis doctoral están compuestas por más de 2000 niños españoles con edades comprendidas entre los 2 y los 18 años. En resumen, los resultados recogidos en la presente Tesis Doctoral indican que 1) la obesidad es un trastorno complejo resultante de la interacción entre factores genéticos y ambientales, 2) la creación de herramientas predictivas basadas en la combinación de polimorfismos genéticos es una estrategia interesante pero sencilla para predecir el desarrollo de obesidad, 3) los enfoques de investigación multi-ómicos en obesidad son necesarios para comprender los complejos mecanismos moleculares subyacentes a la enfermedad, y 4) la aplicación de modelos de aprendizaje automático de Inteligencia Artificial eXplicable (XAI) puede ayudarnos a desentrañar las complejas relaciones existentes entre los elementos moleculares ómicos. La aplicación de enfoques de investigación multi-ómica y el uso de herramientas analíticas complejas (como la bioinformática y la IA) son

el camino correcto hacia una verdadera implementación de una medicina personalizada en la obesidad. En el futuro, deben fomentarse más estudios como los recogidos en la presente Tesis Doctoral, así como proyectos de reclutamiento de cohortes más grandes para validar los hallazgos presentados. Esto requerirá una estrecha colaboración entre clínicos e investigadores básicos, y la creación de equipos multidisciplinares, en los que la presencia de perfiles bioinformáticos mixtos será de gran importancia.

Keywords

Adolescencia; Epigenética; Estudio de Asociación del Epigenoma completo, EWAS; Estudio de Asociación del Genoma completo, GWAS; Expresión génica; Genética; Infancia; Metilación del ADN; Multi-ómicas; Obesidad pediátrica; Pubertad; Resistencia a la insulina; RNAseq.

Abbreviations

Artificial intelligence (AI)
Adiponectin (ADIPOQ)
Body mass index (BMI)
Cardiovascular diseases (CVD)
Coronavirus disease 2019 (COVID-19)
Cytosine followed by a guanine (CpG)
DNA methylation (DNAm)
Epigenome-wide association studies (EWAS)
eXplainable Artificial Intelligence (XAI)
Extracellular matrix (ECM)
Genetic Investigation of ANthropometric Traits consortium (GIANT)
Genome-wide association studies (GWAS)
High-sensitivity CRP (hsCRP)
Insulin resistance (IR)
Interleukin (IL)-6
Machine Learning (ML)
Matrix metalloproteinase-9 (MMP-9)
Maturity onset diabetes of the young (MODY)
Melanocortin-4 receptor (MC4R)
Monocyte chemoattractant protein 1 (MCP-1)
Next-generation transcriptome sequencing (RNA-Seq)
Peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PPAR-γ)
PPAR-γ coactivator 1 (PGC1)
Prohormone convertase 1 (PC1)
Pro-opiomelanocortin (POMC)
P-Selectin, myeloperoxidase (MPO)
Single nucleotide polymorphisms (SNPs)
Soluble intercellular cell adhesion molecule-1 (sICAM-1)
Soluble vascular cell adhesion molecule-1 (sVCAM)
Total plasminogen activator inhibitor-1 (PAI-1)
Transcription factor 7-like 2 (TCF7L2)
Tumor necrosis factor alpha (TNF-α)
Waist circumference (WC)
World Health Organization (WHO)

general introduction

GENERAL INTRODUCTION

OVERWEIGHT AND OBESITY in children are a public health problem that has raised concern worldwide due to the alarming increase in cases observed during the last decades [1]. Many children who are overweight or suffer from obesity before puberty maintain obesity in the early adulthood, which is associated with increased morbidity and mortality. Nowadays, obesity is one of the chronic disorders that most contribute to the worldwide global burden of disease, and one of the most expensive public health problems to be faced by both developed and developing countries. Among its associated comorbidities, Type II Diabetes and cardiovascular disease are the main responsible for the increased rates of mortality observed in obesity. The development of obesity and its associated comorbidities has been attributed to a complex interaction between genetics, epigenetics and environmental factors. Though the environmental factors influencing the development and worsening of obesity are pretty well-known, the complete molecular architecture of obesity is still far from being fully understood. On this sense, one of the most promising fields of research in obesity has involved the identification of early-life predictive molecular biomarkers able to stratify patients according to their risk for developing cardiometabolic complications later in life, or by their expected response after being treated with an anti-obesity agent or intervention. In the last decades, omics technologies have produced a vast amount of molecular data and have helped to draw a first sketch of the main molecules and pathways involved in the development of obesity and the associated metabolic derangement. Nevertheless, the complex interactions and epistatic phenomena existing between genes, RNA molecules, proteins and metabolites affecting obesity still remain unknown. Interestingly, the ideal and most robust biomarker discovery approach would involve the simultaneous analysis of multiple omics data layers at a once, allowing tracking a molecular disturbance from all its possible dimensions. Due to the complexity of omics data, nevertheless, new and innovative analytics approaches have been demanded. Some of the most remarkable drawbacks faced in omics data analysis involve handling with the massive dimensionality of genomics, epigenomics and transcriptomics as well as the need for methods able to mine complex patterns of interactions. Fortunately, bioinformatics and artificial intelligence (AI) are two emerging fields of research that have made available a vast number of resources for facing such methodological issues. Among the best examples of successful research applications derived from the bioinformatics and AI analysis of omics data in chronic diseases highlight; 1) The development of molecular-based tests and expert informatics systems for the stratification of patients according to their risk for disease, 2) Identification of new molecular subgroups of patients susceptible of a differential treatment or intervention, or 3) Identification of new potential

therapeutic molecular targets and biological pathways underlying the pathophysiology of disease.

The present Doctoral Thesis gathers a series of research works in which bioinformatics and AI are conveniently applied to several obesity epidemiological research projects with the above-mentioned applications as main aims. The research works collected in the present Doctoral Thesis are a series of complex omics data analyses performed in Spanish children with obesity and involving the application of programming languages and advance algorithms or software tools. As an ultimate end, through the application of such bioinformatics and AI tools, the research works collected in the present Doctoral Thesis pursues approaching to a true implementation of a personalized care in obesity, which would allow to drastically reduce the associated deaths and economic costs of disease.

With the aim of giving the reader a broader knowledge of each of the topics covered by the works of the Doctoral Thesis, the present general introduction contains a series of sections detailing; 1) the epidemiology of obesity around the world, 2) the main cardiometabolic comorbidities and the derived health cost consequences associated with obesity, 3) the aetiology of obesity as a complex network of interactions between genetics, epigenetics and environment, and 4) the application of bioinformatics and AI for the analysis of complex omics datasets and the identification of predictive biomarkers with application in personalized medicine.

1 Childhood obesity: the unsolved pandemic of the 21st century

According to the World Health Organization (WHO), overweight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. In adults, a body mass index (BMI) over 25 is considered overweight, and over 30 is obese. In children, BMI is estimated as a Z-Score, a measure of relative weight adjusted for child age and sex, and several cut-offs and criteria are available for defining obesity categories depending on the population under study [2]. The obesity problem has grown to pandemic proportions, with over 4 million people dying each year as a result of being overweight or obese in 2017 according to the global burden of disease.

Rates of overweight and obesity continue to grow in both adults and children. From 1975 to 2016, the prevalence of overweight or obese children and adolescents aged 5–19 years increased more than four-fold from 4% to 18% globally. In 2016, the prevalence of obesity was estimated at 50 million girls and 74 million boys worldwide [3] (**Figure 1**). The prevalence of overweight and obesity in Europe in children aged 2 to 10 years ranges from less than 10% in the northern regions to more than 40% in the southern countries [3–5].

Nowadays, obesity is one side of the double burden of malnutrition, and more people are obese than underweight in every region except sub-Saharan Africa and Asia. Once considered a problem only in high-income countries, overweight and obesity are now dramatically on the rise in low- and middle-income countries, particularly in urban settings. The vast majority of overweight or obese children live in developing countries, where the rate of increase has been more than 30% higher than that of developed countries.

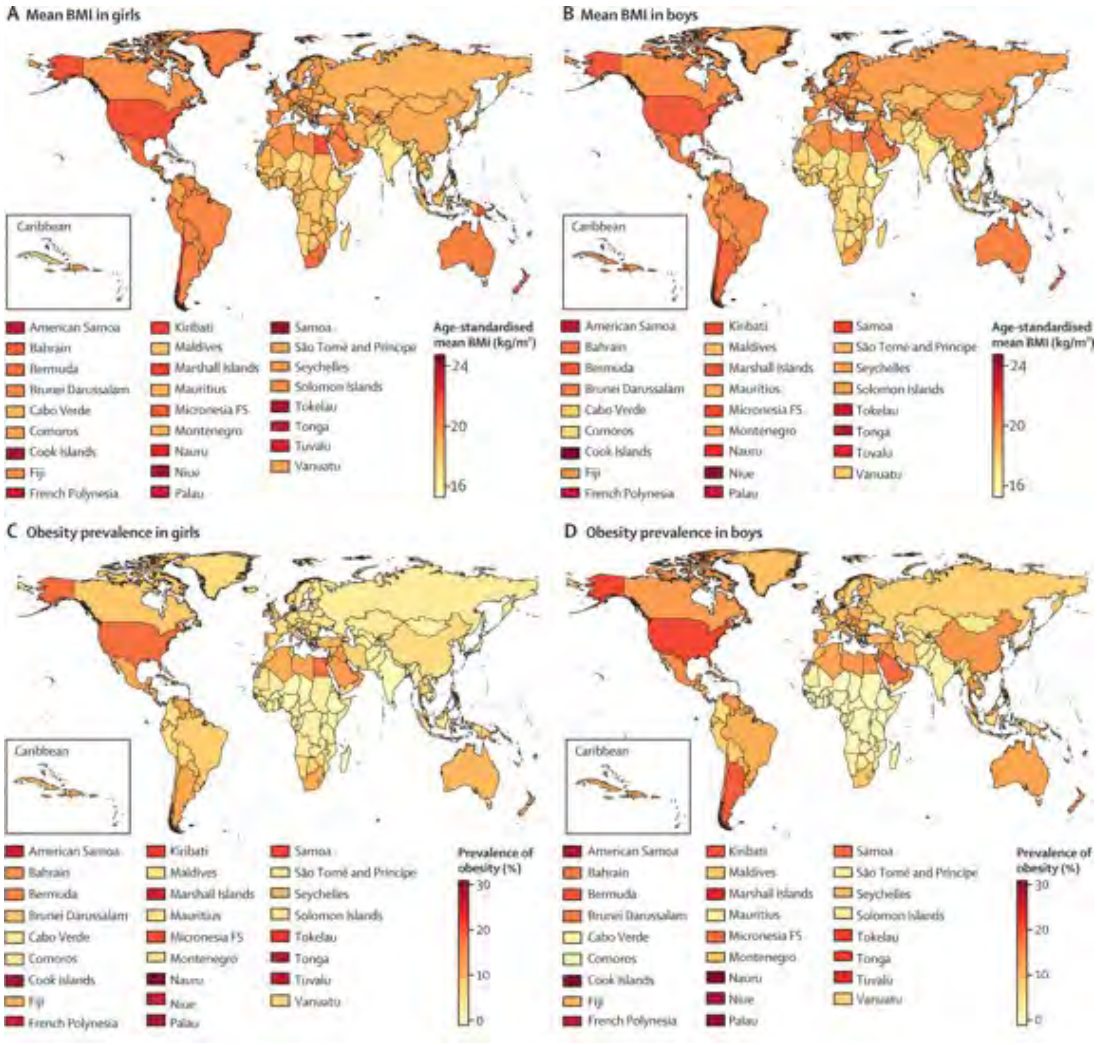


Figure 1. Age-standardised mean BMI, prevalence of obesity, and prevalence of moderate and severe underweight by sex and country in 2016 in children and adolescents Children and adolescents were aged 5–19 years. Obesity was defined as more than 2 SD above the median of the WHO growth reference. Moderate and severe underweight was defined as more than 2 SD below the median. See appendix for results for adults. BMI=body-mass index. ***Source****: Abarca-Gómez, L. et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet 390, 2627–2642 (2017).*

2 Childhood obesity comorbidities and health costs consequences.

Overweight and obesity in children are major risk factors for a number of chronic diseases during adulthood, including cardiovascular diseases (CVD) such as hypertension, heart disease and stroke, which are the leading causes of death worldwide [6]. Being overweight during childhood can also lead to diabetes and its associated conditions. Furthermore, obesity is associated with higher risk of suffering some cancers, including endometrial, breast, ovarian, prostate, liver, gallbladder, kidney and colon [7]. The risk of these and other noncommunicable diseases increases even when a person is only slightly overweight and grows more serious as the BMI rise.

Reduced insulin sensitivity, or insulin resistance (IR), is a pathological condition in which cells fail to respond properly to insulin [8]. IR is one of the metabolic comorbidities of obesity that shows

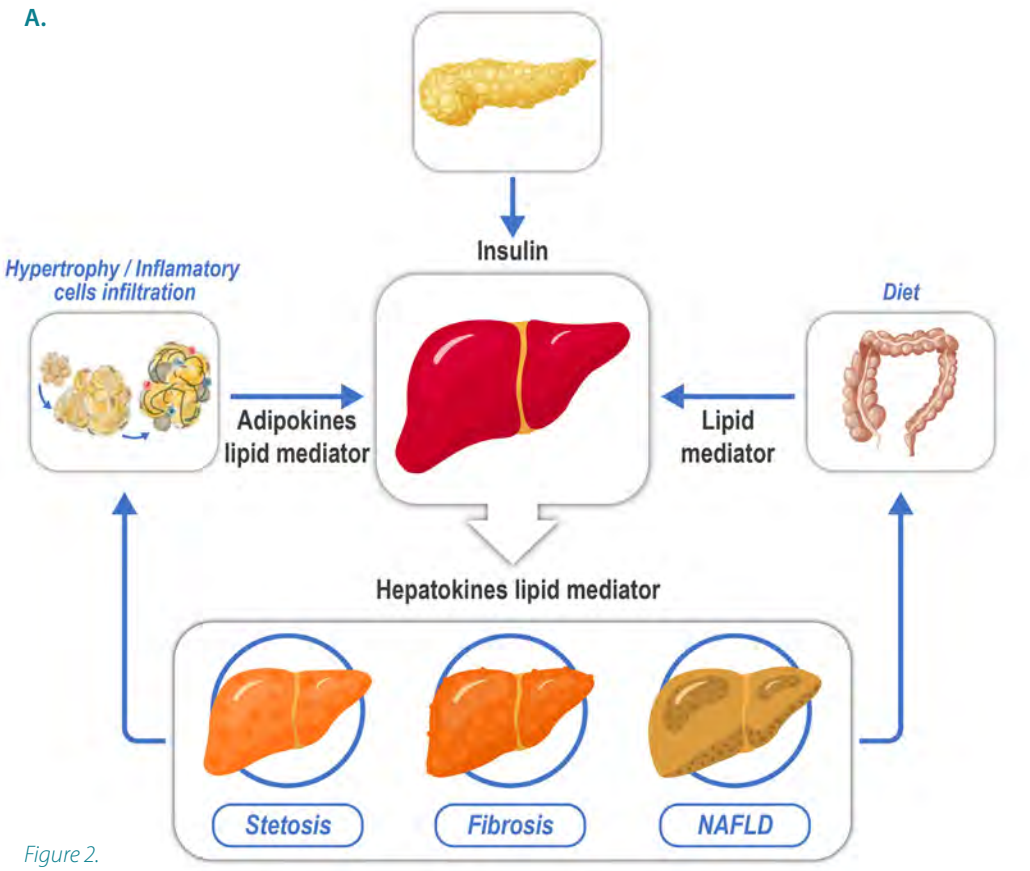


Figure 2.

A. Interorgan crosstalk underlying the development of insulin resistance. Lipid spillover from adipocytes whose storage capacity has been overloaded can result in the inappropriate deposition of lipid in a range of tissues, including muscle, liver and pancreas, a phenomenon referred to as 'lipotoxicity'. Collectively, this results in the development of muscle and liver IR as well as adipocyte IR. This results in further hyperinsulinaemia, which can further exacerbate the development of IR.

an earliest appearance in patients consequence of interrelated stimuli from at least the liver, the pancreas, the gut and the adipose tissue (**Figure 2**). At the molecular level, IR results from the release, by adipose tissue, of increased amounts of non-esterified fatty acids, glycerol, hormones and pro-inflammatory adipocytokines. IR is the main pathophysiological mechanism linking obesity and metabolic disorders such as type 2 diabetes and CVD. Therefore, IR has become a cornerstone in preventing obesity-associated morbimortality [8]. The main aim of the present Doctoral Thesis was the identification of molecular markers of IR that can be measured non-invasively early in life for preventing the appearance of cardiometabolic alterations in children with obesity. This objective was addressed from a multi-omics perspective, reporting new genetics variants, DNA methylation patterns, RNA molecules and proteins associated with IR.

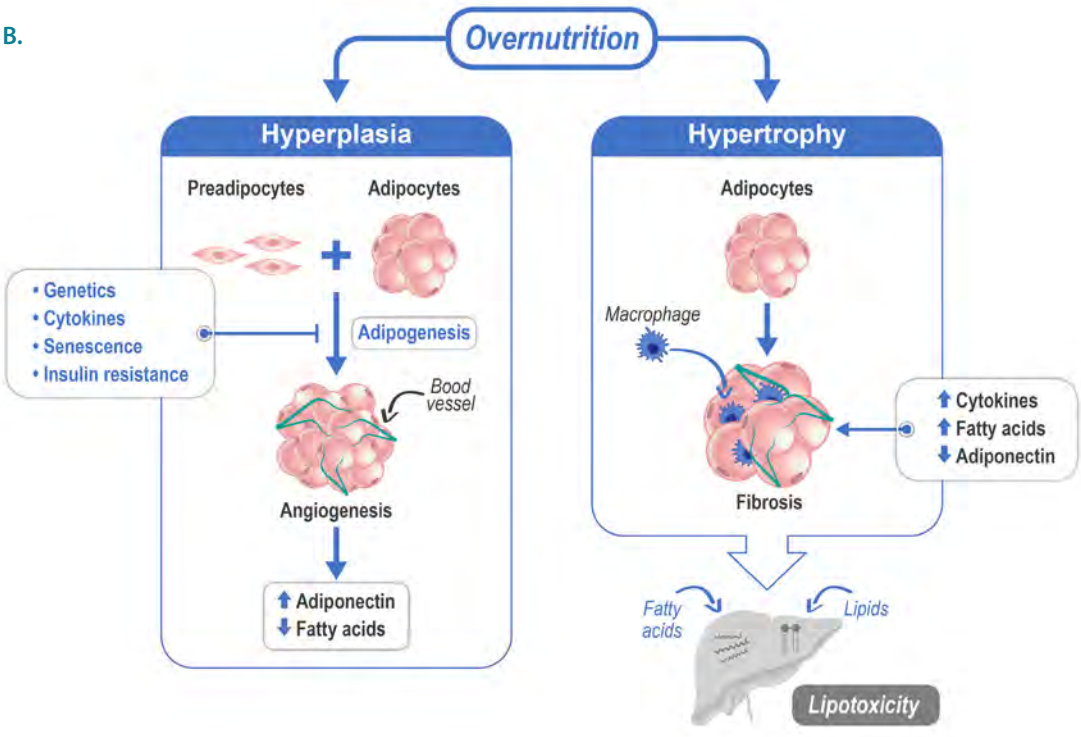


Figure 2.

B. Role of adipose tissue expansion in IR. Adipocytes can expand in response to overnutrition by either hypertrophy (increased cell size, right) or hyperplasia (increased cell number, left). The inability to expand the tissue via hyperplasia due to genetic factors, inflammation or cell senescence is thought to play a major role in the onset of metabolic disease, as it reduces the capability of the tissue to store dietary lipids, causing increased levels of circulating free fatty acids that eventually lead to lipotoxicity. Adipocyte hypertrophy is also thought to cause inflammation by recruiting proinflammatory macrophages into the tissue, resulting in tissue fibrosis, reduced lipid storage capacity and maladaptive changes in the secretion of factors that contribute to IR in muscle and liver. Defective formation of new blood vessels in adipose tissue has also been suggested to be a contributor to systemic IR, since adipogenesis, which is implicated in IR, is intricately linked to angiogenesis.

As a consequence of all these cardiometabolic comorbidities, it is quite clear that obesity is associated with an increased health-care use [9]. Adults with a BMI of 35–40 kg/m^2 had 63% higher general practitioner costs than normal-weight adults, a figure that rise to 116 % in adults with a BMI >40 kg/m^2. The same trend has been reported for the cost of medication use and indirect costs. A meta-analysis reported that the annual medical spending attributable to obesity was $1,900 per person in the year 2014, accounting for $149.4 million at the national level in the USA [10].

The last evidence of obesity as a serious risk factor for developing critical illness and premature death has derived from the coronavirus disease 2019 (COVID-19) pandemic [11]. According to numerous studies; having obesity, and particularly severe obesity, increases the risk of severe illness from COVID-19, and people who are overweight may also be at increased risk. Moreover, having obesity has been reported to triple the risk of hospitalization due to a COVID-19 infection.

All these costs and disease consequences are substantial and demand an urgent response from health professionals and policymakers. Identifying new treatment approaches, such as drugs for weight loss, is therefore a must, especially for children. Today, there are no approved and effective drugs for children and adolescents with obesity [12,13]. The available and approved drugs, such as metformin or orlistat, have only very modest effects on body weight otherwise. In the present Doctoral Thesis, there will be a whole study focused in the identification of new genetics variants with utility as pharmacogenetics biomarkers for metformin response. Thanks to the study of such markers, children with obesity could be stratified into responders or not responders before beginning with a metformin treatment, in order to increase drug efficacy.

3 Puberty, obesity and insulin resistance: the perfect time to act

Regarding adolescence, data published in the past few years indicate that 17% of the adolescents in the USA have obesity (defined as a BMI >95th percentile) and 15% of the adolescents in Europe have obesity (defined as a BMI >97th percentile) [14–16]. Studies have found that without intervention, children and adolescents with obesity will probably continue being obese into adulthood [5,17].

Puberty is a time period characterized by dynamic physiological changes, including activation of the reproductive axis and subsequent secretion of sex steroids, acceleration in growth, and accumulation of both lean and fat mass [18]. Besides physiological events, puberty has also been associated with differential disease prognosis for conditions such as IR, reinforcing the relevance of this development period to the life-long health. Nevertheless, pubertal changes seem not to affect all individuals equally [9,19–22]. While in healthy normal-weight youth, there is a bottom in insulin sensitivity in mid-puberty, which recovers at puberty completion, there is evidence that IR does not resolve in youth who are obese going into puberty, which may result in increased cardio-metabolic risk. Accordingly, the incidence of youth-onset type 2 diabetes is also tightly linked

with pubertal development [23]. Understanding the molecular and biological processes underlying metabolic changes during puberty and the additional impact of obesity on these changes is therefore crucial to prevent type 2 diabetes.

The PUBMEP project is one of the study populations analysed in the present Doctoral Thesis. It is a longitudinal obesity study in which children are followed (from prepuberty to puberty) evaluating the prevalence of metabolic syndrome and the progression of the cardio metabolic risk factors related to it. In this population, a series of multi-omics analyses have been conducted with the aim of discovering new and promising molecular biomarkers of IR during the metabolically critical period of puberty. Results from these approaches will be exposed in details in two complete studies of the thesis.

4 Aetiology of obesity, insulin resistance and diabetes: a complex cocktail of genes and environment

Common obesity and its associated comorbidities constitute a complex phenotype arising from interactions between an at-risk genetic profile and environmental risk factors (**Figure 3**), such as physical inactivity, excessive caloric intake, the intrauterine environment, medications,

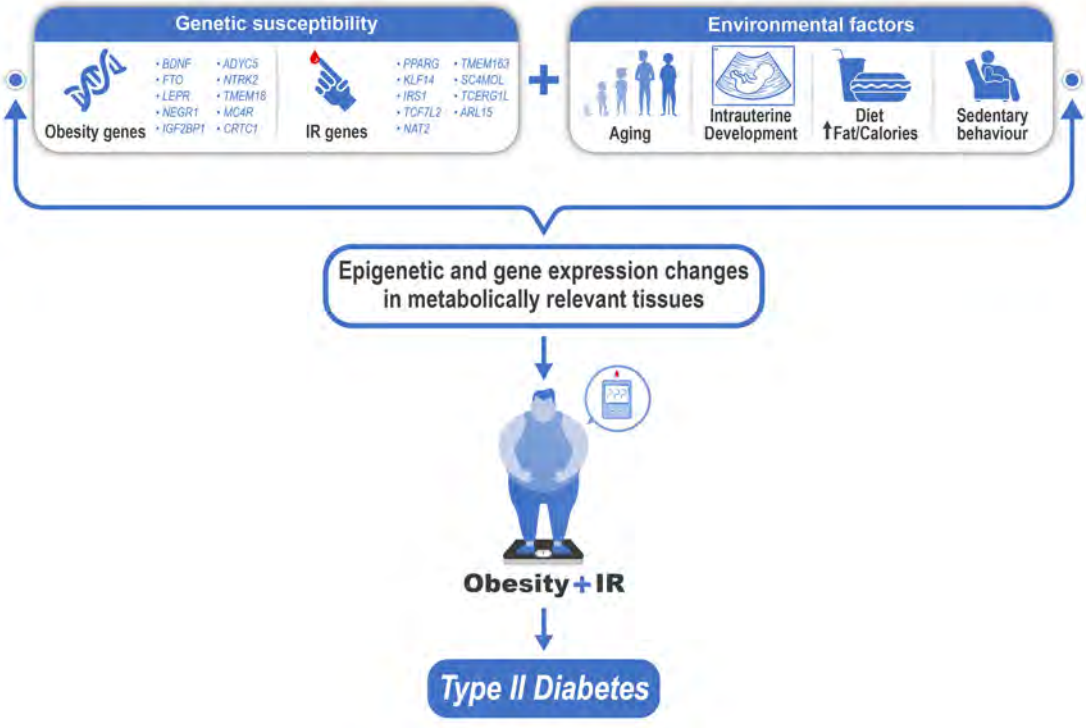


Figure 3. Obesity and its metabolic comorbidities constitute a complex phenotype arising from interactions between an at-risk genetic profile and environmental risk factors.

socioeconomic status, and possibly novel factors such as insufficient sleep, endocrine disruptors, and the gastrointestinal microbiome.

Regarding genetics, in obesity we can differentiate between the existence of rare mutations, which cause cases of severe and monogenic obesity, and predisposing common single nucleotide polymorphisms (SNPs), which contribute with small but cumulative risk effects on disease, leading to the well-known phenotype of common polygenic obesity. In the case of rare mutations, they derive into a rare and severe early-onset obesity with abnormal feeding behaviour and endocrine disorders. Contrarily, in polygenic common obesity, the increase in body weight derives from the simultaneous effect of many polygenic variants. This polygenic basis of obesity implies indeed that the specific set of polygenic variants relevant for obesity in one individual is unlikely to be the same in another obese subject. For monogenic obesity, the most frequent variant is that in the melanocortin-4 receptor (*MC4R*), which accounts for up to 4% of cases of severe obesity [24]. Other rare causes of monogenic severe obesity also include mutations in leptin and the leptin receptor, prohormone convertase 1 (*PC1*) and pro-opiomelanocortin (*POMC*) [24]. Before 2007, candidate gene approaches examined these and hundreds of other genes, but few have been finally confirmed as genetic risk factors for common polygenic obesity in the era of genome-wide association studies (GWAS). Exceptions include variants in the *MC4R* and *BDNF*. Moreover, in 2007, four reports associated SNPs in the first intron of the gene *FTO* (fat mass and obesity associated gene) with obesity-related traits: a GWAS for anthropometric traits [25], a GWAS for early-onset severe obesity, a GWAS for type 2 diabetes, and a population stratification study that incidentally discovered *FTO* [26]. Nowadays, GWAS for BMI, waist-to-hip ratio, and other adiposity traits have identified more than 300 SNPs associated, and *FTO* still remains the strongest and most cross-validated signal across multiple ancestries. Till date, the most remarkable GWAS conducted in obesity is the Genetic Investigation of ANthropometric Traits consortium (GIANT) meta-analysis (comprising more than 339,000 individuals), which identified 97 loci for BMI, 56 of which were novel [27]. Genes near these loci showed expression enrichment in the central nervous system, suggesting that BMI is mainly regulated by processes such as hypothalamic control of energy intake.

Furthermore, the gene variant most commonly associated with insulin sensitivity is the P12A polymorphism in *PPARγ*, which is related to an increased risk of developing diabetes [28,29]. A number of genes associated with β-cell dysfunction have also been identified, and include hepatocyte nuclear factor-4α and 1α — genes known to cause the monogenic disorder maturity onset diabetes of the young (MODY) — the E23K polymorphism in the islet ATP-sensitive potassium channel Kir6.2 (encoded by *KCNJ11*), non-coding SNPs in the transcription factor 7-like 2 (*TCF7L2*) and mutations in the mitochondrial genome [28,29]. Work is also ongoing on many candidate genes, including calpain 10, adiponectin (*ADIPOQ*), PPAR-γ coactivator 1 (*PGC1*) and the glucose transporter *GLUT2* [28,29]. Interestingly, a great bulk of currently known IR-associated genes, such is the case of *TCF7L2*, are related to the extracellular matrix (ECM) in metabolically relevant tissues [30].

Healthy adipose tissue expansion in obesity depends on ECM remodelling and reorganization to provide enough space for the enlargement of adipocytes (hypertrophy) and for the generation of new ones through adipogenesis from the precursor cells (hyperplasia). Otherwise, a failure of this process results in necrotic adipocytes, and hypoxia, which triggers chronic, low-grade inflammation, fibrosis, and lastly IR [31].

Despite all these great advances, in the case of common polygenic obesity, identified loci from the GIANT study explain only 2·7% of the variance in BMI [27]. Simulation studies have suggested that SNPs should account for around 30% of variance in BMI otherwise [32]. This lack in phenotype variance explained by currently known genetic variants is a phenomenon termed as 'missing heritability' [33,34]. Potential sources explaining this missing heritability might include epigenetic alterations, interaction between genetics and environmental factors, the existence of low frequency and rare variants yet to be discovered as well as the presence of X chromosome genetic variation [33,34]. On this regard, the first part of the Present Doctoral Thesis will focus in the study of X chromosome genetics variants and other mentioned obesity candidate-SNPs as potential biomarkers of IR and other alterations of glucose metabolism in children with obesity.

Environmental factors are the flip side of the coin in the development of obesity and type 2 diabetes [35]. Increased caloric availability and fat intake in the framework of decreased physical activity lead to over-nutrition, increased nutrient storage and finally to obesity. Long-term increased dietary fat consumption is also associated with reductions in insulin release. This effect has important consequences if adipocyte or β-cell function are already inherently abnormal owing to mentioned genetic susceptibility. Another proposed environmental mechanism is thought to occur in utero and/or during the early postnatal period when poor nutrition alters metabolism, resulting in a tissue adaptation that favours the storage of nutrients. The end result of these environmental changes is always a deleterious interaction with genes that predispose to the development of obesity and type 2 diabetes. This interactive phenomenon will be also reviewed during the Thesis, with one study focused in the investigation of interactions between health family history and other environmental variables and inherited genetic predisposition to obesity.

Although there are reasons to hope that identified genetics variants of obesity and diabetes will eventually lead to new preventive and therapeutic agents, this will take time because such developments require detailed mechanistic understanding of how an SNP influences phenotype. This involves identification of the gene or genes whose expression is affected by alleles at the variant, and the mechanism (e.g., enhancer, repressor, epigenetic alteration) whereby the variant's alleles differentially affect expression. Consequently, there is growing interest in understanding the role of epigenetic mechanisms surrounding obesity and diabetes.

5 DNA methylation as an epigenetic link between environment and obesity

DNA methylation (DNAm) is a heritable epigenetic mark consisting of the covalent addition of a methyl-group to a cytosine followed by a guanine (CpG). DNAm is potentially reversible and can be altered by environmental factors, resulting in alterations of gene expression and providing an interactive connection between genetics and the environment. In epidemiological studies, DNAm is the most widely studied epigenetic mechanism, partly due to the fact that it can be measured at large scale in epigenome-wide association studies (EWAS).

Differential DNAm in certain loci has been related to obesity [36], systemic IR [37–45], and type 2 diabetes [36,38,39,46–50] in adults, either in blood or in other metabolically relevant tissues (**Figure 4**). The dynamics of DNAm during puberty has also been investigated in one or both genders, emphasising how DNAm is stable at some CpG sites and varies at others [51,52]. On the other hand, transcriptional dysregulation of genes has been reported as a key molecular mechanism associated with IR and obesity, possibly connected to DNAm alterations [53,54]. Some of most replicated genes

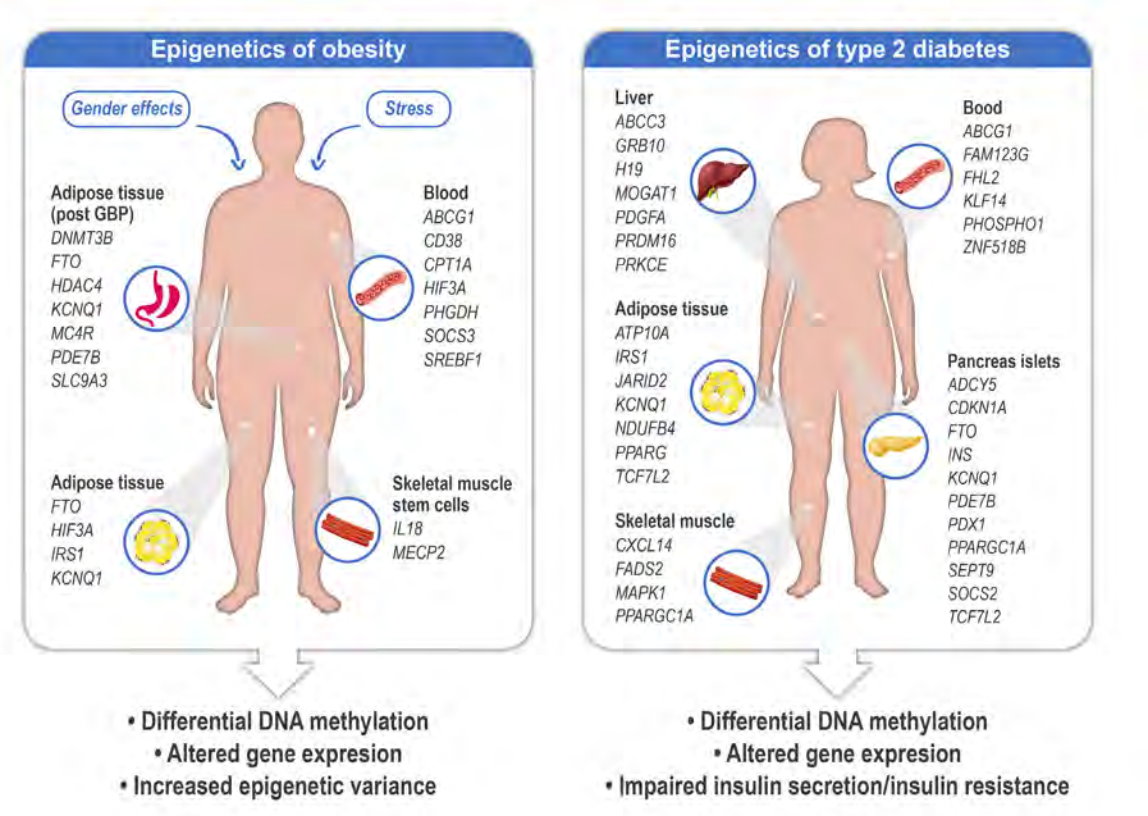


Figure 4. Epigenetic architecture of obesity and type II diabetes.

with altered DNAm in obesity and diabetes include (e.g., *ABCG1, ADCY5, CPT1A, FTO, HCCA2, HDAC4, HIF3A, IGF-1, KCNQ1, PPARG*, and *TCF7L2*). Among them, the most strongly associated are the *HIF3A*, the *CPT1A* and the *ABCG1*[55–57]. *HIF3A* encodes hypoxia-inducible factor 3 subunit alpha, which is part of a group of heterodimeric transcription factors that regulate responses to low oxygen (hypoxia). *CPT1A* encodes the enzyme carnitine palmitoyl transferase 1A, which takes Section In carnitine-dependent transport across the mitochondrial membrane when oxidation of long-chain fatty acids is initiated and is important for several metabolic processes. Of interest for obesity, the protein encoded by *ABCG1* is involved in macrophage cholesterol and phospholipids transport.

For many of these genes, EWASs studies in children have confirmed obesity associations [58,59]. Particularly, recent reports from two robust longitudinal cohorts have found a strong association between their methylation in human fetal tissue and the subsequent development of childhood adiposity. Thus, epigenetic analysis at birth and childhood may have utility in identifying future risk of obesity. Otherwise, EWASs for IR in children and adolescents are still scarce. In the present Doctoral Thesis, EWAS analyses have been performed in hundreds of children with the aim of unveiling the epigenetic regulatory mechanisms underlying the appearance of IR and cardiometabolic disturbances in obesity.

6 Omics, bioinformatics and Artificial Intelligence as the effective way to a personalized care in obesity

The term "omics" refers to the comprehensive characterization, quantitation and quantification of a large number of molecules, grouped according to the fundamental structural or functional biological similarities that they demonstrate. Across years, omics sciences have helped to unveil new and promising obesity predictive biomarkers and therapeutics targets. As we have mentioned in the previous sections, most of these biomarkers include SNPs and DNAm patterns at certain loci, but can also extend to the abundance of certain RNA molecules, serum proteins, or even multi-omics signatures (involving the simultaneous identification of several of such molecules in a tissue at a time) [60]. Interestingly, many of these molecules can be measured non-invasively early in life, when children have not yet developed cardiometabolic disturbances and thus allow a better disease risk stratification of patients. Ultimately, these advances could lead to the so desired personalized medicine, in which the diet and clinical cares would be adapted to the needs and individual genetic preferences of each subject.

Before this to happen, nevertheless, we first need to be able to appropriately model all the "big" data sets deriving from high-throughput omics approaches. Fortunately, recent major advances in omics technologies have been accompanied by great innovations in the field of bioinformatics and AI. Within AI, Machine Learning (ML) techniques have experienced a notable boost due to their ability to automatically obtain descriptive or predictive models from massive amounts of data (big

data). These techniques learn models that allow us to characterize, adapt, learn, predict and analyse complex and large datasets, amplifying our understanding of disease and our capacity to predict with unprecedented precision. To date, there have been increasing applications of ML techniques in the field of obesity and omics research [61–70]. Depending on the learning process implemented, we can distinguish between two main approaches in ML; supervised and unsupervised learning [71]. In supervised learning, the algorithm is provided with inputs (e.g., omics data) corresponding to specific outputs (e.g., presence of an obesity comorbidity or not), where the information is used to develop a general rule that will link the input to the output. However, an associated response or output is not always available. Or even if one is available, we may be interested in discovering other types of associations among variables. In these cases, a number of techniques can be used under the umbrella of what is called unsupervised learning. The term unsupervised refers to the fact that this learning is not based on the existence of a previously known response. We are not interested in prediction, rather, the goal is to discover other interesting relationships between the input variables.

ML tasks in obesity and omics research have typically included: a) dimensionality reduction to reduce the input data mass by decreasing the number of random variables under consideration, b) clustering-classification to organize different variables or subjects in groups with common characteristics, c) density estimation to assess distribution of input variables in specific space, and d) regression to estimate the relationships among variables and for developing predictive models[70,72]. Of special interest has been also the application of ML for the integration of multi-omics data sets, in which information from different layers of omics data is combined to discover the coherent biomarkers of a disease [73,74]. Selecting a multi-omics approach compared to a single-omic analysis offers some profound advantages but has some serious challenges. A major advantage of the multi-omics analysis is the breadth of the information that it provides. As we have previously mentioned, the aetiology of obesity and type 2 diabetes is multifactorial. Thus, identification of one specific factor associated with disease will most probably have limited prognostic or therapeutic value. The multi-omics analysis allows for the identification of associated factors from different biological processes, i.e., gene expression, protein synthesis and posttranslational modifications, cellular metabolic processes, glycosylation, etc., maximizing the available information, and thus, increasing the possibility of identifying the root causes of a disease. Some of the best examples of multi-omics approaches involve the study of mQTLs (GWAS with metabolomics), meQTLs (GWAS with DNA methylation), eQTLs (GWAS with gene expression), eQTMs (DNA methylation and gene expression) and mCpG (DNA methylation with metabolomics) [http://www.metabolomix.com] [60]. One the most ambitious studies conforming the present Doctoral Thesis is a large-scale molecular analysis investigating the multi-omics signatures that underlie the appearance and worsening of IR in children with obesity when they enter into puberty. Our results shed light on the molecular mechanisms underlying epigenetic alterations in obesity and propose novel and promising biomarkers for IR and metabolic alterations in children.

Despite the tremendous advances that AI has experienced in recent times there has been also a wave of concern (especially in biomedical applications), as in most cases we do not know how the software learns and makes decisions. Cases such as the autonomous car project that after thousands of tests decided to turn off the road on a bridge, or the more recent case of IBM Watson at the Danish National Hospital, which made a very serious mistake by recommending a "lethal treatment" for cancer patients without being able to motivate why it was making such a recommendation, are generating a lot of controversy regarding the eXplainability of AI. Another example is the case of well-known deep learning techniques that cannot explain how they make their decisions despite their impressive predictive ability. This whole range of issues has been dubbed the "black box paradigm" and has recently been addressed in the prestigious journal Nature: "Although today's AI systems offer many benefits in many applications, their effectiveness is limited by the lack of explanations when interacting with humans". In all cases, there is agreement on the need to make AI explainable, giving rise to what has recently become known as eXplainable Artificial Intelligence (XAI). With this in mind, the present Doctoral Thesis has opted for the use of ML algorithms with high interpretability and eXplainability, such is the case of association rules, which will allow the expert to understand the internal process that leads the algorithm to make decisions.

Other research field of tremendous success for the analysis of omics data has been the area of bioinformatics. Bioinformatics is defined as an interdisciplinary field that combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret biological data, in particular when the data sets are large and complex. As one of the most successful application of bioinformatics in the field of obesity and genomics we can find the use of genetic risk scores [75]. The construction of a genetic risk score consists on the calculation of a sum of the existing risk-increasing alleles in an individual, often weighted by the effect sizes from the studies that discovered them. In the case of BMI, a genetic risk score would be estimated as the sum of BMI-increasing alleles (0, 1, or 2) at each of the single-nucleotide polymorphisms (SNPs) robustly associated with BMI. Many studies constructing genetic risk scores for BMI have used the 32 SNPs reported in the 2010 GIANT meta-analysis. As often observed for genetic risk scores of other traits, the genetic risk score values in obesity follow a bell-shaped distribution and BMI have been estimated to be 3 kg/m^2 higher for those at the top of the distribution (genetic risk score ≥38) than for those at the bottom (genetic risk score ≤21), with each unit increment in genetic risk score associated with nearly 2 kg/m^2 higher BMI [76]. A whole study of the present Doctoral Thesis will focus in the construction of a genetic risk score for BMI and evaluating its performance as a predictive tool for the appearance or worsening of obesity during puberty.

Furthermore, as we have already mentioned, gene-environment interactions, potentially through epigenetic mechanisms, may also affect the pathogenesis of obesity. Interestingly, genetic risk scores have been also employed successfully for the study of complex environment-

gene interactions [77]. Indeed, simulations have found that genetic risk scores have greater power than individual SNPs for detection of such phenomena [77].

The medical, financial and social problems deriving from the unsolved pandemic of obesity in children and adolescents are substantial and require an urgent response from health professionals and governments. Prevention in high-risk groups seems to be a promising strategy in addition to changing the obesogenic environment. Most importantly, treatment barriers in children have to be resolved. Omics technologies have shown potential for identifying effective predictive biomarkers or new therapeutics targets in recent years. New and exciting multi-omics data sources in obesity are requiring innovative and sophisticated mathematical methods for analysis. Existing and emerging methods in ML are meeting the need for sophisticated high-level prediction and description. The right application of such tools in obesity multi-omics longitudinal childhood cohorts will advance the goals of a personalized disease risk estimation and treatment.

References

1. GBD 2015 Obesity Collaborators et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N. Engl. J. Med. 377, 13–27 (2017).
2. M, de O. & T, L. Defining obesity risk status in the general childhood population: which cut-offs should we use? Int. J. Pediatr. Obes. 5, 458–460 (2010).
3. Abarca-Gómez, L. et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet 390, 2627–2642 (2017).
4. Ahrens, W. et al. Cohort Profile: The transition from childhood to adolescence in European children-how I.Family extends the IDEFICS cohort. Int. J. Epidemiol. 46, 1394–1395 (2017).
5. Gepstein, V. & Weiss, R. Obesity as the Main Risk Factor for Metabolic Syndrome in Children. Frontiers in Endocrinology vol. 10 (2019).
6. Jones, R. E., Jewell, J., Saksena, R., Ramos Salas, X. & Breda, J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Front. Public Heal. 5, 1–12 (2017).
7. KI, A., N, S., CS, M. & M, D. Obesity and cancer risk: Emerging biological mechanisms and perspectives. Metabolism. 92, 121–135 (2019).
8. Bornfeldt, K. E. & Tabas, I. Insulin resistance, hyperglycemia, and atherosclerosis. Cell Metab. 14, 575–585 (2011).
9. Reinehr, T. Long-term effects of adolescent obesity: time to act. Nat. Rev. Endocrinol. 14, 183–188 (2018).
10. DD, K. & A, B. Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. Value Health 19, 602–613 (2016).
11. Gao, M. et al. Associations between body-mass index and COVID-19 severity in 6·9 million people in England: a prospective, community-based, cohort study. Lancet Diabetes Endocrinol. 9, 350–359 (2021).
12. JA, Y. Intensive therapies for pediatric obesity. Pediatr. Clin. North Am. 48, 1041–1053 (2001).
13. SZ, Y. & JA, Y. Long-term drug treatment for obesity: a systematic and clinical review. JAMA 311, 74–86 (2014).
14. Jaarsveld, C. H. M. van & Gulliford, M. C. Childhood obesity trends from primary care electronic health records in England between 1994 and 2013: population-based cohort study. Arch. Dis. Child. 100, 214 (2015).
15. LJ, E. et al. Prevalence of severe childhood obesity in England: 2006-2013. Arch. Dis. Child. 100, 631–636 (2015).
16. T, L. & ML, F. Prevalence of overweight among children in Europe. Obes. Rev. 4, 195–200 (2003).
17. Lloyd, L. J., Langley-Evans, S. C. & McMullen, S. Childhood obesity and risk of the adult metabolic syndrome: A systematic review. International Journal of Obesity vol. 36 1–11 (2012).
18. Abbassi, V. Growth and Normal Puberty. www.aappublications.org/news (1998).
19. Kelsey, M. M. & Zeitler, P. S. Insulin Resistance of Puberty. Current Diabetes Reports vol. 16 (2016).
20. Kelsey, M. M. et al. The impact of obesity on insulin sensitivity and secretion during pubertal progression: A longitudinal study. J. Clin. Endocrinol. Metab. 105, (2020).
21. Reinehr, T. & Roth, C. L. Is there a causal relationship between obesity and puberty? The Lancet Child and Adolescent Health vol. 3 44–54 (2019).
22. Reinehr, T., Wolters, B., Knop, C., Lass, N. & Holl, R. W. Strong effect of pubertal status on metabolic health in obese children: A longitudinal study. J. Clin. Endocrinol. Metab. 100, 301–308 (2015).
23. Dabelea, D. et al. Incidence of diabetes in youth in the United States. J. Am. Med. Assoc. 297, 2716–2724 (2007).
24. Goodarzi, M. O. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. The Lancet Diabetes and Endocrinology vol. 6 223–236 (2018).
25. A, S. et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet. 3, 1200–1210 (2007).
26. C, D. et al. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat. Genet. 39, 724–726 (2007).
27. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197–206 (2015).
28. I, B. Genetics of Type 2 diabetes. Diabet. Med. 22, 517–535 (2005).
29. L, A., J, Z., JL, C. & M, L. Common polymorphisms of the PPAR-gamma2 (Pro12Ala) and PGC-1alpha (Gly482Ser) genes are associated with the conversion from impaired glucose tolerance to type 2 diabetes in the STOP-NIDDM trial. Diabetologia 47, 2176–2184 (2004).
30. Anguita-Ruiz, A. et al. Omics Approaches in Adipose Tissue and Skeletal Muscle Addressing the Role of Extracellular Matrix in Obesity and Metabolic Dysfunction. Int. J. Mol. Sci. 2021, Vol. 22, Page 2756 22, 2756 (2021).

31. FJ, R.-O., A, M.-G., CM, A. & J, P.-D. Extracellular Matrix Remodeling of Adipose Tissue in Obesity and Metabolic Diseases. Int. J. Mol. Sci. 20, (2019).
32. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 2015 4710 47, 1114–1120 (2015).
33. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–53 (2009).
34. E, G. Missing heritability of complex diseases: case solved? Hum. Genet. 139, 103–113 (2020).
35. T, T. et al. The many faces of diabetes: a disease with increasing heterogeneity. Lancet (London, England) 383, 1084–1094 (2014).
36. Ling, C. & Rönn, T. Epigenetics in Human Obesity and Type 2 Diabetes. Cell Metabolism vol. 29 1028–1044 (2019).
37. Hidalgo, B. et al. Epigenome-wide association study of fasting measures of glucose, insulin, and homa-ir in the genetics of lipid lowering drugs and diet network study. Diabetes 63, 801–807 (2014).
38. Chambers, J. C. et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: A nested case-control study. Lancet Diabetes Endocrinol. 3, 526–534 (2015).
39. Kulkarni, H. et al. Novel epigenetic determinants of type 2 diabetes in Mexican-American families. Hum. Mol. Genet. 24, 5330–5344 (2015).
40. Sharma, N. K. et al. Integrative analysis of glucometabolic traits, adipose tissue dna methylation, and gene expression identifies epigenetic regulatory mechanisms of insulin resistance and obesity in African Americans. Diabetes 69, 2779–2793 (2020).
41. Ling, C. & Rönn, T. Epigenetic markers to further understand insulin resistance. Diabetologia vol. 59 2295–2297 (2016).
42. Arner, P. et al. The epigenetic signature of systemic insulin resistance in obese women. Diabetologia 59, 2393–2405 (2016).
43. Małodobra-Mazur, M. et al. Obesity-induced insulin resistance via changes in the DNA methylation profile of insulin pathway genes. Adv. Clin. Exp. Med. 28, 1599–1607 (2019).
44. Arpón, A. et al. Insulin sensitivity is associated with lipoprotein lipase (Lpl) and catenin delta 2 (ctnnd2) dna methylation in peripheral white blood cells in non-diabetic young women. Int. J. Mol. Sci. 20, (2019).
45. Arpón, A. et al. Epigenome-wide association study in peripheral white blood cells involving insulin resistance. Sci. Rep. 9, (2019).
46. Juvinao-Quintero, D. L. et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin. Epigenetics 13, 40 (2021).
47. Florath, I. et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia 59, 130–138 (2016).
48. Soriano-Tárraga, C. et al. Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. Hum. Mol. Genet. 25, 609–619 (2016).
49. Al Muftah, W. A. et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clin. Epigenetics 8, 13 (2016).
50. Meeks, K. A. C. et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: Findings from the RODAM study. Int. J. Epidemiol. 48, 58–70 (2019).
51. Han, L. et al. Changes in DNA methylation from pre-to post-adolescence are associated with pubertal exposures. Clin. Epigenetics 11, 1–14 (2019).
52. Suzuki, M. M. & Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. Nature Reviews Genetics vol. 9 465–476 (2008).
53. Sales, V. & Patti, M. E. The Ups and Downs of Insulin Resistance and Type 2 Diabetes: Lessons from Genomic Analyses in Humans. Current Cardiovascular Risk Reports vol. 7 46–59 (2013).
54. Sharma, N. K. et al. Tissue-specific and genetic regulation of insulin sensitivity-associated transcripts in African Americans. J. Clin. Endocrinol. Metab. 101, 1455–1468 (2016).
55. EW, D. et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Hum. Mol. Genet. 24, 4464–4479 (2015).
56. Aslibekyan, S. et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity 23, 1493–1501 (2015).
57. KJ, D. et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet (London, England) 383, 1990–1998 (2014).
58. Fradin, D. et al. Genome-Wide Methylation Analysis Identifies Specific Epigenetic Marks In Severely Obese Children. Sci. Reports 2017 71 7, 1–8 (2017).
59. Huang, R. C. et al. Genome-wide methylation analysis identifies differentially methylated CpG loci associated with severe obesity in childhood. Epigenetics 10, 995–1005 (2015).

60. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. Genome Biol. 2017 181 18, 1–15 (2017).

61. Abdel-Aal, R. E. & Mangoud, A. M. Modeling Obesity Using Abductive Networks. Comput. Biomed. Res. 30, 451–471 (1997).

62. Acharjee, A., Ament, Z., West, J. A., Stanley, E. & Griffin, J. L. Integration of metabolomics, lipidomics and clinical data using a machine learning method. BMC Bioinformatics 17, 37 (2016).

63. TM, D., S, M., A, C. & S, D. Machine Learning Techniques for Prediction of Early Childhood Obesity. Appl. Clin. Inform. 6, 506–520 (2015).

64. K, E., J, K., S, G., J, S. & G, L. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification. Med. Sci. Sports Exerc. 48, 933–940 (2016).

65. Hamad, R., Pomeranz, J. L., Siddiqi, A. & Basu, S. Large-Scale Automated Analysis of News Media: A Novel Computational Method for Obesity Policy Research. Obesity (Silver Spring). 23, 296 (2015).

66. Chen, Z. & Zhang, W. Integrative Analysis Using Module-Guided Random Forests Reveals Correlated Genetic Factors Related to Mouse Weight. PLOS Comput. Biol. 9, e1002956 (2013).

67. Lee, B. J., Kim, K. H., Ku, B., Jang, J. S. & Kim, J. Y. Prediction of body mass index status from voice signals based on machine learning for automated medical applications. Artif. Intell. Med. 58, 51–61 (2013).

68. MA, P. et al. Face morphology: Can it tell us something about body weight and fat? Comput. Biol. Med. 76, 238–249 (2016).

69. Figueroa, R. L. & Flores, C. A. Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures. J. Med. Syst. 2016 408 40, 1–9 (2016).

70. Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M. & Alcalá-Fdez, J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLOS Comput. Biol. 16, e1007792 (2020).

71. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nat. 2015 5217553 521, 436–444 (2015).

72. DeGregory, K. W. et al. A review of machine learning in obesity. Obes. Rev. 19, 668–685 (2018).

73. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front. Genet. 8, 84 (2017).

74. Bonnet, E., Calzone, L. & Michoel, T. Integrative Multi-omics Module Network Inference with Lemon-Tree. PLoS Comput. Biol. 11, (2015).

75. Schrodi, S. J. et al. Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future. Front. Genet. 5, 1–18 (2014).

76. RJ, L. Genetic determinants of common obesity and their value in prediction. Best Pract. Res. Clin. Endocrinol. Metab. 26, 211–226 (2012).

77. Marigorta, U. M. & Gibson, G. A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. Front. Genet. 5, (2014).

aims

AIMS

AIMS

THE OVERALL AIM of this Doctoral Thesis is to identify early-life molecular predictive biomarkers of IR and other metabolic comorbidities in children and adolescents with obesity, as well as to implement and develop new AI-based tools for the study of complex omics longitudinal datasets in obesity. The present International Doctoral Thesis is composed of a total of seven studies. They are classified into three different parts: **Section I** focuses on the study of genetic variants associated with childhood obesity and its associated-alterations in glucose metabolism; **Section II** focuses on the identification of new multi-omics biomarkers of insulin resistance (IR) and cardiometabolic alterations in childhood obesity during the metabolically critical period of puberty; and **Section III** focuses on the implementation of unsupervised machine learning (ML) models for the analysis of longitudinal omics data in obesity.

Section I

General objective 1: To investigate the genetics basis (alterations in the DNA sequence) of childhood obesity and its associated glucose metabolism disturbances.

- Specific objective 1.1: To study the association between TNMD X-chromosome genetic variants and glucose metabolism complications related to childhood obesity, and to evaluate the potential functionality of the TNMD gene in human adipocytes (**Study 1**).
- Specific objective 1.2: To make public and describe a dataset incorporating phenotype and X chromosome genotype data from a cohort of 915 normal-weight, overweight and obese children, and to deeply describe a whole implementation of the special X chromosome analytic process in genetics (Methodological **Study 2**). This study was conducted as an analytic pipeline to solve some methodological drawbacks encountered during the analysis of Study 1.
- Specific objective 1.3: To test whether obesity-related genetics variants can predict the response to metformin intervention in terms of the post-treatment change in glucose metabolism, anthropometry, lipid metabolism, adipokines, and inflammatory markers in children with obesity (**Study 3**).
- Specific objective 1.4: To evaluate the utility of an adult obesity-predisposing genetic risk score (GRS) for the prediction and pharmacological management of obesity in Spanish children, further investigating its implication in the appearance of cardio-metabolic alterations (**Study 4**).

Section II

General objective 2: To discover new biomarkers of IR through the integration of multi-omics data in children and adolescents with obesity.

- Specific objective 2.1: To evaluate, through a multi-omics perspective, the association between the protein S100A4 and IR in children with obesity during pubertal development (**Study 5**).
- Specific objective 2.2: To understand, through a large-scale multi-omics longitudinal analysis, the molecular architecture and biological processes underlying the development of IR in puberty and the additional impact of obesity on these processes (**Study 6**).

Section III

General objective 3: To implement and develop new unsupervised ML algorithms for the analysis of complex longitudinal omics data.

- Specific objective 3.1: To implement a novel rule-based eXplainable Artificial Intelligence (XAI) strategy (including pre-processing, knowledge extraction and functional validation) for finding biologically relevant sequential patterns from longitudinal human gene expression data in obesity research (**Study 7**).

Methodological overview of the studies included

methodological overview of the studies included

THE PRESENT DOCTORAL THESIS employed multi-omics data derived from three ambitious epidemiological research studies conducted by our research group in Spain (involving the analysis of more than 2000 Spanish children with ages ranging from 2-18) as well as other datasets derived from the Public Repository "Gene Expression Omnibus".

1 The GENOBOX Study: a multicenter cross-sectional design.

The GENOBOX study [1–3], is a project started in 2011 at the Aragon Institute for Health Research (IISA) - Lozano Blesa University Clinical Hospital (Zaragoza), the Santiago de Compostela Health Research Institute (IDIS) - University Clinical Hospital of Santiago de Compostela, the Maimonides Biomedical Research Institute of Cordoba (IMIBIC) - Reina Sofía University Hospital, Córdoba, Spain, and the Institute for Biomedical Research IBS-Granada (**Figure 5**). The GENOBOX study follows an observational cross-sectional design for investigating the relationship between genetic variants, markers of oxidative stress and inflammation, lifestyle, and cardiovascular risk in children and adolescents from Spain. All children attending the three recruiting hospitals for diagnosis of minor disorders; that were not confirmed after clinical and laboratory investigations, or suspecting overweight or obesity, were invited to participate. Exclusion criteria were: birth weight <2500 g, Presence of diabetes mellitus type I, presence of congenital, chronic, or inflammatory disease, psychomotor disability, use of hormonal medication or other that modifies blood pressure, glucose or lipid metabolism, having performed intense exercise in the 24 h previous to the examination, and/or having participated in a research study in the previous three months. The GENOBOX study recruited a total population of 1699 children and adolescents (878 girls) aged 2.00-18.10 years. Subjects were assigned to experimental groups according to their obesity status (513 normal-weight, 412 overweight, and 774 children with obesity).

2 The PUBMEP Study: a longitudinal design.

The PUBMEP study ('Puberty and metabolic risk in obese children. Epigenetic alterations and pathophysiological and diagnostic implications') is a longitudinal study based on the follow-up of a cohort of children who previously participated in the GENOBOX study (**Figure 6)**. The main hypothesis of the PUBMEP study lies beneath the fact that puberty might constitute a metabolic risk factor for childhood obesity, and pursues two main objectives: 1) To clarify the relationship

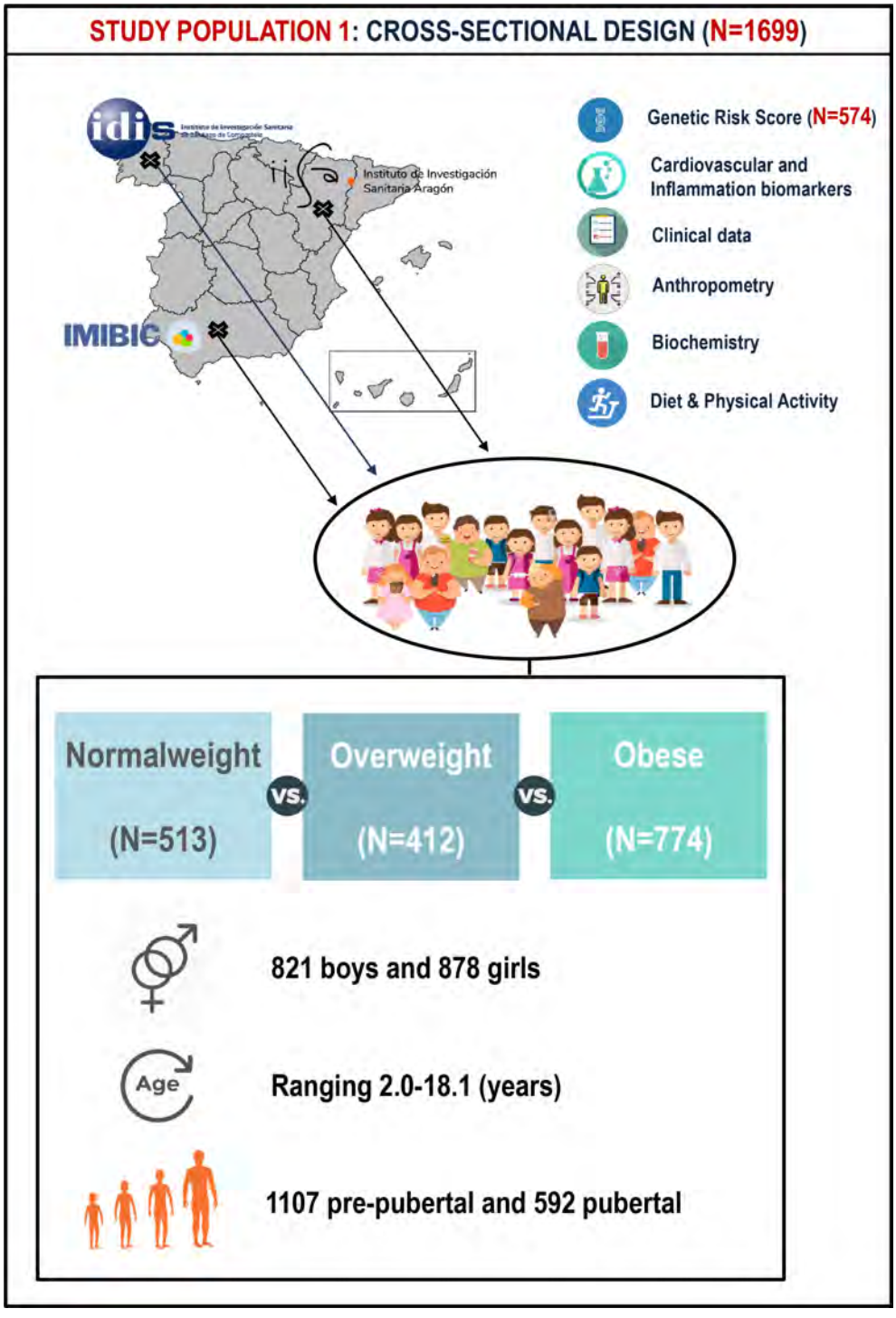


Figure 5. General characteristics of the GENOBOX population, which is based on a previously conducted case-control multicentre cross-sectional design.

between obesity, metabolic alterations and the natural onset of pubertal maturation, and 2) To get a deeper understanding of the underlying epigenetic architecture of obesity and its metabolic complications. In the PUBMEP study, prepubertal boys and girls initially enrolled in the GENOBOX study who had already completed puberty at the time of the PUBMEP study start were invited to participate. During the course of the PUBMEP study (2012–2018), children remained under regular

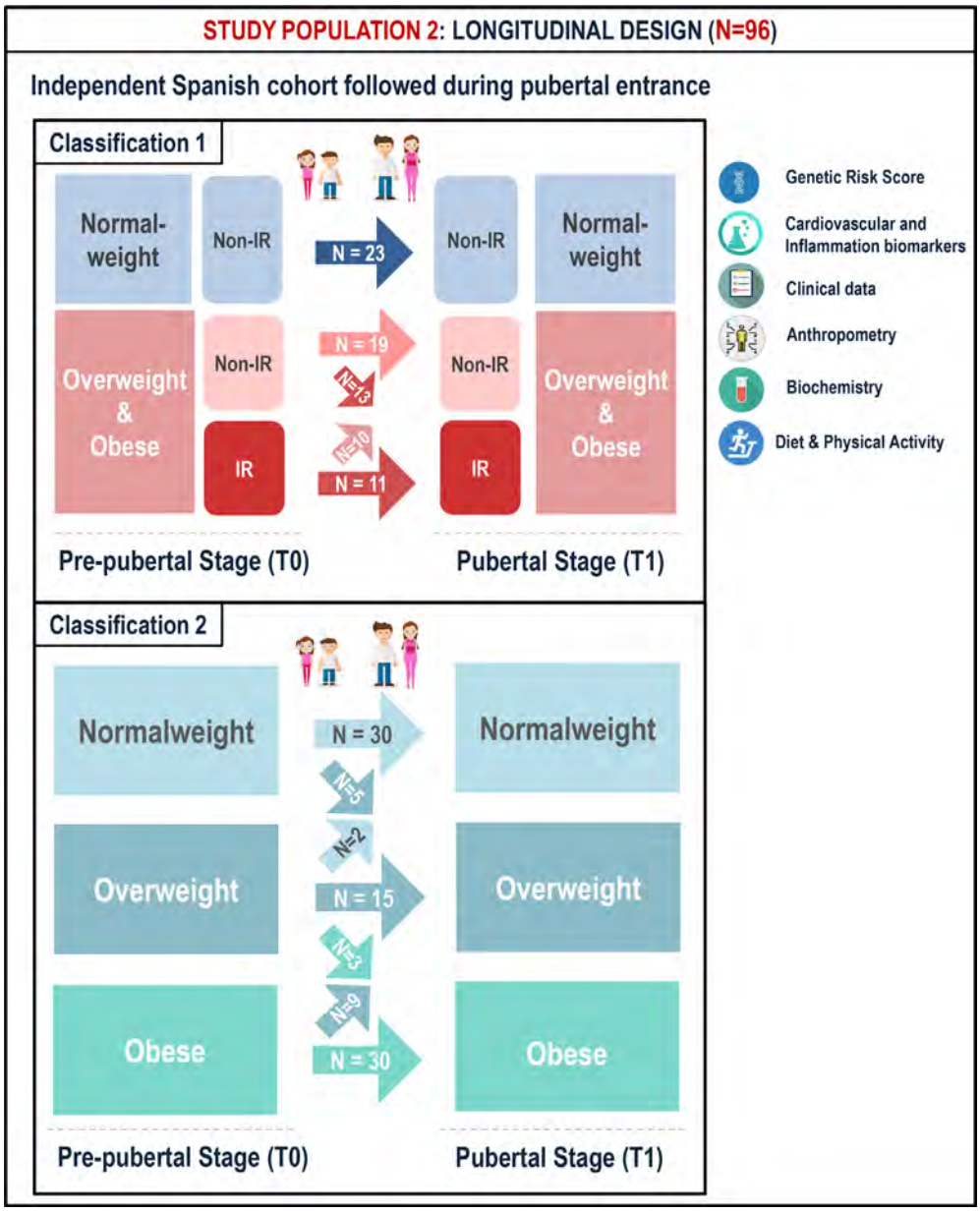


Figure 6. General characteristics of the PUBMEP population, which is based on a longitudinal design on 90 children undergoing puberty.

medical monitoring by the same paediatricians. The assessment of the pubertal stage was carried out following the Tanner classification and confirmed with a hormonal study. Another important inclusion criterion for this project was presenting a good-quality DNA sample in the prepubertal stage for omics analyses. A total of 374 subjects were contacted in the PUBMEP study, of which 49 were not located, 36 could not participate because they had changed their place of residence or met any of the exclusion criteria, and 98 declined the invitation. One hundred ninety-one answered affirmatively, and their parents or legal guardians accepted an appointment to receive all the information related to the PUBMEP study. For omics analyses, a sub-population of 139 children (76 females) from the whole PUBMEP cohort was selected. From them, the 90 Spanish children (47 females) were allocated into five experimental groups according to their obesity and IR status before and after the onset of puberty.

3 The Metformin pharmacogenetics Study: a randomized-control trial.

The Metformin pharmacogenetics study is a multi-centre and double blind randomized controlled trial conducted in 124 children with obesity (59 placebo (30 girls) and 65 treated children (33 girls)) (**Figure 7**). The aim of the present study was to test whether common genetics variants can predict the response to metformin intervention in children in terms of the post-treatment change in glucose metabolism, anthropometry, lipid metabolism, adipokines, and inflammatory markers. A complete workflow detailing the original study design can be found elsewhere[4]. Originally, 160 children with obesity were stratified according to sex and pubertal status and randomly assigned to receive either (1 g/d) metformin or placebo for 6 months after meeting the defined inclusion criteria. The study was registered by European Clinical Trials Database (EudraCT, ID: 2010-023061-21) on 14 November 2011 (URL: https://www.clinicaltrialsregister.eu/ctr-search/trial/2010-023061-21/ES).

4 Gene Expression Omnibus Public Datasets:

With the aim of implementing new unsupervised ML models for the analysis of longitudinal gene expression data, six temporal gene expression datasets available in the Gene expression Omnibus Public Repository were selected (IDs: GSE77962 (N=22), GSE77962 (N=24), GSE70529 (N=9), GSE35411 (N=9), GSE103766 (N=6), GSE103766 (N=13)). All datasets consisted on long-term human interventions for weight loss in individuals with obesity, with transcriptomics array data available in multiple time records (e.g., three or more).

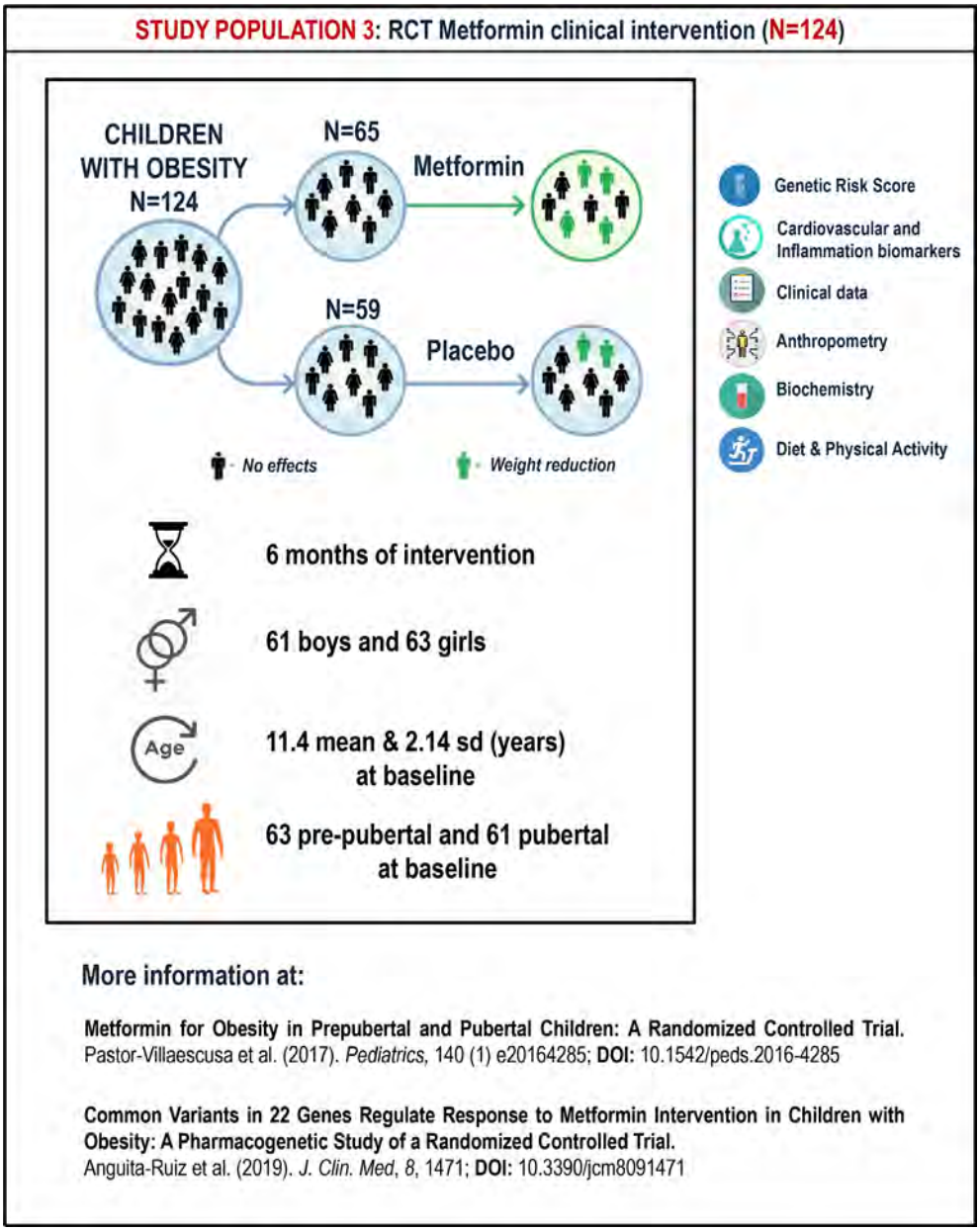


Figure 7. The Metformin pharmacogenetics Study corresponds to a previous multicentre and double blind randomized controlled trial (RCT) conducted in 124 children with obesity.

5 Ethics statement.

All studies were conducted following the Declaration of Helsinki (Edinburgh 2000 revised), and they followed the recommendations of the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legal in-forced Spanish regulation, which regulates the clinical investigation of human beings (RD 223/04 about clinical trials). Accordingly, the corresponding ethics committees approved the study at each of the participating centers (Code IDs GENOBOX: Córdoba01/2017, Santiago 2011/198, Zaragoza 12/2010; and PUBMEP: Córdoba 260/3408, Santiago 2016/522, Zaragoza 22/2016, Granada 01/2017). The metformin pharmacogenetics study was approved by the Ethics Committee for Biomedical Research of Andalusia on 15 January 2012 (acta 1/12) (ID code: 2010-2739).

6 Anthropometry, biochemical measurements, and inflammation and cardiovascular risk biomarkers.

Anthropometric measurements such as body weight (kg), height (cm), hip circumference (cm) and waist circumference (WC) (cm) were measured at each time point using standardized procedures, and BMI (kg/m^2) was calculated. BMI z-score was estimated based on the Spanish reference standards published by Sobradillo *et al.* (2000) [5]. Blood pressure was measured three times for each individual by the same examiner using a mercury sphygmomanometer and following international recommendations [6]. Measures of lipid and glucose metabolism, hormones and classical biochemical parameters were performed at the laboratories of each participating hospital following internationally accepted quality control protocols.

Blood samples from both time points were collected in overnight fasting conditions, centrifuged, and plasma and serum were stored at -80°C. Plasma adipokines, inflammation, and cardiovascular risk biomarkers (adiponectin, leptin, resistin, tumor necrosis factor alpha (TNF-α), high-sensitivity CRP (hsCRP), interleukin (IL)-6, IL-8, total plasminogen activator inhibitor-1 (PAI-1), P-Selectin, myeloperoxidase (MPO), monocyte chemoattractant protein 1 (MCP-1), matrix metalloproteinase-9 (MMP-9), soluble intercellular cell adhesion molecule-1 sICAM-1, and soluble vascular cell adhesion molecule-1 (sVCAM)) were analyzed in all samples and time points using XMap technology (Luminex Corporation, Austin, TX) and human monoclonal antibodies (Milliplex Map Kit; Millipore, Billerica, MA) as previously reported. S100A4 and VASN protein levels were determined in plasma using enzyme-linked immune-absorbent assay kits according to the manufacturers' instructions.

7 Omics analyses.

Genomic DNA was extracted from peripheral white blood cells using two automated kits, the Qiamp DNA Investigator Kit for coagulated samples and the Qiamp DNA Mini & Blood Mini Kit for non-coagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). All extractions were purified using the DNA Clean and Concentrator kit from Zymo Research (Zymo Research, Irvine, CA, USA).

a. Candidate-Gene Genotyping Analysis

Genotyping on a selection of candidate SNPs mapping genes previously associated with Obesity and Diabetes was performed by TaqMan allelic discrimination assay using the QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA).

b. Genome-wide association study (GWAS)

Whole-genome genotyping analysis was performed on the i-SCan platform using the Infinium HTS Assay (Illumina, San Diego, CA, USA). The Bead Chip selected for the project was the Infinium Global Screening Array-24 v3.0 Kit, which includes ~ 654,000 genetic markers associated with complex diseases. After quantification of DNA samples by fluorimetry, they were normalized to 200-400 ng of DNA per sample in deep well plates, as established in the Infinium HTS Assay Protocol.

c. Next-generation transcriptome sequencing (RNA-Seq)

RNA was extracted from peripheral blood using the PAXgene® Blood RNA Kit (PreAnalytiX/QIACUBE) according to the manufacturer's instructions. The concentration and quality of extracted RNA were measured using the Qubit 4 Fluorometer (Thermo Fisher Scientific, MA, USA) and the 2100 Bioanalyzer Instrument (Agilent Technologies, CA, USA). Libraries from mRNA were prepared using 1µg of RNA starting material and the TruSeq Stranded mRNA Library Prep Kit (Illumina, CA, USA) according to the manufacturer's protocol. This protocol captures poly-adenylated RNA by transcription by oligo-dT primer, after which the RNA is fragmented. The sample is back transcribed to generate the cDNA, both in the first and second strands. The 3' ends are adenylated, the adapters and barcodes are ligated, and finally, the sample is enriched by PCR. Adapters and sample codes (index-barcodes) are added to the libraries to be simultaneously sequenced. mRNA libraries were sequenced on the Next-Seq 500 system (Illumina, CA, USA) using the highest output mode and paired-end 75 bp read lengths with a depth of 20 million reads for each sample. To get a depth of 20 million reads per sample 2 runs with 4 lanes for each run were conducted.

d. Epigenome-wide association study (EWAS)

High-quality DNA samples (≥ 500 ng) were treated with bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research Corporation, Irvine, CA). DNA methylation was measured with the Infinium Methylation EPIC array using bead chip technology, which analyzes the methylation status of ~850.000 CpGs across the human genome (Illumina, San Diego, CA, USA).

Table 1. Studies' methodology overview.

Study	General Aim	Design	Cohort and number of participants	Main Study outcomes	Omic approach
Study 1	To study the association between TNMD X-chromosome genetic variants and glucose metabolism complications related to childhood obesity, and to evaluate the potential functionality of the TNMD gene in human adipocytes	Cross-sectional	GENOBOX (N=915)	Childhood obesity and Insulin Resistance	Candidate-Gene Genotyping Analysis
Study 2	To make public and describe a dataset incorporating phenotype and X chromosome genotype data from a cohort of 915 normal-weight, overweight and obese children, and to deeply describe a whole implementation of the special X chromosome analytic process in genetics (Methodological Study 2). This study was conducted as an analytic pipeline to solve some methodological drawbacks encountered during the analysis of Study 1	Cross-sectional	GENOBOX (N=915)	Childhood obesity and Insulin Resistance	Candidate-Gene Genotyping Analysis
Study 3	To test whether obesity-related genetics variants can predict the response to metformin intervention in terms of the post-treatment change in glucose metabolism, anthropometry, lipid metabolism, adipokines, and inflammatory markers in children with obesity	Randomized Controlled Trial The Metformin	Pharmacogenetics Study (N=124)	Childhood obesity, Insulin Resistance, lipid metabolism, adipokines, and inflammatory markers	Candidate-Gene Genotyping Analysis
Study 4	To evaluate the utility of an adult obesity-predisposing genetic risk score for the prediction and pharmacological management of obesity in Spanish children, further investigating its implication in the appearance of cardio-metabolic alterations	• Cross-Sectional • Repeated Measures • Randomized Controlled Trial	• GENOBOX (N=574) • PUBMEP (N=96) • The Metformin Pharmacogenetics Study (N=124)	Childhood obesity, Insulin Resistance, lipid metabolism, adipokines, and inflammatory markers	Candidate-Gene Genotyping Analysis
Study 5	To evaluate, through a multi-omics perspective, the association between the protein S100A4 and IR in children with obesity during pubertal development	• Cross-Sectional • Repeated Measures	• GENOBOX (N=279) - PUBMEP (N=53)	Childhood obesity and Insulin Resistance	EWAS, Trans-criptomics Array, Protein levels
Study 6	To understand, through a large-scale multi-omics longitudinal analysis, the molecular architecture and biological processes underlying the development of IR in puberty and the additional impact of obesity on these processes	Repeated Measures	• PUBMEP (N=139)	Childhood obesity and Insulin Resistance	EWAS, GWAS, RNAseq, Protein levels
Study 7	To implement a novel rule-based XAI strategy (including pre-processing, knowledge extraction and functional validation) for finding biologically relevant sequential patterns from longitudinal human gene expression data in obesity research	Repeated Measures	Gene Expression Omnibus (IDs: GSE77962 (N=22), GSE77962 (N=24), GSE70529 (N=9), GSE35411 (N=9), GSE103766 (N=6), GSE103766 (N=13))	Adult Obesity (weight-loss in response to ca-loric restriction)	Transcripto-mics Array

The column "omic approach" refers to the type of omic data analysed in each study.

References

1. Llorente-Cantarero, F. J. et al. Changes in physical activity patterns from childhood to adolescence: genobox longitudinal study. Int. J. Environ. Res. Public Health 17, 1–13 (2020).
2. Leis, R. et al. Cluster analysis of physical activity patterns, and relationship with sedentary behavior and healthy lifestyles in prepubertal children: Genobox cohort. Nutrients 12, (2020).
3. Latorre-Millán, M. et al. Dietary patterns and their association with body composition and cardiometabolic markers in children and adolescents: Genobox cohort. Nutrients 12, 1–18 (2020).
4. Pastor-Villaescusa B, et al. Evaluation of differential effects of metformin treatment in obese children according to pubertal stage and genetic variations: study protocol for a randomized controlled trial. Trials 17:323 (2016).
5. Sobradillo, B. et al. Fernández Lizárraga A, R. I. Fernández Lizárraga A, . Curvas y tablas de crecimiento (estudios longitudinal y transversal). Fundación Faustino Orbegozo Eizaguirre Madrid, Spain. (2004).
6. McCrindle, B. W. Assessment and management of hypertension in children and adolescents. Nat. Rev. Cardiol. 7, 155–163.

results

RESULTS

Section I

STUDY OF GENETIC VARIANTS ASSOCIATED WITH CHILDHOOD OBESITY AND ALTERATIONS IN THE GLUCOSE METABOLISM

- Study 1
- Study 2
- Study 3
- Study 4

Section II

IDENTIFICATION OF NEW MULTI-OMICS BIOMARKERS OF INSULIN RESISTANCE (IR) AND CARDIOMETABOLIC ALTERATIONS IN CHILDHOOD OBESITY DURING THE METABOLICALLY CRITICAL PERIOD OF PUBERTY

- Study 5
- Study 6

Section III

Section III Implementation of unsupervised machine learning (ML) models for the analysis of longitudinal omics data in obesity

- Study 7

Section I

Section I

STUDY OF GENETIC VARIANTS ASSOCIATED WITH CHILDHOOD OBESITY AND ITS ASSOCIATED-ALTERATIONS IN GLUCOSE METABOLISM

Sci Rep. 2019;9(1):3979. doi:10.1038/s41598-019-40482-0. IF: 3.998, Q1 at MULTIDISCIPLINARY SCIENCES.

Study 1 | Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism

Francisco Javier Ruiz-Ojeda[1,2,*,#], **Augusto Anguita-Ruiz**[1,2*], Azahara I. Rupérez[1], Carolina Gomez-Llorente[1,2,3], Josune Olza[1,2,3], Rocío Vázquez-Cobela[4], Mercedes Gil-Campos[3,6], Gloria Bueno[3,5], Rosaura Leis[3,4], Ramón Cañete[3,6], Luis A. Moreno[5], Ángel Gil[1,2,3], Concepción María Aguilera[1,2,3#].

Abstract Tenomodulin (TNMD) is a type II transmembrane glycoprotein that has been recently linked to obesity, and it is highly expressed in obese adipose tissue. Several sex-dependent associations have been observed between single-nucleotide polymorphisms (SNPs) of the *TNMD* gene, which is located in the X-chromosome, and obesity, type 2 diabetes mellitus (T2DM), and metabolic syndrome in adults. On the other hand, results are lacking for children. We aimed i) to study the association between *TNMD* genetic variants and metabolic complications related to childhood obesity and ii) to investigate the function of TNMD in human adipocytes. We conducted a case-control, multicenter study in 915 Spanish children and demonstrated significant positive associations between *TNMD* genetic variants and BMI z-score, waist circumference, fasting glucose, and insulin

Affiliations 1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n. 18016 Armilla, Granada, Spain. / 2. Instituto de Investigación Biosanitaria IBS.GRANADA, Complejo Hospitalario Universitario de Granada, Granada 18014, Spain. / 3. CIBEROBN (Physiopathology of Obesity and Nutrition Network CB12/03/30038), Instituto de Salud Carlos III (ISCIII), Madrid 28029, Spain. / 4. Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department (USC). Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complexo Hospitalario Universitario de Santiago, Santiago de Compostela, Spain. / 5. Growth, Exercise, NUtrition and Development (GENUD) Research Group, Universidad de Zaragoza, Zaragoza, Spain; Instituto Agroalimentario de Aragón (IA2), Instituto de Investigación Sanitaria de Aragón (IIS Aragón) and Centro de Investigación Biomédica en Red de Fisiopatología de la Nutrición y la Obesidad (CIBEROBN), Zaragoza, Spain. / 6. Department of Paediatrics, Reina Sofia University Hospital, Institute Maimónides of Biomedicine Investigation of Córdoba (IMIBIC), University of Córdoba, Avda Menéndez Pidal s/n, 14004 Córdoba, Spain.

* Equal contributors

\# Corresponding authors

resistance in boys, highlighting the SNP rs4828038. Additionally, we showed a BMI-adjusted inverse association with waist circumference in girls. Second, *in vitro* experiments revealed that TNMD is involved in adipogenesis, along with glucose and lipid metabolism in differentiated adipocytes, and these effects may be mediated through AMPK activation. Hence, these results suggest that *TNMD* genetic variants could be potentially useful as early life risk indicators for obesity and T2DM. In addition, we support the fact that TNMD exhibits significant metabolic functions in adipocytes.

Keywords: Obesity; adipocytes; tenomodulin; polymorphism, single-nucleotide polymorphisms; childhood.

Introduction

Childhood obesity is a major health problem (GBD 2015 Obesity Collaborators) characterized by an expansion of the adipose tissue (AT)[1]. Many children who are overweight or suffer from obesity before puberty maintain obesity in early adulthood, which is associated with increased morbidity and mortality [2]. The expansion of AT implies metabolic alterations that are mainly related to glucose and lipid metabolism [3]. White adipose tissue (WAT) is the main site for energy storage, but it is also an endocrine organ that secretes cytokines and adipokines [4]. White subcutaneous adipose tissue (SAT) and white visceral fat depots (VAT) represent 80% and 20% of total body fat storage, respectively. VAT size is strongly associated with insulin resistance, and it is well established that VAT and SAT are different with respect to adipocyte size and metabolic activity [5].

Tenomodulin (TNMD) was identified as a novel gene in 2001 by Brandau *et al.*[6] (2001) and Shukunami *et al.* [7](2001), and it is located in the human Xq22 region, where it spans approximately 15 kb [6,7]. TNMD is a type II transmembrane protein; it is described as an angiogenesis inhibitor and is highly expressed in hypovascular connective tissues such as tendons and cartilage [6,8]. Indeed, TNMD contains a putative proteinase cleavage and two glycosylation sites where the C-terminus of the protein is cleaved in those tissues [9–12]. Furthermore, its expression in human AT has been recently observed to be higher in obesity and lower after diet-induced weight loss [13]. Analyses of AT *TNMD* expression in obese and lean subjects have also shown that *TNMD* mRNA is correlated with body mass index (BMI) in adults [14–16]. In line with these results, our research group previously found that TNMD was five-fold upregulated in the VAT of prepubertal children with obesity, compared with their normal-weight counterparts [17]. Furthermore, *TNMD* is known to promote human adipocyte differentiation and to act as a protective factor against insulin resistance in obese VAT [18].

Likewise, several studies have indicated that single-nucleotide polymorphisms (SNPs) in the *TNMD* gene are associated with BMI, serum low-density lipoprotein cholesterol (LDL-c) levels, and

inflammatory factors in adults in a sex-specific manner [19]. Specifically, the SNPs rs2073162 and rs2073163 have been associated with type 2 diabetes mellitus (T2DM) in men, central obesity in women and inflammation in men and women [19–23]. On the other hand, results are lacking for children.

At the GWAS level, none of the analyses that have been conducted on obesity traits have reported associations for *TNMD* SNPs. Since the X-chromosome has often been less scrutinized because of the unique statistical challenges it presents [24,25], the X-chromosomal location of *TNMD* could be one of the reasons why its genetic variants have not been widely studied in the genetic context of obesity. Despite this, the X-chromosome has been proposed as a potential source of missing heritability and an important genomic region to be included into analyses [26]. Considering all this and the availability of new tools to overcome these complexities [25,27–30], the present work was undertaken to study the effects of *TNMD* genetic variants in children with obesity and to evaluate the potential metabolic function of this gene in human adipocytes. First, we studied the association between *TNMD* genetic variants and metabolic complications related to childhood obesity. Second, through gene silencing, we aimed to demonstrate that TNMD is required for adipocyte metabolism in fully differentiated adipocytes. To the best of our knowledge, this is the first study to report an association between *TNMD* SNPs and childhood obesity while supporting the implication of *TNMD* in adipocyte metabolism.

Results

TNMD genetic variants are associated with BMI z-score in boys

The anthropometric, clinical, and metabolic characteristics of the children participating in the present study are shown in the supplementary material according to obesity status (Supplementary Table S1). Minor allele frequencies (MAFs) of all markers studied are listed in **Table 1**. All SNPs showed MAFs above 5% regardless of the obesity class. Given the location of *TNMD* in a sex chromosome, all genetic analyses were conducted separately for boys and girls. The linkage disequilibrium (LD) pattern of the region of *TNMD* that was studied is presented in **Fig 1**; two previous literature-reported blocks were also identified in our population in a sex-stratified manner: haploblock-1 (rs11798018, rs5966709, and rs4828037) and haploblock-2 (rs2073162, rs2073163, rs4828038, and rs1155974) [20]. All SNPs within the haploblock-2 showed significant and positive association with the BMI z-score in boys but not in girls (**Table 1**). Conversely, no association was identified between variants of the haploblock-1 and BMI z-score in any sex group. Among the associated SNPs within the haploblock-2, the rs2073162 and the rs4828038 exhibited the highest effect sizes and the most significant P values. All mentioned associations remained statistically significant after applying multiple-test correction by False Discovery Rate (FDR). Instead, only the rs2073162 association stood multiple-test correction by *Bonferroni* adjustment (**Table 1**).

No effects were reported for the haploblock-1 variants on any of the studied phenotypes, and rs4828038 was identified as a tag SNP within the haploblock-2 according to the Bakker's method[31]; therefore, the paper will focus on all associations and findings that have been reported for this marker. A conditional joint multiple-SNP analysis for BMI z-score in boys further revealed there are no independent effects on the phenotype between all linked markers of the haploblock-2 (Supplementary Figure S1). On this matter, our tag SNP rs4828038 might be a good representative marker for the region. Additionally, haplotype-based tests were performed to determine whether the reported associations remained statistically significant when each *TNMD* haploblock was analyzed as an allelic phase and not as independent single variants (Supplementary Table S2). As expected, the association between the haploblock-2 and the BMI z-score in boys remained statistically significant, even after applying a multiple-test correction.

SNP	MAF A1/ A2	Normal-weight	Overweight	Obese	β (95%CI)	*P*-value	*FDR adjusted P*-value	*BONF adjusted P*-value
rs11798018	A/C							
Females		0.272	0.269	0.263	0.11 (−0.13, 0.36)	0.369	0.4696	1
Males		0.283	0.242	0.280	0.14 (−0.34, 0.63)	0.561	0.6545	1
rs5966709	T/G							
Females		0.289	0.294	0.355	0.14 (−0.07, 0.35)	0.202	0.404	1
Males		0.325	0.286	0.316	0.07 (−0.39, 0.52)	0.762	0.8206	1
rs4828037	C/T							
Females		0.297	0.332	0.379	0.15 (−0.06, 0.36)	0.161	0.404	1
Males		0.350	0.281	0.332	0.01 (−0.43, 0.46)	0.952	0.952	1
rs2073162	A/G							
Females		0.447	0.468	0.463	0.11 (−0.08, 0.30)	0.267	0.4262	1
Males		0.395	0.328	0.469	**0.65 (0.22, 1.07)**	**0.003**	**0.0233**	**0.042**
rs2073163	C/T							
Females		0.445	0.479	0.476	0.11 (−0.09, 0.30)	0.274	0.4262	1
Males		0.402	0.345	0.471	**0.59 (0.15, 1.03)**	**0.008**	**0.028**	0.110
rs4828038	T/C							
Females		0.447	0.459	0.463	0,12 (−0.06, 0.31)	0.193	0.404	1
Males		0.395	0.311	0.465	**0.64 (0.21, 1.06)**	**0.004**	**0.0233**	0.056
rs1155974	T/C							
Females		0.447	0.454	0.443	0.09 (−0.10, 0.29)	0.339	0.4696	1
Males		0.400	0.311	0.466	**0.62 (0.19, 1.04)**	**0.005**	**0.0233**	0.070

Table 1. *Association between TNMD SNPs and BMI z-score in children. BMI, body mass index; A1, minor allele; A2 major allele; MAF, minor allele frequency; β, Beta obtained under an additive model; CI, confidence interval. Linear regression analyses stratified by sex were performed under an additive model assuming TNMD locus escapes from the X-chromosome inactivation process. That is, while the female genotypes were coded 0, 1, or 2 according to 0, 1, or 2 TNMD SNP alleles, the genotypes for males were coded 0 or 1 according to 0 or 1 alleles.*

The tag SNP rs4828038 is associated with central adiposity and impaired glucose metabolism in boys, while it correlates with lower waist circumference in girls.

To further explore the implication of *TNMD* in obesity and metabolic alterations, we studied the association between *TNMD* genetic variants and a range of additional anthropometric measurements and metabolic features including cardiovascular disease (CVD) and inflammation biomarkers (**Table 2** and Supplementary Table S3).

Concerning anthropometric indicators of central obesity, a statistically significant and risky correlation was observed between the rs4828038-T-allele and waist circumference (WC) in boys, which disappeared after adjusting the model for BMI confounding. This finding did not remain statistically significant after applying FDR multiple-test correction neither. Conversely, we identified a protective association between the rs4828038-T-allele and WC in girls. The finding remained statistically significant after adjusting the model for BMI confounding (**Table 2**). Interestingly, this result also stood multiple-test correction and showed a 2.4% FDR value. Other central-adiposity indicators such as the waist-to-height ratio (WHR) reported equal findings in girls but did not reach multiple-test significance.

Regarding glucose metabolism, the rs4828038-T-allele was associated with higher levels of fasting glucose and higher values of homeostatic model assessment for insulin resistance (HOMA-IR) in boys. In the same way, the rs4828038-T-allele was also associated with a lower quantitative insulin sensitivity check index (QUICKI) (**Table 2**). All results were obtained under an additive model adjusted for age. Specifically, the correlation between our tag SNP and fasting glucose levels reached both nominal and multiple-test significance, exhibiting an FDR value of 4.9%. Significant results were also obtained after adjusting the model for BMI confounding but with only nominal statistical significance (P=0.009). In relation to HOMA-IR, each rs4828038-T-allele copy increased the index value by 0.33 units in comparison to the rs4828038-C-allele. For the QUICKI index, we reported a risky and statistically significant correlation in boys that stood multiple-test FDR significance. The same results were obtained after adjusting the model for BMI confounding but with only nominal statistical significance (P=0.039). No additional associations were reported regarding our tag SNP and any other glucose metabolism phenotype in neither boys nor girls.

According to previous studies relating *TNMD* SNPs to inflammatory traits and diseases such as age-related macular degeneration (AMD) or T2DM [21,32], we investigated the association between the rs4828038 SNP and inflammation, CVD risk markers and adipokines (**Table 2** and Supplementary Table S3). Interestingly, we reported a significant and positive correlation between the rs4828038-T-allele and interleukin (IL)-6 levels in girls, which remained significant after multiple-test correction (FDR=4.7%). Concordant results were obtained after adjusting the model for BMI confounding but with only nominal significance (P=0.021). Associations in line with this result have been previously identified in adult females in relation to *TNMD* haploblock-1 SNPs [21], which suggests that *TNMD*

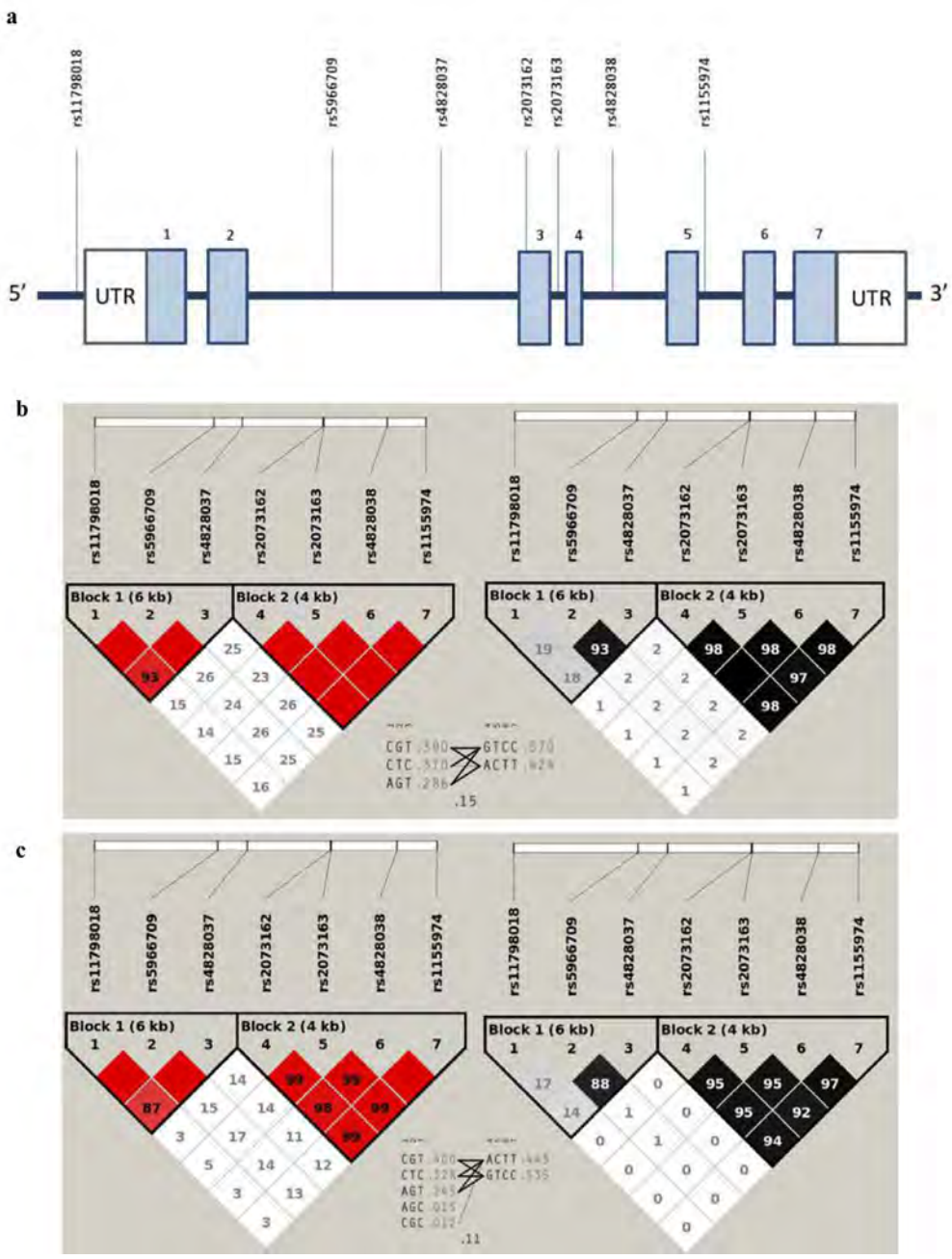


Figure 1. Location of selected markers in the TNMD gene and linkage disequilibrium (LD) analyses. (a) Light blue boxes represent exons, while the connecting blue lines are introns. Abbreviations: rs, reference SNP code; UTR, untranslated region. (b) and (c) show the LD pattern of the region in boys and girls, respectively. Left red triangles represent D′ values while right the black/gray triangles indicate R2 values. Triangle frames indicate observed haploblocks according to the solid spine of LD; the first consists of rs11798018, rs5966709, and rs4828037, and the second consists of rs2073162, rs2073163, rs4828038, and rs1155974. Between triangles, we listed each haplotype in a block along with its population frequency and connections from one block to the next. In the crossing areas, a value of multiallelic D′ is shown. This represents the level of recombination between the two haploblocks.

could mediate its putative effects on obesity via low-grade inflammation. Nonsignificant results were observed for the rest of the analyzed traits of the block regardless of the sex.

Other metabolic features such as blood pressure and lipid traits were also investigated. We detected a significant association between the rs4828038-T-allele and diastolic blood pressure (DBP) in girls when adjusting the model for BMI confounding (**Table 2**). The result did not remain statistically significant after applying multiple-test correction, showing an FDR value of 10.2%.

Phenotype	Genotypes T/T	$n_{T,T}$	T/C	$n_{T,C}$	C/C	$n_{C,C}$	β (95%CI)	P	FDR adjusted P	β_{BMI} (95%CI)	P_{BMI}	FDR adjusted P_{BMI}
Waist circumference (cm)												
Females	75.68 (15.11)	124	77.13 (16.68)	168	77.15 (15.13)	168	−0.42 (−2.05, 1.04)	0.611[a]	0.694	**−1.22 (−2.06, −0.38)**	**0.005[a]**	**0.024**
Males	80.9 (17.4)	157	NA	0	77.04 (17.3)	220	**3.76 (0.55, 6.98)**	**0.022[a]**	0.088	−0.1 (−1.54, 1.35)	0.897[a]	0.962
Waist to Height Ratio												
Females	0.526 (0.09)	109	0.54 (0.10)	151	0.53 (0.09)	142	−0.002 (−0.01, 0.009)	0.701[a]	0.876	**−0.01 (−0.01, 0.001)**	**0.030[a]**	0.172
Males	0.54 (0.10)	130	NA	0	0.53 (0.10)	194	0.015 (−0.007, 0.04)	0.183[a]	0.735	−0.002 (−0.01, 0.01)	0.678[a]	0.753
Systolic BP (mmHg)												
Females	104.8 (14.43)	108	106.9 (12.73)	148	104.9 (13.81)	136	0.006 (−1.65, 1.66)	0.995[b]	0.995	−0.46 (−1.95, 1.03)	0.542[b]	0.995
Males	107.7 (15.06)	131	NA	0	105.9 (16.31)	193	1.01 (−2.15, 4.17)	0.531[b]	0.995	0.07 (−2.91, 3.05)	0.963[b]	0.995
Diastolic BP (mm Hg)												
Females	62.82 (11.99)	108	65.17 (10.46)	148	65.46 (10.77)	136	−1.3 (−2.68, 0.09)	0.067[b]	0.332	**−1.60 (−2.91, −0.30)**	**0.016[b]**	0.102
Males	65.55 (11.31)	131	NA	0	63.82 (10.49)	193	1.53 (−0.82, 3.88)	0.204[b]	0.446	0.96 (−1.32, 3.23)	0.410[b]	0.708
Glucose (mg/dl)												
Females	84.7 (7.93)	132	83.64 (6.89)	170	84.37 (7.02)	172	0.16 (−0.66, 0.97)	0.707[a]	0.980	0.18 (−0.62, 0.99)	0.658[a]	0.906
Males	85.9 (8.45)	171	NA	0	83.71 (8.43)	235	**2.22 (0.57, 3.87)**	**0.009[a]**	**0.049**	**2.21 (0.55, 3.88)**	**0.009[a]**	0.054
Insulin (mU/dl)												
Females	12.85 (10.15)	131	13.5 (9.54)	164	11.63 (7.10)	167	0.81 (−0.14, 1.76)	0.094[a]	0.196	0.72 (−0.15, 1.59)	0.104[a]	0.443
Males	10.86 (7.67)	166	NA	0	9.50 (7.56)	224	1.39 (−0.02, 2.79)	0.053[a]	0.196	0.74 (−0.55, 2.02)	0.261[a]	0.443
HOMA-IR												
Females	2.76 (2.36)	131	2.82 (2.04)	163	2.44 (1.55)	166	0.2 (−0.01, 0.41)	0.064[a]	0.186	0.18 (−0.02, 0.38)	0.075[a]	0.386
Males	2.34 (1.79)	165	NA	0	2.02 (1.72)	224	**0.33 (0.009, 0.66)**	**0.044[a]**	0.186	0.19 (−0.11, 0.50)	0.207[a]	0.386
QUICKI												
Females	0.35 (0.05)	131	0.34 (0.04)	163	0.35 (0.04)	165	0.0001 (−0.004, 0.004)	0.953[a]	0.952	0.0005 (−0.003, 0.004)	0.787[a]	0.810
Males	0.35 (0.04)	165	NA	0	0.37 (0.05)	224	**−0.01 (−0.02, −0.003)**	**0.006[a]**	**0.038**	**−0.008 (−0.02, −0.0004)**	**0.039[a]**	0.232
TAG (mg/dl)												
Females	72.57 (36.18)	132	71.58 (32.59)	170	67.1 (27.12)	172	2.86 (−0.74, 6.46)	0.120[a]	0.615	2.63 (−0.87, 6.13)	0.142[a]	0.878
Males	65.53 (35.37)	172	NA	0	64.32 (35.51)	234	1.35 (−5.59, 8.29)	0.703[a]	0.915	−1.00 (−7.74, 5.73)	0.771[a]	0.905
HDL-c (mg/dl)												
Females	52.32 (14.03)	130	51.95 (13.77)	168	53.39 (13.69)	171	−0.6 (−2.17, 0.96)	0.451[a]	0.683	−0.35 (−1.76, 1.06)	0.625[a]	0.960
Males	55.25 (15.1)	169	NA	0	56.69 (16.53)	231	−1.53 (−4.67, 1.60)	0.338[a]	0.683	0.20 (−2.64, 3.04)	0.892[a]	0.960
IL-6 (ng/l)												
Females	3.86 (7.22)	106	2.78 (4.51)	152	2.21 (3.83)	151	**0.80 (0.17, 1.44)**	**0.014[a]**	**0.047**	**0.75 (0.12, 1.38)**	**0.021[a]**	0.073
Males	3.25 (6.29)	142	NA	0	3.53 (5.80)	205	−0.30 (−1.59, 0.98)	0.642[a]	0.818	−0.38 (−1.68, 0.92)	0.568[a]	0.727

Table 2. *Association between rs4828038 TNMD and anthropometric, biochemical, and inflammation characteristics (mean (SD)) in children. BMI, body mass index; BP, blood pressure; HOMA-IR, homeostasis model assessment for insulin resistance; QUICKI, quantitative insulin sensitivity check index; TAG, triglycerides; HDL-C, high-density lipoprotein cholesterol; IL, interleukin; MAF, minor allele frequency; βBMI, Beta obtained under an additive model adjusted for BMI; P-value BMI, P (P value) obtained under an additive model adjusted for BMI; CI, confidence interval; NA, not applicable. Linear regression analyses stratified by sex were performed under an additive model assuming TNMD locus escapes from the X-chromosome inactivation process. That is, while the female genotypes were coded 0, 1, or 2 according to 0, 1, or 2 TNMD SNP alleles, the genotypes for males were coded 0 or 1 according to 0 or 1 alleles. a Adjusted for age. b Adjusted for age and height.*

Remaining phenotypes of the block did not exhibit any significant correlation with analyzed markers (**Table 2** and Supplementary Table S3).

TNMD is associated with adipogenesis and lipid metabolism in human adipocytes

Based on the data regarding the relationship between *TNMD* genetic variants and the BMI z-score and, especially, WC, together with our previously published results that described a highly significant upregulation of *TNMD* expression in the VAT of children with obesity [33], we performed a functional *in vitro* study. The aim of this study was to elucidate the role of *TNMD* in human adipocytes, as the major constituent cells in AT. To assess *TNMD* gene and protein expression in human adipocytes from adipose-derived stem cells (ADSCs), we determined the mRNA and protein levels at various times during adipogenic differentiation. In agreement with previous reports [13,18], we found that *TNMD* expression and protein levels were significantly upregulated in human differentiated adipocytes at day 14 compared with ADSCs at day 0. However, we did not find any significant differences between days 7, 10, and 14. Immunofluorescence images also showed the differences between *TNMD* expression at days 0 and 14 (**Fig. 2**).

In addition, it has been demonstrated that *TNMD* inhibition blocks adipogenesis in Simpson-Golabi-Behmel syndrome (SGBS) preadipocytes and benefits VAT expansion in mice. Since TNMD is required for adipocyte differentiation in SGBS adipocytes [18], we first confirmed that TNMD was also required for human adipogenesis in ADSCs (Supplementary Fig. S2). Additionally, *TNMD* inhibition expression in fully differentiated adipocytes at day 14, downregulated the gene and protein expression of peroxisome proliferator-activated receptor gamma (PPARG), CCAAT/enhancer-binding protein alpha (CEBPA) and angiopoietin-like 4 (ANGPTL4) in TNMD-knocked-down adipocytes (**Fig. 3**).

Regarding lipid metabolism, knock-down of *TNMD* led to reduced lipolysis as observed by decreased extracellular glycerol levels in cell supernatants (**Fig. 3g**). In addition, the expression levels of lipases such as hormone-sensitive lipase (HSL), adipose triglyceride lipase (ATGL), and perilipin (PLIN) were significantly downregulated upon shRNA-TNMD treatment (**Fig. 3d-f**), as was ANGPTL4, which is a mediator of intracellular lipolysis in adipocytes (**Fig. 3c**).

Since TNMD is involved in the regulation of tenocyte proliferation, tendon development, and angiogenesis inhibition [12], we investigated vascular endothelial growth factor A (VEGFA) gene and protein expression in shRNA-mediated knocked-down *TNMD* cells. However, we did not observe significant changes (Supplementary Fig. S3). In this sense, no differences in blood vessel morphology and density have been previously found in TNMD transgenic mice [18].

TNMD knock-down impairs glucose metabolism in human adipocytes

To test whether TNMD plays a role in glucose metabolism in adipocytes, *TNMD* was inhibited in fully differentiated human adipocytes at day 14. In this experiment, we observed that glucose

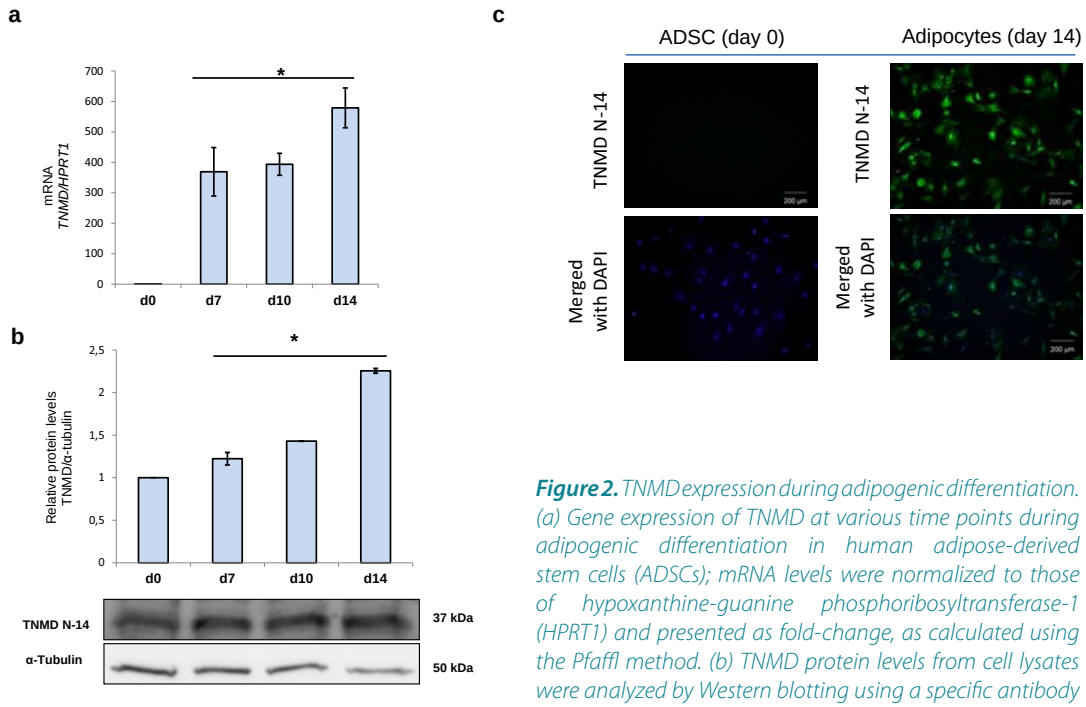


Figure 2. *TNMD expression during adipogenic differentiation. (a) Gene expression of TNMD at various time points during adipogenic differentiation in human adipose-derived stem cells (ADSCs); mRNA levels were normalized to those of hypoxanthine-guanine phosphoribosyltransferase-1 (HPRT1) and presented as fold-change, as calculated using the Pfaffl method. (b) TNMD protein levels from cell lysates were analyzed by Western blotting using a specific antibody against TNMD (N-14), normalized to the internal control (α-tubulin), and expressed as fold-change; the lower section presents a representative crop blot. (c) Immunofluorescent staining of ADSCs (d0) and differentiated adipocytes at day 14 N-14 terminal domains of TNMD (green) and 4.6-diamidino-2-phenylindole (DAPI; blue; scale bar, 200 μm). All values are expressed as the means ± SEM of three independent experiments. Significant differences were identified using the nonparametric Mann-Whitney U test; P-value: *<0.05.*

transporter 4 *(GLUT4)* gene and protein expression were downregulated when *TNMD* expression was inhibited (**Fig. 4. a, b**). This finding was supported by immunofluorescence images that showed reduced GLUT4 protein levels in shRNA-*TNMD*-treated adipocytes (**Fig. 4g**). However, *TNMD* inhibition did not affect basal glucose uptake in a significant manner; however, there was a tendency toward a reduction by approximately 1.5-fold (P-value: 0.08) (**Fig. 4c**). Adiponectin mRNA and protein levels were also decreased in *TNMD*-knocked-down adipocytes (**Fig. 4d**).

Regarding the activation of kinases involved in glucose metabolism, we observed a lower activation of AMP-activated protein kinase-alpha (AMPKα) and no differences in AKT (phosphor-AKT, Ser473) in *TNMD*-inhibited adipocytes. Thus, the association between TNMD and glucose metabolism in human adipocytes could be mediated by AMPK.

TNMD knock-down triggers inflammation in human adipocytes.

Many studies have reported that inflammation occurs in adipocytes associated with obesity, which is further related to metabolic dysfunction and insulin resistance [34,35]. Thus, we next studied the inflammatory status of *TNMD*-inhibited human differentiated adipocytes. In our study, *TNMD*-

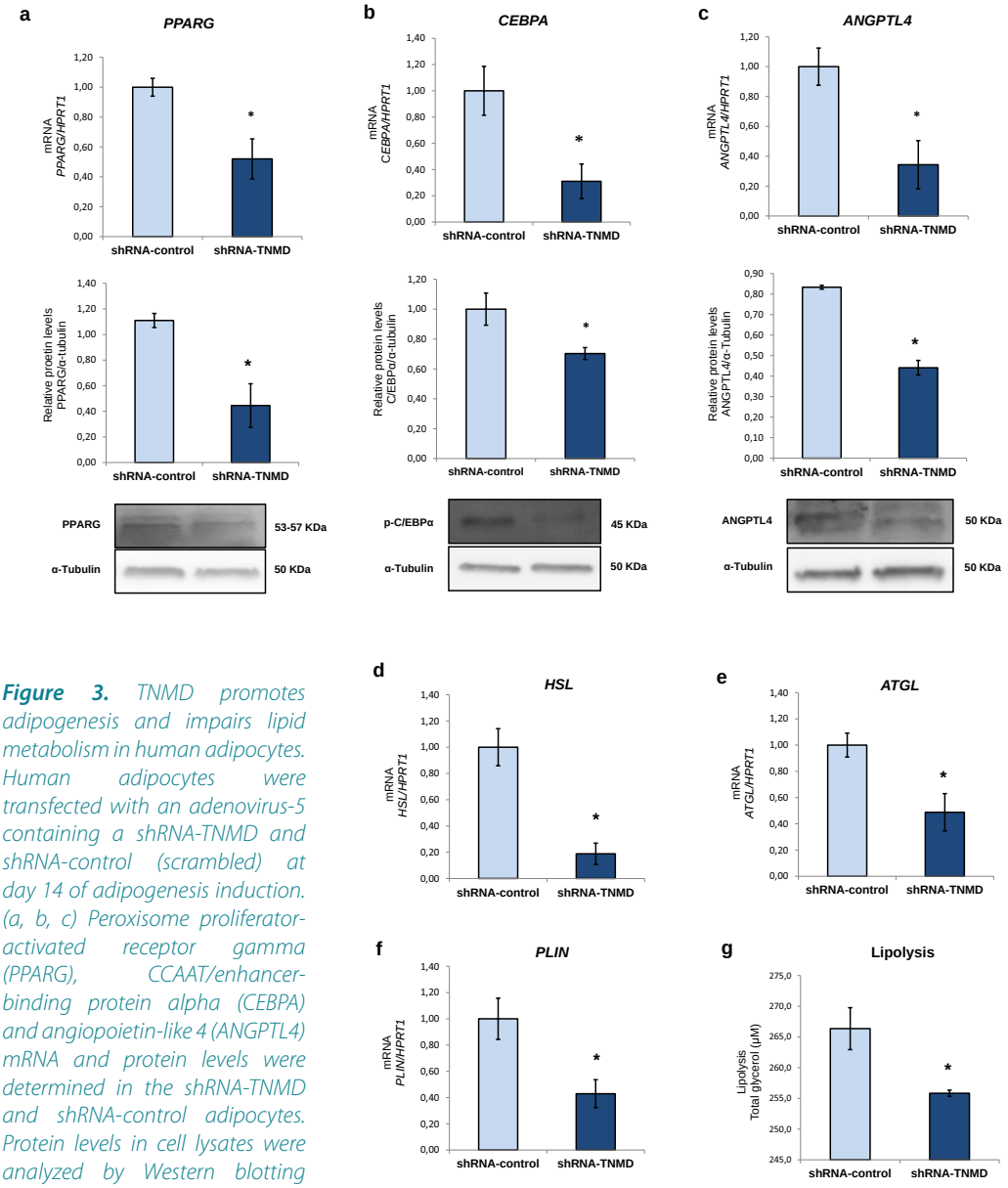


Figure 3. *TNMD promotes adipogenesis and impairs lipid metabolism in human adipocytes. Human adipocytes were transfected with an adenovirus-5 containing a shRNA-TNMD and shRNA-control (scrambled) at day 14 of adipogenesis induction. (a, b, c) Peroxisome proliferator-activated receptor gamma (PPARG), CCAAT/enhancer-binding protein alpha (CEBPA) and angiopoietin-like 4 (ANGPTL4) mRNA and protein levels were determined in the shRNA-TNMD and shRNA-control adipocytes. Protein levels in cell lysates were analyzed by Western blotting using specific antibodies against PPARG, CEBPA and ANGPTL4, normalized to the internal control (α-tubulin), and expressed as fold-change; the lower section shows a representative crop blot. (d, e, f) Hormone-sensitive lipase (HSL) gene expression, adipose triglyceride lipase (ATGL), and perilipin (PLIN) gene expression were determined in the shRNA-TNMD and shRNA-control adipocytes. (g) Glycerol levels (µM) in cell supernatants after treatment with shRNA-TNMD. All values are expressed as the means ± SEM of three independent experiments. Significant differences were identified using the nonparametric Mann-Whitney U test; P-value: *<0.05.*

knocked-down cells showed an upregulation of inflammatory markers such as *IL1-β* and tumor necrosis factor-α *(TNF-α)* mRNA. Furthermore, we observed a significant upregulation of the protein levels in the shRNA-*TNMD*-treated adipocytes compared with that in cells transfected with shRNA-control (P-value <0.05) (**Fig. 5**). However, we did not observe significant differences in the activation of the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) pathway through p65 subunit phosphorylation. Although TNMD could play a role against inflammation in adipocytes, more studies are needed to elucidate the underlying mechanism.

Discussion

In the present study, we show that X-chromosome *TNMD* genetic variants are associated with childhood obesity and metabolic alterations in a cohort of Spanish children. Particularly, we show that the tag SNP rs4828038 is associated with anthropometry, glucose metabolism alterations, and increased levels of pro-inflammatory biomarkers in a sex-specific manner. Furthermore, by *in vitro* gene silencing, we demonstrate that TNMD is required for an adequate glucose and lipid metabolism, and it plays a role in the control of inflammation in cultured human adipocytes.

Genetic association studies are commonly focused on autosomal variants, and genetic polymorphisms in sex chromosomes are often neglected [25]. Notwithstanding, the study of sex chromosomes might help to clarify the role of several genes in the development of many diseases, especially complex human traits that exhibit gender disparity in risk or symptoms [28,36,37]. Studies that employ mouse models and allow the distinction of gonadal from chromosomal effects have revealed that X-chromosome dosage influences food intake, which in turn affects adiposity and the occurrence of adverse metabolic conditions such as hyperinsulinemia, hyperlipidemia, and fatty liver [38].

To date, the present study is the first to analyze and detect associations between *TNMD* *X*-chromosome genetic variants and obesity and its metabolic complications in a cohort of children. For that purpose, we quantified our power in 96.12% to detect small GWAS-size genetic effects (estimated in F^2=0.02) at an alpha level of 0.05.

Specifically, we found a risky correlation between the rs4828038, tag SNP of the *TNMD* haploblock-2, and the BMI z-score and WC in boys. Interestingly, findings related to BMI z-score remained statistically significant after applying multiple-test correction. Regarding these anthropometric measurements, others authors such as Tolppanen *et al.* (2007), have also reported *TNMD* associations [20]. Surprisingly, they have found a protective association between variants located in the haploblock-1 (rs11798018, rs5966709 and rs4828037) and BMI and weight in adult European men. Although these findings are reported for a different haploblock than our associated haploblock-2 SNP, such controversy merits special attention. A possible explanation might rely on the fact that both haploblock-1 and haploblock-2 could elicit contrasting effects on

TNMD expression through, for example, the alteration of microRNA targets sites or the generation of different splicing patterns. Other sources of variability might also rely on the fact that we are studying a cohort of children, while Tolppanen *et al*. focus on adult population with an advanced status of impaired glucose tolerance (IGT) and T2DM. In this regard, further *TNMD* functional genetic studies are needed to clarify such an issue.

Regarding girls, we showed a protective correlation between our tag SNP rs4828038 and WC and WHR. In both associations, BMI was controlled for as a confounder. Specifically, the association between our tag SNP and WC further stood multiple-test correction (FDR=2.4%). Interestingly, our findings are in accordance with prior works of Tolppanen et al. (2007), who detected an association between the rs2073162-A-allele (marker in complete LD with our tag SNP) and smaller horizontal diameters in adult females when adjusting the model by BMI.

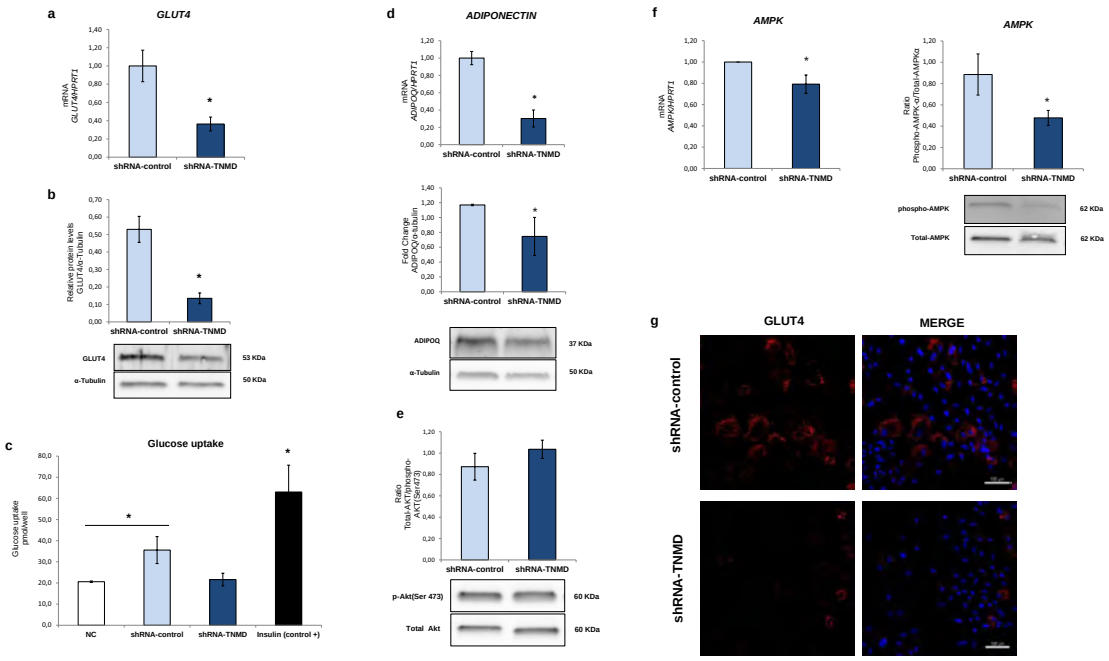


Figure 4. *TNMD is involved in glucose metabolism in human differentiated adipocytes. Human adipocytes were transfected with an adenovirus-5 containing a shRNA-TNMD and shRNA-control (scrambled) at day 14 of adipogenesis induction. (a) Glucose transporter 4 (GLUT4) mRNA levels were normalized to those of hypoxanthine-guanine phosphoribosyltransferase-1 (HPRT1), and the data from three independent experiments are presented as the fold-change, which was calculated using the Pfaffl method. (b) GLUT4 protein levels from cell lysates were analyzed by Western blot using a specific antibody against GLUT4, normalized to the internal control (a-tubulin) and expressed as fold-change; the lower section shows a representative crop blot. (c) Glucose uptake levels in shRNA-TNMD-treated adipocytes compared with the shRNA-control or insulin (1 μM, 30 min) as a positive control. (d) Adiponectin (ADIPOQ) mRNA and protein levels expressed as fold-change. (e) Ratio phosphor-AMPKa/total-AMPKa. (f) Ratio phosphor-AKT/total-AKT. (g) Immunofluorescent staining of adipocytes at day 14 with GLUT4 (red) and 4.6-diamidino-2-phenylindole (DAPI; blue; scale bar, 200 μm) in the shRNA-control and shRNA-TNMD-treated adipocytes. All values are expressed as the means ± SEM of three independent experiments. Significant differences were identified using the nonparametric Mann-Whitney U test; P-value: *<0.05.*

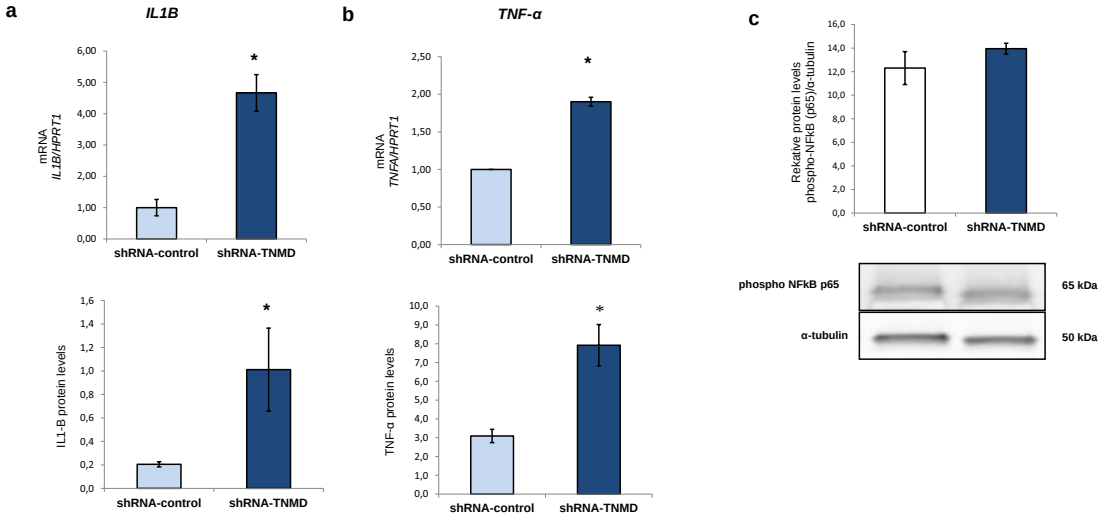


Figure 5. *TNMD triggers inflammation in human differentiated adipocytes. (a) mRNA expression and protein levels of interleukin 1-β (IL-1B); (b) mRNA expression and protein levels of tumor necrosis factor-α (TNF-α), and mRNA levels were normalized to those of hypoxanthine-guanine phosphoribosyltransferase-1 (HPRT1); TNF-α and IL1B protein levels were analyzed by XMap technology (Luminex) as indicated in the methods section. (c) Phospho-NFκB p65 protein levels were analyzed by Western blot using a specific antibody against phospho-NFκB p65, normalized to the internal control (α-tubulin), and expressed as the fold-change. The lower section shows a representative crop blot. The data from three independent experiments are presented as the means ± SEM. Significant differences were identified using the Mann-Whitney U test; P-value: *<0.05.*

Considering all this, we can see how the same region (haploblock-2) appears as a risk factor for obesity in boys, while at the same time, it acts as a protective element for central obesity parameters in girls. Such sex-specific behavior in our study reflects the typical sexual dimorphism of the X-chromosome very well, and it could arise from some X-chromosome particularities including differential gene dosage, the escape from the X-chromosome inactivation (XCI) and the existence of distinct genomic imprint mechanisms [38–42]. To account for all these X-chromosome particularities, several specific steps and procedures have been implemented following published recommendations [25,27–30] (see the method section).

Regarding glucose metabolism, several risky correlations were found for the rs4828038 SNP in boys. Particularly, we identified that the rs4828038-T-allele was associated with higher levels of fasting glucose and HOMA-IR, as well as lower values of QUICKI index. Interestingly, QUICKI and glucose associations remained statistically significant also after controlling for BMI confounding. In analyses without BMI-confounding adjustment, QUICKI and glucose insights further reached FDR multiple-test significance. Altogether, these associations are in concordance with previous findings that have been obtained by Tolppanen *et al.* (2007) during a 3-year follow-up study. For two *TNMD* SNPs (the rs2073163 and the rs1155974) in strong LD with our tag SNP, they showed that men carrying the C and T alleles (respectively) presented an altered oral glucose tolerance

test in comparison to individuals with the T and C alleles. These markers, along with the rs2073162, were also associated with an increased risk to develop T2DM during a 5-year follow-up study conducted in men [20]. Considering all this, we could hypothesize that the small alterations detected in the glucose metabolism of our boys according to the *TNMD* genotypes may be a premature signal of future complications during adulthood such as IGT or even its progression to T2DM. On this matter, *TNMD* genetic variants could be potentially useful as early life risk indicators for T2DM in male subjects. For girls we did not observe significant results in any glucose metabolism outcome. Previous studies in adults have reported contradictory findings in this regard.

In summary, although some of our genetic observations show a multiple-test level of statistical significance, it is important to stress, however, that they have not been corrected for between-trait multiple-test error. Thus, showed FDR corrected values are not study-wide robust and should be interpreted with caution. On the other hand, although we have taken some steps to account for main X-chromosome particularities, not all available suggestions were possible to incorporate in our study since this is a candidate-gene analysis instead of a GWAS approach. This, along with the fact that previous *TNMD* studies are statistically weak and barely accounted for X-chromosome specifications [20–22], indicates that our study should be viewed as hypothesis-generating instead of a replication approach. On this matter, more detailed characterization in bigger and independent children samples as well as additional follow-up studies during adulthood are needed.

According to these results in children, *TNMD* SNPs are associated with impaired glucose metabolism and we previously found *TNMD* overexpression in VAT from prepubertal obese children [17]. Other studies have also described that *TNMD* expression is highly upregulated in human AT, increased in obesity [14–16] and downregulated after diet-induced weight loss [13] and that *TNMD* expression is predominant in adipocytes compared with stromal vascular fraction (SVF) cells [18]. Furthermore, *TNMD* expression promotes preadipocyte proliferation and adipogenesis in SGBS adipocytes, and they improved insulin sensitivity in Tnmd transgenic mice, which suggesting the protective role of TNMD in VAT to alleviate insulin resistance in obesity [18]. Consistent with these results, *TNMD* knock-down led to lower gene expression and protein levels of important transcription factors that are involved in adipogenesis such as *PPAR-γ and C/EBP-α*. On the other hand, as *TNMD* is expressed in dense connective tissues as tendons and ligaments, and the C-terminal domain could be processed as a soluble factor, this fraction could reach the adipose tissue and promote the adipogenic differentiation *in vivo*. However, further studies are needed to clarify this possible effect. Therefore, our results confirm the fact that TNMD promotes adipogenic differentiation, and it could be implicated as a protective factor that contributes to AT expansion.

The reduced lipolysis observed in *TNMD*-knocked-down adipocytes could be explained by the reduced gene expression and protein levels of PPAR-γ, as well as by the reduced AMPK activation because AMPK is the master regulator of metabolism. Indeed, it has been described that *PPARG2*-knocked-out adipocytes exhibit reduced lipolysis [43]; this could be explained by the lower

expression of *HSL* [44] and *ATGL* [45], since they are both transcriptional targets of PPAR-γ. On the other hand, it has been reported that *ANGPTL4* promotes the expression of genes involved in lipolysis in adipocytes [46], and since *ANGPTL4* gene and protein levels were significantly downregulated when *TNMD* was inhibited, the results indicate that TNMD could be directly associated with lipid metabolism through ANGPTL4.

The association found between *TNMD* SNPs and fasting glucose levels in this study in children together with the higher 2-hour plasma glucose levels that was found in adults suggested a potential role for TNMD in adipocyte glucose metabolism. The confirmation of this hypothesis is another key finding in this study. We observed a down-regulation in gene expression and protein levels of the insulin-regulated glucose transporter GLUT4 in *TNMD*-knocked-down adipocytes and a tendency toward lower basal glucose uptake. Indeed, GLUT4 plays a critical role in the regulation of glucose metabolism and the maintenance of body glucose homeostasis [47]. Moreover, these results are in agreement with the observed lower adiponectin expression and lower AMPK activation. Adiponectin is the major secreted molecule of adipocytes and exerts multiple functions in regulation of energy homeostasis and glucose and lipid metabolism [48]. Adiponectin acts by increasing AMPK activity and stimulating *GLUT4* expression [49] and improves insulin sensitivity through inhibiting inflammatory signaling [50]. Upon activation, AMPK promotes *GLUT4* expression and its translocation to the plasma membrane, thus favoring glucose uptake independent of insulin [51,52].

On the other hand, it was reported that adiponectin knock-down did not affect the activation of AKT and p38MAPK (phosphorylation form/total form) but significantly decreased AMPK activation in insulin-responsive adipocytes [53]. In accordance with this finding, when *TNMD* expression was downregulated, we observed a reduced AMPK activation, lower adiponectin protein levels, and lower *GLUT4* expression. However, we did not observe differences in AKT activation. These results suggest that the mechanism underlying the link between TNMD and glucose metabolism involves activation of AMPK. We also observed downregulation of C/EBP-α, which could directly bind and activate the GLUT4 gene promoter. It has been demonstrated that insulin and dexamethasone activate GLUT4 gene expression through C/EBP-α gene expression in brown adipose tissue [54]. Moreover, exogenous expression of C/EBP-α in C/EBP-null cells with PPAR-γ overexpression resulted in an increase in GLUT4 mRNA levels.

Obesity is also associated with an increased expression of pro-inflammatory mediators in AT, and this inflammation has been shown to interfere with glucose metabolism [55]. More specifically, TNF-α has been proposed as a link between adiposity and the development of insulin resistance, given its high expression in the AT of subjects with obesity [56,57]. TNF-α is mainly produced in adipocytes and induces tissue-specific inflammation and insulin resistance through a reduction in *GLUT4* expression [58–60]. Furthermore, TNF-α upregulates the expression of IL-6, IL1-β, and protein phosphatase 2C (PP2C), which, in turn, suppresses AMPK activity. In addition, TNF-α downregulates

the expression of other important genes such as *adiponectin*, *C/EBP-α, PPAR-γ* and *PLIN* [61]. These circumstantial evidence support the fact that TNMD might be directly implicated in the protection against inflammation in the differentiated adipocytes since we found increased levels of TNF-α and IL1-β when *TNMD* was silenced. Additionally, IL-1β has been suggested to be involved in the development of insulin resistance 64. Collectively, these data suggest that lowering the expression of *TNMD* in adipocytes leads to a pro-inflammatory status, which contributes to the dysregulation of glucose metabolism. Nevertheless, these results warrant further studies to elucidate the precise mechanisms.

Finally, in the present work, we find that *TNMD* is highly expressed in human ADSCs and that it is involved in their differentiation into mature adipocytes. This finding is in line with the study performed by Senol-Cosar *et al.* (2016), where TNMD was reported to be involved in human adipogenesis in preadipocytes [18]. However, depending on the cell line, the effects of TNMD on adipogenic differentiation are not completely clear. Shi *et al.* (2017) [62] reported that *TNMD* overexpression did not affect the adipogenic differentiation in ASCs, which suggested that the endogenous *TNMD* gene is already expressed compared to other vascularized soft tissues [6]. Additionally, *TNMD* overexpression showed an inhibitory effect on the adipogenic differentiation of C3H10T1/2 and mMSC cells [63]. Interestingly, Lin *et al*. (2017) demonstrated the same results where *TNMD* knockout in mice exhibited significantly higher adipocyte accumulation, and *TNMD*$^{-/-}$ TSPCs exhibited a higher rate of differentiation into adipocytes [64]. The diverse regulatory mechanism of TNMD is involved in different cell types, and further studies are needed to elucidate the specific TNMD function *in vivo*.

In conclusion, our data show that *TNMD* genetic variants, specifically rs4828038, which is a tag SNP within the presented haploblock 2, are associated with obesity and alterations in glucose metabolism in children. These results replicate previous findings that have been observed in adults and suggest that these markers could be potentially useful as early life risk factor indicators for obesity and the occurrence of alterations in glucose metabolism during adulthood. Additionally, we found a novel paradigm for TNMD in human adipocytes, which plays a role in adipogenesis and glucose and lipid metabolism, and report that these effects might be mediated through AMPK activation. Recent studies have indicated that TNMD is not only a glycoprotein that is expressed in the connective tissue with antiangiogenic properties but also beneficial for VAT expansion [18]. Thus, we demonstrated and supported the fact that TNMD presents significant metabolic functions in adipocytes and that it might be a potential therapeutic target to improve the glucose metabolic status.

Methods

Study population

In this case-control multicenter study, 915 Spanish children (438 boys and 477 girls) were included from three health institutions: Lozano Blesa University Clinical Hospital, Santiago de Compostela University Clinical Hospital, and Reina Sofia University Hospital. Childhood obesity status was defined according to the International Obesity Task Force (IOTF) reference for children [65]. There were 480 children in the obesity group, 177 in the overweight group, and 258 in the normal-BMI group. Inclusion criteria were European-Caucasian heritage and the absence of congenital metabolic diseases. The exclusion criteria were non-European Caucasian heritage; the presence of congenital metabolic diseases (e.g., diabetes or hyperlipidemia); under-nutrition; and the use of medication that alters blood pressure, glucose or lipid metabolism.

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki (Edinburgh 2000 revised), and it followed the recommendations of the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legally enforced Spanish regulation, which regulates the clinical investigation of human beings (RD 223/04 about clinical trials). The Ethics Committee on Human Research of the University of Granada, the Ethics Committee of the Reina Sofía University Hospital of Cordoba, the Bioethics Committee of the University of Santiago de Compostela, and the Ethics Committee in Clinical Research of Aragon approved all experiments and procedures. All parents or guardians provided written informed consent, and the children gave their assent.

Anthropometric and biochemical measurements

Body weight (kg), height (cm), and WC (cm) were measured using standardized procedures, and the BMI z-score was calculated based on the Spanish reference standards published by Sobradillo *et al.* (2004) [66]. Blood pressure was measured three times by the same examiner using a mercury sphygmomanometer and following international recommendations [67]. The biochemical analyses were performed at the participating university hospital laboratories following internationally accepted quality control protocols, including routine measures of lipid and glucose metabolism. QUICKI and HOMA-IR were calculated using fasting plasma glucose and insulin values. Adipokines, CVD risk, and pro-inflammatory biomarkers [adiponectin, leptin, resistin, TNF-α, IL-6, IL-8, total plasminogen activator inhibitor-1 (PAI-1), myeloperoxidase (MPO), matrix metalloproteinase-9 (MMP-9), soluble intercellular cell adhesion molecule-1 sICAM-1, and soluble vascular cell adhesion molecule-1 (sVCAM)] were analyzed on a Luminex 200 system (Luminex Corporation, Austin, Tex., USA) with human monoclonal antibodies (EMD MilliporeCorp, Billerica, MA) using MILLIplexTM kits (HADK1MAG-61K, HADK2MAG-61K and HCVD2MAG-67K), as previously described [68]. High-sensitivity C-reactive protein (hsCRP) was determined using a particle-enhanced turbidimetric immunoassay (Dade Behring Inc., Deerfield, Ill, USA).

Genotyping

Genomic DNA was extracted from peripheral white blood cells using two kits, the Qiamp® DNA Investigator Kit for coagulated samples and the Qiamp® DNA Mini & Blood Mini Kit for noncoagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). All extractions were purified using a DNA Clean and Concentrator kit from Zymo Research (Zymo Research, Irvine, CA, USA).

Based on previously reported associations in adults [19–21] and according to the *Tagger* program [31], which was used to capture (at r2 = 0.8) common (MAF>=5%) variants in European (CEU) HapMap population, we selected seven SNPs located at the TNMD locus for the present association analysis. The seven selected SNPs are distributed through all TNMD sequences and are representative of the region (**Figure 1a**). Genotyping was performed by TaqMan allelic discrimination assay using the QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA). The call rate exceeded 95% for all tested SNPs, except for rs11798018 (92.7%). Minor allele frequencies (MAF) of all SNPs were > 2% and were similar to those reported for Iberian populations in Spain, in phase 3 of the 1000 Genomes Project. The Haploview software [69] was used with specific sex chromosome settings to assess the LD between SNPs in a sex-stratified manner.

Given the number of markers, we considered several parallel approaches to correct for multiple hypothesis testing based on the number of SNPs [70]. Specifically, we employed correction based on the methods proposed by Holm (1979) [70], Hommel (1988) [71], Benjamini and Yekutieli (2001) [72,73] and classical Bonferroni. To estimate the expected proportion of type I errors among the rejected hypotheses, we further computed false discovery rates (FDRs) as described in Benjamini and Hochberg [72]. Given the presence of linkage disequilibrium (LD), the FDR method is a proper approach that does not assume independence between markers.

X-chromosome Inactivation (XCI) assumptions

XCI is one of the main X-chromosome particularities that affect the analytical process. Varieties of statistical tests are available for performing genetic analysis of the X-chromosome, and the choice will mainly depend on the XCI model assumed for each target-study gene.

After revising the literature, we found that the *TNMD* genetic region is barely covered by current studies and that there is a lack of XCI data in adipose-tissue-derived samples [74,75,76]. In this sense, a recent study [77] reported that the XCI status of the *TNMD* region remains unknown. On the other hand, *TNMD* transcript levels have been reported to be two times higher in women than in men [78].

Given this controversy and lack of evidence, both possibilities ('escape' and 'XCI') were tested in our work (see the method section 'X-chromosome particularities and analyses' for more details).

X-chromosome particularities and analyses

Differential gene dosage between sexes, the escape from the XCI and the existence of distinct genomic-imprint mechanisms are some issues that make the X-chromosome a special region for genetic analyses. These particularities will determine important decisions that affect genotype calling, data imputation, quality control and statistics selection. The whole QC process was implemented in PLINK v1.07 [79].

Concerning genotype calling, algorithms that apply different procedures to male and female samples (e.g., Illuminus and CRLMM) have been proven to generally perform better than methods that do not (e.g., GenCall and GenoSNP) [80]. In our work, genotypes were called from fluorescence data files using the Applied Biosystems qPCR app module (ThermoFisher Cloud software) and the autocalling method. According to literature recommendations, sex information for each sample was supplied to the software and genotype calling was performed separately in both sexes. In this regard, although genotyped plates did not consist of only boys or girls, the balanced sex ratio of our population (477f/438m) (Table S1) favored a better performance. Five signal clusters were identified (three in the case of females and two in the case of males). Next, sex information and scatter of the clusters were used to call the genotypes (AA, AB and BB for females, and A- and B- for males). Since the employed software also allows the option of using user-definable boundaries for data analysis, those samples classified as undetermined by the autocalling method were recalled using the manual option. A set of controls were used to deduce these questionable genotype calls. Outliers were omitted from the analysis.

Next, we checked TNMD SNPs for sex-specific allele frequencies, which can induce type I errors in some statistical analyses (especially in the case of unbalanced designs). Tested by means of the Fisher exact test, all SNPs showed nonsignificant P values and thus equal allele frequencies across sex groups (Table S4). Two criteria concerning missing frequency were also employed (sex-specific missing frequencies and the differential missingness between sexes) [29,81]. As shown in table S5, our tag SNP passed the recommended filter in females (Missing Freq<= 2%) but not in males. Regarding the differential missingness test instead, only the rs11798018, rs4828037 and the rs2073163 passed the quality recommended filter ($P \geq 10\text{-}7$). This test was performed in the PLINK software using the flag "test-missing" and replacing the phenotype column of the .ped file by sex information. Regarding additional MAF quality checks, all SNPs showed appropriated frequencies > 1% by sex groups (Table S4). When analyzing the Hardy Weinberg equilibrium (HWE) in girls belonging to the normal-BMI group, all SNPs reported proper values ($P \geq 10\text{-}4$) (Table S6). According to this QC process, we ensured that there are no important genotyping errors and that our genetic data are reliable for further analyses.

Regarding high-level statistical analysis, both ('escape' and XCI) possibilities were tested as previously stated. At an initial phase, we assumed TNMD escapes from XCI and, thus, employed

linear and logistic regressions (stratified by sex) under an additive model. That is, females were coded as 0, 1, or 2, according to the presence of 0, 1, or 2 risk alleles, while males were coded as 0 or 1 according to the presence of 0 or 1 risk allele. This codification was achieved in the PLINK software from the binary file using the flag "*--dosage*". Additionally, we rerun all performed analyses using the X-chromosome specific version of common autosomal tests, developed by Clayton *et al.* (2008) [82]. Clayton's test explicitly accounts for random XCI and allowed the inclusion of females and males together, thereby, increasing the statistical power. This secondary analysis was performed using the snpStats R package [83]. Major significant associations that were reported during the initial phase were further replicated using Clayton's secondary approach (Table S7).

Data Records

The complete genetic data set in the present study complies with the requirements, and it has been uploaded into the European Genome-Phenome archive (EGA). Using the title *"X chromosomal genetic variants are associated with childhood obesity"*, the reference identifier of the project is EGAS00001002738 (2018). Data were uploaded according to obesity classes. The affected group (cases) was composed of children with obesity and those who were overweight (EGA EGAD00010001482 (2018)), and the control group was composed of normal-weight children (EGA EGAD00010001481 (2018)). Three by-experimental condition files are available online (.bed, .bim and .fam files). The .bed file contains the raw genotype data, while the .bim file describes the genotyped SNPs showing information related to chromosome number, SNP identifier, genetic distance in morgans (set as 0 for all markers), base-pair position (bp units) and allele letters. Finally, available fam files contain information related to the study population (sample identifier, family and paternal identifiers (here set as 0), sex (1 for males and 2 for females) and phenotype group (1 for control and 2 for cases)). As previously stated, data are available online according to each experimental condition.

Cell culture and adipogenic differentiation

Human adipose-derived stem cells (ADSCs) were purchased directly from Invitrogen (Gibco, Thermo Fisher Scientific, Carlsbad, CA, USA) (GibcoTM Lot 2117, StemPro Human ADSCs). These commercially available ADSCs are isolated from normal (nondiabetic) women subcutaneous lipoaspirates that are collected during elective surgical liposuction procedures. ADSCs have been reported to differentiate into many different lineages, including chondrogenic, osteogenic, adipogenic, and neural lineages. We cultured, expanded, and differentiated ADSCs into adipocytes according to the manufacturer's recommendations. Briefly, ADSCs were grown and expanded in appropriate sterile plastic dishes in complete Advanced-DMEM (Gibco, Thermo Fisher Scientific, Carlsbad, CA, USA) that was supplemented with 2 mM L-glutamine (25030, Gibco, Thermo Fisher Scientific, Carlsbad, CA, USA), 10% fetal bovine serum (FBS, PT-9000 H, Lonza, Basel, Switzerland), 100 U ml^{-1} penicillin and 100 µg ml^{-1} streptomycin (10378-016, Gibco, Thermo Fisher Scientific,

Carlsbad, CA, USA). We incubated cells at 37°C in a humidified atmosphere containing 5% CO2. The cell culture medium was replaced twice per week, and the cells were passaged up to a maximum of 6 times. To induce differentiation, we seeded cells in 35-mm dishes at a density of 30,000 cells/cm2, and we cultured them in MesenPRO RSTM medium (12746-012, Gibco, Thermo Fisher Scientific, Carlsbad, CA, USA). At 90% confluency, the growth medium was replaced with StemPRO RSTM adipogenic differentiation medium (A1007001, Gibco, Thermo Fisher Scientific, Carlsbad, CA, USA). ADSCs were incubated with differentiation medium for 14 days. We monitored and quantified adipogenesis through morphological examination of the cellular accumulation of lipid droplets via Oil Red O staining (234117, Sigma-Aldrich, St. Louis, MO, USA; Supplementary Figure S4A) and by spectrophotometric determination of washed Oil Red O staining (Supplementary Figure S4B). All treatments were performed on differentiated adipocytes at day 14.

Adenoviral transduction

Briefly, knock-down of TNMD was performed simultaneously using four different shRNA (TR300905 from Santa Origene, Rockville, Maryland, USA) packed in an adenovirus-5 vector (Ad-5). The production of ad-5 PacI-linearized plasmids (6 µg) containing the adenovirus genomes, as well as TNMD shRNA, or null sequences were transfected into 1×10^6 HEK293 cells and the viruses were recovered 8-10 days post-transfection. Next, the viruses were sequentially amplified until the infection of 4×10^8 HEK293 cells. Viruses were purified via two consecutive rounds of CsCl isopycnic density ultracentrifugation and desalted using a Sephadex PD-10 column (Amersham Biosciences, Uppsala, Sweden). The viral particles were measured via absorbance of disrupted virions at 260 nm where one O.D. equals 1×10^{12} particles per mL, while infective particles were measured via end-point dilution assay through counting the number of hexon-producing cells in triplicate [84]. The production of the vectors was conducted at Unitat de Producció de Vectors Virals-Cbateg, Barcelona, Spain. For adenovirus-shRNA experiments, human differentiated adipocytes were transfected with an Ad-5 containing shRNA-TNMD or shRNA-scrambled as a control using hexadimethrine bromide according to the manufacturer's protocol. First, to characterize the toxicity of adenovirus transduction in human adipocytes, we monitored the cellular viability in adipocytes that were exposed to different multiplicities of infection (MOI) (0, 10, 50, 100, 300, 500 and 1000 for 48 h) using a Neubauer chamber and trypan blue (4%). No toxicity was observed for the tested range of adenovirus. Subsequently, based on *TNMD* gene inhibition (approximately 90%), the MOI selected was 300 in all subsequent experiments. Forty-eight hours after transfection, the cells were collected.

RNA isolation and qRT-PCR

Total RNA was extracted from cells using the PeqGOLD HP Total RNA kit (Peqlab, Germany). Isolated RNA was treated with Turbo DNase (Ambion, Life Technologies, Carlsbad, CA, USA). We determined the final RNA concentration and quality, according to the 260/280 ratio, using

a NanoDrop2000 (NanoDrop Technologies, Winooski, Vermont, USA). Total RNA (500 ng) was transcribed into cDNA using the iScript cDNA Synthesis Kit (Bio-Rad Laboratories, California, USA). Next, we determined the differential gene expression levels of *TNMD* (330001 PHH12206A, Qiagen, Hilden, Germany), peroxisome proliferator-activated receptor gamma *(PPARγ)*, leptin (*LEP*), and adiponectin (*ADIPOQ*) during the adipogenic differentiation via qPCR using specific primer sequences (Table S8). The specific primer sequences were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/). Primers for glucose transporter 4 (*GLUT4*), interleukin 1-beta (IL1B), CCAAT/enhancer-binding protein alpha (CEBPA), angiopoietin-like 4 (*ANGPTL4*), tumor necrosis factor alpha (*TNF-α*), hormone-sensitive lipase (*HSL*), adipose triglyceride lipase (*ATGL*), perilipin (*PLIN*), and 5' AMP-activated protein kinase (AMPK) were obtained from Bio-Rad Laboratories, California, USA. qPCR was performed using an ABI Prism 7900HT instrument (Applied Biosystems, Foster City, CA, USA) and SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA, USA). Hypoxanthine-guanine phosphoribosyltransferase-1 (*HPRT1*) was used as a reference gene for the differentiation experiments. Quantification was performed using the Pfaffl method [85]. Compliance with the minimum information for publication of quantitative real-time PCR experiments (MIQE) was made possible using Bio-Rad's PrimePCR assays. We calculated the statistical validation of the stability of the reference genes in each sample. Bio-Rad recommends using a <0.5 value, which is the most stable expression in the tested samples. The results are expressed as the fold-change calculated.

Western blot assays

Protein samples from cell lysates that contain 2.5 µg of protein were mixed with 3X SDS-PAGE sample buffer (100 mM Tris-HCl, pH 6.8, 25% SDS, 0.4% bromophenol blue, 10% β-mercaptoethanol and 2% glycerol), separated via SDS-PAGE using a TGX Any kD gel (Bio-Rad Laboratories, California, USA), and transferred to a nitrocellulose membrane (Bio-Rad Laboratories, California, USA). After incubation in blocking buffer [5% nonfat milk and 0.1% Tween 20 in Tris-buffered saline (TBS)], the membranes were probed with one of the following antibodies: anti-TNMD-N14 (SC-49325; 1:200 in 5% nonfat milk), anti-GLUT4 (H61; 1:100 in 5% nonfat milk), and anti-Angptl4 (sc-373762; 1:500 in 5% BSA), which were acquired from Santa Cruz Biotechnology, CA, USA. Anti-adiponectin (AF1065, R&D Systems, Inc, USA; 1:500 in 5% bovine serum albumin, BSA), anti-PPAR-γ (D69; 1:1000 in 5% BSA), anti-phospho-C/EBP-α (Ser21) (1:1000 in 5% BSA), anti-total AMPK-α, anti-phosphorylated AMPK-α (phospho-AMPKα T172) (both 1:1000 in 5% BSA), anti-AKT (C67E7), and anti-phospho-AKT (Ser473, D9E) (1:1000 in 5% BSA), and anti-phospho-NF-kB p65 (Ser536) (1:500 in 5% BSA) were acquired from Cell Signaling Technologies (Beverly, MA, USA). We purchased anti-α-tubulin (internal control, 1:4000 in 5% nonfat milk) from Sigma. Immunoreactive signals were detected via enhanced chemiluminescence (Super-Signal West Dura Chemiluminescent Substrate, 34075, Thermo Fisher Scientific, Carlsbad, CA, USA). The membrane images were digitally captured and the densitometric analyses were conducted using the ImageJ software. The results were expressed as the fold-change in expression relative to the control. The graph shows a representative crop blot.

Immunofluorescence analysis

Human ADSCs were seeded on cover glasses and cultured for 2 days. Subsequently, adipogenic differentiation was performed over 14 days. We washed the adipocytes twice with PBS and fixed them with 4% paraformaldehyde for 30 minutes. Next, we incubated the cells with a permeabilization solution (0.5% saponin) for 10 minutes and washed them twice with PBS. Subsequently, the cells were incubated with working buffer (WB) containing 0.05% saponin and 1% bovine serum albumin, for 1 hour. The primary antibodies were anti-GLUT4 (1:100 in WB, H-61) and anti-TNMD-N-14 (1:50 in WB, SC-49325). We incubated the samples at 4°C overnight and washed the cover glasses three times with a working buffer for 5 min per wash. Next, the secondary antibodies were added, and GLUT4 and TNMD were visualized using an Alexa 488-conjugated chicken anti-goat IgG and Alexa 594-conjugated chicken anti-rabbit IgG at 1:1000 dilutions (Molecular Probes, Thermo Fisher Scientific, Carlsbad, CA, USA). Finally, we used ProLong Gold Antifade Mountant with DAPI (P36931, Molecular Probes, Thermo Fisher Scientific, Carlsbad, CA, USA) to fix cells with cover slips (Menzel-Glaser, 24 × 60 mm #1, Denmark). Image acquisition was performed with cells examined under a Nikon A1 confocal microscope equipped with a 20X immersion objective. Z-series optical sections were collected using a 1-micron-step-size and displayed as maximum z-projections using the NIS Elements/ImageJ software. Image acquisition was additionally performed using a fluorescence microscope (Olympus IX2).

Intracellular IL-1β and TNF-α protein levels

The intracellular IL-1β and TNF-α levels were determined in cell lysates in the shRNA-TNMD- and shRNA-control-treated adipocytes. Samples were harvested with protein lysis buffer, diluted in the appropriate buffer diluents and added to the wells with the rest of the reagents. IL-1β and TNF-α were determined using a MILLI*plex*™ kit (HSTCMAG-28SK) on a Luminex 200 system (Luminex Corporation, Austin, Tex., USA).

Glucose-uptake assays

Glucose uptake was determined using a colorimetric assay kit (MAK083, Sigma-Aldrich, St. Louis, MO, USA). Briefly, we differentiated ADSCs in 12-well plates, as described in the "Cell culture and incubation" section. After adenovirus transfection at day 14, we washed the differentiated adipocytes twice with PBS and starved them overnight in a serum-free medium. Next, we washed the cells 3 times with PBS and glucose starved them by incubating for 40 min in KRPH buffer (5 mM Na2HPO4, 20 mM HEPES, pH 7.4, 1 mM MgSO4, 1 mM CaCl2, 137 mM NaCl and 4.7 mM KCl) containing 2% BSA. Glucose uptake was assessed with 1 mM 2-deoxy-D-glucose in KRPH for 20 min at 37°C and 5% CO2. As a positive control, the cells were stimulated with insulin (1 µM) for 20 min. Glucose uptake levels were expressed in pmol/well.

Statistical analysis

The results in the tables are presented as the mean (SD). The one-way ANOVA test and Tukey post hoc test were performed to compare phenotype data between obese, overweight, and normal-BMI children. P values < 0.05 were considered statistically significant. These statistical analyses were conducted in R environment [86]. A specific genetic analysis design was implemented in PLINK v1.07 [78] and R environment to handle the X-chromosomal location of TNMD; the respective codes are available upon request.

In vitro experiments were repeated at least three times. In each experiment, two replicates were performed. Data are expressed as the mean ± standard error of the mean (SEM). Significant differences in the levels of gene and protein expression and glucose uptake were determined using the nonparametric Mann-Whitney U test; statistical significance was defined as P-value *< 0.05, P-value **< 0.01. Statistical analyses were performed using SPSS version 22, for Windows (SPSS, Chicago, IL, USA).

Acknowledgments

This work was supported by Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), Instituto de Salud Carlos III-Fondo de Investigación Sanitaria (Projects numbers PI020826, PI051968, PI1102042, PI1600871), RETIC (Red SAMID RD12/0026/0015), Fondo Europeo De Desarrollo Regional (FEDER) and the Junta de Andalucía (project number CTS-6770); Secretaría General de Universidades, Investigación y Tecnología. Consejería de Economía, Innovación y Ciencia). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This paper will be part of Augusto Anguita-Ruiz's doctorate, which is being performed under the "Nutrition and Food Sciences Program" at the University of Granada.

Author contributions

Responsible for child recruitment and anthropometric measures: R. V.-C., M. G.-C., G.B., R.L., R.C. and L.M. Conception and design of experiments: F.J.R.-O., C.G.-L., A.G. and C.M.A. Performed in vitro experiments: F.J.R.-O. Performed the SNP analysis results: A.A.-R. and A.I.R. Analyzed the results: F.J.R.-O., A.A.-R., A.I.R., J.O., C.G.-L., A.G., and C.M.A. Wrote the manuscript: F.J.R.-O., A.A.-R., and C.M.A.

Competing Interests:

The authors declare that they have no competing interests.

Data Availability:

The authors confirm that all data underlying the findings are fully available without restriction. All raw data underlying the findings described in the manuscript are freely available in Figshare. doi: 10.6084/m9.figshare.5258980.

Electronic supplementary material

Supplementary information is available at https://doi.org/10.1038/s41598-019-40482-0

References

1. Effects, H. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N. Engl. J. Med. 377, 13–27 (2017).
2. Jones, R. E., Jewell, J., Saksena, R., Ramos Salas, X. & Breda, J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Front. Public Heal. 5, 1–12 (2017).
3. Morigny, P., Houssier, M., Mouisel, E. & Langin, D. Adipocyte lipolysis and insulin resistance. Biochimie 125, (2016).
4. Baraban, E. et al. Anti-inflammatory properties of bone morphogenetic protein 4 in human adipocytes. Int. J. Obes. 40, 319–327 (2016).
5. Shoelson, S. E., Lee, J. & Goldfine, A. B. Review series Inflammation and insulin resistance. J. Clin. Invest. 116, 1793–1801 (2006).
6. Brandau, O., Meindl, A., Fässler, R. & Aszódi, A. A novel gene, tendin, is strongly expressed in tendons and ligaments and shows high homology with chondromodulin-I. Dev. Dyn. 221, 72–80 (2001).
7. Shukunami, C., Oshima, Y. & Hiraki, Y. Molecular cloning of tenomodulin, a novel Chondromodulin-I related gene. Biochem. Biophys. Res. Commun. 280, 1323–1327 (2001).
8. Oshima, Y. et al. Expression and localization of tenomodulin, a transmembrane type chondromodulin-I-related angiogenesis inhibitor, in mouse eyes. Investig. Ophthalmol. Vis. Sci. 44, 1814–1823 (2003).
9. Hiraki, Y. et al. Identification of Chondromodulin I as a Novel. Biochemistry 272, 32419–32426 (1997).
10. Docheva, D., Hunziker, E. B., Fässler, R. & Brandau, O. Tenomodulin is necessary for tenocyte proliferation and tendon maturation. Mol. Cell. Biol. 25, 699–705 (2005).
11. Naritaka Kimura, Chisa Shukunami, Daihiko Hakuno, Masatoyo Yoshioka, Shigenori Miura, Denitsa Docheva, Tokuhiro Kimura, Yasunori Okada, Goki Matsumura, Toshiharu Shin'oka, Ryohei Yozu, Junjiro Kobayashi, Hatsue Ishibashi-Ueda, Yuji Hiraki, K. F. Local tenomodulin absence, angiogenesis, and matrix metalloproteinase activation are associated with the rupture of the chordae tendineae cordis. Circulation 118, 1737–1747 (2008).
12. Alexandrov, V. P. & Naimov, S. I. A Prospectus of Tenomodulin. Folia Med. (Plovdiv). 58, 19–27 (2016).
13. Saiki, A. et al. Tenomodulin is highly expressed in adipose tissue, increased in obesity, and down-regulated during diet-induced weight loss. J. Clin. Endocrinol. Metab. 94, 3987–3994 (2009).
14. Kolehmainen, M. et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. Int. J. Obes. 2005 32, 292–303 (2008).
15. Johansson, L. E. et al. Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance. Am J Clin Nutr. 96, 196-207 (2012).
16. González-Muniesa, P., Marrades, M. P., Martínez, J. A. & Moreno-Aliaga, M. J. Differential proinflammatory and oxidative stress response and vulnerability to metabolic syndrome in habitual high-fat young male consumers putatively predisposed by their genetic background. Int. J. Mol. Sci. 14, 17238–17255 (2013).
17. Aguilera, C. M. et al. Genome-wide expression in visceral adipose tissue from obese prepubertal children. Int. J. Mol. Sci. 16, 7723–7737 (2015).
18. Senol-Cosar, O. et al. Tenomodulin promotes human adipocyte differentiation and beneficial visceral adipose tissue expansion. Nat. Commun. 7, 10686 (2016).
19. Tolppanen, A.-M. et al. The genetic variation in the tenomodulin gene is associated with serum total and LDL cholesterol in a body size-dependent manner. Int. J. Obes. 2005 32, 1868–1872 (2008).
20. Tolppanen, A.-M. et al. Tenomodulin is associated with obesity and diabetes risk: the Finnish diabetes prevention study. Obes. Silver Spring Md 15, 1082–1088 (2007).
21. Tolppanen, A.-M. et al. The genetic variation of the tenomodulin gene (TNMD) is associated with serum levels of systemic immune mediators--the Finnish Diabetes Prevention Study. Genet. Med. Off. J. Am. Coll. Med. Genet. 10, 536–544 (2008).
22. Tolppanen, A.-M., Kolehmainen, M., Pulkkinen, L. & Uusitupa, M. Tenomodulin gene and obesity-related phenotypes. Ann. Med. 42, 265–275 (2010).
23. Dex, S., Lin, D., Shukunami, C. & Docheva, D. Tenogenic modulating insider factor: Systematic assessment on the functions of tenomodulin gene. Gene 587, 1–17 (2016).
24. Accounting for sex in the genome. Nat. Med. 23, 1243 (2017).
25. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion : Toward Integrating the X Chromosome in Genome-wide Association Analyses. Am. J. Hum. Genet. 92, 643–647 (2013).
26. Loley, C., Erdmann, J. & Ziegler, A. How to Include Chromosome X in Your Genome-Wide Genetic Epidemiology. (2014). doi:10.1002/gepi.21782
27. Hickey PF, B. M. X chromosome association testing in genome wide association studies. Genet Epidemiol. 35, 664–70 (2011).
28. Chang, D., Gao, F., Slavney, A., Ma, L. & Waldman, Y. Y. Accounting for eXentricities : Analysis of the X

Chromosome in GWAS Reveals X- Linked Genes Implicated in Autoimmune Diseases. 1–31 (2014). doi:10.1371/journal.pone.0113684

29. König, I.R., Loley C., Erdmann J., Ziegler A. How to Include Chromosome X in Your Genome-Wide Association Study. Genet. Epidemiol. 38, 97-103 (2014).
30. Gao, F. et al. Computer Note XWAS : A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. 666–671 (2015). doi:10.1093/jhered/esv059
31. Bakker, P. I. W. De et al. Efficiency and power in genetic association studies. 37, 1217–1223 (2005).
32. Tolppanen, A. et al. Single nucleotide polymorphisms of the tenomodulin gene (TNMD) in age-related macular degeneration. 762–770 (2009).
33. Aguilera, C. M., Gomez-llorente, C., Tofe, I. & Gil-campos, M. Genome-Wide Expression in Visceral Adipose Tissue from Obese Prepubertal Children. 7723–7737 (2015).
34. Olefsky, J. M. & Glass, C. K. and Insulin Resistance. Annual review of physiology 72, (2010).
35. Lumeng, C. N. & Saltiel, A. R. Review series Inflammatory links between obesity and metabolic disease. Life Sci. 121, 2111–2117 (2011).
36. Maher, B. The case of the missing heritability. 456, (2008).
37. Manolio, T. A. et al. NIH Public Access. 461, 747–753 (2010).
38. Chen, X. et al. The Number of X Chromosomes Causes Sex Differences in Adiposity in Mice. 8, (2012).
39. Carruth, L. L., Reisert, I. & Arnold, A. P. brief communications Sex chromosome genes directly affect brain Auditory midbrain. Nature Neuroscience. 5, 933–934 (2002).
40. Dewing, P., Shi, T., Horvath, S. & Vilain, E. Sexually dimorphic gene expression in mouse brain precedes gonadal differentiation. Brain Res Mol Brain Res. 118, 82–90 (2003).
41. Reisert, I. & Pilgrim, C. Sexual differentiation of monoaminergic neurons--genetic or epigenetic? Trends Neurosci. 14, 468–73 (1991).
42. Burgoyne PS, Thornhill AR, Boudrean SK, Darling SM, Bishop CE, E. E. Mechanisms in vertebrate sex determination - The genetic basis of XX-XY differences present before gonadal sex differentiation in the mouse. Philos Trans R Soc L. B Biol Sci. 350, 253–60 (1995).
43. Rodriguez-Cuenca, S., Carobbio, S. & Vidal-Puig, A. Ablation of Pparg2 impairs lipolysis and reveals murine strain differences in lipolytic responses. FASEB J. 26, 1835–1844 (2012).
44. Yajima, H., Kobayashi, Y., Kanaya, T. & Horino, Y. Identification of peroxisome-proliferator responsive element in the mouse HSL gene. Biochem. Biophys. Res. Commun. 352, 526–531 (2007).
45. Kershaw, E. E. et al. NIH Public Access. Am. J. Physiol. 293, (2010).
46. Mandard, S. et al. The fasting-induced adipose factor/angiopoietin-like protein 4 is physically associated with lipoproteins and governs plasma lipid levels and adiposity. J. Biol. Chem. 281, 934–944 (2006).
47. Manna, P., Achari, A. E. & Jain, S. K. Vitamin D supplementation inhibits oxidative stress and upregulate SIRT1/AMPK/GLUT4 cascade in high glucose-treated 3T3L1 adipocytes and in adipose tissue of high fat diet-fed diabetic mice. Arch. Biochem. Biophys. 615, 22–34 (2017).
48. Yamauchi, T. et al. Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase. Nat. Med. 8, 1288–95 (2002).
49. Tanabe, H. et al. Crystal structures of the human adiponectin receptors. Nature 520, 312+ (2015).
50. Maeda, N. et al. Diet-induced insulin resistance in mice lacking adiponectin/ACRP30. Nat Med 8, 731–737 (2002).
51. R, G. Molecular mechanisms of GLUT4 regulation in adipocytes. Diabetes Metab 40, 400--10 (2014).
52. Bolsoni-Lopes, A. et al. Palmitoleic acid (n-7) increases white adipocytes GLUT4 content and glucose uptake in association with AMPK activation. Lipids Health Dis. 13, 199 (2014).
53. Chang, E. et al. Adiponectin deletion impairs insulin signaling in insulin-sensitive but not insulin-resistant 3T3-L1 adipocytes. Life Sci. 132, 93–100 (2015).
54. Im, S.-S., Kwon, S.-K., Kim, T.-H., Kim, H.-I. & Ahn, Y.-H. Regulation of glucose transporter type 4 isoform gene expression in muscle and adipocytes. IUBMB Life 59, 134–145 (2007).
55. Hotamisligil, G. S. Inflammation, metaflammation and immunometabolic disorders. Nature 542, 177–185 (2017).
56. Hotamisligil, G. S. Inflammatory pathways and insulin action. Int J Obes Relat Metab Disord. (2003).
57. Nieto-Vazquez I, Fernández-Veledo S, Krämer DK, Vila-Bedmar R, Garcia-Guerra L, L. M. Insulin resistance associated to obesity: the link TNF-alpha. Arch Physiol Biochem. 114, (2008).
58. Akash, M. S. H., Rehman, K. & Liaqat, A. Tumor Necrosis Factor-Alpha: Role in Development of Insulin Resistance and Pathogenesis of Type 2 Diabetes Mellitus. J. Cell. Biochem. (2017). doi:10.1002/jcb.26174
59. Olson, A. L. Regulation of GLUT4 and Insulin-Dependent Glucose Flux. ISRN Mol. Biol. 2012, (2012).
60. Guilherme, A., Virbasius, J. V, Vishwajeet, P. & Czech, M. P. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. Nat. Rev. Mol. … 9, 367–377 (2008).

61. Gao, D. et al. Interleukin-1β mediates macrophage-induced impairment of insulin signaling in human primary adipocytes. Am. J. Physiol. Endocrinol. Metab. 307, E289-304 (2014).
62. Shi, Y. et al. Conditional tenomodulin overexpression favors tenogenic lineage differentiation of transgenic mouse derived cells. Gene 598, 9–19 (2017).
63. Yongkang Jiang Yuan Shi Jing He Zhiyong Zhang Guangdong Zhou Wenjie Zhang Yilin Cao Wei Liu. Enhanced tenogenic differentiation and tendon-like tissue formation by tenomodulin overexpression in murine mesenchymal stem cells. J. Tissue Eng. Regen. Med. 11, (2016).
64. Lin, D. et al. Tenomodulin is essential for prevention of adipocyte accumulation and fibrovascular scar formation during early tendon healing. Nat. Publ. Gr. 8, e3116-12 (2017).
65. Cole, T. J., Bellizzi, M. C., Flegal, K. M. & Dietz, W. H. and obesity worldwide : international survey. 1–6 (2000).
66. Sobradillo B, Aguirre A, Aresti U, Bilbao A, Fernández Ramos C, Lizárraga A, Lorenzo H, Madariaga L, Rica I, R. I. Curvas y tablas de crecimiento (estudios longitudinal y transversal). Fundación Faustino Orbegozo Eizaguirre Madrid, Spain. (2004).
67. Mccrindle, B. W. Assessment and management of hypertension in children and adolescents. Nat. Rev. Cardiol. 7, 155–163 (2010).
68. Olza, J. et al. A gene variant of 11β-hydroxysteroid dehydrogenase type 1 is associated with obesity in children. Int. J. Obes. 1558–1563 (2012). doi:10.1038/ijo.2012.4
69. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview : analysis and visualization of LD and haplotype maps. 21, 263–265 (2005).
70. Dudoit, S., Shaffer, J. P. & Boldrick, J. C. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science 18, 71–103
71. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, (1988).
72. Benjamini, Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. (2014). doi:10.2307/2346101
73. Benjamini, Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29, 1165-1188 (2001).
74. Cotton, A. M. et al. Landscape of DNA methylation on the X chromosome re fl ects CpG density , functional chromatin state and X-chromosome inactivation. 24, 1528–1539 (2015).
75. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434, 400–404 (2005).
76. Slavney, A., Arbiza, L., Clark, A. G. & Keinan, A. Strong Constraint on Human Genes Escaping X-Inactivation Is Modulated by their Expression Level and Breadth in Both Sexes. Mol. Biol. Evol. 33, 384–393 (2016).
77. Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248 (2017).
78. Kolehmainen, M. et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. Int. J. Obes. 32, 292–303 (2008).
79. Purcell, S. et al. REPORT PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. 81, 559–575 (2007).
80. Ritchie, M. E., Liu, R., Carvalho, B. S., Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), T. A. and N. Z. M. S. G. C. & Irizarry, R. A. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics 12, 68 (2011).
81. Ling, H., Hetrick, K., Bailey-wilson, J. E. & Pugh, E. W. BMC Proceedings. 5, 1–5 (2009).
82. Clayton, D. Testing for association on the X chromosome. Biostatistics 9, 593–600 (2008).
83. Clayton, D. SnpMatrix and XSnpMatrix classes and methods. (2015). doi:10.18129/B9.bioc.snpStats
84. Kay, M., A., Glorioso, J., C., Naldini, L. Viral Vectors for Gene Therapy. Nat Med 7, 33–40. (2001).
85. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 29, (2001).
86. R Development Core Team. R: a language and environment for statistical computing. (2011). doi:3-900051-07-0
87. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. 76, 887–893 (2005).

Sci Data. 2019;6(1):130. doi:10.1038/s41597-019-0109-3. IF: 5.541, Q1 at MULTIDISCIPLINARY SCIENCES.

Study 2 | X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline

Augusto Anguita-Ruiz[1,2,3,4]*, Julio Plaza-Diaz[1,2,3], Francisco Javier Ruiz-Ojeda[1,3], Azahara I. Rupérez[1,6], Rosaura Leis[4,5], Gloria Bueno[4,6], Mercedes Gil-Campos[4,7], Rocío Vázquez-Cobela[4,5], Ramón Cañete[4,7], Luis A. Moreno[6], Ángel Gil[1,2,3,4] & Concepción María Aguilera[1,2,3,4].

Abstract X chromosome genetic variation has been proposed as a potential source of missing heritability for many complex diseases, including obesity. Currently, there is a lack of public available genetic datasets incorporating X chromosome genotype data. Although several X chromosome-specific statistics have been developed, there is also a lack of readily available implementations for routine analysis. Here, we aimed: 1) to make public and describe a dataset incorporating phenotype and X chromosome genotype data from a cohort of 915 normal-weight, overweight and obese children, and 2) to deeply describe a whole implementation of the special X chromosome analytic process in genetics. Datasets and pipelines like this are crucial to get familiar with the steps in which X chromosome requires special attention and may raise awareness of the importance of this genomic region.

Affiliations 1. Department of Biochemistry and Molecular Biology II, School of Pharmacy, University of Granada, Spain. / 2. Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada. Avda. del Conocimiento s/n. 18016 Armilla, Granada, Spain. / 3. Instituto de Investigación Biosanitaria IBS.GRANADA, University Clinical Hospital, Granada 18014, Spain. / 4. CIBEROBN, (Physiopathology of Obesity and Nutrition CB12/03/30038), Institute of Health Carlos III (ISCIII), Madrid 28029, Spain. / 5. Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department (USC). Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), University Clinical Hospital, Santiago de Compostela, Spain. / 6. Growth, Exercise, Nutrition and Development (GENUD) Research Group, Universidad de Zaragoza, Zaragoza, Spain; Instituto Agroalimentario de Aragón (IA2) and Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain. / 7. Department of Pediatric Endocrinology, Reina Sofia University Clinical Hospital, Institute Maimónides of Biomedicine Investigation of Córdoba (IMIBIC), University of Córdoba, Avda. Menéndez Pidal s/n, 14004 Córdoba, Spain.
* Corresponding authors

Background & Summary

Overweight and obesity in children are a public health problem that has raised concern worldwide[1]. Childhood obesity is characterized by an expansion of the adipose tissue (AT)[1] and plays an important role in the development of cardiometabolic alterations during early adulthood, thereby increasing morbidity and mortality[2]. According to twin and family studies, around 40–70% of the interindividual variability in body mass index (BMI) has been attributed to genetic factors[3-5]. Despite this, known single-nucleotide polymorphisms (SNPs) explain < 2% of BMI variation[6], a phenomenon termed 'missing heritability'. Potential sources explaining this missing heritability include epigenetic components, the existence of low frequency and rare variants as well as the presence of X chromosome genetic variation.

Analysis in current genetic association studies is usually focused on autosomal variants while the sex chromosomes, and specially the X chromosome, are often neglected. Among the reasons, it highlights a lower gene density on the X chromosome, a lower coverage of the region in current genotyping platforms and a number of technical hurdles including complications in genotype calling, imputation and selection of test statistics[7]. According to a previous report, only 242 out of all 743 GWAS conducted from 2005 to 2011 considered the X chromosome in their analyses[7]. The proportion was similar when only family-based GWAS were considered. There is therefore a lack of available public datasets including X chromosome genotype data for analysis. On the other hand, although several X chromosome-specific statistical tests and guidelines have now become available, there is also a lack of readily available implementations and user-friendly apps incorporating them for routine analysis[8,9].

The majority of the technical hurdles faced when analysing X chromosomal data rise from two of its main particularities. The first one is the fact of women having two allele copies while males having only one. As a consequence, if males are included in the analysis, special caution must be taken. Particularly, the study design process should be performed carefully, trying to maintain a balanced female/male ratio across experimental conditions. Otherwise, many available statistical tests will suffer from type I errors as soon as sex-specific allele frequencies occur, which is typically observed for a great number of variants. Other problems derived from an unbalanced sex ratio in the study sample include problems during the genotype calling process, as the signal intensities obtained from standard array genotyping platforms will be always lower in males than for females (who carry two alleles).

The second uniqueness motivating X-chromosome specific analyses lies in the X chromosome inactivation (XCI) process, through which most of the cells of females express only one X chromosome allele in order to compensate the genetic dosage with regard to males. Before selecting a particular statistical approach, it should be mandatory to carefully investigate the concrete XCI model to assume for a gene in a particular tissue. Depending on the XCI model

assumed, we should proceed one way or another during the selection of the test statistics. These and other particularities must be addressed as long as X chromosomal data are included into genetic studies.

In relation to obesity, only a few studies have reported association with markers on the X chromosome. One of the most remarkable findings involves the tenomodulin (*TNMD*) gene, a Xq22 located locus encoding a type II transmembrane glycoprotein. First time associated with adult obesity at the genetic level[10], its presence in adult human AT has been demonstrated showing higher expression in obesity and lower expression after diet-induced weight loss. Regarding children population, our research group found that *TNMD* expression was five fold-times up-regulated in visceral adipose tissue (VAT) of children with obesity, compared with their normal-weight counterparts (Gene Expression Omnibus GSE9624)[11,12]. Recently, we have reported new associations between TNMD SNPs and childhood obesity and metabolic alterations in a Spanish children population[13]. Interestingly, our study has been the first to analyse and detect associations between X chromosome *TNMD* genetic variants and obesity in a children cohort.

Similarly, SNPs in the *SLC6A14* gene, also located in the X chromosome, have shown evidence of association with obesity[14]. As a whole, these *TNMD* and *SLC6A14* reports support the fact that X chromosome genetic variants may be not only useful early life risk indicators of obesity but also an interesting source of missing heritability[13].

Given the lack of public available genetic datasets incorporating X chromosome genetic variants and the still prevalent statistical hurdles that make the X chromosome a difficult region to be tested in functional genetics, we here aimed: 1) to make public and describe a dataset incorporating X chromosome genotype data from a children cohort[13,15], and 2) to outline a whole implementation of the special X chromosome analytic process in genetics. The presented research dataset includes X-chromosomal SNP data (mapping the genes *TNMD* and *SLC6A14*) from a children cohort composed of 915 normal-weight, overweight and obese subjects. Some topics covered in this paper include dataset sharing and description, explanation of sample design, genotype calling, quality control, and test statistics selection procedures. Additionally, a short section explaining and interpreting findings obtained after analysing the dataset with a specific X chromosome analytic approach is presented.

Methods

Experimental design and study population

These methods are an expanded version of descriptions in our related work and general characteristics of the dataset have been previously described[13]. Briefly, in this case-control multicentre study, 915 Spanish children (438 males and 477 females) were recruited from three

national health institutions: Lozano Blesa University Clinical Hospital, Santiago de Compostela University Clinical Hospital and Reina Sofía University Clinical Hospital. According to specific X-chromosomal analytic requirements, the female/male ratio of the study sample was perfectly balanced.

Childhood obesity status was defined according to the International Obesity Task Force (IOTF) reference for children[16] which is based on the application, on children population, of the widely used cut-off points of BMI for adults (25 and 30 kg/m^2, for overweight and obesity respectively). Particularly, these criteria constitute a range of age and sex specific cut-off points for children that have been extracted from solid percentile tables constructed on 97876 boys and 94851 girls (ranging from 2 to 18 years). After the application these specific cut-off points, the dataset was composed of 480 children in the obesity group, 177 in the overweight group and 258 in the normal-BMI group. Children were allocated into two experimental groups according to their obesity status; the affected group (cases) composed of both children with obesity or overweight and the control group composed of normal-weight children. An unbalanced female/male ratio across cases and controls has been proven to heavily affect the power of some specific X chromosome association tests[17]. In our study, a balanced female/male ratio was maintained across each experimental condition (122/136 in controls and 355/302 in cases) (**Figure 1**).

Inclusion criteria were European-Caucasian heritage and the absence of congenital metabolic diseases. Otherwise, the exclusion criteria were non-European Caucasian heritage, the presence of congenital metabolic diseases (e.g., diabetes or hyperlipidemia), undernutrition, and the use of medication that alters blood pressure, glucose or lipid metabolism.

Ethical statement

All procedures in the study were conducted in accordance with the Declaration of Helsinki (Edinburgh 2000 revised), and followed the recommendations of both the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legally enforced Spanish regulation for clinical investigation of human beings (RD 223/04 about clinical trials). The Ethics Committees on Human Research of all participant institutions approved all experiments and analyses with registration Code: "2011/198". All parents or guardians provided written informed consent and children gave their assent.

DNA extraction, processing and analysis.

The presented dataset consists on genotype data for eight target SNPs mapping the X-chromosomal genes *TNMD* and *SLC6A14* in the study population. Details regarding SNP selection and molecular analyses are briefly covered here since they have already been fully detailed in our previous work[13]. On the contrary, we pay special attention in the explanation of X chromosomal particularities, data description as well as in the summarization of each data analysis and processing step.

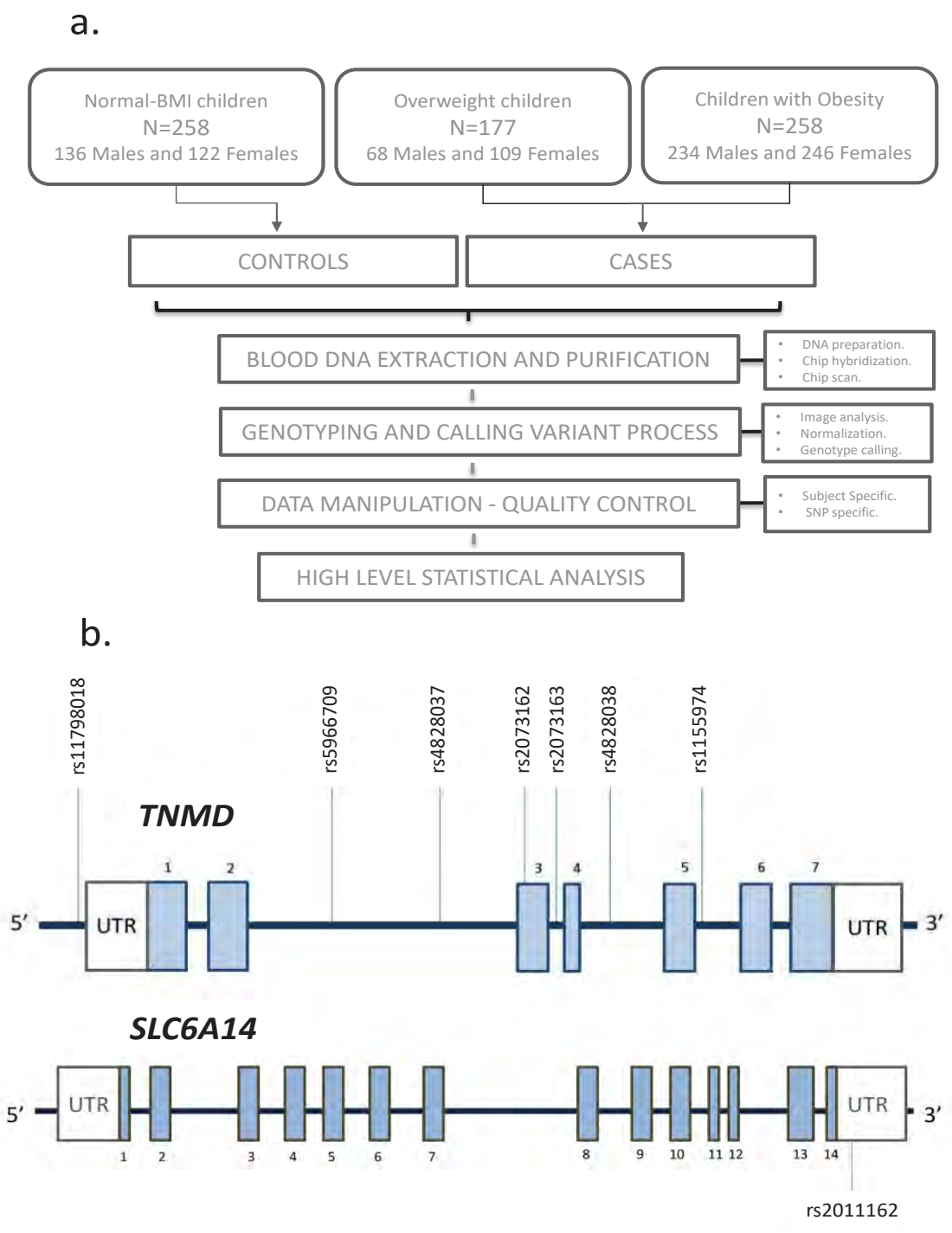


Figure 1. Study design and characteristics. (a) Experimental workflow used to generate and analyse the data output. (b) Genomic context of selected markers; light blue boxes represent exons, while the connecting lines are introns. Abbreviations; rs, reference SNP; UTR, untranslated region.

CHR	SNP	BP	A1	MAF (All)	MAF (Females)	MAF (Males)	A2	P	OR
23	rs11798018	100584572	A	0.26	0.26	0.27	C	0.84	0.97
23	rs5966709	100589508	T	0.32	0.32	0.31	G	0.75	1.05
23	rs4828037	100590686	C	0.34	0.34	0.33	T	0.53	1.08
23	rs2073162	100594019	A	0.45	0.45	0.42	G	0.37	1.12
23	rs2073163	100594053	C	0.45	0.46	0.43	T	0.35	1.13
23	rs4828038	100596678	T	0.44	0.45	0.42	C	0.31	1.13
23	rs1155974	100598283	T	0.44	0.44	0.42	C	0.59	1.07
23	**rs2011162**	**116459132**	**C**	**0.34**	**0.37**	**0.30**	**G**	**0.02***	**1.36**

Table 1. Allele frequencies in the whole study population and by sex group. P and OR columns correspond to P-values and Odd Ratios obtained by means of the Fisher exact test for sex-specific allele frequencies. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; BP, Base Pair; A1, Minor Allele; MAF, Minor Allele Frequency; A2, Alternative Allele; P, P-value; and OR, Odd Ratio.

CHR	SNP	GROUP	A1	MAF
23	rs11798018	OVERWEIGHT	A	0.26
23	rs11798018	OBESITY	A	0.27
23	rs11798018	NORMAL-BMI	A	0.28
23	rs5966709	OVERWEIGHT	T	0.29
23	rs5966709	OBESITY	T	0.34
23	rs5966709	NORMAL-BMI	T	0.30
23	rs4828037	OVERWEIGHT	C	0.32
23	rs4828037	OBESITY	C	0.36
23	rs4828037	NORMAL-BMI	C	0.31
23	rs2073162	OVERWEIGHT	A	0.44
23	rs2073162	OBESITY	A	0.46
23	rs2073162	NORMAL-BMI	A	0.43
23	rs2073163	OVERWEIGHT	C	0.45
23	rs2073163	OBESITY	C	0.47
23	rs2073163	NORMAL-BMI	C	0.43
23	rs4828038	OVERWEIGHT	T	0.43
23	rs4828038	OBESITY	T	0.46
23	rs4828038	NORMAL-BMI	T	0.43
23	rs1155974	OVERWEIGHT	T	0.42
23	rs1155974	OBESITY	T	0.45
23	rs1155974	NORMAL-BMI	T	0.43
23	rs2011162	OVERWEIGHT	C	0.34
23	rs2011162	OBESITY	C	0.36
23	rs2011162	NORMAL-BMI	C	0.34

Table 2. Allele frequencies in the study population stratified by experimental condition. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; BMI, Body Mass Index; A1, Minor Allele; and MAF, Minor Allele Frequency.

Seven SNPs located at the *TNMD* locus and one located at the *SLC6A14* were selected for genotyping analysis. Genomic DNA was extracted from peripheral white blood cells using two automated kits, the Qiamp DNA Investigator Kit for coagulated samples and the Qiamp DNA Mini & Blood Mini Kit for non-coagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). All extractions were purified using the DNA Clean and Concentrator kit from Zymo Research (Zymo Research, Irvine, CA, USA). Genotyping was performed by TaqMan allelic discrimination assay using the QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA). Given the X-chromosomal location, it is recommendable to analyse females and males in separate plates during the genotyping process or, at least, maintain a balanced female/male ratio by plate.

Once genotyping was accomplished, we checked candidate SNPs for sex-specific allele frequencies, which can induce type I errors in some statistical X-chromosome analyses (especially in the case of unbalanced designs). Tested by means of the Fisher exact test, all SNPs in the *TNMD* showed no significant P-values and thus equal allele frequencies across sex groups (**Table 1**). On the contrary, the SNP in the *SLC6A14* did not (P=0.01). This fact should be taken into consideration when selecting an appropriate test for high-level statistical analyses unless a balanced sex ratio across experimental conditions is presented in the population (which is our case). Information regarding minor allele frequencies (MAFs) stratified by experimental condition for all candidate markers is presented in (**Table 2**). Linkage disequilibrium (LD) status of the TNMD gene was studied using the Haploview Software separately in males and females[13,18].

Data Records

The complete research dataset (genotype and phenotype data) has been uploaded into the European Genome-Phenome archive (EGA). The work can be found online with the title "X chromosomal genetic variants are associated with childhood obesity" or with the identifier EGAS00001002738 (2018)[15]. Online data are sorted and presented according to obesity status; the affected group (cases) composed of both children with obesity or overweight (EGA reference EGAD00010001482 (2018)) and the control group composed of normal-weight children (EGA reference EGAD00010001481 (2018)). Three files by-experimental condition (a total of six) are available online (*.bed*, *.bim* and *.fam* files). The .bed files contain raw genotype data while the *.bim* files describe information relative to target SNPs (chromosome number, SNP identifier, genetic distance in morgans (set as 0 for all markers), base-pair position and coding alleles). Instead, the *.fam* files contain information relative to subjects (sample identifiers, family and paternal identifiers (here set as 0), sex (1 for males and 2 for females) and experimental group (1 for control and 2 for cases)). All presented formats can be easily readable in PLINK 1.9 software using the *–bfile* command option and further transformed to a more standard file format with the *–dosage* option[22].

The complete data set in the current study complies with the requirements of the EGA archive. Detailed information about each sample and shared data files is presented in online-only

tables 1, 2 and 3. Specifically, DOI and descriptions for each shared file are provided in the online-only **table 2**.

Technical Validation

X chromosome particularities

Before introducing further steps, we here list two issues making the X chromosome a difficult region for genetic analyses. These particularities will determine important decisions related to genotype calling, data imputation and statistical analysis. It is important to note, however, that all here-described particularities are only applicable to those X chromosomal loci outside the pseudo-autosomal region of the X chromosome (which is the case of *TNMD* and *SLC6A14*).

The first noticeable uniqueness of the X chromosome is the fact of women having two allele copies while males having only one. As a result, while females can present the standard three possible allele combinations (AA, AB and BB), males are homozygous and have only two distinct possible genotypes (A- and B-). For this reason, standard autosomal association tests, such as the Cochran-Armitage trend test[23,24], are not immediately applicable to X chromosome data. The second particularity affecting the X-chromosome analysis lies in the X chromosome inactivation (XCI) process, through which the transcription from one of the two X chromosome copies in female mammalian cells is silenced in order to balance the expression dosage between XX females and XY males. XCI is, however, incomplete in humans: with up to one-third of the X-chromosomal genes escaping from this silencing epigenetic mechanism. The degree of 'escape' from inactivation has been reported to strongly vary between genes, tissues and individuals[25,26], with three possible scenarios at the gene level: complete XCI, partial XCI or total escape from XCI[27,28]. Depending on the XCI model assumed for a certain gene, we should proceed one way or another during the selection of the test statistics (see section 'High-Level Analysis: Statistical Analysis' for further details). The assumption of a particular XCI model is therefore a process that must be performed carefully.

Until date, the extent to which XCI is shared between cells and tissues remains poorly characterized and there is a lack of standardized criteria nor well-established databases to check if a gene escapes or not from XCI in a concrete situation. In order to do so, an exhaustive search in PUBMED and other scientific databases should be performed looking for particular studies supporting a certain XCI hypothesis. Currently, the most similar resource to a standardized database on this regard is the initiative carried out by the Genotype-Tissue Expression (GTEx) consortium[9] in 2018, which describes a systematic survey of XCI, integrating over 5500 transcriptomes from 449 Individuals, spanning 29 tissues from the GTEx (v6p release) and 940 single-cell transcriptomes, combined with genomic sequence data. Particularly, they show that XCI at 683 X-chromosomal genes is generally uniform across human tissues and that incomplete XCI affects at least 23% of X-chromosomal genes. Overall, this work presents an updated catalogue of XCI across human

tissues which may be of great utility during the selection of a particular XCI model for a gene. Other available resources also include the work of Slavney *et al.* (2016)[29], which gathers the main insights from previous studies on X-chromosome gene expression datasets.

By way of example, we here illustrate the whole process followed for the identification of the optimal XCI model in the case of *TNMD*. First, we interrogated the Slanvey et al. (2016) work[29], where no evidence of escape from XCI was reported for this locus. In order to get more information about this fact, we further studied in detail the three works summarized in the Slanvey *et al.* (2016)[29] paper. The first work on which the paper is based is a study from Carrel *et al.* (2005)[25], in which we could not identify any probe covering the *TNMD* region. Instead, a few surrounding regions were mapped; among which the *SRPX2, ZD89B07* and the *SYTL4* reported escaping from the XCI process. In spite of it, this study was based on a fibroblast cell model and thus not applicable to our adipose tissue context. Regarding the second revised article[30], again, there were not available probes covering *TNMD*. Thus, neither conclusions nor new information could be extracted. In relation to the third included article[31], we were not able to find any table or supplemental material showing an output list of the analysed regions.

Next, we investigated the well-established work from the GTEx consortium[9] and found that the XCI status of the *TNMD* region remains catalogued as unknown (supplementary tables S2 and S13 of this paper).

As a complementary approach, we performed a search in PUBMEP looking for individual studies focused on the gene expression status of *TNMD* from different sexes. As a result, we found a work reporting higher basal expression of *TNMD* in women than in men[32], which could indicate that *TNMD* escapes from the XCI.

Taking all this into consideration and given the lack of agreement, both possibilities ('escape from XCI' and 'XCI') should be tested in the case of *TNMD*. A searching process like this is highly recommendable to be done for any X chromosome locus before the selection of a particular statistical approach.

Raw data processing

The primary step of the data analysis consisted on the extraction of genotype calls from fluorescence array data and the construction of work data files for data manipulation and analysis. Details regarding the exact procedure for genotype calling, which is an important procedure in X-chromosomal analyses, are listed below (`Genotype Calling´ section).

Once we obtained genotype calls for the 915 individuals, we generated standard format files (*.ped* and *.map*) transforming the ThermoFisher cloud-derived outputs from long to wide format using an own script in R environment[33]. Finally, data were imported into PLINK 1.9 software[18] and converted into binary format files using the *--make-bed* flag. These binary formats (*.bed*, *.bim* and

.fam) are a more compact representation of the data that saves space and speeds up subsequent analyses.

Genotype calling

This is the first step of any primary genotype analysis and consists on the extraction of genotype calls from fluorescence array data at the SNP and individual level. Along with the test statistics selection procedure, the genotype calling process is an analytical step heavily affected by X chromosome particularities. Specifically, the main X chromosome uniqueness affecting this process is the dosage imbalance between males and females. Since males carry only one X allele, signal intensities obtained from the Real-Time PCR System are lower in males than for females and thus a correction should be implemented. On this matter, calling algorithms which apply different models to male and female samples (e.g. Illuminus and CRLMM) have been proven to generally perform better than methods which do not (e.g. GenCall and GenoSNP)[34].

Here, we employed the Applied Biosystems qPCR app module (ThermoFisher Cloud software) and the autocalling method for genotype calling. According to literature recommendations, the sex information for each sample was supplied to the software and genotype calling was performed separately in both sexes. In this regard, although genotyped plates did not consist on only boys or girls, the balanced sex ratio of our population (477 females and 438 males) favoured a better performance of the algorithm. Five signal clusters were identified (three in the case of females and two in the case of males). Then, sex information and scatter of the clusters were used to call the genotypes (AA, AB and BB for females, and A- and B- for males). Since the employed software also allows the option of applying user-definable boundaries for data analysis, those samples classified as undetermined by the autocalling method were recalled using the manual option. A set of controls were used to deduce these questionable genotype calls. Outliers were omitted from the analysis.

Table 3. Quality control (QC) for missing frequency in the selected markers stratified by sex. P column correspond to P-value obtained in a differential missingness test between sex groups. Asymptotic P-values were obtained by means of Fisher's exact test. SNPs in bold did pass the QC recommended filters. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; and MISS FREQ, Missing Frequency.

CHR	SNP	MISS FREQ (Males)	MISS FREQ (Females)	MISS FREQ (Males-Females)	P
23	rs11798018	0.11	0.04	0.07	**2.01e-05**
23	rs5966709	0.06	**0.004**	0.06	3.11e-07
23	rs4828037	0.06	**0.01**	0.05	**3.46e-05**
23	rs2073162	0.08	**0.008**	0.07	2.86e-08
23	rs2073163	0.14	0.09	0.05	**0.02**
23	rs4828038	0.07	**0.002**	0.07	4.19e-10
23	rs1155974	0.08	**0.002**	0.07	9.73e-11
23	rs2011162	0.07	**0.02**	0.05	9.30e-05

CHR	SNP	TEST	A1	GENO	O(HET)	E(HET)	P
23	rs11798018	ALL	A	34/177/249	0.38	0.39	0.72
23	rs11798018	AFF	A	25/133/187	0.38	0.39	0.89
23	rs11798018	UNAFF	A	9/44/61	0.39	0.40	0.81
23	rs5966709	ALL	T	67/175/234	0.37	0.44	0.0005
23	rs5966709	AFF	T	55/128/171	0.36	0.45	0.0005
23	rs5966709	UNAFF	T	12/46/63	0.38	0.41	0.38
23	rs4828037	ALL	C	75/178/219	0.38	0.45	0.0003
23	rs4828037	AFF	C	63/129/158	0.37	0.46	0.0001
23	rs4828037	UNAFF	C	12/48/61	0.40	0.42	0.66
23	rs2073162	ALL	A	132/172/170	0.36	0.50	4.59e-09
23	rs2073162	AFF	A	99/128/124	0.36	0.50	6.78e-07
23	rs2073162	UNAFF	A	33/43/46	0.35	0.49	0.002
23	rs2073163	ALL	C	129/148/156	0.34	0.50	6.92e-011
23	rs2073163	AFF	C	98/112/113	0.35	0.50	3.83e-08
23	rs2073163	UNAFF	C	31/35/43	0.32	0.49	0.0002
23	rs4828038	ALL	T	133/171/173	0.36	0.50	1.56e-09
23	rs4828038	AFF	T	100/128/127	0.36	0.50	2.49e-07
23	rs4828038	UNAFF	T	33/43/46	0.35	0.49	0.002
23	rs1155974	ALL	T	127/172/178	0.36	0.49	4.18e-09
23	rs1155974	AFF	T	94/128/132	0.36	0.49	4.02e-07
23	rs1155974	UNAFF	T	33/43/46	0.35	0.49	0.002
23	rs2011162	ALL	C	84/179/207	0.38	0.47	0.0001
23	rs2011162	AFF	C	62/136/150	0.39	0.47	0.003
23	rs2011162	UNAFF	C	22/42/57	0.35	0.46	0.01

Table 4. Genotype counts and Hardy-Weinberg test statistics for each SNP in the female group. Each SNP has three entries showing results for either ALL, AFF (overweight and children with obesity) or UNAFF (normal-BMI children only) individuals. Hardy Weinberg analysis was performed with the exact test described and implemented by Wigginton et al. 42. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; A1, minor allele; GENO, genotype counts; O(HET), observed heterozygosity; E(HET), expected heterozygosity; and P, Hardy Weinberg obtained P-value.

Data QC

Prior to high-level statistical analyses, the quality control (QC) process is an important step in any genetic analysis and especially in the X-chromosome analysis. Specific QC guidelines for X chromosome genotype data have been previously reviewed in detail[8]. All these criteria can help us to detect genotype errors or not reliable SNPs which should be excluded from analysis.

Here, the whole QC process was implemented in PLINK 1.9 software[22]. According to literature, two criteria concerning missing frequency were employed (the sex-specific missing frequency and the differential missingness between sexes)[35,36]. As genotype calling was performed separately

in males and females (that is, no heterozygote calls in males were allowed), the proportion of heterozygote calls in males, proposed as a filter criterion by Ling and Ziegler *et al.*[35,36], was not considered in our QC process. All SNPs (with exception of the rs11798018 and the rs2073163 from the *TNMD* gene) passed the recommended missing frequency filter in females (<= 2 %) (**Table 3**). On the other hand, none SNP passed the filter in males. Regarding the differential missingness test, the SNPs (rs11798018, rs4828037 and rs2073163) from the TNMD and the rs2011162 from the *SLC6A14*, passed the recommended filter ($P \geq 10^{-7}$). The other SNPs, instead, evidenced a marked differential missingness between sex groups. This test was performed in PLINK software using the flag "test-missing" and replacing the phenotype column of the *.ped* file by the sex information (**Table 3**).

Regarding additional MAF quality checks, all SNPs showed appropriated frequencies > 1% by sex groups (**Table 1**). When analysing the Hardy Weinberg equilibrium (HWE) in girls belonging to the normal-BMI group, all SNPs reported proper values ($P \geq 10^{-4}$) (**Table 4**). According to this QC process, we ensured that there were not important genotyping errors and that our genetic data were reliable for further analyses.

On this point, it is important to note that since genotyping array technologies are not specially designed for sexual chromosomes, quality is always hoped to be lower on X chromosome genetic variants compared to autosomal data.

High-Level Analysis: Statistical Analysis

As we previously mentioned, most of available test statistics for performing genetic association analyses are designed for autosomal variants and thus not applicable to X chromosome data (especially when dealing with mixed-sex samples). In these cases, testing for association on the X chromosome raises unique challenges that have motivated the development of X-specific statistical tests in the literature[20,37]. Association tests on the X chromosome should incorporate into their models not only the fact of dosage imbalance between males and females but also, depending on the analysed locus, a specific XCI model. Some of available approaches include:

- Clayton Tests (2008)[20]. Clayton tests are two X chromosome specific versions of the common autosomal tests that explicitly account for the XCI process and allow the inclusion of males and females together. In the case of different allele frequencies in males and females, Clayton statistics have inflated type I error frequencies. These tests are available in the R package snpMatrix[21]:
    - S1[20]: It is analogous to a Cochran-Armitage trend test of a combined male and female genotype contingency table; it follows a Chi^2 distribution on one degree freedom (df) under the null hypothesis.

- S2[20]: It is analogous to a Pearson's Chi^2 test on 2 df of a combined male and female genotype contingency table, it follows a Chi^2 distribution on 2 df under the null hypothesis.

- Zheng tests (2007)[37]. They are a set of six different statistics that apply to the same SNP and from which a minimum P-value is computed, needing to be adjusted according to the correlation between the test statistics. Zheng *et al.*[37] showed that the optimal choice of statistic among the six tests depends on whether HWE holds at the locus and whether males and females have the same risk allele. For example, in the case there is departure from HWE in females, the Zheng (Z^2_{mfG}) test has been presented a good choice. For further information regarding test statistic selection, we recommend to read next works[8,37]. Of note is that the Zheng's tests do not explicitly account for the XCI process.

As previously mentioned, an unbalanced female/male ratio between cases and control would affect the relative power of both Zheng and Clayton statistics. If combined with sex-specific allele frequencies, these tests will suffer from increased type I errors.

- Traditional methods easily implementable in PLINK 1.9 or R environment:

  - Ignore males entirely and analyse female data using conventional autosomal tests (a genotypic-based Cochran-Armitage trend test or an allele-based Chi^2 by Pearson with 1 df). The problem related to this approach is that we are missing all data from male subjects and therefore losing statistical power. The Cochran-Armitage trend test is the default test employed when a naive analysis of X chromosome data is run in PLINK using the flag *–model*[22]. Regarding males, an allele-based test accounting for the number of A- and B-alleles between experimental conditions should be employed apart.

  - Linear or logistic regression analyses on all the samples adjusting by sex. This approach has the advantage of adjusting the model by covariates of interest. Here, if we assume the locus of interest escapes from XCI, females should be coded as 0, 1, or 2, according to 0, 1, or 2 number of SNP risk alleles, and males should be coded as 0 or 1 according to 0 or 1 allele copies. On the contrary, if XCI is assumed to occur, females should be coded as 0, 1, or 2, according to 0, 1, or 2 number of SNP risk alleles, and males should be coded as 0 or 2 according to 0 or 1 allele copies. By default, the application of the "*--dosage*" flag to X chromosome input data files (*.bed*, *.bim* and *.fam*) in PLINK will produce a codification which assumes escape from XCI. For XCI to be considered, new allele code numbers should be manually replaced in male samples with a standard text editor (e.g: gedit software).

In general, the selection of the most suitable test among the presented choices will depend on three different criteria; the XCI model assumed for the locus of interest, deviation from HWE of analysed markers and the existence of sex-specific allele frequencies in the study population, which would be a substantial problem in the case of an unbalanced female/male ratio. Regarding XCI, if inactivation is assumed to occur, then either the Clayton's statistics or regression models

(with males coded as 0 and 2 (for 0 and 1 risk allele, respectively) would be tests of choice. On the contrary, in the case of a locus 'escaping' from XCI, Zheng's tests or regression models (with males coded as 0 and 1 (for 0 and 1 risk allele, respectively) should be employed. In the case of sex-specific allele frequencies, independently of the XCI assumed model, the Zheng's test (Z^2_{mfG}) has been presented a better choice over the Clayton approach. On the other hand, in the case of an adjustment for covariates is required, only regression models can be applied. Of note is that most of the test statistics and analysis considerations covered here are available to implement by means of the command-line toolset XWAS developed by Keinan A. and collaborators[38-40].

Although for the analysis of our dataset both possibilities ('escape from XCI' and 'XCI') were tested in the original work[13], we here only present results under the XCI assumption. As we have

SNP	N	Chi.squared.1.df	Chi.squared.2.df	P.1df	P.2df
HOMA-IR					
rs11798018	811	0.10	1.68	0.74	0.43
rs5966709	849	0.14	2.71	0.70	0.25
rs4828037	844	0.35	2.91	0.55	0.23
rs2073162	841	5.48	6.34	**0.01**	**0.04**
rs2073163	773	4.78	5.93	**0.02**	0.05
rs4828038	849	6.00	7.24	**0.01**	**0.02**
rs1155974	844	4.22	6.68	**0.03**	**0.03**
rs2011162	839	0.48	0.91	0.48	0.63
Glucose (mg/dl)					
rs11798018	844	0.004	1.06	0.94	0.58
rs5966709	881	0.55	0.59	0.45	0.74
rs4828037	876	1.22	1.25	0.26	0.53
rs2073162	873	5.17	8.13	**0.02**	**0.01**
rs2073163	804	2.8	4.006	0.09	0.13
rs4828038	880	4.78	6.42	**0.02**	**0.04**
rs1155974	876	3.94	4.74	**0.04**	0.09
rs2011162	871	0.92	3.55	0.33	0.16
BMI z-score					
rs11798018	845	0.97	1.15	0.32	0.56
rs5966709	881	0.77	1.21	0.37	0.54
rs4828037	877	0.51	0.51	0.47	0.77
rs2073162	872	8.61	9.59	**0.003**	**0.008**
rs2073163	803	7.09	8.60	**0.007**	**0.01**
rs4828038	877	9.02	10.38	**0.002**	**0.005**
rs1155974	875	7.75	8.69	**0.005**	**0.01**
rs2011162	871	3.31	5.33	0.06	0.06

Table 5. Association between X chromosome SNPs and HOMA-IR, Glucose and BMI Z-Score phenotypes in our dataset. SNPs in bold showed statistically significant associations with tested attributes under the Clayton Statistics. These tests explicitly accounted for random X-inactivation and allowed the inclusion of females and males together, increasing thereby the statistical power. P.1df and Chi.squared.1.df columns corresponds to Clayton S1 statistic results while P.2df and Chi.squared.2.df corresponds to Clayton S2 results. Abbreviations; SNP, Single Nucleotide Polymorphism; N, number of included subjects in the analysis; HOMA-IR, homeostasis model assessment for insulin resistance; and BMI ZSCORE, body mass index adjusted by sex and age.

previously seen, selected markers in our sample did not exhibit HWE deviations nor sex-specific allele frequencies. Moreover, the female/male ratio was balanced across experimental groups. For these reasons, and following published recommendations[8], Clayton test was selected to perform the main statistical analysis[17,20,41]. According to an in silico simulation work, the Clayton's S1 statistic showed the best performance among all X-specific introduced tests across a wide range of disease models, sex ratios and allele frequencies[41]. Moreover, it allows the inclusion of females and males together, increasing thereby the statistical power.

In **Table 5**, results derived from the application of Clayton's S1 and S2 statistics to three different continuous phenotypes of the population are presented. All these phenotype data have also been shared and are available in the metadata file (Online-only **table 1**). The implementation of this process was performed in R, using the snpStats R package and the code have been shared online[19]. All reported associations in our previous work[13] were here replicated under XCI assumption. These findings support therefore a good performance of the Clayton statistics as well as ensure the reliability of the present dataset.

In conclusion, we here share a genetic dataset and present a whole implementation of the special X chromosome analytic process in genetics. Altogether, the pipeline and the shared data will allow researchers to get familiar with the X chromosome particularities and should encourage them to include X chromosome into their genetic studies. Closing this gap is crucial to elucidate the genetic background of complex diseases, especially of those with sex-specific features.

Code availability

All custom R codes employed in this work have been shared online in a GitHub repository (http://doi.org/10.5281/zenodo.2578182)[19]. Two short scripts are available online; *"script_from_long_to_wide.r"* and *"Clayton_analysis_code.r"*.

The first one (named *"script_from_long_to_wide.r")* is a short script designed for loading a genetic dataset (genotype calls) derived from OpenArray technology and transforming it into a handy-format file, which can be further imported into PLINK software. Basically, this script carries out a dataset manipulation and transformation from long to wide format. In order to run the script, users will need an input file derived from OpenArray technology containing information in the long format arranged into three columns (NCBI_SNP_Reference, Sample_ID and Genotype_Call).

The second script shared (named "*Clayton_analysis_code.r*") gathers functions and R commands required for the application of the X-chromosome specific statistical tests developed by Clayton and collaborators[20,21] (see section 'High-Level Analysis: Statistical Analysis' for further details).

Acknowledgements

This paper will be part of Augusto Anguita-Ruiz's doctorate, which is being completed as part of the "Nutrition and Food Sciences Program" at the University of Granada, Spain. The authors would like to thank the children and parents who participated in the study. This work was supported by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I + D + I), Instituto de Salud Carlos III-Fondo de Investigación Sanitaria (FONDOS FEDER) Projects numbers PI051968, PI1102042 and PI1600871, Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015) and the Mapfre Foundation. The authors also acknowledge the Institute of Health Carlos III for personal funding: "Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016, Project number: IFI17/00048".

Author contributions

AG and CAG contributed to the study concept and design. RL, GB, MGC, RVC, LM and RC participated in the child recruitment and anthropometric measures. AIR and CAG selected the SNPs and revised the DNA extraction. AAR did the all data processing steps and shared the dataset in the EGA repository. All authors took part in the interpretation of data, the drafting of the manuscript and the critical revision of the manuscript. AG, CAG, RL, GB and RC obtained funding. AAR, FJRO and JPD wrote the manuscript.

Competing interests

The authors disclose no conflicts.

Additional Information

Supplementary Information is available for this paper at https://doi.org/10.1038/s41597-019-0109-3.

References

1 Collaborators, G. B. D. O. et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. The New England journal of medicine 377, 13-27, https://doi.org/10.1056/NEJMoa1614362 (2017).

2 Jones, R. E., Jewell, J., Saksena, R., Ramos Salas, X. & Breda, J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Frontiers in public health 5, 58, https://doi.org/10.3389/fpubh.2017.00058 (2017).

3 Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. Behavior genetics 27, 325-351 (1997).

4 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. American journal of human genetics 90, 7-24, https://doi.org/10.1016/j.ajhg.2011.11.029 (2012).

5 Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS genetics 9, e1003520, https://doi.org/10.1371/journal.pgen.1003520 (2013).

6 Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206, https://doi.org/10.1038/nature14177 (2015).

7 Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. American journal of human genetics 92, 643-647, https://doi.org/10.1016/j.ajhg.2013.03.017 (2013).

8 Konig, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to include chromosome X in your genome-wide association study. Genetic epidemiology 38, 97-103, https://doi.org/10.1002/gepi.21782 (2014).

9 Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. Nature 550, 244-248, https://doi.org/10.1038/nature24265 (2017).

10 Tolppanen, A. M. et al. Tenomodulin is associated with obesity and diabetes risk: the Finnish diabetes prevention study. Obesity 15, 1082-1088, https://doi.org/10.1038/oby.2007.613 (2007).

11 Aguilera, C. M. et al. Genome-wide expression in visceral adipose tissue from obese prepubertal children. International journal of molecular sciences 16, 7723-7737, https://doi.org/10.3390/ijms16047723 (2015).

12 Aguilera, C. M. et al. Differential gene expression in omental adipose tissue from obese children. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9624. (2018).

13 Ruiz-Ojeda, F. J. et al. Effects of X-chromosome Tenomodulin genetic variants on obesity in a children's cohort and implications of the gene in adipocyte metabolism. Scientific Reports, https://doi.org/10.1038/s41598-019-40482-0. (2019).

14 Suviolahti, E. et al. The SLC6A14 gene shows evidence of association with obesity. The Journal of clinical investigation 112, 1762-1772, https://doi.org/10.1172/JCI17491 (2003).

15 Anguita-Ruiz, A., Ruiz-Ojeda, F. J. & Aguilera, C. M. The European Genome-phenome Archive (EGA). X chromosomal genetic variants are associated with childhood obesity. https://ega-archive.org/studies/EGAS00001002738. (2018).

16 Cole, T. J., Bellizzi, M. C., Flegal, K. M. & Dietz, W. H. Establishing a standard definition for child overweight and obesity worldwide: international survey. Bmj 320, 1240-1243, https://doi.org/10.1136/bmj.320.7244.1240 (2000).

17 Loley, C., Ziegler, A. & Konig, I. R. Association tests for X-chromosomal markers--a comparison of different test statistics. Human heredity 71, 23-36, https://doi.org/10.1159/000323768 (2011).

18 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-265, https://doi.org/10.1093/bioinformatics/bth457 (2005).

19 Anguita-Ruiz, A. R scripts for the manipulation, transformation and statistical analysis of Openarray genotype datasets. (Version v1.0.2). Zenodo. http://doi.org/10.5281/zenodo.2578182 (2019).

20 Clayton, D. Testing for association on the X chromosome. Biostatistics 9, 593-600, https://doi.org/10.1093/biostatistics/kxn007 (2008).

21 Clayton, D. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.32.0. (2018).

22 Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575, https://doi.org/10.1086/519795 (2007).

23 Cochran, W. G. Some methods for strengthening the common x^2 tests. Biometrics 10, 417-451, http://dx.doi.org/10.2307/3001616 (1954).

24 Armitage, P. Tests for Linear Trends in Proportions and Frequencies. Biometrics 11, 375-386, https://doi.org/10.2307/3001775 (1955).

25 Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434, 400-404, https://doi.org/10.1038/nature03479 (2005).

26 Cotton, A. M. et al. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome biology 14, R122, https://doi.org/10.1186/gb-2013-14-11-r122 (2013).

27 Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. Annual review of genomics and human genetics 6, 69-92, https://doi.org/10.1146/annurev.genom.6.080604.162350 (2005).

28 Amos-Landgraf, J. M. et al. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. American journal of human genetics 79, 493-499, https://doi.org/10.1086/507565 (2006).

29 Slavney, A., Arbiza, L., Clark, A. G. & Keinan, A. Strong Constraint on Human Genes Escaping X-Inactivation Is Modulated by their Expression Level and Breadth in Both Sexes. Molecular biology and evolution 33, 384-393, https://doi.org/10.1093/molbev/msv225 (2016).

30 Cotton, A. M. et al. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Human molecular genetics 24, 1528-1539, https://doi.org/10.1093/hmg/ddu564 (2015).

31 Schultz, M. D. et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523, 212-216, https://doi.org/10.1038/nature14465 (2015).

32 Kolehmainen, M. et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. International journal of obesity 32, 292-303, https://doi.org/10.1038/sj.ijo.0803718 (2008).

33 R Development Core Team. R: a language and environment for statistical computing. https://doi.org/3-900051-07-0 (2011).

34 Ritchie, M. E. et al. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC bioinformatics 12, 68, https://doi.org/10.1186/1471-2105-12-68 (2011).

35 Ling, H., Hetrick, K., Bailey-Wilson, J. E. & Pugh, E. W. Application of sex-specific single-nucleotide polymorphism filters in genome-wide association data. BMC proceedings 3 Suppl 7, S57, https://doi.org/10.1186/1753-6561-3-S7-S57 (2009).

36 Ziegler, A. Genome-wide association studies: quality control and population-based measures. Genetic epidemiology 33 Suppl 1, S45-50, https://doi:10.1002/gepi.20472 (2009).

37 Zheng, G., Joo, J., Zhang, C. & Geller, N. L. Testing association for markers on the X chromosome. Genetic epidemiology 31, 834-843, https://doi.org/10.1002/gepi.20244 (2007).

38 Gao, F. et al. XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. The Journal of heredity 106, 666-671, https://doi.org/10.1093/jhered/esv059 (2015).

39 Chang, D. et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PloS one 9, e113684, https://doi.org/10.1371/journal.pone.0113684 (2014).

40 Ma, L., Hoffman, G. & Keinan, A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. BMC genomics 16, 241, https://doi.org/10.1186/s12864-015-1463-y (2015).

41 Hickey, P. F. & Bahlo, M. X chromosome association testing in genome wide association studies. Genetic epidemiology 35, 664-670, https://doi.org/10.1002/gepi.20616 (2011).

42 Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. American journal of human genetics 76, 887-893, https://doi.org/10.1086/429864 (2005).

J Clin Med. 2019;8(9):1471. doi:10.3390/jcm8091471. IF: 3.303, Q1 at MEDICINE, GENERAL & INTERNAL.

Study 3 | Common Variants in 22 Genes Regulate Response to Metformin Intervention in Children with Obesity: A Pharmacogenetic Study of a Randomized Controlled Trial

Augusto Anguita-Ruiz[1,2,3], Belén Pastor-Villaescusa[1,4,*], Rosaura Leis[3,5], Gloria Bueno[3,6], Raúl Hoyos[7], Rocío Vázquez-Cobela[5], Miriam Latorre-Millán[6,8], M. Dolores Cañete[9], Javier Caballero-Villarraso[10], Ángel Gil[1,2,3], Ramón Cañete[9,11] and Concepción M. Aguilera [1,2,3] .

Abstract Metformin is a first-line oral antidiabetic agent that has shown additional effects in treating obesity and metabolic syndrome. Inter-individual variability in metformin response could be partially explained by the genetic component. Here, we aimed to test whether common genetic variants can predict the response to metformin intervention in obese children. The study was a multicenter and double-blind randomized controlled trial that was stratified according to sex and pubertal status in 160 children with obesity. Children were randomly assigned to receive either metformin (1g/d) or

Affiliations 1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n. Armilla, 18016, Granada, Spain. augustoanguitaruiz@gmail.com (A.A.-R.); Belen.Pastor@med.uni-muenchen.de (B.P.-V.); agil@ugr.es (Á.G.); caguiler@ugr.es (C.M.A.) / 2. Instituto de Investigación Biosanitaria IBS.GRANADA, Complejo Hospitalario Universitario de Granada, Granada, 18014, Spain. / 3. CIBEROBN (Physiopathology of Obesity and Nutrition Network CB12/03/30038), Institute of Health Carlos III (ISCIII), Madrid, 28029, Spain. mariarosaura.leis@usc.es (R.L.); gbuenoloz@yahoo.es (G.B.) / 4. LMU – Ludwig-Maximilians-University of Munich, Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, Munich, 80337, Germany / 5. Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department, Clinic University Hospital of Santiago, University of Santiago de Compostela, Santiago de Compostela, 15706, Spain. cobela.rocio@gmail.com (R.V.-C.) / 6. Pediatric Department, Lozano Blesa University Clinical Hospital, University of Zaragoza, Zaragoza, 50009, Spain. latorremiriam0@gmail.com (M.L.-M.) / 7. Pediatric Department, Virgen de las Nieves University Hospital, Andalusian Health Service, Granada, 18014, Spain. raul_hoyosgurrea@yahoo.es (R.H.) / 8. Health Sciences Institute in Aragon, Zaragoza, 50009, Spain / 9. PAIDI CTS-329, Maimonides Institute of Biomedical Research of Córdoba (IMIBIC), Córdoba, 14004, Spain. mdcanete@hotmail.com (M.D.C.); em1caesr@uco.es (R.C.) / 10. Clinical Analysis Services, IMIBIC/Reina Sofía Hospital, University of Córdoba, Córdoba, 14004, Spain. bc2cavij@uco.es (J.C.-V.) / 11. Unit of Pediatric Endocrinology, Reina Sofia University Hospital, Córdoba, 14004, Spain
* Corresponding authors

placebo for six months after meeting the defined inclusion criteria. We conducted a post hoc genotyping study in 124 individuals (59 placebo, 65 treated) comprising finally 231 genetic variants in candidate genes. We provide evidence for 28 common variants as promising pharmacogenetics regulators of metformin response in terms of a wide range of anthropometric and biochemical outcomes, including body mass index (BMI) Z-score, and glucose, lipid, and inflammatory traits. Although no association remained statistically significant after multiple-test correction, our findings support previously reported variants in metformin transporters or targets as well as identify novel and promising loci, such as the *ADYC3* and the *BDNF* genes, with plausible biological relation to the metformin's action mechanism. Trial Registration: Registered on the European Clinical Trials Database (EudraCT, ID: 2010-023061-21) on 14 November 2011 [1].

Keywords: metformin; obesity; pediatrics; SNP; pharmacogenetics; clinical trials

1. Introduction

The prevalence of overweight and obese children is a serious worldwide issue and one of the major health challenges of the 21st century [2]. Childhood obesity plays an important pathophysiologic role in the development of insulin resistance, dyslipidemia, and hypertension[3], leading to type 2 diabetes mellitus (T2DM) and enhanced risk of cardiovascular disease during adulthood [2]. Several investigations have confirmed that intensive lifestyle interventions can increase weight-loss as well as reduce the later risk of developing T2DM in children with obesity[3]. Nevertheless, lifestyle changes alone are not always effective [5]. On the other hand, there are no approved weight-loss medications for children under 12 years of age [5]. Metformin is the first-line oral anti-hyperglycemic agent approved by the US Food Drug Administration to treat T2DM in adults and children aged >10 years. Beyond its antidiabetic effects, metformin has been considered a promising compound for the amelioration of adolescent and childhood obesity; especially through the reduction of body mass index (BMI) Z-score and waist circumference (WC) [5-7].

According to the literature, there is considerable inter-individual variability in response to metformin. In relation to the glycemic response, although an important heritable component has been described (20–34%), there is only a few available genome-wide association studies (GWAS) and yet no consistently replicated genetic variants [8,9]. Fewer efforts have been made in the context of the anti-obesity action of metformin in children, with only two available pharmacogenetic studies [10,11]. The first one, which is focused on the metformin organic cation transporter 1 (OCT1) (*SLC22A1* gene), yielded controversial findings and reported the need for additional obesity targets to be analyzed in future approaches [10]. The other study is a pharmacokinetic approach in children with obesity, and could not identify any influence of genetic variants from the OCT1 and multidrug and toxin extrusion protein 1 (MATE1) transporters on the pharmacokinetics of metformin [11].

GWAS and other genetic studies for BMI, waist-to-hip ratio, and other adiposity measures have identified more than 300 single-nucleotide polymorphisms (SNPs) that are strongly associated with obesity risk [12]. Interestingly, a pharmacogenetic approach focusing on these and other candidate genes might shed light on the action mechanism of metformin as a weight-reduction drug in children. At the same time, it might identify genetic variants that may be useful clinically to predict metformin efficacy.

On a previous work, we conducted a randomized control trial (RCT) in children with obesity and demonstrated that a six-month intervention with metformin decreases the BMI Z-score and improves inflammatory and cardiovascular-related obesity parameters [6]. Here, we conduct a genotyping study comprising hundreds of obesity and metformin candidate genes in 124 children, which are part of our previous RCT, with the aim to test whether common variants can predict the response to metformin intervention in terms of the post-treatment change in glucose metabolism, anthropometry, lipid metabolism, adipokines, and inflammatory markers. To our knowledge, this is the first candidate-gene pharmacogenetic approach focused on the effects of metformin in children with obesity. Pharmacogenetic studies such as this are crucial to provide new insight into the mechanisms regulating metabolic dysfunction and may point the way toward novel therapeutic targets for more precise interventions in childhood obesity.

2. Experimental Section

2.1. Study Design, Participants, and Intervention

The study was a multicenter and double-blind RCT, stratified according to sex and pubertal status in 160 children with obesity. Pubertal stage was determined according to Tanner criteria [13], and obesity was defined according to BMI by using the age and sex-specific cutoff points proposed by Cole *et al.* [14]. Children were randomly assigned to receive either (1 g/d) metformin or placebo for six months after meeting the defined inclusion criteria [15]. Figure S1 shows the flow diagram of participants throughout the study. All the details regarding study protocol, design, sample size, intervention, and participants (participant's data collection and processing, samples codification, randomization method, double-blind condition, and adverse effects assessment) have been previously described [6,15]. The CONSORT statement (Consolidated Standards of Reporting Trials) has been considered in the study design report and the flow diagram (Figure S1).

2.2. Informed Consent and Ethics

All the patients and their parents/guardians were previously informed about the characteristics of the trial. The informed consent, read and signed, was mandatory to participate in this study. The study was conducted in accordance with the Declaration of Helsinki and received ethics approval. It was approved by the Ethics and Investigation Committees of the hospitals (Hospital Universitario

Reina Sofía, Hospital Universitario de Santiago de Compostela, Hospital Clínico Universitario Lozano-Blesa, Hospital Universitario Virgen de las Nieves) at which the study was developed, whose reference was provided by the Ethics Committee for Biomedical Research of Andalusia on 15 January 2012 (acta 1/12) (ID code: 2010-2739). The study was registered by the European Clinical Trials Database (EudraCT, ID: 2010-023061-21) on 14 November 2011 [1].

2.3. Blood Samples Collection

Blood samples were obtained between 08:30 and 10:30, and collected in overnight fasting conditions at the beginning and at the end of the trial, as previously reported [15]. For DNA extraction, peripheral white blood cells (buffy coat) were taken. All the samples were collected and stored frozen at −80 °C until analysis.

2.4. Anthropometric and Biochemical Measurements

Anthropometry, blood pressure, and serum concentrations of glucose, insulin, lipids (total cholesterol, triglycerides, high-density lipoprotein cholesterol (HDLc), low-density lipoprotein cholesterol (LDLc)), apolipoprotein A1 (Apo A1), and apolipoprotein B (Apo B) were measured, as previously reported [15]. The quantitative insulin sensitivity check index (QUICKI) and homeostasis model assessment for insulin resistance (HOMA-IR) were also calculated. Specific plasma adipokines, inflammation, and cardiovascular risk biomarkers (adiponectin, leptin, resistin, myeloperoxidase (MPO), total plasminogen activator inhibitor-1 (tPAI-1), tumor necrosis factor-alpha (TNF-α), interferon-γ (IFN-γ), C-reactive protein (CRP), monocyte chemoattractant protein-1 (MCP-1), interleukin-8 (IL-8), soluble intercellular adhesion molecule-1 (sICAM-1), and soluble vascular adhesion molecule-1 (sVCAM-1) were analyzed in duplicate by using XMap technology (Luminex Corporation, Austin, TX, USA) and human monoclonal antibodies (Milliplex Map Kit; Millipore, Billerica, MA, USA), as previously detailed [6].

Based on the adiponectin and leptin concentrations, the adiponectin–leptin ratio (ALR) was calculated.

2.5. DNA Extraction and Genotyping

The 140 individuals that completed the study intervention (68 treated children and 72 placebo) were included for the current genetic analyses. Genomic DNA was extracted from peripheral white blood cells using two kits, the Qiamp® DNA Investigator Kit for coagulated samples and the Qiamp® DNA Mini & Blood Mini Kit for non-coagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). Genotyping analysis was performed by TaqMan allelic discrimination assay using the QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA).

2.5.1. Candidate Gene and SNP Selection

For the genotyping in the 140 DNA samples, we selected candidate genes and SNPs according to seven categories (Table S1): (1) SNPs in high-likelihood candidate genes for human obesity according to the literature and previous genotyping studies conducted by our research group; (2) SNPs in brown fat cell differentiation genes; (3) SNPs in differentially expressed genes according to a previous own microarray analysis of visceral adipose tissue in obese and normal-weight children; (4) SNPs identified by ongoing GWAS and big cohort studies for obesity and related metabolic traits in adult European populations; (5) SNPs related to metformin drug-metabolizing enzymes, transporters, and other previously reported pharmacogenetic targets; (6) SNPs predicted as binding sites of microRNAs related to obesity and metabolic dysfunction; and (7) SNPs in inflammation, oxidative stress, and antioxidant defense genes. During the SNP selection procedure, we used the Tagger program to capture (at $r^2 = 0.8$) common (minor allele frequency (MAF) ≥5%) variants in European (CEU) HapMap population in these candidate genetic regions. The candidate genes and the number of SNPs analyzed per category are detailed in Table S1 [9,12,16-28]. Moreover, the genomic information for all the analyzed SNPs is summarized in Table S2. As a result, 255 SNPs on loci strongly associated with obesity as well as previously known metformin pharmacogenetic targets were finally genotyped. Our 255 selected markers map to 181 candidate loci across the human genome.

For the quality control analysis of all the candidate markers, we evaluated the linkage disequilibrium (LD), call rate, Hardy–Weinberg equilibrium (HWE), and MAF by experimental arm. In both the treatment and placebo arms, the MAFs of all SNPs were >5% and similar to those reported for Iberian populations in Spain in phase 3 of the 1000 Genomes Project. To account for the presence of genotyping errors, all SNPs and individuals with a <90% call rate were excluded from the analyses. In relation to HWE, Wigginton's exact test [29] was applied at an alpha of 0.05 in a cohort of 258 normal weight Spanish children [30]. After all quality control checks, 24 SNPs were removed. However, although due to significant deviation from HWE, the SNP rs7943316 should have been excluded, it was forced to analysis and therefore integrated in the SNP selection process (Figure 1; File S1). The reason was that it did not deviate from HWE in the Iberian population in Spain according to the frequency data presented from the 1000 Genomes Project in the Ensembl database (A|A: 0.112, A|T: 0.411 and T|T: 0.477). In addition, 16 individuals (three treated children and 13 placebo) were excluded due to a call rate <90%. This resulted in 232 markers from a final study population of 124 participants (65 treated children (32 boys) and 59 placebo (29 boys)) available for the statistical analysis. A complete workflow detailing the SNP selection procedure in the study population can be found in **Figure 1**. Additionally, all lists of SNPs excluded at each step of the selection process and more detailed information are available as a supplementary file (File S1).

Among them, there are genes strongly associated with several forms of obesity (including monogenic obesity), T2DM, as well as known drug targets or drug-metabolizing/transporting

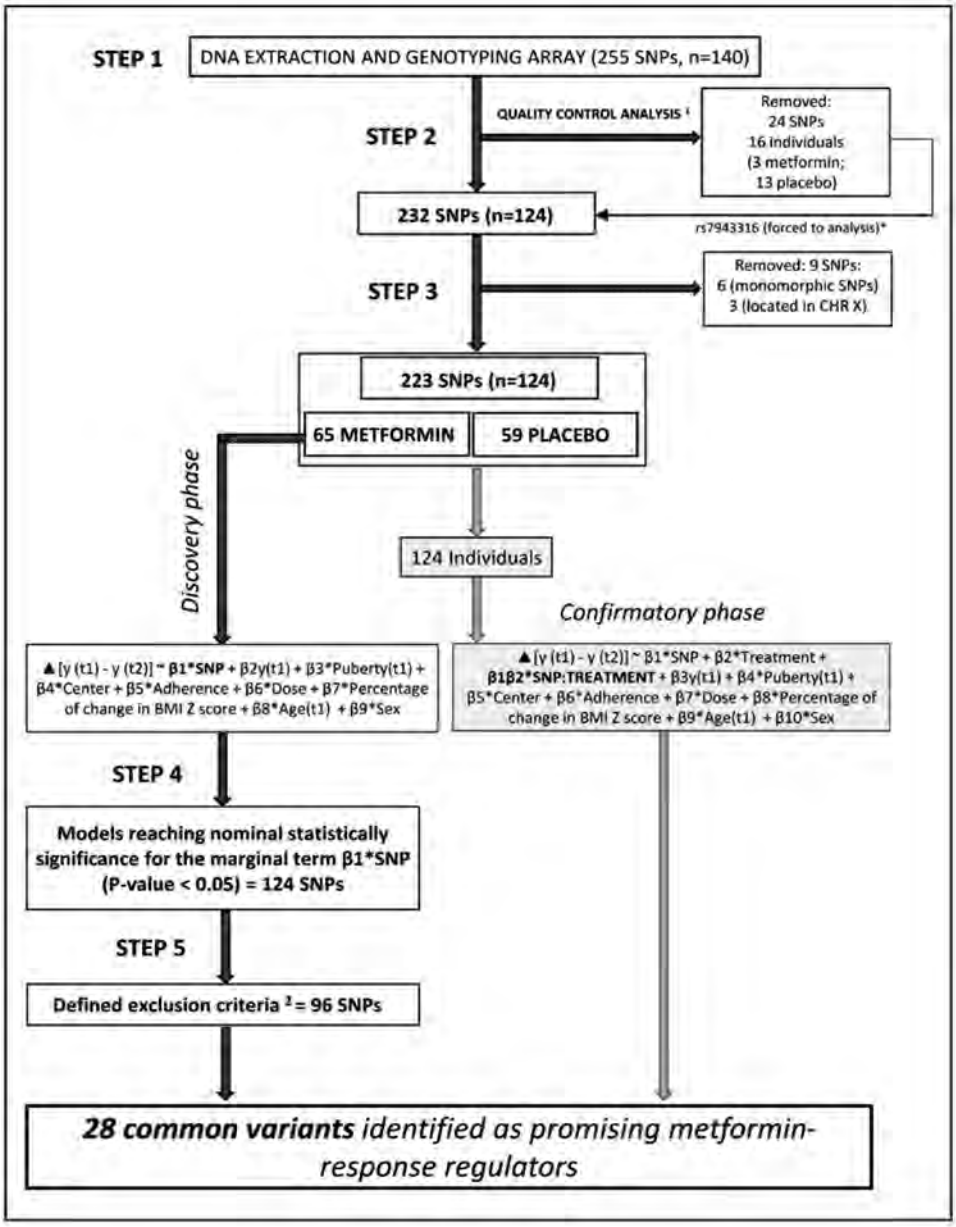


*Figure 1. Workflow of the entire SNPs selection process for the statistical approach to identify the genetic variants as promising candidates of metformin-response regulation. 1 Quality control analysis, removed if: call rate per SNP <90%, call rate per subject <90%, HWE p-value < 0.05, MAF <5%, and LD is observed. * The SNP rs7943316 was forced to analysis and therefore integrated in the SNP selection process (more details in File S1). 2 Defined exclusion criteria: (a) SNPs presenting a weak p-value (defined as significant p-value ≥ 0.045) by trait analysis (removed = 11 SNPs); (b) SNPs only associated with one of the outcomes studied and not previously evidenced as metformin pharmacogenetic targets (removed = 57 SNPs); and (c) SNPs not showing a coherent behavior in their association across different phenotypes (regarding the direction of their beta estimates) and not previously evidenced as metformin pharmacogenetic targets (removed = 28 SNPs). Abbreviations: HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; MAF, minor allele frequency; SNP, single nucleotide polymorphism.*

enzymes. A functional enrichment analysis (FEA) performed with the GeneTerm Linker R package revealed that they participate in important cellular processes and functions. Functional meta groups identified in the FEA analysis comprise: brown fat cell differentiation, cellular response to insulin stimulus, glucose homeostasis, regulation of blood pressure, response to oxidative stress, cellular component movement, response to glucocorticoid stimulus, protein kinase binding, cytokine-mediated signaling pathways, activation of adenylate cyclase activity, and respiratory electron transport chain.

Previously associated metformin pharmacogenetic variants that were not genotyped in our study comprise the *CAPN10*-rs3792269, the *OCT1*-rs628031, the *OCT1*-rs36056065, the *KCNQ1*-rs163184, and the *SP1*-rs784888. Thus, neither information nor new knowledge has been reported here for these variants.

2.6. Statistical Analysis

Pharmacogenetic analyses of metformin response were performed in the treatment arm as part of a discovery phase. However, since the application alone of a single-arm design could skip important pharmacogenetic behaviors (especially in the case of weight-loss interventions), a complementary phase to our treatment-arm approach was conducted including an SNP*treatment interaction term and placebo individuals (confirmatory phase). Especially for the case of weight-loss interventions, the inclusion of a secondary phase such as this is of special importance, helping in the confirmation of true pharmacogenetic regulators of metformin-induced weight-loss. That is to say, it will allow the final confirmation of genetic loci with effects seen in the treatment arm but not the control arm. **Figure 1** also describes all the steps for the statistical analysis.

In both phases, we applied multiple linear regression models to test the effect of each SNP on metformin response under an additive genetic model of inheritage, where gi $\in$ {0,1,2} is the number of minor alleles for the ith individual. Delta changes (T1–T2) for each outcome were calculated and used as dependent variables in the analyses. To address potential confounding, we implemented a variable selection procedure. Some variables were included in the models based on previous findings [6] or expert knowledge, while other variables were selected based on the backwards selection approach and the Bayesian information criterion. Final employed models after covariate selection can be found in File S2. Therefore, the covariates included in all the models were: the corresponding outcome, pubertal stage (prepubertal/pubertal), exact age (years) (all of them at baseline (T1)), center of recruitment (Hospital Universitario Reina Sofía, Hospital Universitario de Santiago de Compostela, Hospital Clínico Universitario Lozano-Blesa, Hospital Universitario Virgen de las Nieves), adherence (((pills ingested – pills returned)/pills predicted) × 100), dose (mg metformin or placebo/kg body weight), and sex. Additionally, models for outcomes strongly correlated to BMI Z-score (glucose metabolism, blood pressure, lipid metabolism, fat mass, adipokines, inflammation, and cardiovascular risk biomarkers) were further adjusted by the

percentage of BMI Z-score change as a confounder. Furthermore, height was also considered as another confounder for the blood pressure outcomes [31]. Continuous variables and calculated deltas were tested for normality using the Shapiro–Wilk test and transformed when necessary by means of the natural log or the rank-based inverse normal transformation. All regression models were evaluated by model control (investigating the linearity of effects on outcome(s), consistency with a normal distribution, and variance homogeneity). All residuals versus fitted, normal Q-Q, scale location, and residuals versus leverage plots are available upon request.

We quantified the statistical power of our approach to detect modest genetic effects (F^2 = 0.30) according to an alpha value of 0.05, a sample size of 65 metformin-treated children, and up to nine independent variables.

Correction for multiple tests requires special attention in genetic association studies. Given the high number of markers and collected measures, we considered several parallel approaches to correct for multiple hypothesis testing based on the number of SNPs and outcomes examined. Specifically, we employed multiple-test correction based on the methods proposed by Holm (1979), Hommel (1988), and Benjamini and Yekutieli (2001). To estimate the expected proportion of type I errors among the rejected hypotheses, we further computed false discovery rates (FDRs) as in Benjamini and Hochberg [32]. Given the presence of LD, the FDR method is a proper approach that does not assume independence between markers. Here, none of our findings underwent strict statistical correction for multiple hypotheses testing by FDRs. In this regard, the novel findings reported here should be viewed as hypothesis generating.

3. Results

3.1. Identification of 28 Common Variants as promising metformin pharmacogenetic markers

Among all the models that reached nominal statistically significance for the pharmacogenetic associations in the discovery phase (step 4: 124 SNPs; **Figure 1**), we removed 96 SNPs according to the exclusion criteria defined in the legend of the **Figure 1**. Although we tried to parameterize the process (more details are given in File S1), expert knowledge and a strict scientific criterion were the cornerstones during the SNP selection. Hence, as special conditions to maintain important SNPs in the analysis, we established SNPs that were considered as metformin pharmacogenetic targets [9,25,27,33,34] as well as SNPs that had presented a high p-value in their associations, as analyzed in step 4. Altogether, the selected markers represented the set of 28 common variants (**Figure 1**) distributed in 22 genes and mainly represented by intronic-like SNPs. Beyond the discovery phase, all associations were further interrogated for confirmation including a SNP–treatment interaction term and placebo individuals in the models.

For the 28 selected SNPs, information related to SNP-type (exonic, ncRNA-intronic, promoter, UTR3'-5', and intergenic variants), chromosome number, HWE, and MAF by experimental arm is

presented in Table S3. In relation to HWE, all associated SNPs, except for rs7943316, hold equilibrium according to Wigginton's exact test.

For both intervention groups, the general and clinical characteristics of the 124 children at the baseline and post-treatment stages, as well as the details of the statistical analysis in relation to the differences at baseline and post-treatment between groups are reported in the Table S4.

Among the reported pharmacogenetic associations, the results highlight a simultaneous effect of certain individual SNPs on several phenotypes as metformin-response regulators (**Figure 2**).

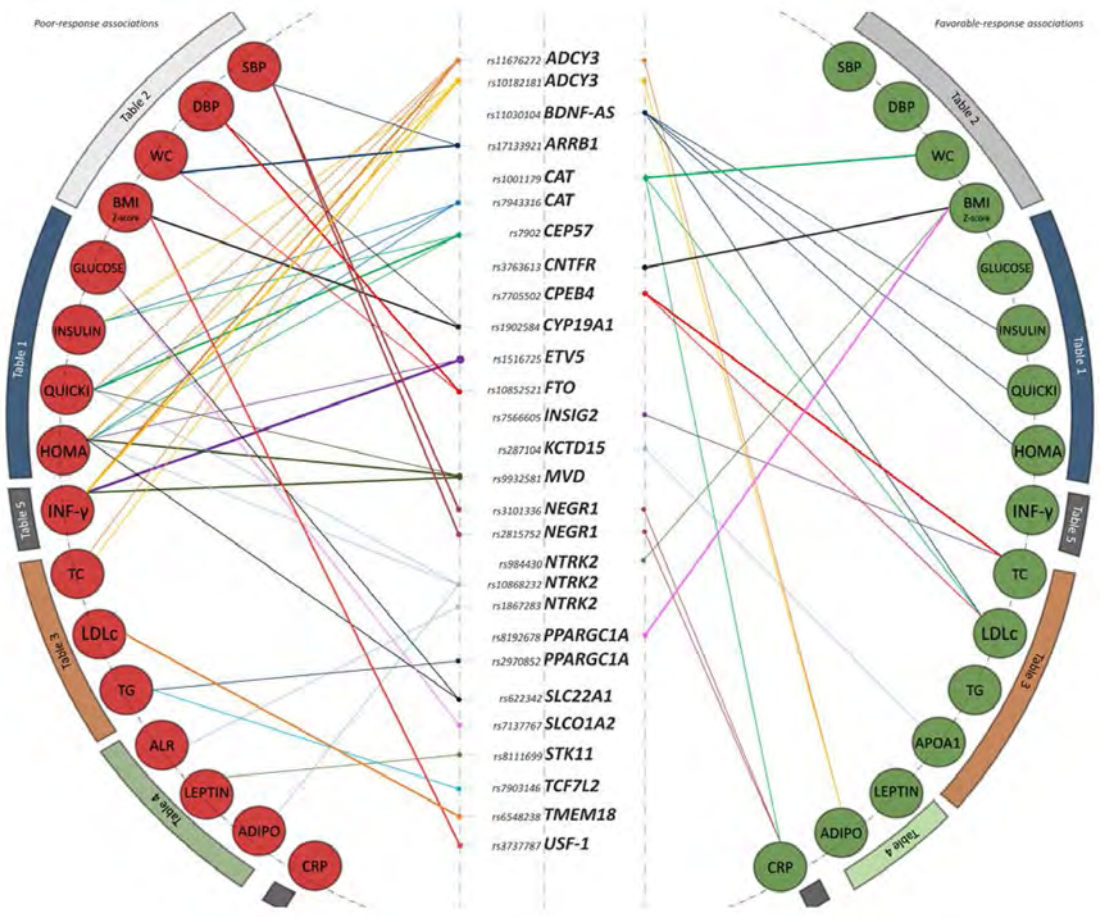


Figure 2. Interaction graph comprising all reported statistically significant associations in the discovery phase. Associations are clustered by phenotype block. The right half of the plot represents favorable-response associations, while the left half of the plot represents poor-response associations. Graph edges are weighted by the level of significance reported for each association. Abbreviations: ADIPO, adiponectin; ALR, adiponectin–leptin ratio; CRP, C-reactive protein; DBP, diastolic blood pressure; HOMA-IR, homeostasis model assessment for insulin resistance; INF-γ, interferon-γ; LDLc, low-density lipoproteins-cholesterol; QUICKI, quantitative insulin sensitivity check index; TC, total cholesterol; TG, triglycerides; SBP, systolic blood pressure; WC, waist circumference.

3.2. Glucose Metabolism

Most results and metformin pharmacogenetic targets were identified within the axis of glucose-related phenotypes, which includes fasting glucose, insulin levels, and the HOMA-IR and QUICKI indexes. **Table 1** gathers all the results obtained for the 28 selected common variants in this phenotype block. Genetic variants in the loci *ADCY3, CAT, CEP57, ETV5, MVD, NTRK2, SLC01A2,* and *SLC22A1* behaved as poor-response markers after the six-month intervention (**Figure 2**). The most significant finding of the present block corresponded to the *CEP57*-rs7902 SNP and the QUICKI index change as outcome. This SNP stood out as a poor-response marker associated with a worsening in the ability of metformin to ameliorate the QUICKI index after the intervention (β = 0.49, confidence interval (CI) = (0.2, 0.78), p-value = 0.002). The result was consistent with a study-wide 62% FDR. Other poor-response associations were reported between the *MVD*-rs9932581, *SLC22A1*-rs622342, and *ADCY3*-rs11676272, and the HOMA-IR change as outcome (β = −0.45, CI = (−0.74, −0.16), p-value = 0.004; β = −0.38, CI = (−0.72, −0.04), p-value = 0.03 and β = −0.31, CI = (−0.58, −0.05), p-value = 0.03, respectively). On the contrary, the BDNF-AS-rs11030104 was the only variant underlined as a favorable pharmacogenetic marker in the block. Specifically, children carrying the G minor allele experienced an enhanced effect of metformin on fasting insulin levels (β = 0.48, CI = (0.11, 0.85), p-value = 0.01), HOMA-IR (β = 0.48, CI = (0.12, 0.83), p-value = 0.01), and QUICKI index (β = −0.47, CI = (−0.82, −0.11), p-value = 0.01) after the six-month intervention. All reported associations were independent of BMI Z-score.

3.3. Anthropometry and Blood Pressure

The results for anthropometry and blood pressure outcomes are presented in **Table 2**. Here, while genetic variants in the *ARRB1, CYP19A1, FTO, NEGR1* and *USF-1* genes behaved as poor-response markers, SNPs in the *CAT, CNTFR, NTRK2,* and *PPARGC1A* were reported as favorable pharmacogenetic targets (**Figure 2**). The top significant result of this phenotype block belonged to the *USF-1*-rs3737787 marker and the BMI Z-score change as outcome. The association implied a worsening in the response to metformin estimated in a less decrease, per the A allele copy, of BMI Z-score after the six-month intervention (β = −0.57, CI = (−0.91, −0.24), p-value = 0.001). The result was consistent with a study-wide 39% FDR. Similar direction in findings was obtained for the *FTO*-rs10852521 SNP and the outcomes WC and diastolic blood pressure (DBP) (β = −2.41, CI = (−4.44, −0.38), p-value = 0.02 and β = −4.79, CI = (−8.07, −1.52), p-value = 0.007, respectively). In relation to favorable-response pharmacogenetic markers, there was a remarkable association among the variants *PPARGC1A*-rs8192678, *CNTFR*-rs3763613, and *NTRK2*-rs984430, and the BMI Z-score change as outcome (β = 0.51, CI = (0.16, 0.87), p-value = 0.007; β = 0.55, CI = (0.19, 0.91), p-value = 0.004; and β = 0.56, CI = (0.05, 1.08), p-value = 0.03, respectively); as well as between the *CAT*-rs1001179 and the WC change (β = 3.03, CI = (1.24, 4.82), p-value = 0.002). All reported associations in blood pressure outcomes were independent of BMI Z-score and height. Further data obtained for weight

Table 1. Summary of association data for the 28 selected common variants in glucose metabolism outcomes.

			Δ Fasting Glucose		Δ Fasting Insulin		Δ HOMA-IR		Δ QUICKI	
SNP	**Nearest Gene**	**Effect (Other) Allele**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**
rs11676272	*ADCY3*	A (G)	0.08 (−2.57, 2.74)	0.95	−0.36 (−0.63, −0.08)	**0.01**	−0.31 (−0.58, −0.05)	**0.03**	0.31 (0.04, 0.59)	0.03
rs10182181	*ADCY3*	A (G)	−0.32 (−3.03, 2.39)	0.82	−0.34 (−0.62, −0.06)	**0.02**	−0.31 (−0.58, −0.04)	**0.03**	0.27 (−0.01, 0.55)	0.07
rs17133921	*ARRB1*	A (G)	0.70 (−4.32, 5.71)	0.79	0.33 (−0.2, 0.86)	0.23	0.32 (−0.18, 0.83)	0.22	−0.02 (−0.54, 0.51)	0.95
rs11030104	*BDNF-AS*	G (A)	2.43 (−1.08, 5.95)	0.18	0.48 (0.11, 0.85)	0.01	0.48 (0.12, 0.83)	**0.01**	−0.47 (−0.82, −0.11)	**0.01**
rs1001179	*CAT*	T (C)	0.43 (−2.89, 3.75)	0.80	0.27 (−0.1, 0.64)	0.16	0.24 (−0.12, 0.59)	0.20	−0.29 (−0.65, 0.06)	0.11
rs7943316	*CAT*	A (T)	−1.57 (−5.94, 2.8)	0.49	−0.63 (−1.07, −0.19)	**0.009**	−0.59 (−1.05, −0.13)	**0.02**	0.46 (0.04, 0.88)	**0.04**
rs7902	*CEP57*	A (G)	−1.02 (−3.82, 1.78)	0.49	−0.37 (−0.7, −0.05)	**0.03***	−0.38 (−0.69, −0.08)	**0.02***	0.49 (0.2, 0.78)	**0.002***
rs3763613	*CNTFR*	T (G)	−0.78 (−4.35, 2.8)	0.67	0.17 (−0.2, 0.55)	0.36	0.15 (−0.2, 0.51)	0.41	−0.06 (−0.42, 0.3)	0.73
rs7705502	*CPEB4*	A (G)	−0.7 (−4.85, 3.47)	0.75	0.08 (−0.38, 0.54)	0.74	0.02 (−0.42, 0.47)	0.92	−0.06 (−0.51, 0.39)	0.8
rs1902584	*CYP19A1*	T (A)	2.6 (−3.66, 8.85)	0.42	0.45 (−0.29, 1.18)	0.24	0.35 (−0.36, 1.06)	0.34	−0.66 (−1.35, 0.03)	0.07
rs1516725	*ETV5*	T (C)	−3.85 (−9.04, 1.35)	0.15	−0.58 (−1.24, 0.07)	0.09	−0.70 (−1.3, −0.1)	**0.03**	0.57 (−0.05, 1.19)	0.08
rs10852521	*FTO*	T (C)	−2.60 (−5.98, 0.78)	0.14	−0.25 (−0.61, 0.12)	0.19	−0.25 (−0.59, 0.1)	0.17	0.15 (−0.21, 0.52)	0.40
rs7566605	*INSIG2*	C (G)	−2.03 (−5.13, 1.08)	0.21	−0.11 (−0.45, 0.23)	0.53	−0.19 (−0.51, 0.14)	0.26	−0.02 (−0.35, 0.32)	0.93
rs287104	*KCTD15*	G (A)	0.69 (−2.11, 3.48)	0.64	−0.11 (−0.45, 0.23)	0.53	−0.17 (−0.46, 0.13)	0.28	0.10 (−0.2, 0.4)	0.53
rs9932581	*MVD*	T (C)	−1.86 (−4.82, 1.1)	0.23	−0.43 (−0.75, −0.12)	**0.01**	−0.45 (−0.74, −0.16)	**0.004**	0.32 (0.01, 0.64)	0.05
rs3101336	*NEGR1*	T (C)	−0.90 (−3.85, 2.05)	0.55	−0.05 (−0.39, 0.28)	0.76	−0.08 (−0.4, 0.24)	0.64	−0.01 (−0.34, 0.31)	0.94
rs2815752	*NEGR1*	G (A)	−0.90 (−3.85, 2.05)	0.55	−0.05 (−0.39, 0.28)	0.76	−0.08 (−0.4, 0.24)	0.64	−0.01 (−0.34, 0.31)	0.94
rs984430	*NTRK2*	T (C)	1.88 (−2.71, 6.48)	0.43	0.06 (−0.46, 0.57)	0.83	0.16 (−0.33, 0.65)	0.52	0.10 (−0.38, 0.59)	0.67
rs10868232	*NTRK2*	G (A)	−2.51 (−6.78, 1.76)	0.26	−0.43 (−0.89, 0.02)	0.07	−0.52 (−0.94, −0.09)	**0.02**	0.49 (0.06, 0.93)	**0.03**
rs1867283	*NTRK2*	G (A)	0.84 (−2.08, 3.77)	0.57	−0.18 (−0.5, 0.13)	0.26	−0.18 (−0.49, 0.13)	0.26	0.07 (−0.24, 0.39)	0.64
rs8192678	*PPARGC1A*	T (C)	2.01 (−1.17, 5.19)	0.22	0.08 (−0.28, 0.44)	0.66	0.11 (−0.23, 0.45)	0.53	0.17 (−0.17, 0.52)	0.33
rs2970852	*PPARGC1A*	T (C)	1.01 (−1.92, 3.94)	0.50	−0.07 (−0.39, 0.26)	0.70	−0.05 (−0.36, 0.26)	0.76	0.01 (−0.3,0. 32)	0.95
rs622342	*SLC22A1*	C (A)	−3.70 (−6.95, −0.46)	**0.03**	−0.31 (−0.69, 0.06)	0.11	−0.38 (−0.72, −0.04)	**0.03***	0.11 (−0.27, 0.49)	0.57
rs7137767	*SLCO1A2*	A (C)	−3.75 (−6.57, −0.93)	**0.01**	−0.06 (−0.43, 0.3)	0.74	−0.14 (−0.48, 0.2)	0.43	−0.05 (−0.41, 0.3)	0.77
rs8111699	*STK11*	C (G)	−0.65 (−3.71, 2.42)	0.68	0.09 (−0.26, 0.45)	0.62	0.12 (−0.22, 0.46)	0.48	−0.04 (−0.39, 0.3)	0.80
rs7903146	*TCF7L2*	T (C)	−1.93 (−5.1, 1.24)	0.24	0.18 (−0.18, 0.54)	0.33	0.05 (−0.24, 0.4)	0.77	−0.06 (−0.41, 0.29)	0.72
rs6548238	*TMEM18*	T (C)	−0.83 (−4.93, 3.27)	0.69	0.18 (−0.27, 0.64)	0.43	0.08 (−0.36, 0.51)	0.73	−0.16 (−0.6, 0.28)	0.47
rs3737787	*USF-1*	A (G)	0.37 (−2.84, 3.58)	0.82	0.25 (−0.12, 0.61)	0.19	0.22 (−0.12, 0.57)	0.21	−0.02 (−0.38, 0.34)	0.92

All analyses were adjusted for baseline age, sex, pubertal stage, center of recruitment, adherence to treatment, supplied dosage, and percentage of change in BMI Z-score (see Figure S1 for more details regarding employed regression models). Specific allele effects in the treatment arm are reported here (discovery phase). Listed p-values are not adjusted for multiple comparisons. Asterisks (*) indicate which associations reached statistically significance also as treatment–SNP interactions in the confirmatory phase. Abbreviations: B, beta; CI, confidence interval; HOMA-IR, homeostasis model assessment for insulin resistance; QUICKI, quantitative insulin sensitivity check index; SNP, single-nucleotide polymorphism.

and height outcomes did not show any significant association and consequently these are not presented here, but are available upon request.

3.4. Lipid Metabolism

Regarding lipid metabolism outcomes, variants in the *ADCY3, PPARGC1A, TCF7L2,* and *TMEM18* genes were associated with a worse response to metformin intervention (**Table 3**). Otherwise, variants in the B*DNF-AS, CAT, CPEB4, INSIG2,* and *KCTD15* were identified as favorable-response markers (**Table 3** and **Figure 2**). The top findings of the present block corresponded to the SNP TMEM18-rs6548238, which was identified as a poor-response marker for the change in LDLc levels (β = −14.44, CI = (−23.88, −5), p-value = 0.005), and to the *CPEB4*-rs7705502, which was highlighted as a favorable-response marker for the change in total cholesterol levels (β = 13.81, CI = (4.77, 22.85), p-value = 0.005). These results were consistent with a study-wide FDR of 67% and 49%, respectively. The SNP *CPEB4*-rs7705502 was further identified as a favorable metformin-response marker for the change in LDLc levels. On the other hand, the SNPs *ADCY3*-rs11676272 and *ADCY3*-rs10182181 were again identified as poor-response pharmacogenetic targets, correlating with a worse ability of metformin to decrease total cholesterol levels in children carrying the effective A alleles (**Table 3**). In the same way, genetic variants in *TCF7L2* and *PPARGC1A* loci showed statistically significant results in relation to change in triglycerides levels. All the reported associations were independent of BMI Z-score. Data in relation to HDLc and Apo B did not show any significant association and consequently these are not presented here, but are available upon request.

3.5. Adipokines and Inflammatory Biomarkers

With regard to adipokines levels, we found the *STK11*-rs8111699 SNP as a poor-response marker for the change of leptin levels after the intervention (**Table 4**). Other poor-response associations of the block were reported between the *NTRK2* SNPs and the outcomes adiponectin and ALR changes.

Finally, findings related to inflammatory biomarkers are presented in **Table 5**. Outstanding results from this block were reported for the INF-γ change as outcome and the poor-response markers *ETV5*-rs1516725 and *MVD*-rs9932581 (β = −1.13, CI = (−1.68, −0.59), p-value < 0.001 and β = −0.51, CI = (−0.82, −0.21), p-value = 0.002, respectively). These results were consistent with a study-wide 6% and 22% FDR respectively. Both SNPs also presented concordant associations as negative regulators of HOMA-IR metformin-response in previous blocks (**Table 1** and **Figure 2**). Getting back to the INF-γ outcome, the *ADCY3*-rs11676272 and the *ADCY3*-rs10182181 were also underlined as poor-response markers. In this regard, children carrying the effective A alleles experienced a lower amelioration of their INF-γ levels after the intervention in comparison to major-allele carriers (β = −0.45, CI = (−0.73, −0.17), p-value = 0.003, and β = −0.45, CI = (−0.74, −0.16), p-value = 0.004, respectively). In relation to favorable-response markers, there was a remarkable

Table 2. Summary of association data for the 28 selected common variants in anthropometry and blood pressure outcomes.

			Δ BMI Z Score		Δ WC (cm)		Δ DBP		Δ SBP	
SNP	**Nearest Gene**	**Effect (other) Allele**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**
rs11676272	*ADCY3*	A (G)	−0.27 (−0.58, 0.04)	0.09	−1.35 (−3.2, 0.5)	0.16	−2.58 (−5.39, 0.22)	0.08	−0.57 (−4.61, 3.47)	0.78
rs10182181	*ADCY3*	A (G)	−0.26 (−0.59, 0.07)	0.13	−1.43 (−3.31, 0.45)	0.14	−2.26 (−5.08, 0.56)	0.12	−1.13 (−5.29, 3.03)	0.60
rs17133921	*ARRB1*	A (G)	−0.14 (−0.76, 0.47)	0.65	−4.38 (−7.19, −1.56)	**0.004***	−4.06 (−9.56, 1.44)	0.16	−8.22 (−15.58, −0.86)	0.03
rs11030104	*BDNF−AS*	G (A)	0.40 (−0.02, 0.81)	0.07	0.97 (−1.28, 3.22)	0.40	2.47 (−1.32, 6.26)	0.21	−2.59 (−8.1, 2.92)	0.36
rs1001179	*CAT*	T (C)	0.08 (−0.32, 0.49)	0.68	3.03 (1.24, 4.82)	**0.002***	0.31 (−3.34, 3.96)	0.87	−1.59 (−6.51, 3.34)	0.53
rs7943316	*CAT*	A (T)	0.002 (−0.53, 0.54)	0.99	−1.92 (−4.58, 0.75)	0.17	0.69 (−4.12, 5.50)	0.78	1.68 (−4.66, 8.02)	0.61
rs7902	*CEP57*	A (G)	−0.21 (−0.57, 0.15)	0.26	1.07 (−0.52, 2.66)	0.19	0.79 (−2.43, 4.01)	0.63	−0.03 (−4.52, 4.46)	0.99
rs3763613	*CNTFR*	T (G)	0.55 (0.19, 0.91)	**0.004***	0.87 (−1.17, 2.91)	0.41	0.27 (−3.31, 3.84)	0.88	2.27 (−2.81, 7.36)	0.39
rs7705502	*CPEB4*	A (G)	0.25 (−0.82, 0.26)	0.32	2.87 (0.02, 5.71)	0.05	−2.92 (−7.31, 1.47)	0.20	−2.62 (−8.99, 3.75)	0.43
rs1902584	*CYP19A1*	T (A)	−0.96 (−1.59, −0.32)	**0.005***	1.63 (−2.14, 5.4)	0.40	−8.93 (−15.2, −2.67)	**0.008**	−5.67 (−15.70, 4.36)	0.27
rs1516725	*ETV5*	T (C)	0.54 (−0.05, 1.14)	0.08	0.52 (−2.51, 3.55)	0.74	−4.48 (−9.70, 0.73)	0.10	−8.49 (−15.99, −0.99)	0.03
rs10852521	*FTO*	T (C)	−0.27 (−0.67, 0.11)	0.17	−2.41 (−4.44, −0.38)	**0.02**	−4.79 (−8.07, −1.52)	**0.007**	−0.08 (−5.2, 5.04)	0.98
rs7566605	*INSIG2*	C (G)	−0.08 (−0.47, 0.32)	0.71	−0.30 (−2.3, 1.7)	0.77	−1.05 (−4.51, 2.42)	0.56	−1.72 (−6.63, 3.20)	0.50
rs287104	*KCTD15*	G (A)	−0.01 (−0.36, 0.35)	0.97	0.8 (−0.9, 2.5)	0.36	1.56 (−1.38, 4.51)	0.30	0.29 (−3.90, 4.47)	0.89
rs9932581	*MVD*	T (C)	0.12 (−0.22, 0.47)	0.49	−1.52 (−3.1, 0.06)	0.07	−0.64 (−4.01, 2.73)	0.71	−0.52 (−5.18, 4.15)	0.83
rs3101336	*NEGR1*	T (C)	−0.07 (−0.44, 0.3)	0.72	0.33 (−1.51, 2.16)	0.73	−2.51 (−5.55, 0.53)	0.11	−6.22 (−10.28, −2.15)	**0.005**
rs2815752	*NEGR1*	G (A)	−0.07 (−0.44, 0.3)	0.72	0.33 (−1.51, 2.16)	0.73	−2.51 (−5.55, 0.53)	0.11	−6.22 (−10.28, −2.15)	**0.005**
rs984430	*NTRK2*	T (C)	0.56 (0.05, 1.08)	**0.03***	0.28 (−2.57, 3.12)	0.85	−0.67 (−5.67, 4.33)	0.80	−0.93 (−8.16, 6.29)	0.80
rs10868232	*NTRK2*	G (A)	0.10 (−0.4, 0.61)	0.69	0.36 (−2.25, 2.96)	0.79	−3.73 (−8.38, 0.93)	0.12	0.16 (−7.36, 7.68)	0.97
rs1867283	*NTRK2*	G (A)	−0.05 (−0.39, 0.3)	0.78	0.80 (−0.95, 2.55)	0.38	−1.79 (−4.89, 1.31)	0.27	−2.38 (−6.87, 2.10)	0.30
rs8192678	*PPARGC1A*	T (C)	0.51 (0.16, 0.87)	**0.007***	−0.02 (−2.07, 2.04)	0.99	−0.26 (−3.76, 3.25)	0.89	1.57 (−3.73, 6.86)	0.57
rs2970852	*PPARGC1A*	T (C)	−0.33 (−0.68, 0.01)	0.06	−0.20 (−1.98, 1.59)	0.83	−0.90 (−4.02, 2.22)	0.57	−1.43 (−5.99, 3.12)	0.54
rs622342	*SLC22A1*	C (A)	0.14 (−0.25, 0.54)	0.47	2.14 (0.16, 4.12)	0.05	1.05 (−2.53, 4.63)	0.57	−0.57 (−5.44, 4.30)	0.82
rs7137767	*SLCO1A2*	A (C)	0.15 (−0.22, 0.52)	0.44	−0.44 (−2.46, 1.58)	0.67	0.19 (−3.18, 3.56)	0.91	−4.21 (−8.74, 0.32)	0.08
rs8111699	*STK11*	C (G)	−0.26 (−0.62, 0.1)	0.17	−1.27 (−2.94, 0.4)	0.14	−0.94 (−4.24, 2.35)	0.58	0.36 (−4.37, 5.09)	0.88
rs7903146	*TCF7L2*	T (C)	0.24 (−0.14, 0.62)	0.22	−0.15 (−2.17, 1.86)	0.88	−1.20 (−4.68, 2.28)	0.50	−1.71 (−6.68, 3.26)	0.50
rs6548238	*TMEM18*	T (C)	−0.43 (−0.93, 0.06)	0.09	1.57 (−0.9, 4.04)	0.22	1.86 (−2.47, 6.18)	0.41	0.40 (−5.84, 6.65)	0.90
rs3737787	*USF−1*	A (G)	−0.57 (−0.91, −0.24)	**0.001***	−0.1 (−2.03, 1.83)	0.92	−2.80 (−6.11, 0.51)	0.11	−3.59 (−8.36, 1.18)	0.15

All analyses were adjusted for baseline age, sex, pubertal stage, center of recruitment, adherence to treatment, and supplied dosage. Additionally, the percentage of change in BMI and height for blood pressure outcomes (see Figure S1 for more details). Specific allele effects in the treatment arm are reported here (discovery phase). Listed p-values are not adjusted for multiple comparisons. Asterisks (*) indicate which associations reached statistically significance also as treatment–SNP interactions in the confirmatory phase. Abbreviations: B, beta; BMI, body mass index; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure; SNP, single-nucleotide polymorphism; WC, waist circumference.

Table 3. Summary of association data for the 28 selected common variants in lipid metabolism outcomes.

			Δ LDLc		Δ Total Cholesterol		Δ Triglycerides		Δ Apo A1	
SNP	Nearest Gene	Effect (other) Allele	B (95%CI)	*p*-Value	B (95%CI)	*p*-Value	B (95%CI)	*p*-Value	B (95%CI)	*p*-Value
rs11676272	*ADCY3*	A (G)	−5.77 (−12.17, 0.63)	0.08	−8.57 (−14.91, −2.22)	**0.01***	0.09 (−0.25, 0.42)	0.61	2.28 (−6.84, 11.4)	0.63
rs10182181	*ADCY3*	A (G)	−5.77 (−12.4, 0.85)	0.09	−8.77 (−15.03, −2.51)	**0.009***	0.09 (−0.24, 0.42)	0.61	−0.44 (−9.48, 8.59)	0.92
rs17133921	*ARRB1*	A (G)	−4.93 (−16.57, 6.71)	0.41	−7.44 (−18.77, 3.89)	0.20	−0.25 (−0.82, 0.32)	0.40	2.33 (−11.17, 15.83)	0.74
rs11030104	*BDNF–AS*	G (A)	9.47 (1.07, 17.87)	**0.03**	8.38 (0.12, 16.63)	0.05	−0.38 (−0.8, 0.04)	0.09	0.58 (−11.45, 12.62)	0.92
rs1001179	*CAT*	T (C)	9.99 (2.34, 17.64)	**0.01**	7.73 (−0.13, 15.6)	0.06	−0.16 (−0.6, 0.29)	0.49	−5.44 (−16.15, 5.27)	0.33
rs7943316	*CAT*	A (T)	−0.98 (−11.82, 9.86)	0.86	1.02 (−9.48, 11.51)	0.85	0.12 (−0.41, 0.65)	0.66	2.35 (−10.96, 15.66)	0.73
rs7902	*CEP57*	A (G)	2.74 (−3.99, 9.47)	0.43	−0.62 (−7.37, 6.12)	0.86	0.13 (−0.21, 0.47)	0.47	−0.59 (−10.2, 9.01)	0.90
rs3763613	*CNTFR*	T (G)	2.50 (−5.86, 10.86)	0.56	3.33 (−4.8, 11.47)	0.43	−0.13 (−0.53, 0.27)	0.52	−3.52 (−14.83, 7.79)	0.55
rs7705502	*CPEB4*	A (G)	12.87 (3.31, 22.43)	**0.01**	13.81 (4.77, 22.85)	**0.005**	0.30 (−0.19, 0.79)	0.23	6.05 (−6.82, 18.91)	0.36
rs1902584	*CYP19A1*	T (A)	−1.27 (−16.78, 14.23)	0.87	0.62 (−15.15, 16.4)	0.94	−0.37 (−1.12, 0.37)	0.33	−14.67 (−37.91, 8.56)	0.23
rs1516725	*ETV5*	T (C)	7.95 (−4.99, 20.9)	0.24	1.54 (−10.95, 14.02)	0.81	0.25 (−0.32, 0.82)	0.39	−5.14 (−21.37, 11.08)	0.54
rs10852521	*FTO*	T (C)	5.27 (−2.78, 13.33)	0.21	3.68 (−4.31, 11.66)	0.37	−0.10 (−0.49, 0.29)	0.63	−4.86 (−15.38, 5.65)	0.37
rs7566605	*INSIG2*	C (G)	5.19 (−2.3, 12.67)	0.18	7.74 (0.67, 14.82)	**0.03**	0.26 (−0.1, 0.63)	0.17	4.71 (−6.65, 16.07)	0.42
rs287104	*KCTD15*	G (A)	−3.35 (−10.17, 3.46)	0.34	−4.06 (−11.22, 3.09)	0.27	0.08 (−0.26, 0.42)	0.65	−9.81 (−17.56, −2.06)	**0.02**
rs9932581	*MVD*	T (C)	−5.45 (−12.19, 1.28)	0.12	−5.45 (−12.45, 1.54)	0.13	0.12 (−0.25, 0.49)	0.52	−1.70 (−9.6, 6.2)	0.67
rs3101336	*NEGR1*	T (C)	2.22 (−5.03, 9.47)	0.55	1.45 (−5.63, 8.53)	0.69	−0.34 (−0.68, 0)	0.06	−2.85 (−12.17, 6.47)	0.55
rs2815752	*NEGR1*	G (A)	2.22 (−5.03, 9.47)	0.55	1.45 (−5.63, 8.53)	0.69	−0.34 (−0.68, 0)	0.06	−2.85 (−12.17, 6.47)	0.55
rs984430	*NTRK2*	T (C)	2.61 (−9.08, 14.3)	0.66	3.52 (−7.69, 14.73)	0.54	−0.03 (−0.58, 0.52)	0.91	5.15 (−8.86, 19.15)	0.48
rs10868232	*NTRK2*	G (A)	−0.77 (−11.43, 9.89)	0.89	2.05 (−8.37, 12.47)	0.70	0.08 (−0.45, 0.62)	0.76	11.95 (−0.08, 23.98)	0.06
rs1867283	*NTRK2*	G (A)	−0.47 (−8.03, 7.08)	0.90	−0.94 (−8.15, 6.27)	0.80	−0.10 (−0.45, 0.26)	0.59	−2.87 (−11.91, 6.17)	0.54
rs8192678	*PPARGC1A*	T (C)	2.96 (−5.14, 11.06)	0.48	4.43 (−3.32, 12.18)	0.27	0.06 (−0.33, 0.45)	0.77	4.69 (−5.55, 14.93)	0.38
rs2970852	*PPARGC1A*	T (C)	−0.60 (−8.01, 6.81)	0.87	−1.57 (−8.77, 5.64)	0.67	−0.36 (−0.7, −0.03)	**0.04**	1.80 (−7.26, 10.85)	0.70
rs622342	*SLC22A1*	C (A)	−0.99 (−8.87, 6.89)	0.81	−1.27 (−9.3, 6.77)	0.76	0.36 (−0.04, 0.76)	0.08	−2.15 (−11.77, 7.47)	0.66
rs7137767	*SLCO1A2*	A (C)	3.45 (−4.22, 11.12)	0.39	1.34 (−6.13, 8.8)	0.73	0.23 (−0.15, 0.6)	0.24	−7.87 (−17.37, 1.62)	0.12
rs8111699	*STK11*	C (G)	−5.23 (−12.3, 1.83)	0.15	−4.07 (−11.02, 2.88)	0.26	0.02 (−0.35, 0.39)	0.93	6.43 (−2.94, 15.81)	0.19
rs7903146	*TCF7L2*	T (C)	2.98 (−5.0, 11.0)	0.47	1.49 (−6.74, 9.73)	0.72	−0.44 (−0.8, −0.07)	**0.02***	−10.90 (−20.81, −1)	**0.04**
rs6548238	*TMEM18*	T (C)	−14.44 (−23.88, −5)	**0.005**	−9.01 (−18.81, 0.8)	0.08	−0.08 (−0.59, 0.43)	0.76	0.18 (−11.8, 12.16)	0.98
rs3737787	*USF–1*	A (G)	2.51 (−5.4, 10.42)	0.54	0.71 (−7.05, 8.47)	0.86	−0.07 (−0.46, 0.32)	0.73	−2.64 (−12.02, 6.75)	0.59

All analyses were adjusted for baseline age, sex, pubertal stage, center of recruitment, adherence to treatment, supplied dosage, and percentage of change in BMI Z-score (see supplementary Figure S1 for more details regarding employed regression models). Specific allele effects in the treatment arm are reported here (discovery phase). Listed p-values are not adjusted for multiple comparisons. Asterisks (*) indicate which associations reached statistically significance also as treatment–SNP interactions in the confirmatory phase. Abbreviations: B, beta; CI, confidence interval; LDLc, low-density lipoprotein cholesterol; SNP, single-nucleotide polymorphism.

Table 4. Summary of association data for the 28 selected common variants in relation to adipokines and fat mass levels.

			Δ Adiponectin		Δ Leptin		Δ ALR		Δ Fat Mass	
SNP	**Nearest Gene**	**Effect (other) Allele**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**
rs11676272	*ADCY3*	A (G)	−0.43 (−0.76, −0.10)	**0.01***	−1.82 (−4.21, 0.57)	0.14	−0.12 (−0.47, 0.22)	0.50	−0.23 (−1.84, 1.38)	0.78
rs10182181	*ADCY3*	A (G)	−0.47 (−0.79, −0.14)	**0.008***	−2.17 (−4.59, 0.25)	0.09	−0.14 (−0.50, 0.22)	0.44	−0.21 (−1.86, 1.43)	0.80
rs17133921	*ARRB1*	A (G)	0.20 (−0.42, 0.82)	0.53	−2.32 (−6.78, 2.15)	0.32	0.36 (−0.27, 0.98)	0.27	−0.39 (−3.2, 2.43)	0.79
rs11030104	*BDNF–AS*	G (A)	0.14 (−0.33, 0.62)	0.56	−0.62 (−3.95, 2.71)	0.72	0.10 (−0.36, 0.56)	0.68	−0.48 (−2.6, 1.64)	0.66
rs1001179	*CAT*	T (C)	−0.16 (−0.61, 0.30)	0.51	0.17 (−2.89, 3.23)	0.91	−0.23 (−0.66, 0.21)	0.31	1.40 (−0.62, 3.41)	0.18
rs7943316	*CAT*	A (T)	0.24 (−0.34, 0.83)	0.42	0.81 (−3.55, 5.17)	0.72	0.36 (−0.26, 0.97)	0.27	−0.69 (−3.27, 1.89)	0.60
rs7902	*CEP57*	A (G)	−0.08 (−0.45, 0.30)	0.69	0.25 (−2.22, 2.71)	0.85	−0.22 (−0.62, 0.17)	0.28	0.22 (−1.52, 1.96)	0.80
rs3763613	*CNTFR*	T (G)	−0.05 (−0.50, 0.40)	0.84	0.92 (−2.37, 4.21)	0.59	0.05 (−0.42, 0.53)	0.83	0.19 (−1.79, 2.17)	0.85
rs7705502	*CPEB4*	A (G)	−0.51 (−1.05, 0.04)	0.07	−0.09 (−4.09, 3.92)	0.97	−0.21 (−0.76, 0.34)	0.46	−0.13 (−2.65, 2.4)	0.92
rs1902584	*CYP19A1*	T (A)	−0.21 (−1.05, 0.62)	0.62	−0.75 (−6.75, 5.24)	0.81	−0.53 (−1.38, 0.31)	0.22	2.67 (−0.92, 6.27)	0.15
rs1516725	*ETV5*	T (C)	0.06 (−0.74, 0.87)	0.88	−2.14 (−7.03, 2.75)	0.40	−0.03 (−0.82, 0.77)	0.94	1.07 (−2.14, 4.29)	0.52
rs10852521	*FTO*	T (C)	−0.06 (−0.50, 0.38)	0.79	−2.26 (−5.24, 0.72)	0.14	0.09 (−0.37, 0.54)	0.72	−1.38 (−3.27, 0.52)	0.16
rs7566605	*INSIG2*	C (G)	−0.07 (−0.49, 0.35)	0.75	0.06 (−2.85, 2.96)	0.97	0.14 (−0.28, 0.56)	0.52	−0.03 (−1.86, 1.8)	0.97
rs287104	*KCTD15*	G (A)	0.17 (−0.20, 0.54)	0.36	0.46 (−2.17, 3.08)	0.74	0.22 (−0.14, 0.58)	0.24	0.62 (−0.59, 1.83)	0.32
rs9932581	*MVD*	T (C)	0.001 (−0.41, 0.40)	0.99	−2.08 (−4.60, 0.44)	0.11	0.10 (−0.27, 0.47)	0.59	0.50 (−1.22, 2.22)	0.57
rs3101336	*NEGR1*	T (C)	0.16 (−0.23, 0.55)	0.42	−1.98 (−4.66, 0.69)	0.15	−0.02 (−0.41, 0.36)	0.90	0.72 (−1, 2.45)	0.41
rs2815752	*NEGR1*	G (A)	0.16 (−0.23, 0.55)	0.42	−1.98 (−4.66, 0.69)	0.15	−0.02 (−0.41, 0.36)	0.90	0.72 (−1, 2.45)	0.41
rs984430	*NTRK2*	T (C)	−0.03 (−0.66, 0.61)	0.94	−4.09 (−8.31, 0.12)	0.06	0.12 (−0.47, 0.71)	0.70	−0.08 (−2.77, 2.61)	0.95
rs10868232	*NTRK2*	G (A)	0.60 (0.06, 1.13)	**0.03**	1.39 (−2.60, 5.37)	0.50	0.29 (−0.24, 0.82)	0.29	1.63 (−0.87, 4.14)	0.21
rs1867283	*NTRK2*	G (A)	0.33 (−0.04, 0.70)	0.09	0.44 (−2.26, 3.14)	0.75	0.43 (0.08, 0.77)	**0.02***	0.52 (−1.19, 2.22)	0.56
rs8192678	*PPARGC1A*	T (C)	−0.05 (−0.51, 0.42)	0.85	2.87 (−0.07, 5.81)	0.06	−0.10 (−0.54, 0.34)	0.67	−0.15 (−2.05, 1.75)	0.88
rs2970852	*PPARGC1A*	T (C)	0.13 (−0.25, 0.52)	0.50	−0.36 (−3.16, 2.44)	0.80	0.11 (−0.25, 0.48)	0.54	1.04 (−0.65, 2.72)	0.23
rs622342	*SLC22A1*	C (A)	−0.17 (−0.64, 0.30)	0.49	0.77 (−2.18, 3.72)	0.61	0.03 (−0.39, 0.44)	0.90	−0.02 (−1.93, 1.89)	0.98
rs7137767	*SLCO1A2*	A (C)	0.20 (−0.23, 0.64)	0.37	0.12 (−2.90, 3.13)	0.94	0.07 (−0.37, 0.50)	0.77	0.56 (−1.17, 2.29)	0.53
rs8111699	*STK11*	C (G)	−0.30 (−0.69, 0.09)	0.14	−2.76 (−5.17, −0.34)	**0.03**	−0.16(−0.54, 0.23)	0.43	−0.09 (−1.87, 1.68)	0.92
rs7903146	*TCF7L2*	T (C)	0.16 (−0.26, 0.59)	0.46	−1.26 (−4.22, 1.71)	0.41	0.13 (−0.28, 0.55)	0.54	0.98 (−0.9, 2.85)	0.31
rs6548238	*TMEM18*	T (C)	0.03 (−0.53, 0.60)	0.91	0.57 (−3.30, 4.44)	0.77	−0.33 (−0.86, 0.21)	0.24	0.94 (−1.46, 3.35)	0.45
rs3737787	*USF–1*	A (G)	−0.25 (−0.71, 0.20)	0.28	0.15 (−3.09, 3.39)	0.93	−0.45 (−0.86, −0.04)	0.05	0.70 (−1.17, 2.58)	0.46

All analyses were adjusted for baseline age, sex, pubertal stage, center of recruitment, adherence to treatment, supplied dosage, and percentage of change in BMI Z-score (see Figure S1 for more details regarding employed regression models). Specific allele effects in the treatment arm are reported here (discovery phase). Listed p-values are not adjusted for multiple comparisons. Asterisks (*) indicate which associations reached statistically significance also as treatment–SNP interactions in the confirmatory phase. Abbreviations: ALR, adiponectin–leptin ratio; B, beta; CI, confidence interval; SNP, single-nucleotide polymorphism.

Table 5. Summary of association data for the 28 selected common variants in relation to inflammatory biomarkers.

			Δ INF−γ		Δ CRP	
SNP	**Nearest Gene**	**Effect (other) Allele**	**B (95%CI)**	***p*-Value**	**B (95%CI)**	***p*-Value**
rs11676272	*ADCY3*	A (G)	−0.45 (−0.73, −0.17)	**0.003**	−0.26 (−0.64, 0.12)	0.19
rs10182181	*ADCY3*	A (G)	−0.45 (−0.74, −0.16)	**0.004**	−0.18 (−0.57, 0.20)	0.36
rs17133921	*ARRB1*	A (G)	0.21 (−0.33, 0.75)	0.44	−0.46 (−1.23, 0.31)	0.24
rs11030104	*BDNF−AS*	G (A)	0.11 (−0.34, 0.56)	0.63	0.25 (−0.24, 0.75)	0.32
rs1001179	*CAT*	T (C)	0.12 (−0.26, 0.50)	0.54	0.49 (0.05, 0.93)	**0.03***
rs7943316	*CAT*	A (T)	−0.13 (−0.63, 0.36)	0.61	0.12 (−0.51, 0.75)	0.72
rs7902	*CEP57*	A (G)	−0.13 (−0.46, 0.20)	0.45	0.07 (−0.34, 0.49)	0.72
rs3763613	*CNTFR*	T (G)	−0.08 (−0.50, 0.33)	0.69	0.23 (−0.26, 0.71)	0.37
rs7705502	*CPEB4*	A (G)	−0.24 (−0.73, 0.24)	0.33	0.05 (−0.56, 0.66)	0.87
rs1902584	*CYP19A1*	T (A)	−0.10 (−0.88, 0.69)	0.81	0.50 (−0.44, 1.43)	0.31
rs1516725	*ETV5*	T (C)	−1.13 (−1.68, −0.59)	**<0.001**	0.06 (−0.73, 0.84)	0.89
rs10852521	*FTO*	T (C)	−0.13 (−0.50, 0.24)	0.51	0.18 (−0.27, 0.64)	0.43
rs7566605	*INSIG2*	C (G)	−0.27 (−0.63, 0.08)	0.14	−0.31 (−0.73, 0.11)	0.16
rs287104	*KCTD15*	G (A)	−0.14 (−0.47, 0.20)	0.43	0.40 (0.02, 0.79)	0.05
rs9932581	*MVD*	T (C)	−0.51 (−0.82, −0.21)	**0.002**	−0.11 (−0.52, 0.30)	0.61
rs3101336	*NEGR1*	T (C)	−0.06 (−0.40, 0.28)	0.72	0.57 (0.15, 1.00)	**0.01**
rs2815752	*NEGR1*	G (A)	−0.06 (−0.40, 0.28)	0.72	0.57 (0.15, 1.00)	**0.01**
rs984430	*NTRK2*	T (C)	0.09 (−0.46, 0.64)	0.75	0.46 (−0.19, 1.11)	0.17
rs10868232	*NTRK2*	G (A)	−0.24 (−0.74, 0.26)	0.36	−0.10 (−0.71, 0.50)	0.75
rs1867283	*NTRK2*	G (A)	−0.01 (−0.36, 0.34)	0.96	0.24 (−0.18, 0.65)	0.27
rs8192678	*PPARGC1A*	T (C)	0.02 (−0.35, 0.40)	0.91	−0.09 (−0.54, 0.36)	0.69
rs2970852	*PPARGC1A*	T (C)	−0.05 (−0.39, 0.29)	0.77	0.02 (−0.38, 0.42)	0.92
rs622342	*SLC22A1*	C (A)	−0.27 (−0.68, 0.13)	0.20	0.12 (−0.35, 0.60)	0.62
rs7137767	*SLCO1A2*	A (C)	−0.18 (−0.54, 0.19)	0.35	0.08 (−0.36, 0.53)	0.71
rs8111699	*STK11*	C (G)	0.21 (−0.14, 0.55)	0.25	−0.11 (−0.53, 0.31)	0.61
rs7903146	*TCF7L2*	T (C)	0.06 (−0.32, 0.45)	0.75	0.16 (−0.31, 0.62)	0.51
rs6548238	*TMEM18*	T (C)	0.43 (−0.04, 0.89)	0.08	−0.10 (−0.74, 0.53)	0.75
rs3737787	*USF−1*	A (G)	0.14 (−0.26, 0.53)	0.50	−0.37 (−0.83, 0.09)	0.13

All analyses were adjusted for baseline age, sex, pubertal stage, center of recruitment, adherence to treatment, supplied dosage, and percentage of change in BMI Z-score (see Figure S1 for more details regarding employed regression models). Specific allele effects in the treatment arm are reported here (discovery phase). Listed p-values are not adjusted for multiple comparisons. Asterisks (*) indicate which associations reached statistically significance also as treatment–SNP interactions in the confirmatory phase. Abbreviations: B, beta; CI, confidence interval; CRP, C-reactive protein; INF-γ, interferon-γ; SNP, single-nucleotide polymorphism.

association reported between the variant *CAT*-rs1001179 and the change in CRP levels (β = −0.49, CI = (0.05, 0.93), p-value = 0.03). All associations reported here were independent of BMI Z-score. No significant association was found for the following outcomes: resistin, MPO, tPAI-1, TNF-α, MCP-1, IL-8, sICAM-1 and sVCAM-1. Hence, they are not presented here, but are available upon request.

3.6. Confirmatory Phase

In order to confirm our findings, as we previously mentioned, all reported associations were evaluated as SNP–treatment interactions after the inclusion of placebo individuals in the analyses. As a result, up to 18, among all the previously described associations, remained statistically significant (marked with an asterisk in **Tables 1–5**). These associations can be understood as true pharmacogenetic phenomena and drug–gene interactions exclusive of the individuals belonging to the metformin arm.

4. Discussion

Our results show that the well-known variability in metformin response might have a genetic origin, also in the context of weight-loss and childhood obesity. We provide evidence for 28 common variants as promising pharmacogenetic regulators of metformin response in terms of a wide range of anthropometric and biochemical outcomes including glucose, lipid, and inflammatory traits (**Figure 2**). Our results not only support previously reported associations of variants in metformin transporters or targets *(SLC22A1*, *TCFL2* and *PPARGC1A*) but also identify novel and promising loci such as the *ADYC3* and the *BDNF-AS* genes, with biological relevance in the AMP kinase (AMPK) route and other metformin-related pathways. Despite the study initially being focused on the effect of metformin as an anti-obesity agent, the bulk of the findings and metformin pharmacogenetic targets were identified within the axis of glucose-related phenotypes (**Figure 2**). This, although striking, is to be expected, taking into account the well-reported glucose-lowering effect of metformin in T2DM European adult populations [9,25,27,33,34] and the improvement of insulin status in patients with hyperinsulinemia or insulin resistance [35,36]. In this regard, it might happen that some of the beneficial effects of metformin in childhood obesity could be mediated through an improvement of the impaired glucose metabolism.

Among the novel and promising reported targets in our study, the *ADCY3* (adenylate cyclase 3) locus is especially interesting. Two SNPs within this gene were identified as poor metformin-response markers in the glucose, lipid, and inflammatory phenotype blocks (**Figure 2**). The *ADCY3* protein is a member of the mammalian adenylyl cyclase family responsible for generating the second messenger cyclic adenosine monophosphate (cAMP) in human tissues. Several lines of evidence suggest the interesting possibility that the *ADCY3* protein may play an important role in the regulation of adiposity as well as crucial physiological roles in mice muscle and liver [37], which are all target tissues of metformin. Likewise, it has been proposed that *ADCY3* dysfunction

in peripheral tissues could be related to metabolic disorders by inducing adipocyte dysfunction and insulin resistance in mice [37]. The molecular mechanism of this relation might underlie the dysregulation of the ATP/cAMP cellular balance and the resulting disruption of the PKA-induced AMPK activation. Taking it into account and given that AMPK is the main target by which metformin elicits its effects in the body, our *ADCY3* pharmacogenetic report merits special attention as a candidate gene for consideration in other genotyping and functional studies.

Other interesting findings involved two loci related to the brain-derived neurotrophic factor (BDNF) protein. These were the *BDNF-AS* region, which is an antisense RNA gene upstream the *BDNF*, and the *NTRK2* locus, which encodes the BDNF receptor protein. SNPs within these loci were robustly associated as favorable and poor-response markers respectively in all the analyzed phenotype blocks (**Figure 2**). The *BDNF-AS*–rs11030104 discovery is especially noticeable according to previous works indicating that the *BDNF-AS* intron region has a key role in regulating *BDNF* expression in humans [38]. Furthermore, our *BDNF-AS*–rs11030104 and other *BDNF* SNPs have been strongly associated with obesity risk [39] and weight response after intensive lifestyle modification [40]. BDNF is a neurotrophin that plays important functions in the central nervous system and systemic or peripheral inflammatory conditions such as acute coronary syndrome and T2DM. Interestingly, BDNF has been demonstrated to have strong anti-hyperglycemic and anti-inflammatory effects against the progression of T2DM [41]. Some studies have also revealed a strong effect of metformin as a *BDNF*-expression enhancer in mice [42,43]. On this matter, a recent review suggested that the correlation between BDNF and metformin might be the reason for metformin-induced insulin action by insulin receptor binding, metformin-induced high BDNF levels due to increasing AMPK, and enhanced tyrosine kinase receptor activity, which may amplify BDNF signaling [44]. Altogether, these findings suggest that the BDNF product could be a key element for the successful action of the drug against both obesity and T2DM conditions. Therefore, the *BDNF-AS, NTRK2,* and the *BDNF* locus could be good candidate pharmacogenetic targets to be studied in future human and *in vitro* studies.

On the other hand, we also provide evidence for exclusive and robust pharmacogenetic associations within anthropometric traits (**Table 2** and **Figure 2**). Top findings within the block involved well-known obesity genes such as the *FTO*, the *CYP19A1* and the *USF-1*. Similar results for SNPs in the *FTO* gene have been reported in a previous metformin pharmacogenetic study for the BMI Z-score change in girls with androgen excess [45]. Given that *FTO* is a key obesity-associated gene and an important factor controlling feeding behavior and energy expenditure, it could be likely that metformin elicits direct actions on obesity via adiposity reduction.

The most significant report in our study belonged to the poor-response marker *ETV5*-rs1516725 and the INF-γ change as outcome. With a study-wide FDR of 6%, this finding almost reached multiple testing correction significance. This genetic variant also presented concordant association as a negative regulator of the HOMA-IR response (**Table 1** and **Figure 2**). According to literature,

the *ETV5* gene has been associated with BMI in multiple GWAS studies [46,47] and functionally linked to obesity [48]. Specifically, *ETV5* seems to have a critical role in regulating insulin secretion and glucose metabolism in mice, which might support our strong association as a metformin-response regulator [48]. Other novel genetic regions identified in our study comprised the loci *CEP57*, *CPEB4*, *CAT*, or the *SLC01A2*, which showed concordant associations across different phenotype blocks (**Figure 2**). Interestingly, the encoded proteins of these loci participate in molecular processes that are strongly related to the action mechanism of metformin via AMPK-independent pathways [49,50].

Regarding previously reported genes in the literature, our study identified some well-known pharmacokinetic and pharmacodynamic targets of metformin such as the metformin transporter *SLC22A1*-rs622342 and the transcription factors *TCFL2*-rs7903146 and *PPARGC1A* (rs2970852 and rs8192678), for which we here replicate all previously reported associations [9,51-54]. Other literature variants also announced in our study map within the loci *STK11, FTO, INSIG2*, and the *KCTD15*. For these variants, although we do not replicate exact results, we provide findings in line with those previously presented [45,55-57], thereby strengthening our proposal and broadening previous knowledge.

We are aware of some limitations in the current study: 1) First, our observations are from a setting of multiple hypotheses testing, which only reach a nominal level of statistical significance. 2) Regarding null associations, there are variants such as the *SLC47A1*-rs2289669 or the *ATM*-rs11212617 which, in spite the wide backup of association in previous GWAS and candidate studies [25,34,51,58-60], have not reached nominally statistically significance in any of our analyses. Although the association of the *ATM*-rs11212617 as a pharmacogenetic marker remains controversial in the literature [26], the lack of significance for this and other markers in our study requires special attention. One reason for that could be a lack of statistical power in our design. Although we have reported enough statistical power 83.36% to detect previously described modest effects ($F^2 = 0.30$), we actually have an inadequate power for detecting such small effect sizes such as those identified in GWAS studies. Notwithstanding, considering the number of variants likely to influence the phenotypes under study, even a submaximal power is likely to provide a number of true positive associations. On this matter, reported associations in the *ADCY3* locus and *BDNF*-related regions still merit consideration as true pharmacogenetic associations.

5. Conclusions

In conclusion, we propose novel mechanisms by which genetics might contribute to variation in response to metformin as an anti-obesity agent in different traits. Both poor-responses and favorable-responses were identified, relying upon the allele copy to achieve an effect of metformin on glucose levels and insulin sensitivity, anthropometric parameters, blood pressure, lipid profile, adipokines, and inflammatory biomarkers. Genetic variants in promising loci such as the *ADYC3*

and the *BDNF-AS* could explain the inter-individual variability in metformin response, and therefore clinically predict the metformin efficacy based on genetics. Although interesting, none of the reported associations remained statistically significant after multiple-test correction, and thus should be interpreted with caution. Certainly, these and other generated hypotheses require more detailed characterization in bigger and independent samples. Pharmacogenetic approaches such as this might provide new insight into mechanisms regulating metabolic dysfunction and may point the way toward novel therapeutic targets for more precise interventions in childhood obesity.

Supplementary Materials

The following are available online at https://www.mdpi.com/2077-0383/8/9/1471#supplementary : Figure S1: Flow diagram of participants.; File S1: Description of the steps for the SNPs selection; File S2: Final employed statistical models for discovery phase and confirmatory phase; Table S1: Candidate genes and number of SNPs analyzed per selection category. Table S2: Genomic information for all SNPs analyzed. Table S3: Quality control parameters and genomic information for the 28 associated common variants. Table S4: Clinical characteristics of the study population at baseline and post-treatment stages.

Author Contributions

Conceptualization, A.A.-R., B.P.-V. and C.M.A.; Methodology, A.A.-R., B.P.-V., C.M.A.; Software, A.A.-R.; Validation, A.A.-R., B.P.-V. and C.M.A.; Formal analysis, A.A.-R.; Investigation, B.P.-V., R.V.-C., R.L., M.L.-M., G.B., R.H., M.D.C., A.G., R.C. and C.M.A.; Resources, R.L, G.B, R.H, Á.G, R.C, C.M.A.; Data curation, A.A.-R.; Writing—original draft preparation, A.A.-R.; Writing—review and editing, A.A.-R., B.P.-V. and C.M.A.; Visualization, A.A.-R.; Supervision, C.M.A.; Project administration, R.C.; Funding acquisition, J.C.-V., R.L., G.B., Á.G., R.C. All authors read and approved the final manuscript and take full responsibility for the manuscript content.

Funding

This research was funded by the Spanish Ministry of Health, Social and Equality, General Department for Pharmacy and Health Products (codes and beneficiaries: EC10-243, Ramón Cañete, Reina Sofía Hospital, Córdoba; EC10-056, Ángel Gil, University of Granada and Virgen de las Nieves University Hospital, Granada; EC10-281, Rosaura Leis, Clinic University Hospital of Santiago, Santiago de Compostela; and EC10-227, Gloria Bueno, Lozano Blesa University Clinical Hospital, Zaragoza).

Acknowledgments

The authors acknowledge the Spanish Ministry of Health, Social and Equality, General Department for Pharmacy and Health Products for financing this study and the Instituto de Salud Carlos III-Fondo de Investigación Sanitaria (FONDOS FEDER), Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015). The authors also thank all the children and their parents/guardians for their participation in the current RCT. This paper will be part of Augusto Anguita-Ruiz's doctorate, which is being performed under the "Nutrition and Food Sciences" program at the University of Granada.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Clinical Trials register - Search for 2010-023061-21 Available online: https://www.clinicaltrialsregister.eu/ctr-search/search?query=2010-023061-21 (accessed on Sep 12, 2019).
2. GBD 2015 Obesity Collaborators; Afshin, A.; Forouzanfar, M.H.; Reitsma, M.B.; Sur, P.; Estep, K.; Lee, A.; Marczak, L.; Mokdad, A.H.; Moradi-Lakeh, M.; et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N. Engl. J. Med. 2017, 377, 13–27.
3. Olza, J.; Gil-Campos, M.; Leis, R.; Bueno, G.; Aguilera, C.M.; Valle, M.; Cañete, R.; Tojo, R.; Moreno, L.A.; Gil, A. Presence of the metabolic syndrome in obese children at prepubertal age. Ann. Nutr. Metab. 2011, 58, 343–350.
4. Knowler, W.C.; Barrett-Connor, E.; Fowler, S.E.; Hamman, R.F.; Lachin, J.M.; Walker, E.A.; Nathan, D.M.; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N. Engl. J. Med. 2002, 346, 393–403.
5. Mead, E.; Atkinson, G.; Richter, B.; Metzendorf, M.-I.; Baur, L.; Finer, N.; Corpeleijn, E.; O'Malley, C.; Ells, L.J. Drug interventions for the treatment of obesity in children and adolescents. Cochrane Database Syst. Rev. 2016, 11, CD012436.
6. Pastor-Villaescusa, B.; Cañete, M.D.; Caballero-Villarraso, J.; Hoyos, R.; Latorre, M.; Vázquez-Cobela, R.; Plaza-Díaz, J.; Maldonado, J.; Bueno, G.; Leis, R.; et al. Metformin for Obesity in Prepubertal and Pubertal Children: A Randomized Controlled Trial. Pediatrics 2017, 140, e20164285.
7. Warnakulasuriya, L.S.; Fernando, M.M.A.; Adikaram, A.V.N.; Thawfeek, A.R.M.; Anurasiri, W.-M.L.; Silva, R.R.; Sirasa, M.S.F.; Rytter, E.; Forslund, A.H.; Samaranayake, D.L.; et al. Metformin in the Management of Childhood Obesity: A Randomized Control Trial. Child. Obes. 2018, 14, 553–565.
8. Zhou, K.; Donnelly, L.; Yang, J.; Li, M.; Deshmukh, H.; Van Zuydam, N.; Ahlqvist, E.; Spencer, C.C.; Groop, L.; Morris, A.D.; et al. Heritability of variation in glycaemic response to metformin: A genome-wide complex trait analysis. Lancet Diabetes Endocrinol. 2014, 2, 481–487.
9. Florez, J.C. The pharmacogenetics of metformin. Diabetologia 2017, 60, 1648–1655.
10. Sam, W.J.; Roza, O.; Hon, Y.Y.; Alfaro, R.M.; Calis, K.A.; Reynolds, J.C.; Yanovski, J.A. Effects of SLC22A1 Polymorphisms on Metformin-Induced Reductions in Adiposity and Metformin Pharmacokinetics in Obese Children with Insulin Resistance. J. Clin. Pharmacol. 2017, 57, 219–229.
11. van Rongen, A.; van der Aa, M.P.; Matic, M.; van Schaik, R.H.N.; Deneer, V.H.M.; van der Vorst, M.M.; Knibbe, C.A.J. Increased Metformin Clearance in Overweight and Obese Adolescents: A Pharmacokinetic Substudy of a Randomized Controlled Trial. Pediatr. Drugs 2018, 20, 365–374.
12. Goodarzi, M.O. Genetics of obesity: What genetic association studies have taught us about the biology of obesity and its complications. Lancet Diabetes Endocrinol. 2018, 6, 223–236.
13. Tanner, J.M.; Whitehouse, R.H. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. Arch. Dis. Child. 1976, 51, 170–179.
14. Cole, T.J.; Bellizzi, M.C.; Flegal, K.M.; Dietz, W.H. Establishing a standard definition for child overweight and obesity worldwide: International survey. BMJ 2000, 320, 1240–1243.
15. Pastor-Villaescusa, B.; Caballero-Villarraso, J.; Cañete, M.D.; Hoyos, R.; Maldonado, J.; Bueno, G.; Leis, R.; Gil, Á.; Cañete, R.; Aguilera, C.M. Evaluation of differential effects of metformin treatment in obese children according to pubertal stage and genetic variations: Study protocol for a randomized controlled trial. Trials 2016, 17, 323.
16. Olza, J.; Gil-Campos, M.; Leis, R.; Rupérez, A.I.; Tojo, R.; Cañete, R.; Gil, A.; Aguilera, C.M. A gene variant of 11β-hydroxysteroid dehydrogenase type 1 is associated with obesity in children. Int. J. Obes. 2012, 36, 1558–1563.
17. Olza, J.; Gil-Campos, M.; Leis, R.; Rupérez, A.I.; Tojo, R.; Cañete, R.; Gil, Á.; Aguilera, C.M. Influence of variants in the NPY gene on obesity and metabolic syndrome features in Spanish children. Peptides 2013, 45, 22–27.
18. Rupérez, A.I.; Olza, J.; Gil-Campos, M.; Leis, R.; Mesa, M.D.; Tojo, R.; Cañete, R.; Gil, Á.; Aguilera, C.M. Association of Genetic Polymorphisms for Glutathione Peroxidase Genes with Obesity in Spanish Children. J. Nutrigenet. Nutri. 2014, 7, 130–142.
19. Olza, J.; Rupérez, A.; Gil-Campos, M.; Leis, R.; Cañete, R.; Tojo, R.; Gil, Á.; Aguilera, C. Leptin Receptor Gene Variant rs11804091 Is Associated with BMI and Insulin Resistance in Spanish Female Obese Children: A Case-Control Study. Int. J. Mol. Sci. 2017, 18, 1690.
20. Rupérez, A.I.; Olza, J.; Gil-Campos, M.; Leis, R.; Mesa, M.D.; Tojo, R.; Cañete, R.; Gil, Á.; Aguilera, C.M. Are Catalase –844A/G Polymorphism and Activity Associated with Childhood Obesity? Antioxid. Redox Signal. 2013, 19, 1970–1975.
21. Aguilera, C.M.; Gomez-Llorente, C.; Tofe, I.; Gil-Campos, M.; Cañete, R.; Gil, Á. Genome-wide expression in visceral adipose tissue from obese prepubertal children. Int. J. Mol. Sci. 2015, 16, 7723–7737.

22. Deloukas, P.; Kanoni, S.; Willenborg, C.; Farrall, M.; Assimes, T.L.; Thompson, J.R.; Ingelsson, E.; Saleheen, D.; Erdmann, J.; Goldstein, B.A.; et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. 2013, 45, 25–33.
23. Locke, A.E.; Kahali, B.; Berndt, S.I.; Justice, A.E.; Pers, T.H.; Day, F.R.; Powell, C.; Vedantam, S.; Buchkovich, M.L.; Yang, J.; et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015, 518, 197–206.
24. Warrington, N.M.; Howe, L.D.; Paternoster, L.; Kaakinen, M.; Herrala, S.; Huikari, V.; Wu, Y.Y.; Kemp, J.P.; Timpson, N.J.; Pourcain, B.S.; et al. A genome-wide association study of body mass index across early life and childhood. Int. J. Epidemiol. 2015, 44, 700–712.
25. Jablonski, K.A.; McAteer, J.B.; De Bakker, P.I.W.; Franks, P.W.; Pollin, T.I.; Hanson, R.L.; Saxena, R.; Fowler, S.; Shuldiner, A.R.; Knowler, W.C.; et al. Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program. Diabetes 2010, 59, 2672–2681.
26. Florez, J.C.; Jablonski, K.A.; Taylor, A.; Mather, K.; Horton, E.; White, N.H.; Barrett-Connor, E.; Knowler, W.C.; Shuldiner, A.R.; Pollin, T.I.; et al. The C allele of ATM rs11212617 does not associate with metformin response in the Diabetes Prevention Program. Diabetes Care 2012, 35, 1864–1867.
27. Zhou, K.; Yee, S.W.; Seiser, E.L.; van Leeuwen, N.; Tavendale, R.; Bennett, A.J.; Groves, C.J.; Coleman, R.L.; van der Heijden, A.A.; Beulens, J.W.; et al. Variation in the glucose transporter gene SLC2A2 is associated with glycemic response to metformin. Nat. Genet. 2016, 48, 1055–1059.
28. Rupérez, A.I.; Gil, A.; Aguilera, C.M. Genetics of oxidative stress in obesity. Int. J. Mol. Sci. 2014, 15, 3118–3144.
29. Wigginton, J.E.; Cutler, D.J.; Abecasis, G.R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. 2005, 76, 887–893.
30. Ruiz-Ojeda, F.J.; Anguita-Ruiz, A.; Rupérez, A.I.; Gomez-Llorente, C.; Olza, J.; Vázquez-Cobela, R.; Gil-Campos, M.; Bueno, G.; Leis, R.; Cañete, R.; et al. Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism. Sci. Rep. 2019, 9, 3979.
31. Fan, H.; Liu, Y.; Zhang, X. Validation of recommended definition in identifying elevated blood pressure in adolescents. J. Clin. Hypertens. 2019, 1–7, doi:10.1111/jch.13640.
32. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B 1995, 57, 289–300.
33. Pawlyk, A.C.; Giacomini, K.M.; McKeon, C.; Shuldiner, A.R.; Florez, J.C. Metformin pharmacogenomics: Current status and future directions. Diabetes 2014, 63, 2590–2599.
34. Zhou, K.; Bellenguez, C.; Spencer, C.C.A.; Bennett, A.J.; Coleman, R.L.; Tavendale, R.; Hawley, S.A.; Donnelly, L.A.; Schofield, C.; Groves, C.J.; et al. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. Nat. Genet. 2011, 43, 117–120.
35. Clarson, C.; Mahmud, F.; Baker, J. Metformin in combination with structured lifestyle intervention improved body mass index in obese adolescents, but did not improve insulin resistance. Endocrine 2009, 36, 141–146.
36. Yanovski, J.A.; Krakoff, J.; Salaita, C.G.; McDuffie, J.R.; Kozlosky, M.; Sebring, N.G.; Reynolds, J.C.; Brady, S.M.; Calis, K.A. Effects of metformin on body weight and body composition in obese insulin-resistant children: A randomized clinical trial. Diabetes 2011, 60, 477–485.
37. Tong, T.; Shen, Y.; Lee, H.-W.; Yu, R.; Park, T. Adenylyl cyclase 3 haploinsufficiency confers susceptibility to diet-induced obesity and insulin resistance in mice. Sci. Rep. 2016, 6, 34179.
38. Pruunsild, P.; Kazantseva, A.; Aid, T.; Palm, K.; Timmusk, T. Dissecting the human BDNF locus: Bidirectional transcription, complex splicing, and multiple promoters. Genomics 2007, 90, 397–406.
39. Akbarian, S.A.; Salehi-Abargouei, A.; Pourmasoumi, M.; Kelishadi, R.; Nikpour, P.; Heidari-Beni, M. Association of Brain-derived neurotrophic factor gene polymorphisms with body mass index: A systematic review and meta-analysis. Adv. Med. Sci. 2018, 63, 43–56.
40. Delahanty, L.M.; Pan, Q.; Jablonski, K.A.; Watson, K.E.; McCaffery, J.M.; Shuldiner, A.; Kahn, S.E.; Knowler, W.C.; Florez, J.C.; Franks, P.W. Genetic predictors of weight loss and weight regain after intensive lifestyle modification, metformin treatment, or standard care in the Diabetes Prevention Program. Diabetes Care 2012, 35, 363–366.
41. Yamanaka, M.; Itakura, Y.; Tsuchida, A.; Nakagawa, T.; Taiji, M. Brain-derived neurotrophic factor (BDNF) prevents the development of diabetes in prediabetic mice. Biomed. Res. 2008, 29, 147–153.
42. Yoo, D.Y.; Kim, W.; Nam, S.M.; Yoo, K.-Y.; Lee, C.H.; Choi, J.H.; Won, M.-H.; Hwang, I.K.; Yoon, Y.S. Reduced Cell Proliferation and Neuroblast Differentiation in the Dentate Gyrus of High Fat Diet-Fed Mice are Ameliorated by Metformin and Glimepiride Treatment. Neurochem. Res. 2011, 36, 2401–2408.
43. Ma, J.; Liu, J.; Yu, H.; Chen, Y.; Wang, Q.; Xiang, L. Effect of metformin on Schwann cells under hypoxia condition. Int. J. Clin. Exp. Pathol. 2015, 8, 6748–6755.

44. Eyileten, C.; Kaplon-Cieslicka, A.; Mirowska-Guzel, D.; Malek, L.; Postula, M. Antidiabetic Effect of Brain-Derived Neurotrophic Factor and Its Association with Inflammation in Type 2 Diabetes Mellitus. J. Diabetes Res. 2017, 2017, 2823671.
45. Díaz, M.; López-Bermejo, A.; Sánchez-Infantes, D.; Bassols, J.; De Zegher, F.; Ibáñez, L. Responsiveness to metformin in girls with androgen excess: Collective influence of genetic polymorphisms. Fertil. Steril. 2011, 96, 208–213.
46. Willer, C.J.; Speliotes, E.K.; Loos, R.J.F.; Li, S.; Lindgren, C.M.; Heid, I.M.; Berndt, S.I.; Elliott, A.L.; Jackson, A.U.; Lamina, C.; et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat. Genet. 2009, 41, 25–34.
47. Thorleifsson, G.; Walters, G.B.; Gudbjartsson, D.F.; Steinthorsdottir, V.; Sulem, P.; Helgadottir, A.; Styrkarsdottir, U.; Gretarsdottir, S.; Thorlacius, S.; Jonsdottir, I.; et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat. Genet. 2009, 41, 18–24.
48. Gutierrez-Aguilar, R.; Kim, D.-H.; Casimir, M.; Dai, X.-Q.; Pfluger, P.T.; Park, J.; Haller, A.; Donelan, E.; Park, J.; D'Alessio, D.; et al. The role of the transcription factor ETV5 in insulin exocytosis. Diabetologia 2014, 57, 383–391.
49. Arner, P.; Sahlqvist, A.-S.; Sinha, I.; Xu, H.; Yao, X.; Waterworth, D.; Rajpal, D.; Loomis, A.K.; Freudenberg, J.M.; Johnson, T.; et al. The epigenetic signature of systemic insulin resistance in obese women. Diabetologia 2016, 59, 2393–2405.
50. Hur, K.Y.; Lee, M.-S. New mechanisms of metformin action: Focusing on mitochondria and the gut. J. Diabetes Investig. 2015, 6, 600–609.
51. Becker, M.; Visser, L.; van Schaik, R.; Hofman, A.; Uitterlinden, A.; Stricker, B. Genetic variation in the multidrug and toxin extrusion 1 transporter protein influences the glucose-lowering effect of metformin in patients with diabetes: A preliminary study. Diabetes 2009, 58, 745–749.
52. Becker, M.L.; Visser, L.E.; van Schaik, R.H.N.; Hofman, A.; Uitterlinden, A.G.; Stricker, B.H.C. Interaction between polymorphisms in the OCT1 and MATE1 transporter and metformin response. Pharmacogenet. Genom. 2010, 20, 38–44.
53. Franks, P.W.; Christophi, C.A.; Jablonski, K.A.; Billings, L.K.; Delahanty, L.M.; Horton, E.S.; Knowler, W.C.; Florez, J.C. Common variation at PPARGC1A/B and change in body composition and metabolic traits following preventive interventions: The Diabetes Prevention Program. Diabetologia 2014, 57, 485–490.
54. Becker, M.L.; Visser, L.E.; van Schaik, R.H.N.; Hofman, A.; Uitterlinden, A.G.; Stricker, B.H.C. Genetic variation in the organic cation transporter 1 is associated with metformin response in patients with diabetes mellitus. Pharm. J. 2009, 9, 242–247.
55. Lopez-Bermejo, A.; Diaz, M.; Moran, E.; de Zegher, F.; Ibanez, L. A Single Nucleotide Polymorphism in STK11 Influences Insulin Sensitivity and Metformin Efficacy in Hyperinsulinemic Girls with Androgen Excess. Diabetes Care 2010, 33, 1544–1548.
56. Wing, M.R.; Ziegler, J.M.; Langefeld, C.D.; Roh, B.H.; Palmer, N.D.; Mayer-Davis, E.J.; Rewers, M.J.; Haffner, S.M.; Wagenknecht, L.E.; Bowden, D.W. Analysis of FTO gene variants with obesity and glucose homeostasis measures in the multiethnic Insulin Resistance Atherosclerosis Study cohort. Int. J. Obes. 2011, 35, 1173–1182.
57. Franks, P.W.; Jablonski, K.A.; Delahanty, L.M.; McAteer, J.B.; Kahn, S.E.; Knowler, W.C.; Florez, J.C. Assessing gene-treatment interactions at the FTO and INSIG2 loci on obesity-related traits in the Diabetes Prevention Program. Diabetologia 2008, 51, 2214–2223.
58. Christensen, M.M.H.; Brasch-Andersen, C.; Green, H.; Nielsen, F.; Damkier, P.; Beck-Nielsen, H.; Brosen, K. The pharmacogenetics of metformin and its impact on plasma metformin steady-state levels and glycosylated hemoglobin A1c. Pharmacogenet. Genom. 2011, 21, 837–850.
59. Tkáč, I.; Klimčáková, L.; Javorský, M.; Fabianová, M.; Schroner, Z.; Hermanová, H.; Babjaková, E.; Tkáčová, R. Pharmacogenomic association between a variant in SLC47A1gene and therapeutic response to metformin in type 2 diabetes. Diabetes Obes. Metab. 2013, 15, 189–191.
60. Mousavi, S.; Kohan, L.; Yavarian, M.; Habib, A. Pharmacogenetic variation of SLC47A1 gene and metformin response in type2 diabetes patients. Mol. Biol. Res. Commun. 2017, 6, 91–94.

J Clin Med. 2020;9(6):1705. doi:10.3390/jcm9061705. IF: 4.241, Q1 at MEDICINE, GENERAL & INTERNAL.

Study 4

Evaluation of the Predictive Ability, Environmental Regulation and Pharmacogenetics Utility of a BMI-Predisposing Genetic Risk Score during Childhood and Puberty

Augusto Anguita-Ruiz [1,2,3], Esther M. González-Gil [1,3,4], Azahara I. Rupérez [4], Francisco Jesús Llorente-Cantarero [3,5,6], Belén Pastor-Villaescusa [1,7], Jesús Alcalá-Fdez [8], Luis A. Moreno [3,4], Ángel Gil [1,2,3], Mercedes Gil-Campos [3,6,9], Gloria Bueno [3,4,10], Rosaura Leis [3,11,*] and Concepción M. Aguilera [1,2,3]

Abstract Polygenetic risk scores (pGRSs) consisting of adult body mass index (BMI) genetic variants have been widely associated with obesity in children populations. The implication of such obesity pGRSs in the development of cardio-metabolic alterations during childhood as well as their utility for the clinical prediction of pubertal obesity outcomes has been barely investigated otherwise. In the present study, we evaluated the utility of an adult BMI predisposing pGRS for the prediction and pharmacological management of obesity in Spanish children, further investigating its implication in the

Affiliations 1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n. Armilla, 18016 Granada, Spain; augustoanguita@ugr.es (A.A.-R.); esthergg@ugr.es (E.M.G.-G.); Belen.Pastor@med.uni-muenchen.de (B.P.-V.); agil@ugr.es (Á.G.); caguiler@ugr.es (C.M.A.) / 2. Instituto de Investigación Biosanitaria ibs.GRANADA, 18014 Granada, Spain / 3. CIBEROBN (Physiopathology of Obesity and Nutrition Network CB12/03/30038), Institute of Health Carlos III (ISCIII), 28029 Madrid, Spain; llorentefj@yahoo.es (F.J.L.-C.); lmoreno@unizar.es (L.A.M.); mercedes_gil_campos@yahoo.es (M.G.-C.); gbuenoloz@yahoo.es (G.B.) / 4. Growth, Exercise, Nutrition and Development (GENUD) Research Group, Instituto Agroalimentario de Aragón (IA2), Universidad de Zaragoza, Instituto de Investigación Sanitaria de Aragón (IIS Aragón), 50009 Zaragoza, Spain; airuperez@unizar.es / 5. Department of Specific Didactics, Faculty of Education, University of Córdoba, 14004 Córdoba, Spain / 6. PAIDI CTS-329, Maimonides Institute of Biomedical Research of Córdoba (IMIBIC), 14004 Córdoba, Spain / 7. LMU—Ludwig-Maximilians-University of Munich, Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, 80337 Munich, Germany / 8. Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain; jalcala@decsai.ugr.es / 9. Unit of Pediatric Endocrinology, Reina Sofia University Hospital, 14004 Córdoba, Spain / 10. Pediatric Department, Lozano Blesa University Clinical Hospital, University of Zaragoza, 50009 Zaragoza, Spain / 11. Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department, Clinic University Hospital of Santiago, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain
* Corresponding authors

appearance of cardio-metabolic alterations. For that purpose, we counted on genetics data from three well-characterized children populations (composed of 574, 96 and 124 individuals), following both cross-sectional and longitudinal designs, expanding childhood and puberty. As a result, we demonstrated that the pGRS is strongly associated with childhood BMI Z-Score (B = 1.56, SE = 0.27 and p-value = 1.90×10^{-8}), and that could be used as a good predictor of obesity longitudinal trajectories during puberty. On the other hand, we showed that the pGRS is not associated with cardio-metabolic comorbidities in children and that certain environmental factors interact with the genetic predisposition to the disease. Finally, according to the results derived from a weight-reduction metformin intervention in children with obesity, we discarded the utility of the pGRS as a pharmacogenetics marker of metformin response.

Keywords: obesity; childhood obesity; metabolic syndrome; genetics; genetic risk score; pharmacogenetics; predictive ability; gene-environment interactions; puberty; childhood; Spanish children

1. Introduction

Among noncommunicable common diseases, overweight and obesity in children are a public health problem that has raised concern worldwide [1]. Characterized by an expansion of the adipose tissue, childhood obesity plays an important role in the development of cardio-metabolic alterations during adulthood, further increasing morbidity and mortality [2]. The early-life identification of high-risk individuals for severe obesity or cardio-metabolic alterations during adulthood is therefore indispensable to tackle down the obesity-associated mortality. A wide range of clinical and molecular factors have proven useful for obesity prediction. Among them, genetic markers are of special importance, since they allow a risk assessment from the moment of childbirth. This, combined with the strong modulatory effects of some environmental exposures, such as diet or physical activity (PA), would allow the design of personalized lifestyle plans that effectively prevent the appearance of severe obesity and cardio-metabolic alterations later in life.

Twin studies have proven a strong heritable component of body mass index (BMI), and genome-wide association studies (GWAS) have shown that adult BMI is influenced by hundreds of common genetic variants [3]. Evidence from cross-sectional and longitudinal studies has further indicated that some of these adult loci also affect BMI in childhood and puberty [4–8]. For many of these BMI-associated single-nucleotide polymorphisms (SNPs), significant pleiotropic genetic effects for adult cardio-metabolic traits have also been reported [9], and there is a strong evidence of a regulatory impact of environmental factors [10–12].

Although initial expectations for obesity GWASs were high, the results derived after two decades of research have not met the previsions (e.g., mentioned SNPs individually account for only small proportions (1–2%) of the BMI heritability [3]). Consequently, the practice of utilizing individual SNPs to predict disease is now considered a limited approach [13] and other innovative perspectives have emerged to take advantage of available GWAS insight [14]. For example, several genomic studies have proposed to study multiple common SNPs collectively to improve the estimation of disease predisposition [15]. Based on the construction of polygenic risk scores (pGRSs), that include multiple genetic variants at the same time, these approaches have recently gathered considerable interest [16] and have proven utility to identify groups of individuals who could benefit from the knowledge of their probabilistic susceptibility to disease. In brief, a pGRS is usually calculated as a weighted sum of the number of risk alleles carried by an individual, where the risk alleles and their weights are defined by the loci and their measured effects as detected by GWAS in a particular trait [17].

The inclusion of adult-BMI SNPs under a pGRS could serve therefore as an excellent predictive, and preventive, tool for facing the obesity-associated morbidity and mortality from the early periods of life. Although some previous studies have already investigated the utility of adult-BMI pGRSs for the management of obesity in children [18–22], no study to date has addressed the question focusing in cardio-metabolic alterations, and never under a longitudinal design comprising the metabolically risky period of puberty. In fact, puberty has been designated as the life stage where the majority of obesity-associated cardio-metabolic derangements arise [23]. The exact mechanisms connecting puberty and metabolic alterations in obesity remain unknown otherwise [24]. Furthermore, it would be interesting to elucidate to which extent BMI is due to heritable genetic factors and lifestyle behaviors; studying how the environment modulates the genetic susceptibility to disease during childhood. Beyond prognostic utility, some pGRSs have also proven to have pharmacological utility. For example, a coronary artery disease pGRS is not only able to stratify individuals by risk for disease but also by the potential clinical benefit of statin therapy [25]. However, unlike heart disease, pGRS pharmacogenetics evidences for obesity have not yet been investigated in neither children nor adults. Considering all this, we decided to evaluate the utility of an adult BMI pGRS for the prediction and pharmacological management of obesity in children, further investigating its implication in the appearance of cardio-metabolic alterations. The study design consisted of three well-characterized children populations following both cross-sectional and longitudinal approaches. For all these analyses, we employed a pGRS based on the top 44 SNPs that have previously been associated with adult BMI in the most comprehensible GWAS performed to date [3].

2. Objectives

1. To demonstrate how a pGRS can quantify inherited susceptibility to obesity and its cardio-metabolic comorbidities in children.
2. To evaluate the effects of genetic predisposition for obesity during childhood and how they evolve when entering puberty.
3. To describe the plausible modulatory role of environmental factors over inherited genetic susceptibility in children.
4. To investigate the utility of the pGRS for the pharmacological management of obesity in children.

3. Materials and Methods

3.1. Study Design

The present study design consisted of three independent children populations following cross-sectional and longitudinal approaches. A general description for each study population as well as each study design are presented in **Figure 1**.

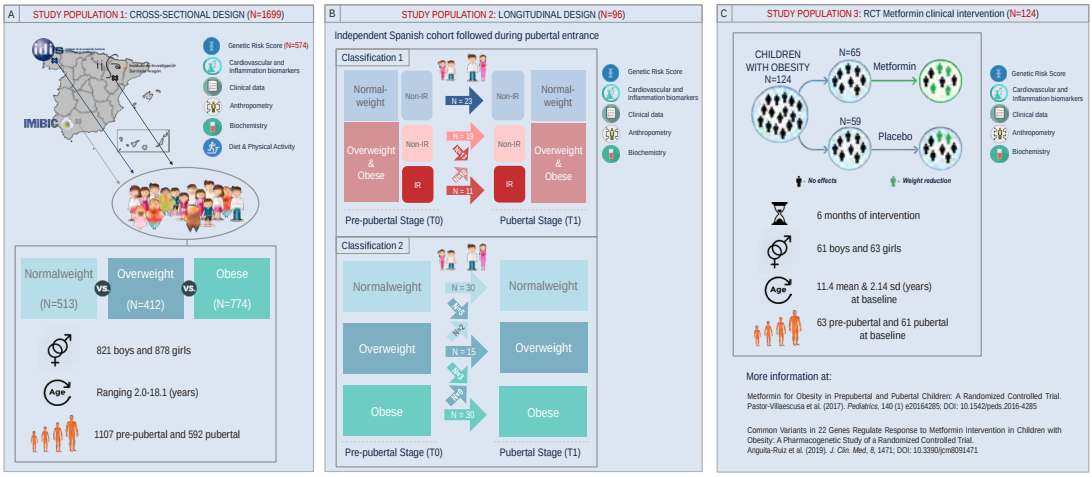


Figure 1. General description of study populations and design. (A) General characteristics of study population 1, which is based on a previously conducted case-control multicentre cross-sectional design. (B) General characteristics of study population 2, which is based on a previously conducted longitudinal design on 96 children undergoing puberty. (C) General characteristics of study population 3, which corresponds to a previous multicentre and double blind randomized controlled trial (RCT) conducted in 124 children with obesity.

3.1.1. Study Population 1: Cross-Sectional Approach

In order to demonstrate how the pGRS can quantify inherited susceptibility to obesity, and its cardio-metabolic comorbidities, we counted on a cross-sectional cohort of Spanish children. This cohort was referred to as study population 1 and is based on a previously conducted case-control multicentre cross-sectional design (**Figure 1A**) [26,27]. Among all available participants from the previous work (N = 1699), current genetic analyses were performed in a subset population of 574 children (293 girls) who had good quality DNA samples. Children were recruited at three Spanish health institutions: Lozano Blesa University Clinical Hospital in Zaragoza, Santiago de Compostela University Clinical Hospital in Santiago de Compostela and Reina Sofia University Clinical Hospital in Córdoba. Obesity status was defined according to BMI by using the age- and sex-specific cut-off points proposed by Cole *et al*. (2000) [28]. For the present analysis, there were 256 children in the obesity group, 131 in the overweight group and 187 in the normal weight group. Inclusion criteria were European-Caucasian heritage and the absence of congenital metabolic diseases. The exclusion criteria were non-European Caucasian heritage; the presence of congenital metabolic diseases (e.g., diabetes or hyperlipidaemia); undernutrition; and the use of medication that alters blood pressure, glucose or lipid metabolism. General characteristics of the 574 participants with genetics data are presented in the Supplementary Table S1.

3.1.2. Study Population 2: Longitudinal Approach

With the aim of studying the effects of the pGRS on BMI changes during the course of childhood and puberty, we also performed a longitudinal analysis using data from 96 boys and girls undergoing sexual maturation (**Figure 1B**) recruited in the PUBMEP project ("Puberty and metabolic risk in obese children. Epigenetic alterations and pathophysiological and diagnostic implications") [29]. Children were allocated into five experimental groups according to their obesity and insulin resistance (IR) status before and after the onset of puberty. Pubertal stage was evaluated by clinicians in all participants according to the Tanner scale (I for prepubertal and II-V for pubertal children) [30]. All details regarding the adopted longitudinal design are illustrated in **Figure 1B**. Obesity status was defined according to BMI by using the age- and sex-specific cut-off points proposed by Cole *et al.* (2000) [28]. On the other hand, the IR status was defined by means of the homeostatic model assessment for insulin resistance (HOMA-IR) index. Since HOMA-IR strongly varies with age, sex and diseases [31], and since no reference values have been yet established in neither children nor adult populations [31,32], cut-off points were extracted from a previous well-described Spanish cohort composed of 1669 children and adolescents [27,33]. For the prepubertal stage, a single cut-off value of HOMA-IR ≥ 2.5 was considered for IR [26,33]. For the pubertal stage instead, sex information was taken into consideration and different cut-off points were adopted for IR according to the 95th HOMA-IR percentile. Extracted from 778 pubertal Spanish children, pubertal IR cut-off values were HOMA-IR ≥ 3.38 in boys and HOMA-IR ≥ 3.90 in girls. Descriptive statistics for baseline data as well as longitudinal within-group and between-group changes in analyzed variables for the 96 participating children are presented in Supplementary Table S2.

3.1.3. Study Population 3: RCT Metformin Clinical Intervention

In order to test whether the constructed pGRS presents utility for the pharmacological management of obesity in children, a third obesity cohort was submitted to genetic analyses in the present work. This cohort corresponded to a previous multicentre and double blind randomized controlled trial (RCT) conducted in 124 children with obesity (**Figure 1C**). A complete workflow detailing the study design can be found elsewhere [34–36]. Briefly, 160 children with obesity were stratified according to sex and pubertal status and randomly assigned to receive either (1 g/d) metformin or placebo for 6 months after meeting the defined inclusion criteria [34,35]. All the details regarding informed consent, ethics, study protocol, sample size, intervention and participants (participant's data collection and processing, samples codification, randomization method, double-blind condition and adverse effects assessment) were previously described [34,35]. The original study was registered by European Clinical Trials Database (EudraCT, ID: 2010-023061-21) on 14 November 2011 (URL: https://www.clinicaltrialsregister.eu/ctr-search/trial/2010-023061-21/ES). Among the 160 subjects participating in the original RCT, 124 (59 placebo (29 boys) and 65 treated children (32 boys)) had an appropriate DNA sample quality for the present genetic analyses. General characteristics of the selected study population at baseline and post-treatment stages are summarized in the Supplementary Table S3. For the present pharmacogenetics analysis, differential drug response was assessed via BMI Z-score reduction after the intervention.

3.2. Ethics Statement

All described projects were conducted in accordance with the Declaration of Helsinki (Edinburgh 2000 revised) and followed the recommendations of the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legally enforced Spanish regulation, which regulates the clinical investigation of human beings (RD 223/04 about clinical trials). The Ethics Committee on Human Research of the University of Granada (ID code: 01/2017), the Ethics Committee of the Reina Sofía University Clinical Hospital of Cordoba (ID code: 260/3408), the Bioethics Committee of the University of Santiago de Compostela (ID codes: 2011/198 and 2016/522), the Ethics Committee in Clinical Research of Aragon (ID codes: 12/2010 and 22/2016) and the Ethics Committee for Biomedical Research of Andalusia on 15 January 2012 (acta 1/12) (ID code: 2010-2739) have approved all experiments and procedures. All parents or guardians provided written informed consent, and the children gave their assent.

3.3. DNA Extraction, Genotyping and pGRS Construction

In all participants of the present study genomic DNA was extracted from peripheral white blood cells using two kits, the Qiamp® DNA Investigator Kit for coagulated samples and the Qiamp® DNA Mini & Blood Mini Kit for noncoagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). All extractions were purified using a DNA Clean and Concentrator kit from Zymo Research (Zymo

Research, Irvine, CA, USA). Genotyping was performed by TaqMan allelic discrimination assay using the QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA).

A total of 56 previously BMI-associated SNPs from the largest and most comprehensive GWAS performed to date in obesity research [3] were considered for genotyping analyses. Among them, twelve SNPs were removed in our population either due to a call rate under 95% or to a deviation from Hardy Weinberg equilibrium (Supplementary Table S4). The remaining 44 SNPs were annotated and are listed in Supplementary Table S5. Raw fluorescence measures for these genetic variants were transformed into a dosage format, where each individual genotype was represented by the number of risk alleles. Next, regression coefficients (beta-estimates) of each SNP were obtained from the GIANT consortium meta-analysis for BMI (particularly from the European population with males and females combined) [3]. The weighted pGRS was finally calculated for each individual by multiplying the number of risk alleles carried for each SNP and the corresponding extracted beta-estimate (further calculating the sum over all SNPs) (1):

$$\mathrm{pGRS}_{\mathrm{BMI}} = \sum_{i=1}^{44} \beta_{\mathrm{SNP}_i} \times \mathrm{SNP}_i \qquad (1)$$

3.4. Phenotypic Measurements and Lifestyle Factors

In all described cohorts, body weight (kg), height (cm) and waist circumference (cm) were measured using standardized procedures, and the BMI Z-score was calculated based on the Spanish standards reference [37]. Blood pressure was measured three times by the same examiner. Biochemical marker analyses were performed for all study populations at participating hospital laboratories following internationally accepted quality control protocols, including routine measures for lipid and glucose metabolism. Quantitative insulin sensitivity check index (QUICKI) and HOMA-IR index were calculated using fasting plasma glucose and insulin values. High-sensitivity C-reactive protein (hsCRP) was determined using a particle-enhanced turbidimetric immunoassay (Dade Behring Inc., Deerfield, IL, USA). In study populations 1 and 3, adipokines, cardiovascular risk and proinflammatory biomarkers (i.e., adiponectin, leptin, resistin, tumour necrosis factor alpha (TNF-α), interleukin (IL)-6, IL-8, total plasminogen activator inhibitor-1 (PAI-1), myeloperoxidase (MPO), matrix metalloproteinase-9 (MMP-9), soluble intercellular cell adhesion molecule-1 sICAM-1 and soluble vascular cell adhesion molecule-1 (sVCAM)) were analyzed using a Luminex 200 system (Luminex Corporation, Austin, TX, USA) with human monoclonal antibodies from Millipore (EMD Millipore Corp, Billerica, MA, USA). Descriptive statistics for all measurements were conducted in each cohort separately and can be found in Supplementary Tables S1–S3.

For study population 1, environmental exposures were further assessed through an interview that focused on PA, sedentary behaviors, disease family history and familial educational level (at mother's and father's levels separately). Among all available environmental data, only quantitative or ordinal variables were selected for interaction analyses.

This resulted in 47 lifestyle questions described in the Supplementary Table S6. The interviews were carried out during the school time or when children attended the consulting room at the hospital, taking approximately 30 min. In the case of PA performance and sedentary habits, they were evaluated by means of a short test based on the Physical Activity Questionnaire for Older Children (PAQ-C) and HELENA questionnaire, respectively, as well as an individual interview.

3.5. Statistical Analysis

3.5.1. General Descriptive Analysis

All continuous variables were tested for normality using the Shapiro–Wilk test and transformed when necessary by means of the natural log or the rank-based inverse normal transformation. Heteroscedasticity between experimental groups was explored by means of the Levene test. One-way Anova, Kruskal-Wallis and the Welch test were employed to assess group differences in measurements according to standard statistical assumptions. Pairwise-t-tests, pairwise Mann–Whitney U-tests and Dunn tests were applied conveniently as post-hoc analyses to determine which experimental groups differ from each other. Values in descriptive tables are expressed as mean and standard deviation, or median and range if not normally distributed. In the descriptive statistics of the longitudinal cohort, within-group changes from baseline to puberty in all continuous measurements were assessed by means of a paired design; employing either a paired t-test or a Wilcoxon signed-rank test. Between-group differences were instead assessed by conveniently applying One-way Anova, Kruskal–Wallis or Welch tests to the computed delta values (T^1–T^0) for each continuous measurement.

3.5.2. Association between the pGRS and Obesity Outcomes and Evaluation of the pGRS Predictive Ability

In the study population 1, logistic regression models were applied in order to test whether higher genetic risk scores were observed for subjects with obesity than for normal weight controls. A logistic regression model was further applied for comparing obesity prevalence among participants presenting a high-risk genetic profile (Q2, Q3 or Q4) versus those belonging to the reference quartile (Q1). Multiple linear regressions were employed instead to investigate the relationship between continuous measurements (including BMI Z-score) and the pGRS. For these analyses, the pGRS was treated both as a continuous and discrete variable (quartiles). To determine which SNPs within the pGRS had an independent contribution in the association with BMI Z-Score, we further performed stepwise linear regression using the "step" function included in the *stats* R package. This function uses the Akaike information criterion (AIC) to select variables for a linear model. Adjusted R^2 and model deviance (D^2) were calculated to assess the amount of outcome variability explained by each model. In all analyses, age, gender, pubertal stage, origin, height and BMI Z-score where adjusted for as confounders when necessary. Linear regression models were evaluated by model control (investigating linearity of effects on outcome(s), consistency with a

normal distribution and variance homogeneity). All residuals- vs.-fitted, normal Q-Q, scale-location and residuals- *vs*.-leverage plots are available upon request. A *p*-value < 0.05 was considered as statistically significant. Given the number of analyzed markers, we also considered the false discovery rate (FDR) as in Benjamini and Hochberg to correct for multiple hypothesis testing in all analyses.

The ability of the pGRS to comprehensively discriminate between normal weight and subjects with obesity was quantified (alone or in combination with other traditional risk factors) using the area under the curve (AUC) of the receiver operating characteristic curve. This plot represents the true positive rate (sensitivity) versus the false-positive rate (specificity) and is equivalent to the overall probability that the predicted risk of an individual with disease is higher than the predicted risk of an individual without disease [38,39]. Models were first constructed based on each risk factor alone and then all models reaching an AUC ≥ 0.6 were combined. For that purpose, all samples from the study population 1 with valid data for each factor were included (not restricted to the 574 children with genetics information). Only subjects with normal weight and obesity were then selected and randomly assigned to training and test subsets in which predictive models were trained and evaluated respectively. All predictive assessments were conducted using the *PredictABEL* and the *pROC* R packages.

In order to study the ability of the pGRS as a predictor of future obesity outcomes following puberty entrance (study population 2), we performed logistic regression models with the dichotomized pGRS as an independent predictor variable (1st and 2nd tertiles vs. 3rd tertil), including the longitudinal experimental group classifications from **Figure 1B** as dependent dummy variables (each category vs. the reference normal weight group). Tertiles, instead of quartiles, were used here due to the low sample size of the cohort. Moreover, these models included a range of confounding factors as independent variables as indicated in the Results section. In study population 2, we further applied multiple linear regressions with deltas for continuous cardio-metabolic measurements as input variables (computed as T_1–T_0).

All statistical analyses were performed in R environment, version 3.6.0 (R Project for Statistical Computing).

3.5.3. Identification of Gene × Environment Modulatory Effects

In the study population 1, and in each pubertal stage of the study population 2 separately, linear regression models were used to estimate the effect of gene-environment interactions (pGRS × *E*) for each collected lifestyle factor (*E*) individually. In addition to the pGRS × *E* interaction term, each tested model also included covariates such as origin and puberty, in accordance with previously published recommendations [40] (2):

$$\text{BMI } Z - Score = \beta_0 + \beta_1 \text{pGRS} + \beta_2 E + \beta_3 (\text{pGRS} \times E) + \beta_4 Origin + \beta_5 Tanner + \varepsilon \quad (2)$$

For assessing statistical significance, we focused our attention on the estimate β_3 (pGRS x *E*) (2) and, more specifically, whether this estimate significantly deviated from zero. The null hypothesis $H_0 = 0$ was either accepted or rejected, depending on the outcome of a two-sided marginal student's *t*-test, which in this case (i.e., one degree-of-freedom difference between the nested models and normal regularity conditions) is equivalent to a likelihood-ratio test of the hypothesis $H_0 = 0$. p-values lower than the significance level $\alpha = 0.05$ were considered as statistically significant after accounting for the family-wise error rate using the FDR method. Calculations were performed in R environment, version 3.6.0 (R Project for Statistical Computing) using the "lm" function included in the stats package.

3.5.4. pGRS-Drug Interaction

Pharmacogenetics analyses of metformin response were performed using two parallel approaches in the study population 3. First, we applied a multiple linear regression to test the effect of the pGRS on BMI Z-Score responses. For that purpose, delta changes of BMI Z-Score (computed as T_0–T_1) were calculated and used as the dependent variable. On the other hand, we applied a linear mixed-effects (LME) model adjusted for confounders such as tanner stage and time as fixed effects and a random intercept for each patient. Test significance for the LME model was evaluated on the pGRS:Time:Group interaction term.

4. Results

4.1. The pGRS Associates with BMI Z-Score and Performs Well in the Identification of High-Risk Children

In order to investigate the general relationship between the pGRS and obesity, we merged the anthropometric baseline data available in the cross-sectional study population 1 (N = 574) and the metformin-RCT study population 3 (N = 124) (**Figure 1A,C**). Descriptive statistics for each study population can be found in Supplementary Tables S1 and S3. In the resulting population (N = 698), a model adjusted by puberty and origin showed a strongly significant association between the pGRS and the BMI Z-score (B = 1.56, SE = 0.27, t value = 5.69 and *p*-value = 1.90×10^{-8}) (**Figure 2**). This association was quantified with an increase of 1.56 Kg of weight by each additional 0.1 of the pGRS (B = 15.6, SE = 3.93, *t* value = 5.69 and *p*-value = 8.06×10^{-5}). The amount of BMI Z-score variance explained by the full model was 14.12%, being 4.5 the percentage of variance explained by the genetic component alone. Supplementary Figure S1 represents the density distribution plot of the constructed pGRS by experimental condition and the Supplementary Figure S2 the observed obesity: overweight: normal weight counts within each quartile of the pGRS. The pGRS followed a normal distribution in the whole study sample (D = 0.03; *p*-value = 0.08 in Lilliefors test). The mean (SD) of the pGRS in the whole sample was 1.18 (0.13), being 1.15 (0.12) in normal weight children, 1.18 (0.13) in overweight children and 1.21 (0.14) in children with obesity. After excluding

overweight subjects, a logistic regression model adjusted for puberty and origin revealed a stronger risk association between the pGRS and the obesity status, so that the odds of having obesity were estimated to increase 4.7 times for each additional tenth in the pGRS (h^2 = 5.6%, OR = 47.36; CI 95% = [9.8,229.38]; p-value = 1.64×10^{-6}). The obesity variability attributable to the genetic component under this model was estimated in 5.6%. When comparing individuals presenting the highest risk scores (Q4 and Q3) to those belonging to the first quartile (Q1), significant associations were also evidenced (OR = 3.33; CI 95% = [1.96, 5.67]; *p*-value = 9.14×10^{-6} and OR = 1.66; CI 95% = [1.02, 2.71]; *p*-value = 0.04 respectively) (Supplementary Figure S2).

Next, we aimed to know which SNPs contribute the most to the pGRS-BMI association. For that purpose, we performed a stepwise linear regression including all 44 tested SNPs and found the genetic variants rs543874-*LOC101928778:SEC16B*, rs7138803-*BCDIN3D:FAIM2*, rs10132280-*STXBP6:NOVA1*, rs1558902-*FTO* and rs12940622-*RPTOR* to be the most determinant polymorphisms for BMI Z-Score (Supplementary Figure S3). This finding was further supported by the univariate association analyses conducted between the BMI Z-score and each individual SNP (Supplementary Table S7).

To demonstrate the validity of the pGRS for the clinical prediction of obesity (alone or in combination with other traditional risk factors), logistic regression models were constructed including different combination of risk factors and further evaluated using AUC (**Table 1**). For each

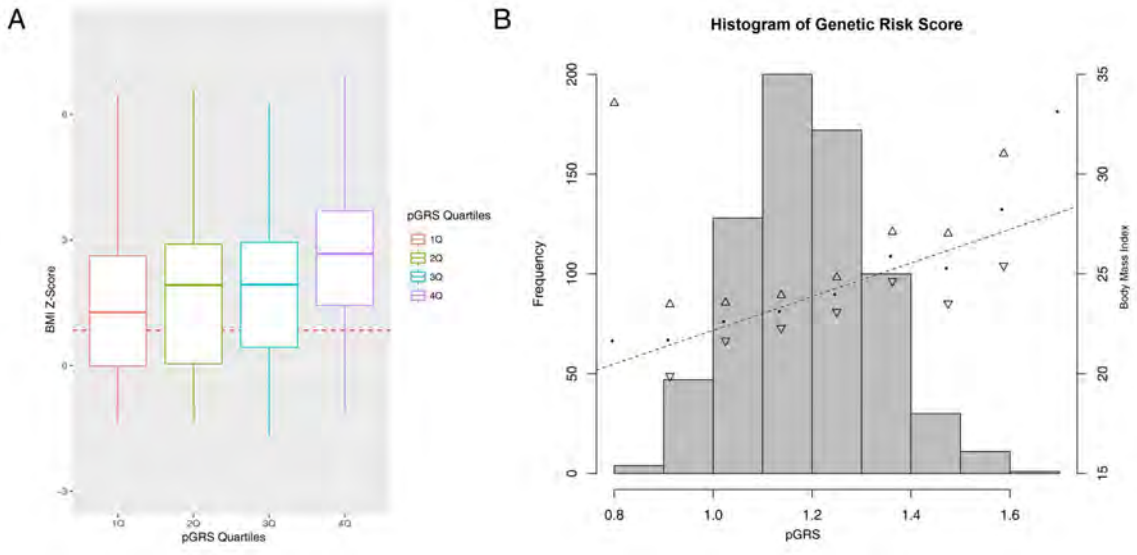


Figure 2. Association between polygenetic risk scores (pGRS) and body mass index (BMI) Z-score in the study population 1; analysis adjusted for origin and pubertal status of children. (A) Boxplot graph for BMI Z-Score according to each quartile of the pGRS; the dashed line in the plot represents the cut-off BMI Z-Score for overweight and obesity in the study population 1. (B) Histogram of genetic risk score values in the study population 1 and their correlation with BMI (R^2 = 0.2).

Table 1. Obesity predictive ability for assessed traditional and genetics risk factors in the study population 1.

		Whole Population			Training Set			Test Set		
Predictors	**AUC [95% CI]**	**n**	**n Normal-Weight**	**n Obese**	**n**	**n Normal-Weight**	**n Obese**	**n**	**n Normal-Weight**	**n Obese**
Tanner, Origin, Sex and Age	0.66 [0.61–0.72]	1285	512	773	901	359	542	384	153	231
pGRS	0.72 [0.63–0.80]	443	187	256	311	131	180	132	56	76
Obesity Family History	0.70 [0.63–0.77]	686	232	454	481	163	318	205	69	136
Maternal Smoking	0.49 [0.43–0.55]	632	218	414	443	153	290	189	65	124
Gestational Diabetes	0.49 [0.45–0.54]	620	214	406	435	150	285	185	64	121
Birthweight	0.60 [0.51–0.69]	610	211	399	428	148	280	182	63	119
Gestational Weight Gain	0.54 [0.45–0.62]	569	206	363	400	145	255	169	61	108
Parents BMI	0.76 [0.68–0.84]	530	199	331	372	140	232	158	59	99
Type of Breastfeeding	0.55 [0.45–0.64]	555	193	362	390	136	254	165	57	108
Tanner, Origin, Sex, Age, pGRS, Obesity Family History, Birthweight, and Parents BMI	0.81 [0.7–0.93]	176	78	98	124	55	69	52	23	29

Models were first constructed based on each risk factor alone and then combined according to the improvements in area under the curve (AUC). For this purpose, subjects with normal weight and with obesity were selected from the study population 1 and randomly assigned to training and test subsets in whichpredictive models were trained and evaluated, respectively. Individual models showing an AUC ≥ 0.60 were combined into the full model presented in the last row. Abbreviations: AUC, area under the curve; BMI, body mass index; CI, confidence interval; n; number of effective individuals for analysis; pGRS, polygenetic risk score.

predictive model, subjects presenting valid data for assayed variables were selected from the study population 1 and divided into a training set (composed of the 75% of total available samples) and a test set (formed by the remaining 25%). Performance statistics from each trained model in the respective test set are presented in Table 1. Among all single-variable predictive models, the model including the pGRS demonstrated one of the greater predictive abilities (only surpassed by the model including parental BMI information). The joint combination of all models individually surpassing an AUC of 0.6 yielded a considerable improvement in the predictive ability (AUC = 0.81 CI 95%= [0.7–0.93]), which could be sufficient for clinical discrimination.

4.2. The pGRS is Associated with Longitudinal Trajectories for Obesity and IR in Children Undergoing Puberty

With the aim of studying the effects of the pGRS on obesity during the course of puberty, we also performed a longitudinal design on 96 boys and girls undergoing sexual maturation (study population 2). All details regarding the adopted longitudinal design are illustrated in **Figure 1B.** The 96 individuals were stratified according to two classification criteria; (1) joint longitudinal trajectories for obesity and IR, and (2) the longitudinal trajectories for obesity. The number of resulting experimental groups per classification as well as the final sample size per group are shown in Figure 1B. Longitudinal within-group and between-group changes for all analyzed biochemical variables are shown in Supplementary Table S2, according to the experimental classification 1. Changes in anthropometric variables showed a coherent behavior according to each experimental condition. In particular, for waist circumference (WC), which is a metabolic health indicator in obesity, we found significant within-group increases accompanying sexual maturation in all groups. The higher increase corresponded to group 4, in which children with obesity become IR with pubertal maturation. The metabolic health derangement observed in groups 4 and 5 for WC was also confirmed by changes in blood pressure, insulin and glucose levels, QUICKI, HOMA-IR and triglycerides.

Regarding the pGRS, findings reported in the Results Section 4.1 (merged study populations 1 and 3) were independently validated here with the longitudinal study population 2 (N = 96), using data from each time point individually. For the prepubertal stage, a multiple linear regression analysis revealed a significant association between the pGRS and the BMI Z-Score after adjusting by origin (B = 2.84, CI 95% = [0.31, 5.37]; *p*-value = 0.03). When excluding overweight individuals from analysis, the odds of obesity were quantified as 8.22 times higher in the children belonging to the 3rd tertile of the pGRS with regard to children belonging to the 1st and 2nd tertiles (CI 95% = [1.95, 34.61]; *p*-value = 0.004). For the pubertal stage, the multiple linear regression model did not find a significant association between the continuous pGRS and the BMI Z-Score after adjusting by origin and pubertal status (B = 0.9, CI 95% = [−1.57, 3.38]; *p*-value = 0.47). Instead, when excluding overweight individuals, the odds of obesity were estimated to be 5.54 times significantly higher in

pubertal children belonging to the third tertile of the pGRS in comparison to those belonging to the first two tertiles (CI 95% = [1.41, 21.52]; *p*-value = 0.01).

In order to study the ability of the pGRS to predict future outcomes after puberty entrance, we next performed logistic regression models with the dichotomized pGRS as an independent predictor variable (1st and 2nd tertiles vs. 3rd tertil), including the longitudinal experimental group classifications from Figure 1B as dependent dummy variables (each category vs. the reference normal weight group). These models also included the tanner stage and origin of children as confounding factors. When modelling the experimental groups based on obesity and IR outcomes together (classification 1), we found higher odds of being "obese or overweight with IR that remain IR after puberty entrance" in children within the 3rd tertil of the pGRS when comparing them to the children in the reference bottom pGRS group (1st and 2nd tertiles) (OR = 54.15, *p*-value = 0.008, FDR = 0.03). Higher odds of being "obese or overweight non-IR that become IR after puberty entrance" were also reported among 3rd tertile children in comparison to 1st and 2nd tertiles children though without statistical significance (OR = 15.52, p-value = 0.05, FDR = 0.12). Nonsignificant results were obtained for the rest of the comparisons performed. Figure 3A represents the boxplots for the continuous pGRS in each of the mentioned experimental groups. On the other hand, when modelling the experimental groups that consider only longitudinal trajectories for obesity (classification 2), we reported higher odds of being "obese remaining obese after puberty entrance" (OR = 31.91, *p*-value = 0.0009, FDR = 0.005), and "normal weight becoming

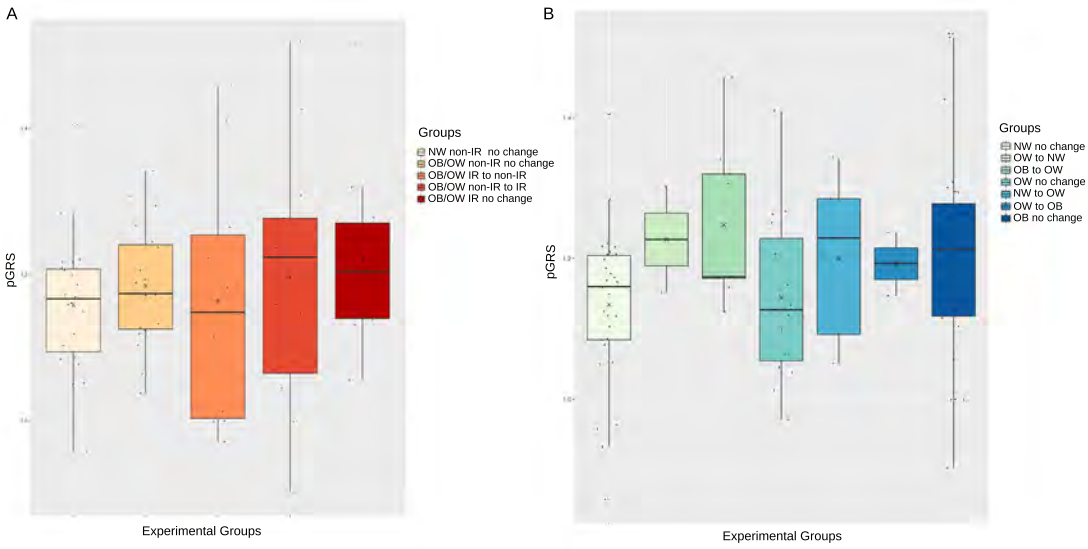


Figure 3. Boxplots for the continuous pGRS according to the two available experimental group classifications of study population 2. (A) pGRS boxplots according to joint longitudinal trajectories for obesity and insulin resistance (IR). (B) pGRS boxplots according to the longitudinal trajectories for obesity. The x symbol in plots represents the mean pGRS for each group. Abbreviations: NW, normal weight; OB, obese; OW, overweight; IR, insulin resistance.

overweight after puberty entrance" (OR = 26.31, *p*-value = 0.02, FDR = 0.07) among children in the 3rd tertile of the pGRS when comparing them to children in the reference bottom pGRS group (1st and 2nd tertiles). Nonsignificant results were obtained for the rest of the comparisons performed. Figure 3B represents the boxplots for the continuous pGRS in each mentioned experimental group.

4.3. The pGRS Does not Correlate with Increased Cardio-Metabolic Alterations in Children and Adolescents

In our cross-sectional cohort study population 1, we studied if the quartilized pGRS was associated with a metabolically unhealthy status as well as with any of its six dichotomized components (high glucose, HOMA-IR, DBP, SBP or triglycerides values or low HDLc levels) according to the criteria we have previously published [33]. In parallel, 30 continuous biochemical markers were tested for potential association with the pGRS. These biomarkers included lipid and glucose metabolism biomarkers, adipokines, as well as cardiovascular risk and proinflammatory biomarkers. From the analyses on the components of metabolic syndrome, models adjusted for BMI Z-Score, sex, age, puberty and origin showed no statistically significant association with pGRS (Supplementary Table S8). Instead, from the analyses on the 30 continuous biochemical outcomes, we found only one significant risk association for the APO B/LDLc ratio (Table 2) (that became nonsignificant after correction for multiple-hypothesis testing).

In order to validate these findings at the longitudinal level, we further applied multiple linear regressions with deltas for continuous cardio-metabolic measurements as input variables (computed as T_1–T_0) in the longitudinal study population 2. All analyses were again adjusted for confounders such as the change in BMI Z-score, sex, elapsed time, age at baseline or origin of the children. We found a significant positive correlation between the pGRS and APO B levels (*p*-value = 0.02, FDR = 0.29) during the course of puberty (Table 3). Moreover, a significant inverse correlation was also reported between the pGRS and the change in HDLc levels (*p*-value = 0.03, FDR = 0.33). Again, no model passed the multiple-hypothesis testing correction.

Table 2. Association between the pGRS and metabolic outcomes in the cross-sectional cohort of 574 children (study population 1), in decreasing order of statistical significance.

Measurement	Beta	SE	CI.LO	CI.HI	T-Value	*p*-Value	FDR
APO B/LDLc Ratio	−0.12	0.05	−0.21	−0.02	−2.29	0.02	0.60
Triglycerides (mg/dL)	−17.69	9.17	−35.67	0.28	−1.93	0.05	0.68
APO B (mg/dL)	−10.94	6.49	−23.66	1.78	−1.69	0.09	0.68
APO A/APO B	0.40	0.25	−0.09	0.88	1.61	0.11	0.68
WC/Height Ratio	0.03	0.02	−0.01	0.06	1.51	0.13	0.68
WC (cm)	3.75	2.52	-1.19	8.70	1.49	0.14	0.68
Adiponectin/Leptin Ratio	0.50	0.36	-0.21	1.22	1.38	0.17	0.68
MCP1 (ng/L)	−21.94	16.22	−53.73	9.84	−1.35	0.18	0.68
aPAI (ug/L)	−3.56	2.90	−9.24	2.12	−1.23	0.22	0.69
IL8 (ng/L)	−0.60	0.54	−1.66	0.45	−1.12	0.26	0.69
QUICKI	0.01	0.01	−0.01	0.03	0.96	0.34	0.69
DBP (mmHg)	3.11	3.35	−3.46	9.68	0.93	0.35	0.69
Adiponectin (mg/L)	−3.09	3.34	−9.63	3.45	−0.93	0.35	0.69
IL6 (ng/L)	2.32	2.57	−2.71	7.35	0.90	0.37	0.69
HC (cm)	1.96	2.19	−2.32	6.25	0.90	0.37	0.69
HOMA-IR index	−0.37	0.42	−1.19	0.46	−0.87	0.38	0.69
WC/HC Ratio	0.02	0.03	−0.03	0.08	0.85	0.39	0.69
Total cholesterol (mg/dL)	−5.95	9.48	−24.53	12.63	−0.63	0.53	0.88
hsCRP (mg/L)	0.42	0.81	−1.16	2.01	0.52	0.60	0.89
HDLc/LDLc Ratio	−0.05	0.09	−0.22	0.13	−0.49	0.62	0.89
Glucose (mg/dL)	−1.14	2.38	−5.81	3.53	−0.48	0.63	0.89
Resistin (ug/L)	1.28	2.88	−4.37	6.93	0.44	0.66	0.89
SBP (mmHg)	−1.74	4.17	−9.92	6.44	−0.42	0.68	0.89
LDLc (mg/dL)	−2.43	8.57	−19.21	14.36	−0.28	0.78	0.89
TNF (ng/L)	0.15	0.55	−0.93	1.24	0.28	0.78	0.89
Leptin (ug/L)	0.77	2.86	−4.84	6.37	0.27	0.79	0.89
APO A (mg/dL)	−2.35	9.74	−21.44	16.75	−0.24	0.81	0.89
MPO (ug/L)	−2.17	10.09	−21.95	17.60	−0.22	0.83	0.89
HDLc (mg/dL)	−0.50	4.22	−8.76	7.77	−0.12	0.91	0.94
MMP9 (ug/L)	−0.05	20.57	−40.38	40.27	0.00	1.00	1.00

Multiple linear regression analyses with the pGRS as independent variable were run adjusted for sex, BMI Z-Score, origin and puberty. When the dependent variable was the blood pressure, we further added the height as a confounder in the model. Abbreviations: APO, apolipoprotein; CI.HI, high confidence interval; CI.LO, low confidence interval; DBP, diastolic blood pressure; FDR, false discovery rate; HC, hip circumference; HDLc, high-density lipoproteins-cholesterol; HOMA-IR, homeostasis model assessment for insulin resistance; hsCRP, high-sensitivity C reactive protein; IL, interleukin; LDLc, low-density lipoproteins-cholesterol; MCP1, monocyte chemoattractant protein 1; MMP9, Matrix metallopeptidase 9; MPO, myeloperoxidase; PAI-1, plasminogen activator inhibitor-1; QUICKI, quantitative insulin sensitivity check index; SBP, systolic blood pressure; SE, standard error; TNF-α, tumour necrosis factor alpha; WC, waist circumference.

Table 3. Association between the pGRS and deltas for continuous cardio-metabolic measurements (computed as T_1–T_0) in the longitudinal study population 2.

Measurement (Delta T_1–T_0)	Beta	SE	CI.LO	CI.HI	T-Value	*p*-Value	FDR
APO B (mg/dL)	57.44	21.63	15.05	99.82	2.66	0.02	0.29
HDLc (mg/dL)	−18.91	8.98	−36.51	−1.32	−2.11	0.03	0.33
Triglycerides (mg/dL)	45.38	31.58	−16.51	107.27	1.44	0.15	0.78
APO A (mg/dL)	−35.22	30.97	−95.92	25.48	−1.14	0.26	0.78
DBP (mmHg)	−13.34	12.98	−38.79	12.10	−1.03	0.31	0.78
Insulin (mU/L)	8.94	8.89	−8.49	26.37	1.00	0.32	0.78
SBP (mmHg)	15.57	16.77	−17.29	48.43	0.93	0.36	0.78
HOMA-IR index	1.88	2.07	−2.17	5.94	0.91	0.37	0.78
HDLc/LDLc Ratio	−0.29	0.43	−1.14	0.56	−0.67	0.51	0.86
Total cholesterol (mg/dL)	−13.63	24.38	−61.42	34.16	−0.56	0.58	0.86
Glucose (mg/dL)	−3.80	8.72	−20.89	13.29	−0.44	0.66	0.86
WC/HC	−0.03	0.08	−0.19	0.12	−0.41	0.68	0.86
WC (cm)	−3.11	9.17	−21.09	14.87	−0.34	0.74	0.86
LDLc (mg/dL)	−5.94	18.41	−42.02	30.15	−0.32	0.75	0.86
QUICKI	0.01	0.04	−0.06	0.09	0.31	0.75	0.86

Multiple linear regression analyses with the pGRS as independent variable were run adjusted for sex, the change in BMI Z-Score, the origin, the elapsed time from baseline to puberty as well as the pubertal stage reached. When the dependent variable was the change in blood pressure, we further added the change in height as a confounder in the model. Abbreviations: APO, apolipoprotein; CI.HI, high confidence interval; CI.LO, low confidence interval; DBP, diastolic blood pressure; FDR, false discovery rate; HC, hip circumference; HDLc, high-density lipoproteins-cholesterol; HOMA-IR, homeostasis model assessment for insulin resistance; hsCRP, high-sensitivity C reactive protein; LDLc, low-density lipoproteins-cholesterol; QUICKI, quantitative insulin sensitivity check index; SBP, systolic blood pressure; SE, standard error; WC, waist circumference.

4.4. Lifestyle Factors Interact with the Inherited Genetic Susceptibility to Obesity in Children

Once we demonstrated the relationship between the pGRS and obesity as well as discarded a direct implication of the pGRS in the development of cardio-metabolic alterations, we next aimed to describe the plausible modulatory role of environmental factors over inherited genetic susceptibility to obesity. For that purpose, we applied multiple linear regression models including an interaction term for the pGRS and each assessed environmental factor in the study population 1. As a result, this approach revealed the pGRS to interact with three lifestyle factors related to parental educational level and physical activity (Table 4 and Figure 4). When we applied FDR adjustment for multiple testing, only two of them remained statistically significant. Interestingly, higher educational level of mothers and fathers were separately associated with lower BMI Z-Score of children depending on the pGRS (*p*-value = 0.0004 and FDR = 0.02 and *p*-value = 0.0008 and FDR = 0.02 respectively). The "protective" effect of mother's and father's educational levels on BMI was only achieved in children presenting low values of the pGRS (Figure 4A,B).

Table 4. Interaction analyses between the pGRS and each assessed environmental factor in the study population 1.

Lifestyle Factor	Beta	SE	CI.LO	CI.HI	T-Value	*p*-Value	FDR
Educational level of the mother	3.41	0.95	1.56	5.27	3.60	3.86×10^{-4}	0.02
Educational level of the father	2.93	0.86	1.24	4.62	3.40	7.92×10^{-4}	0.02
How many minutes per week do you spend exercising at a sport program?	0.02	0.01	0.00	0.04	2.13	0.03	0.47
Presence of AH in father or mother	−4.01	2.01	−7.95	−0.08	−2.00	0.05	0.56
Mother BMI	−0.22	0.11	−0.44	0.00	−1.93	0.06	0.56
How long does it take to get to the school on walk?	−0.24	0.14	−0.52	0.04	−1.66	0.10	0.58
How often do you eat fruit while watching TV?	1.73	1.05	−0.34	3.80	1.64	0.10	0.58
How often do you eat snacks while watching TV?	1.80	1.14	−0.44	4.03	1.58	0.12	0.58
How many hours do you spend doing home activities?	−2.55	1.62	−5.73	0.63	−1.57	0.12	0.58
How much time do you play videogames in a day during weekend?	1.03	0.70	−0.35	2.41	1.46	0.15	0.58
How many hours each day do you spend doing vigorous efforts like training activity?	2.27	1.55	−0.77	5.32	1.46	0.15	0.58
Presence of hypercholesterolemia in father or mother	1.78	1.24	−0.64	4.21	1.44	0.15	0.58
Presence of heart stroke in father or mother	−9.74	6.86	−23.19	3.70	−1.42	0.16	0.58
Presence of vascular problems in father or mother	−39.77	29.84	−98.26	18.72	−1.33	0.18	0.59
How many hours do you spend exercising in a sport club?	0.04	0.03	−0.02	0.10	1.30	0.19	0.59
How often do you eat salted potatoes while watching TV?	2.11	1.65	−1.12	5.34	1.28	0.20	0.59
Presence of diabetes in father or mother	11.59	9.96	−7.94	31.12	1.16	0.25	0.64
How often do you eat nuts while watching TV?	2.21	1.95	−1.62	6.04	1.13	0.26	0.64
How many days per week do you spend walking with vigorous efforts?	0.63	0.56	−0.47	1.73	1.13	0.26	0.64
How many hours each day do you spend practicing activities that do not require physical activity (e.g., reading)	0.41	0.44	−0.44	1.27	0.95	0.34	0.77
How much time do you play videogames in a day during the week?	−0.79	0.85	−2.46	0.87	−0.93	0.35	0.77
How many hours do you usually sleep every day during the week?	0.68	0.80	−0.89	2.26	0.85	0.40	0.77
How many hours do you spend doing physical activity in family?	1.21	1.42	−1.58	4.00	0.85	0.40	0.77
How often do you eat candies while watching TV?	1.50	1.77	−1.96	4.96	0.85	0.40	0.77
How many hours do you usually sleep every day during the weekends?	−0.55	0.67	−1.85	0.76	−0.82	0.41	0.77
Diagnosed hypertriglyceridemia in father or mother	1.06	1.34	−1.57	3.69	0.79	0.43	0.77
How often do you eat sweets while watching TV?	1.37	1.78	−2.12	4.86	0.77	0.44	0.77
How much time per weekend do you usually use the smartphone?	0.69	0.95	−1.17	2.55	0.73	0.47	0.79
Do you usually eat in front of the TV?	0.77	1.15	−1.49	3.02	0.67	0.51	0.83
How many hours per week do you spend on physical education during school hours?	−2.01	3.48	−8.83	4.82	−0.58	0.57	0.85
How often do you eat fruits while playing video games?	−2.07	3.73	−9.38	5.23	−0.56	0.58	0.85
Presence of obesity in the father or mother	−0.45	0.82	−2.07	1.17	−0.54	0.59	0.85
How much time during the week do you usually watch TV?	0.39	0.75	−1.07	1.85	0.52	0.60	0.85
How often do you eat fruits while surfing internet?	1.81	3.88	−5.80	9.43	0.47	0.64	0.88
How much time in per weekend do you usually use internet	0.25	0.63	−0.98	1.48	0.40	0.69	0.92
How many hours a day do you spend walking with vigorous efforts?	−0.35	1.00	−2.32	1.62	−0.35	0.73	0.92

Table 4. *Cont.*

Lifestyle Factor	Beta	SE	CI.LO	CI.HI	T-Value	*p*-Value	FDR
How often do you eat snacks while surfing the internet?	1.26	3.77	−6.13	8.65	0.33	0.74	0.92
How many days per week do you spend doing home activities?	0.20	0.64	−1.05	1.44	0.31	0.75	0.92
How many days per week do you spend doing physical activity in family?	−0.26	0.91	−2.05	1.53	−0.29	0.78	0.92
How many hours each day you spend walking quite a lot without vigorous efforts?	0.17	0.63	−1.07	1.40	0.26	0.79	0.92
Father BMI	−0.04	0.17	−0.38	0.30	−0.25	0.80	0.92
How often do you eat snacks while playing videogames?	−0.73	3.99	−8.56	7.09	−0.18	0.85	0.94
How much time during the weekend do you spend watching TV and DVD?	0.14	0.79	−1.41	1.70	0.18	0.86	0.94
How many days per week do you exercise in a sport club?	0.07	0.69	−1.28	1.43	0.11	0.92	0.98
How many hours do you spend doing homework outside of school hours?	0.05	0.99	−1.88	1.99	0.05	0.96	0.99
How much time per day do you use the internet during the week?	−0.02	0.71	−1.41	1.37	−0.03	0.98	0.99
How many days do you spend doing vigorous efforts like training activity?	−0.01	0.62	−1.22	1.19	−0.02	0.99	0.99

These analyses were adjusted for origin and pubertal status of children. Abbreviations: AH, arterial hypertension; BMI, body mass index; CI.HI, high confidence interval; CI.LO, low confidence interval; FDR, false discovery rate; SE, standard error.

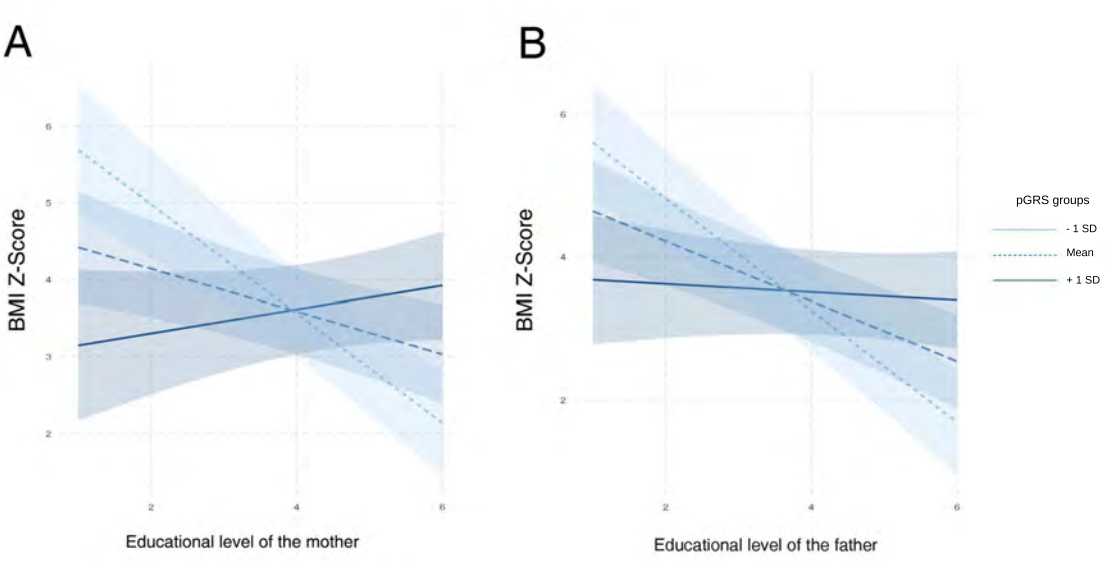


Figure 4. Interaction plots for observed significant modulatory effects of environment over inherited genetic susceptibility to obesity (pGRS-environment interactions) in study population 1; these analyses were adjusted for the origin and pubertal status of children. Subfigure (A) shows the modulatory effect of the educational level of mothers over the pGRS-BMI Z-Score association. Subfigure (B) shows the modulatory effect of the educational level of fathers over the pGRS-BMI Z-Score association. In both subfigures, the pGRS is categorized according to the cut off values -1 standard deviation and + 1 standard deviation. Abbreviations: BMI, body mass index; SD, standard deviation.

4.5. The pGRS is not helpful for the Pharmacogenetics Management of Obesity in Children

On the other hand, we employed the data derived from a previous metformin RCT (study population 3) in order to test whether the constructed pGRS presents utility for the pharmacological management of obesity in children. As a result, we found no differential response (in terms of BMI Z-Score reduction) according to the pGRS (B = 0.39, SE = 0.41, *t* value = 0.97, *p*-value = 0.34 for the interaction term GRS*Treatment in the multiple linear regression, and *p*-value = 0.33 in for the interaction term GRS:Time:Experimental Group under a LME model).

5. Discussion

In the present study, we evaluated the utility of an adult-BMI pGRS for the prediction and pharmacological management of obesity in children, further investigating its implication in the appearance of cardio-metabolic alterations. For that purpose, we counted on data from three well-characterized children populations following both cross-sectional and longitudinal designs. As a result, we demonstrated that the pGRS is associated with childhood BMI Z-Score and could be used as a good predictor of obesity longitudinal trajectories during puberty. On the other hand, we demonstrated that the pGRS is not associated with cardio-metabolic comorbidities in children and

that certain environmental factors interact with the genetic predisposition to the disease. Finally, according to the results derived from a weight-reduction metformin intervention in children with obesity, we discarded the utility of the pGRS as a pharmacogenetics marker of metformin response.

As one of the main findings from this work, it highlights the strongly significant association evidenced between the pGRS and the BMI Z-score in a children population composed of 698 pre- and pubertal subjects with and without obesity (Figure 1). When excluding overweight individuals, significant results were also obtained with even stronger effects sizes and a higher percent of heritability explained (Supplementary Figure S2). When performing logistic regressions based on quartiles, the most significant and strongest result was obtained when comparing children in the 4th quartile of the pGRS vs. those in the reference bottom group. On the other hand, we demonstrated that only 9 over the total 44 analyzed SNPs presented an individual significant association with the BMI Z-Score, showing barely significant p-values (Supplementary Table S7). The *FTO* locus was identified among the most relevant loci, which is in accordance with previous studies pointing out the *FTO* as a central piece within the genetics architecture of obesity [41]. A few conclusions could be extracted from these results. The first remark is the fact that, although all assessed SNPs individually elicit small risk effects for obesity, as shown here (Supplementary Table S7) and in previous studies [3]; it is the accumulation of many of these small-risk effect variants in the same individual which finally triggers the clinical manifestation of the obesity phenotype, leading to a robust significant association (Figure 2A). This is what is known as "concerted polygenetic behavior" and has been previously described for obesity and many other complex diseases [13,42]. Under these circumstances, the use of a weighted pGRS approach is the only way to account for small risk genetics effects on disease that would otherwise remain undetected. Thus the use of pGRSs is an additional way to unravel part of the missing heritability of complex traits [13]. Furthermore, the use of a weighted approach (e.g., instead of a simple sum of the number of risk alleles by individual) improved the robustness of associations and allowed us to create a model with a higher similarity to the real to the real molecular basis of the disease.

The second remark that could be extracted from our results is the fact that the overweight status seems to be a midway phenotype (between normal weight and obesity), in which genetics might not play a determinant role (Supplementary Figure S1). Although both remarks had been previously described in the literature [18,22], our approach reinforces these hypotheses and adds novel insights for Iberian populations in Spain, which is quite important considering the well-known genetic interpopulations variability within the European ancestry [43]. All these findings from our cross-sectional study populations 1 and 3 were independently validated also at each pubertal stage of the study population 2 (please see Results Section 4.2.). This corroborates the robustness of our design and reaffirms the fact that genetic predisposition to obesity starts early in childhood and persists during puberty [2]. Interestingly, the obesity heritability attributable to these genetic markers in our study was estimated in 5.6%, which is far higher than the 1–2% reported in the

adult study from Locke *et al.* (2015) [3]. This could be explained by the fact that the environmental modulatory effects on genetics during childhood may be softer than in adults.

As a secondary aim, we demonstrated that the pGRS is useful for the prediction of obesity in children. Among all trained single-variable predictive models, the one based on the pGRS showed one of the greater predictive abilities (only surpassed by the model including parental BMI information). The joint combination of all models individually surpassing an AUC of 0.6 yielded a considerable improvement in the predictive ability (AUC = 0.81 CI 95% = [0.7–0.93]), which could be sufficient for clinical discrimination. All these results are in concordance with previous insights from Butler *et al.* (2019) [44], who demonstrated that early clinical factors, including maternal age, prepregnancy maternal (and paternal) BMI, birthweight, gestational age, weight gain during early infancy and other easily and measurable factors, do fairly well in predicting childhood obesity. Moreover, these results suggest that the combination of a high-risk genetic profile along with an unhealthy familial environment (represented in terms of parents BMI and obesity family history) could boost the predisposition to the disease. Beyond AUC predictive analyses, we also showed how a higher pGRS is associated with obesity longitudinal trajectories when entering puberty in study population 2. We found higher pGRS in children remaining obese after puberty when compared to children remaining with normal weight when entering puberty (Figure 3). From these results, we can conclude that the pGRS could be a powerful predictive tool, assayable from the moment of childbirth, with application in the risk assessment for future obesity. Besides, since severe obesity is usually accompanied by higher odds for metabolic complications during adulthood, these risk estimations could lead to the application of personalized preventive strategies in order to tackle the relevant problem of obesity-associated morbidity and mortality. Interestingly, as far as we are concerned, this is the first time a study focused on the longitudinal effects of a pGRS during pubertal development. Puberty has been identified as a major influence on cardiovascular risk factors, the impaired glucose tolerance of pubertal adolescents with obesity being the best explanation [24,45,46]. The demonstrated ability of the pGRS to predict, from early childhood, the pubertal obesity status of each child is therefore a tool of great interest for identifying children with higher odds for cardio-metabolic disturbances at this metabolically critical stage of life.

In previous studies performed in adults, obesity pGRSs have also yielded secondary findings for multiple cardio-metabolic outcomes, including a heightened risk of all-cause mortality, diabetes, coronary artery disease, hypertension, stroke, and venous thromboembolism, all of them after correcting for BMI. While we knew the clinical association of obesity with these outcomes and conditions, the pGRS correlation now adds the genomic underpinning. To date, no studies have demonstrated such associations in children populations otherwise. Here, we only found slightly significant associations (BMI-adjusted) between the pGRS and certain lipid metabolism outcomes (**Table 2**; **Table 3**, and Supplementary Table S8), none of them passing multiple-test adjustment. Among the rest of the analyzed outcomes, such as inflammatory and cardiovascular biomarkers,

no additional significant association was found. From these results, we could conclude that the associations reported in adults between the pGRS and cardio-metabolic disturbances could be a consequence of the strong correlation between obesity, the obesogenic environment and these outcomes, rather than a direct consequence of having a higher pGRS. This is not surprising since most of the loci conforming the pGRS, 60% of them, are loci highly expressed in regions of the brain and hypothalamus regulating energy balance, appetite, food preference, and reward-seeking behavior [3], rather than loci involved in inflammatory or glucose metabolism-related processes.

From our gene-environment cross-sectional approach, we found that only the educational level of the parents demonstrated a significant interaction with the pGRS. Compared with other socioeconomic indicators, the educational level of the mother is the one that had presented the strongest association with unhealthy factors in literature, such as adiposity, in both children and adolescents [47,48]. Particularly, in our cohort, we saw how this variable was not able to break the genetic determinism or susceptibility to obesity conditioned by the pGRS. Although no other factor evidenced a genetic risk modulatory capacity in our study (neither PA measurements), this does not mean that there is no influence of the environment in the genetic predisposition to obesity in children.

Although we have previously shown that certain individual obesity-SNPs could act as pharmacogenetics regulators of metformin response in children with obesity [36], we here discarded the utility of the pGRS as a marker for the obesity pharmacological management. Again, this is not surprising given the type of SNPs included in the pGRS, where the metformin target pathways are not included. On this matter, we can conclude that a higher genetic predisposition to obesity (according to the genes involved in satiety and energy balance regulatory mechanisms) does not determine a worse BMI Z-Score response when treating with metformin. For the pharmacological personalized management of children with obesity instead, we recommend the use of individual validated target SNPs [36].

Among the limitations of our current approach, we can highlight the inclusion of only European ancestry individuals, limiting extrapolation for other ancestries and the lack of objectively measured physical activity and diet assessments. These important questions remain unanswered and will define the potential benefit derived from using this obesity pGRS.

Prevention of key medical conditions such as obesity has been a long-standing dream that largely remains unfulfilled. If we are to take advantage of the opportunity, we need to know as much as possible for prediction, acknowledging it will never be deterministic. The obesity pGRS reported in the present study provides an extremely powerful tool for the early risk detection. While there remains uncertainties and practical limitations for making such pGRS results widely available, such as the requirement for considerable education for the medical community and the general population, we are moving in the right direction for someday pre-empting important conditions that would have otherwise been manifest.

Supplementary Materials

The following are available online at https://www.mdpi.com/2077-0383/9/6/1705/s1, Figure S1: Density distribution plots for the constructed pGRS according to the obesity status in study population 1, Figure S2: Bar plot showing the number of normal weight, overweight and children with obesity according to each quartile of the pGRS in the study population 1, Figure S3: Stepwise linear regression including all 44 tested SNPs in order to know which of them contribute the most to the pGRS-BMI association, Table S1: General characteristics, anthropometry, biochemical parameters, adipokines and cardiovascular/proinflammatory biomarkers in the cross-sectional cohort of 574 children (study population 1), Table S2: Descriptive statistics for the longitudinal study population 2, Table S3: Clinical characteristics of the study population 3 at baseline and post-treatment stages, Table S4: Hardy-Weinberg Equilibrium test for all analyzed genetic variants in study population 1, Table S5: List of 44 SNPs passing quality control filters and finally included in the Genetic Risk Score, Table S6: Lifestyle factors assessed in our study for the study population 1, Table S7: Individual single-SNP analyses on BMI Z-Score in the study population 1, Table S8: Association between the pGRS (quartilized) and the metabolic health status of children in the study population 1.

Author Contributions

Conceptualization, Á.G. and C.M.A.; methodology, A.A.-R., E.M.G.-G, A.I.R., F.J.L.-C., B.P.-V., L.A.M, M.G.-C., G.B., R.L.; software, A.A.-R.; validation, A.A.-R.; formal analysis, A.A.-R.; investigation, A.A.-R.; resources, J.A.-F., L.A.M, Á.G., M.G.-C., G.B., R.L., C.M.A.; data curation, A.A.-R.; writing—original draft preparation, A.A.-R.; writing—review and editing, A.A.-R., E.M.G.-G, A.I.R., F.J.L.-C., B.P.-V., J.A.-F., L.A.M, Á.G., M.G.-C., G.B., R.L., C.M.A.; visualization, A.A.-R.; supervision, C.M.A.; project administration, C.M.A.; funding acquisition, J.A.-F., L.A.M, Á.G., M.G.-C., G.B., R.L., C.M.A. All authors have read and agreed to the published version of the manuscript. This paper will be part of Augusto Anguita-Ruiz's doctorate thesis, which is being performed under the "Nutrition and Food Sciences Program" at the University of Granada.

Funding

This research was supported by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I + D + I), Instituto de Salud Carlos III-Health Research Funding (FONDOS FEDER) (PI1102042, PI1102059, PI1601301 and PI1600871); by the Spanish Ministry of Health, Social and Equality, General Department for Pharmacy and Health Products (EC10-243, EC10-056, EC10-281 and EC10-227); by the Regional Government of Andalusia ("Plan Andaluz de investigación, desarrollo e innovación (2018), P18-RT-2248") and by the Mapfre Foundation ("Research grants by Ignacio H. de Larramendi 2017"). The authors also acknowledge Instituto de Salud Carlos III for personal funding: Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016 (IFI17/00048). The authors also acknowledge the University of Granada "Plan Propio de Investigacion 2016-Excellence actions: Unit of Excellence on Exercise and Health (UCEES)".

Acknowledgments

The authors would like to thank the Spanish children and parents who participated in the study.

Conflicts of Interest:

The authors declare no conflict of interest.

References

1. GBD 2015 Obesity Collaborators; Afshin, A.; Forouzanfar, M.H.; Reitsma, M.B.; Sur, P.; Estep, K.; Lee, A.; Marczak, L.; Mokdad, A.H.; Moradi-Lakeh, M.; et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N. Engl. J. Med. 2017, 377, 13–27, doi:10.1056/NEJMoa1614362.
2. Jones, R.E.; Jewell, J.; Saksena, R.; Ramos Salas, X.; Breda, J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Front. Public Health 2017, 5, 1–12, doi:10.3389/fpubh.2017.00058.
3. Locke, A.E.; Kahali, B.; Berndt, S.I.; Justice, A.E.; Pers, T.H.; Day, F.R.; Powell, C.; Vedantam, S.; Buchkovich, M.L.; Yang, J.; et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015, 518, 197–206, doi:10.1038/nature14177.
4. Belsky, D.W.; Moffitt, T.E.; Houts, R.; Bennett, G.G.; Biddle, A.K.; Blumenthal, J.A.; Evans, J.P.; Harrington, H.L.; Sugden, K.; Williams, B.; et al. Polygenic risk, rapid childhood growth, and the development of obesity: Evidence from a 4-decade longitudinal study. Arch. Pediatr. Adolesc. Med. 2012, 166, 515–521, doi:10.1001/archpediatrics.2012.131.
5. Elks, C.E.; Loos, R.J.F.; Hardy, R.; Wills, A.K.; Wong, A.; Wareham, N.J.; Kuh, D.; Ong, K.K. Adult obesity susceptibility variants are associated with greater childhood weight gain and a faster tempo of growth: The 1946 British Birth Cohort Study. Am. J. Clin. Nutr. 2012, 95, 1150–1156, doi:10.3945/ajcn.111.027870.
6. Felix, J.F.; Bradfield, J.P.; Monnereau, C.; van der Valk, R.J.P.; Stergiakouli, E.; Chesi, A.; Gaillard, R.; Feenstra, B.; Thiering, E.; Kreiner-Møller, E.; et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum. Mol. Genet. 2016, 25, 389–403, doi:10.1093/hmg/ddv472.
7. Mäkelä, J.; Lagström, H.; Pitkänen, N.; Kuulasmaa, T.; Kaljonen, A.; Laakso, M.; Niinikoski, H. Genetic risk clustering increases children's body weight at 2 years of age - the STEPS Study. Pediatr. Obes. 2016, 11, 459–467, doi:10.1111/ijpo.12087.
8. Zandoná, M.R.; Sangalli, C.N.; Campagnolo, P.D.B.; Vitolo, M.R.; Almeida, S.; Mattevi, V.S. Validation of obesity susceptibility loci identified by genome-wide association studies in early childhood in South Brazilian children. Pediatr. Obes. 2017, 12, 85–92, doi:10.1111/ijpo.12113.
9. Tekola-Ayele, F.; Lee, A.; Workalemahu, T.; Sánchez-Pozos, K. Shared genetic underpinnings of childhood obesity and adult cardiometabolic diseases. Hum. Genom. 2019, 13, 17, doi:10.1186/s40246-019-0202-x.
10. Moon, J.Y.; Wang, T.; Sofer, T.; North, K.E.; Isasi, C.R.; Cai, J.; Gellman, M.D.; Moncrieft, A.E.; Sotres-Alvarez, D.; Argos, M.; et al. Objectively measured physical activity, sedentary behavior, and genetic predisposition to obesity in U.S. Hispanics/Latinos: Results from the hispanic community health study/study of Latinos (HCHS/SOL). Diabetes 2017, 66, 3001–3012, doi:10.2337/db17-0573.
11. Mead, E.; Brown, T.; Rees, K.; Azevedo, L.B.; Whittaker, V.; Jones, D.; Olajide, J.; Mainardi, G.M.; Corpeleijn, E.; O'Malley, C.; et al. Diet, physical activity and behavioural interventions for the treatment of overweight or obese children from the age of 6 to 11 years. Cochrane Database Syst. Rev. 2017, 2017.
12. Fang, J.; Gong, C.; Wan, Y.; Xu, Y.; Tao, F.; Sun, Y. Polygenic risk, adherence to a healthy lifestyle, and childhood obesity. Pediatr. Obes. 2019, 14, doi:10.1111/ijpo.12489.
13. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. Nature 2009, 461, 747–753, doi:10.1038/nature08494.
14. Schrodi, S.J.; Mukherjee, S.; Shan, Y.; Tromp, G.; Sninsky, J.J.; Callear, A.P.; Carter, T.C.; Ye, Z.; Haines, J.L.; Brilliant, M.H.; et al. Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future. Front. Genet. 2014, 5, 1–18, doi:10.3389/fgene.2014.00162.
15. Khera, A.V.; Chaffin, M.; Aragam, K.G.; Haas, M.E.; Roselli, C.; Choi, S.H.; Natarajan, P.; Lander, E.S.; Lubitz, S.A.; Ellinor, P.T.; et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 2018, 50, 1219–1224, doi:10.1038/s41588-018-0183-z.
16. Torkamani, A.; Wineinger, N.E.; Topol, E.J. The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. 2018, 19, 581–590, doi:10.1038/s41576-018-0018-x.
17. Chatterjee, N.; Shi, J.; García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. 2016, 17, 392–406, doi:10.1038/nrg.2016.27.
18. Khera, A.V.; Chaffin, M.; Wade, K.H.; Zahid, S.; Brancale, J.; Xia, R.; Distefano, M.; Senol-Cosar, O.; Haas, M.E.; Bick, A.; et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell. 2019, 177, 587–596.e9, doi:10.1016/j.cell.2019.03.028.
19. Hohenadel, M.G.; Baier, L.J.; Piaggi, P.; Muller, Y.L.; Hanson, R.L.; Krakoff, J.; Thearle, M.S. The impact of genetic variants on BMI increase during childhood versus adulthood. Int. J. Obes. 2016, 40, 1301–1309, doi:10.1038/ijo.2016.53.

20. Choh, A.C.; Lee, M.; Kent, J.W.; Diego, V.P.; Johnson, W.; Curran, J.E.; Dyer, T.D.; Bellis, C.; Blangero, J.; Siervogel, R.M.; et al. Gene-by-age effects on BMI from birth to adulthood: The fels longitudinal study. Obesity 2014, 22, 875–881, doi:10.1002/oby.20517.
21. Song, M.; Zheng, Y.; Qi, L.; Hu, F.B.; Chan, A.T.; Giovannucci, E.L. Associations between genetic variants associated with body mass index and trajectories of body fatness across the life course: A longitudinal analysis. Int. J. Epidemiol. 2018, 47, 506–515, doi:10.1093/ije/dyx255.
22. Torkamani, A.; Topol, E. Polygenic Risk Scores Expand to Obesity. Cell 2019, 177, 518–520.
23. Hannon, T.S.; Janosky, J.; Arslanian, S.A. Longitudinal study of physiologic insulin resistance and metabolic changes of puberty. Pediatr. Res. 2006, 60, 759–763, doi:10.1203/01.pdr.0000246097.73031.27.
24. Reinehr, T.; Roth, C.L. Is there a causal relationship between obesity and puberty? Lancet Child Adolesc. Health 2019, 3, 44–54.
25. Mega, J.L.; Stitziel, N.O.; Smith, J.G.; Chasman, D.I.; Caulfield, M.; Devlin, J.J.; Nordio, F.; Hyde, C.; Cannon, C.P.; Sacks, F.; et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: An analysis of primary and secondary prevention trials. Lancet 2015, 385, 2264–2271, doi:10.1016/S0140-6736(14)61730-X.
26. Olza, J.; Aguilera, C.M.; Gil-Campos, M.; Leis, R.; Bueno, G.; Martínez-Jiménez, M.D.; Valle, M.; Canête, R.; Tojo, R.; Moreno, L.A.; et al. Myeloperoxidase is an early biomarker of inflammation and cardiovascular risk in prepubertal obese children. Diabetes Care 2012, 35, 2373–2376, doi:10.2337/dc12-0614.
27. Anguita-Ruiz, A.; Plaza-Diaz, J.; Ruiz-Ojeda, F.J.; Ruperez, A.I.; Leis, R.; Bueno, G.; Gil-Campos, M.; Vazquez-Cobela, R.; Canete, R.; Moreno, L.A.; et al. X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline. Sci. Data 2019, 6, 130, doi:10.1038/s41597-019-0109-3.
28. Cole, T.J.; Bellizzi, M.C.; Flegal, K.M.; Dietz, W.H. Establishing a standard definition for child overweight and obesity worldwide: International survey. BMJ 2000, 320, 1240–1243, doi:10.1136/bmj.320.7244.1240.
29. Anguita-Ruiz, A.; Mendez-Gutierrez, A.; Ruperez, A.I.; Leis, R.; Bueno, G.; Gil-Campos, M.; Tofe, I.; Gomez-Llorente, C.; Moreno, L.A.; Gil, Á.; et al. The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity. Metabolism 2020, 105, doi:10.1016/j.metabol.2020.154187.
30. Tanner, J.M.; Whitehouse, R.H. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. Arch. Dis. Child. 1976, 51, 170–179.
31. Tang, Q.; Li, X.; Song, P.; Xu, L. Optimal cut-off values for the homeostasis model assessment of insulin resistance (HOMA-IR) and pre-diabetes screening: Developments in research and prospects for the future. Drug Discov. Ther. 2015, 9, 380–385, doi:10.5582/ddt.2015.01207.
32. Andrade, M.I.S. de; Oliveira, J.S.; Leal, V.S.; Lima, N.M.S. da; Costa, E.C.; Aquino, N.B. de; Lira, P.I.C. de Identification of cutoff points for Homeostatic Model Assessment for Insulin Resistance index in adolescents: Systematic review. Rev. Paul. Pediatr. 2016, 34, 234, doi:10.1016/J.RPPEDE.2016.01.004.
33. Rupérez, A.I.; Olza, J.; Gil-Campos, M.; Leis, R.; Bueno, G.; Aguilera, C.M.; Gil, A.; Moreno, L.A. Cardiovascular risk biomarkers and metabolically unhealthy status in prepubertal children: Comparison of definitions. Nutr. Metab. Cardiovasc. Dis. 2018, 28, 524–530, doi:10.1016/j.numecd.2018.02.006.
34. Pastor-Villaescusa, B.; Caballero-Villarraso, J.; Cañete, M.D.; Hoyos, R.; Maldonado, J.; Bueno, G.; Leis, R.; Gil, Á.; Cañete, R.; Aguilera, C.M. Evaluation of differential effects of metformin treatment in obese children according to pubertal stage and genetic variations: Study protocol for a randomized controlled trial. Trials 2016, 17, 323, doi:10.1186/s13063-016-1403-4.
35. Pastor-Villaescusa, B.; Cañete, M.D.; Caballero-Villarraso, J.; Hoyos, R.; Latorre, M.; Vázquez-Cobela, R.; Plaza-Díaz, J.; Maldonado, J.; Bueno, G.; Leis, R.; et al. Metformin for Obesity in Prepubertal and Pubertal Children: A Randomized Controlled Trial. Pediatrics 2017, 140, e20164285, doi:10.1542/peds.2016-4285.
36. Anguita-Ruiz, A.; Pastor-Villaescusa, B.; Leis, R.; Bueno, G.; Hoyos, R.; Vázquez-Cobela, R.; Latorre-Millán, M.; Cañete, M.D.; Caballero-Villarraso, J.; Gil, Á.; et al. Common Variants in 22 Genes Regulate Response to Metformin Intervention in Children with Obesity: A Pharmacogenetic Study of a Randomized Controlled Trial. J. Clin. Med. 2019, 8, 1471, doi:10.3390/jcm8091471.
37. Sobradillo, B.; Aguirre, A.; Aresti, U.; Bilbao, A.; Fernández-Ramos, C.; Lizárraga, A.; Lorenzo, H.; Madariaga, L.; Rica, I.; Ruiz, I.; et al. Curvas y tablas de crecimiento (estudios longitudinal y transversal); Fundación Faustino Orbegozo Eizaguirre: Madrid, Spain, 2004; ISBN 84-607-9967-0.
38. Wray, N.R.; Yang, J.; Goddard, M.E.; Visscher, P.M. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. PLoS Genet. 2010, 6, e1000864, doi:10.1371/journal.pgen.1000864.

39. Cook, N.R. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. Circulation 2007, 115, 928–935, doi:10.1161/CIRCULATIONAHA.106.672402.
40. Keller, M.C. Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. Biol. Psychiatry 2014, 75, 18–24, doi:10.1016/j.biopsych.2013.09.006.
41. Loos, R.J.F.; Yeo, G.S.H. The bigger picture of FTO - The first GWAS-identified obesity gene. Nat. Rev. Endocrinol. 2014, 10, 51–61.
42. Apalasamy, Y.D.; Mohamed, Z. Obesity and genomics: Role of technology in unraveling the complex genetic architecture of obesity. Hum. Genet. 2015, 134, 361–374.
43. Durbin, R.M.; Altshuler, D.L.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Collins, F.S.; De La Vega, F.M.; Donnelly, P.; et al. A map of human genome variation from population-scale sequencing. Nature 2010, 467, 1061–1073, doi:10.1038/nature09534.
44. Butler, É.M.; Derraik, J.G.B.; Taylor, R.W.; Cutfield, W.S. Prediction Models for Early Childhood Obesity: Applicability and Existing Issues. Horm. Res. Paediatr. 2018, 90, 358–367, doi:10.1159/000496563.
45. Goran, M.I.; Shaibi, G.Q.; Weigensberg, M.J.; Davis, J.N.; Cruz, M.L. Deterioration of insulin sensitivity and beta-cell function in overweight Hispanic children during pubertal transition: A longitudinal assessment. Int. J. Pediatr. Obes. 2006, 1, 139–145, doi:10.1080/17477160600780423.
46. Reinehr, T.; Wabitsch, M.; Kleber, M.; de Sousa, G.; Denzer, C.; Toschke, A.M. Parental diabetes, pubertal stage, and extreme obesity are the main risk factors for prediabetes in children and adolescents: A simple risk score to identify children at risk for prediabetes. Pediatr. Diabetes 2009, 10, 395–400, doi:10.1111/j.1399-5448.2008.00492.x.
47. McCrory, C.; Leahy, S.; Ribeiro, A.I.; Fraga, S.; Barros, H.; Avendano, M.; Vineis, P.; Layte, R.; Alenius, H.; Baglietto, L.; et al. Maternal educational inequalities in measured body mass index trajectories in three European countries. Paediatr. Perinat. Epidemiol. 2019, 33, 226–237, doi:10.1111/ppe.12552.
48. Adler, N.E.; Boyce, T.; Chesney, M.A.; Cohen, S.; Folkman, S.; Kahn, R.L.; Syme, S.L. Socioeconomic Status and Health: The Challenge of the Gradient. Am. Psychol. 1994, 49, 15–24, doi:10.1037/0003-066X.49.1.15.

Section II

Section II

IDENTIFICATION OF NEW MULTI-OMICS BIOMARKERS OF IR AND CARDIOMETABOLIC ALTERATIONS IN CHILDHOOD OBESITY DURING THE METABOLICALLY CRITICAL PERIOD OF PUBERTY

Metabolism. 2020;105:154187. doi:10.1016/j.metabol.2020.154187. IF: 8.694, D1 at ENDOCRINOLOGY & METABOLISM.

Study 5 | The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity

Augusto Anguita-Ruiz[1,2,3,4*], Andrea Méndez-Gutiérrez[1,2,3,4*], Azahara I Ruperez[4,5], Rosaura Leis[4,6], Gloria Bueno[4,5], Mercedes Gil[4,7], Inés Tofe[7], Carolina Gomez-Llorente[1,2,3], Luis A. Moreno[4,5], Ángel Gil[1,2,3,4] & Concepción M Aguilera[1,2,3,4#].

Abstract Background: S100A4 is a metastasis-associated protein also reported as a promising marker for dysfunctional white adipose tissue (WAT) and insulin resistance (IR) in adult and adolescent populations.

Objective: We aimed to evaluate the association between the protein S100A4 and obesity and IR in children and during pubertal development.

Design and Methods: The study design consisted of three cross-sectional populations of 249, 11 and 19 prepubertal children respectively (named study population 1, 2 and 3), and a longitudinal population of 53 girls undergoing sexual maturation (study population 4). All subjects were classified into experimental groups according to their sex, obesity and IR status. All study populations counted on anthropometry, glucose, and lipid metabolism, inflammation and cardiovascular biomarkers as well as S100A4 plasma levels measured.

Affiliations 1. Department of Biochemistry and Molecular Biology II, School of Pharmacy, University of Granada, Spain. / 2. Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada. Avda. del Conocimiento s/n. 18016 Armilla, Granada, Spain. / 3. Instituto de Investigación Biosanitaria ibs.GRANADA, 18012 Granada, Spain. / 4. CIBEROBN, (Physiopathology of Obesity and Nutrition CB12/03/30038), Institute of Health Carlos III (ISCIII), Madrid 28029, Spain. / 5. Growth, Exercise, Nutrition and Development (GENUD) Research Group, University of Zaragoza, Zaragoza, Spain; Instituto Agroalimentario de Aragón (IA2) and Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain1. / 6. Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department (USC). Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), University Clinical Hospital, Santiago de Compostela, Spain. / 7. Department of Pediatric Endocrinology, Reina Sofia University Clinical Hospital, Institute Maimónides of Biomedicine Investigation of Córdoba (IMIBIC), University of Córdoba, Avda. Menéndez Pidal s/n, 14004 Córdoba, Spain.
Corresponding author
* Equal contributors.

The study population 1 was intended as the discovery population in which to elucidate the relationship between Obesity-IR and S100A4 plasma levels in prepubertal children. The cross-sectional populations 2 and 3 further counted on WAT gene expression data for investigating the molecular basis of this association. Instead, the longitudinal study population 4 presented blood whole-genome DNA methylation data at each temporal record, allowing deepening into the Obesity-IR-S100A4 relationship during puberty as well as deciphering plausible epigenetic mechanisms altering S100A4 plasma levels.

Results: S100A4 plasma levels were strongly associated with several metabolic and anthropometric outcomes, namely IR, in prepubertal non-diabetic obese children. We also found highly significant positive associations during the course of puberty between the increase in S100A4 levels and the increase in HOMA-IR (P=0.0003, FDR=0.005) and insulin levels (P=0.0003, FDR=0.005). Methylation in two-enhancer related CpG sites of the *S100A4* region (cg07245635 and cg10447638) was associated with IR biomarkers at the prepubertal stage and with longitudinal changes in these measurements. We further reported an association between visceral WAT (vWAT) *S100A4* expression and HOMA-IR, insulin levels and BMI Z-Score, but not with circulating S100A4.

Conclusions: We report for the first time the association of S100A4 with IR and WAT dysfunction in prepubertal populations as well as how the change in plasma S100A4 levels accompanies longitudinal trajectories of IR in children during pubertal development. Moreover, we propose epigenetic changes in two methylation sites and an altered *S100A4* vWAT expression as plausible molecular mechanisms underlying this disturbance in obesity.

Keywords: Obesity; Longitudinal Study; Children; Puberty; Epigenetics.

Abbreviations: WAT (white adipose tissue), IR (insulin resistance), sWAT (subcutaneous WAT), vWAT (visceral WAT), WC waist circumference), TNF-α (tumour necrosis factor alpha), hsCRP (high-sensitivity CRP), IL (interleukin), PAI-1 (total plasminogen activator inhibitor-1), MPO (myeloperoxidase), MCP-1 (monocyte chemoattractant protein 1), MMP-9 (matrix metalloproteinase-9), sICAM-1 soluble intercellular cell adhesion molecule-1), VCAM (soluble vascular cell adhesion molecule-1), HPRT (hypoxanthine-guanine phosphoribosyltransferase-1), MLR (multiple linear regression), FDR (false discovery rate), LME (linear mixed-effects), AIC (Akaike information criterion).

1. Introduction

The human S100A4 is a member of the soluble calcium (Ca 2+)-binding proteins [1] primarily classified as a metastasis-associated protein [2]. With reported expression in a wide variety of cell types [3–5]), S100A4 seems to participate in cell migration, chemotaxis, angiogenesis and extracellular matrix remodeling [6]. Given its implication in these biological processes, besides cancer, S100A4 has also been related to other non-tumour diseases including multiple sclerosis, psoriasis or rheumatoid arthritis, where it participates in pro-inflammatory pathways [7].

Recently, S100-proteins have been extensively reviewed in the context of adipose tissue and obesity [8]. Regarding the S100A4 protein, it highlights two recent studies postulating it as a promising circulating marker for dysfunctional white adipose tissue (WAT) and insulin resistance (IR) in adult [9] and adolescent populations [10]. Particularly, these studies have shown S100A4 as a novel adipokine associated with subcutaneous WAT (sWAT) inflammation, hepatovisceral fat excess, obesity-related IR as well as with a higher prevalence of type 2 diabetes. These associations have not been yet investigated in childhood, though it is the best period of life for understanding how obesity biomarkers correlate with later BMI changes and cardiometabolic derangement [11]. The early-life identification of high-risk individuals for IR and type 2 diabetes complications is of special importance for reducing obesity-associated mortality. Nowadays, monitoring non-invasive biomarkers such as S100A4 offers a great opportunity for disease prevention. However, today there is not enough evidence regarding the role of this protein as a biomarker of IR or the molecular mechanisms behind it.

In the last years, CpG DNA methylation has been involved in numerous diseases, where it has been established as an important etiological molecular mechanism and a link with environmental exposures [12]. Previous cancer studies have shown the epigenetic regulation of several members of the S100 family through methylation of key CpG sites within their genes or promoters [13,14]. In obesity, previous studies have confirmed that the epigenome is an important regulator of gene expression [15]. Therefore, our research hypothesis was that epigenetic alterations in *S100A4* could be relevant for understanding its role in the IR-obesity axis.

The aim of the present study was to evaluate the association between S100A4 and different obesity and IR parameters making use of multiple well-characterized children populations under both cross-sectional and longitudinal designs. In both prepubertal and pubertal stages, we carried out a multi-omics approach (including gene expression and DNA methylation analyses) for elucidating the potential molecular basis of the S100A4-IR association.

2. Research Design and Methods

2.1. The cross-sectional study design

A total of 249 Spanish prepubertal children (124 girls), aged 3.5–12.6 years, all Caucasian, were recruited from primary care centers and schools in three Spanish cities (Córdoba (southern Spain) (*N*=87), Santiago de Compostela (northwestern Spain) (*N*=99), and Zaragoza (northeastern Spain) (*N*=63)) during the year period 2012-2015. This sample was named as study population 1 and constituted the discovery population for investigating the relationship between S100A4 and childhood obesity. Inclusion criteria were the absence of congenital metabolic diseases, prepubertal status (Tanner 0) and European-Caucasian heritage. Exclusion criteria were pubertal stage, the presence of congenital metabolic diseases or undernutrition, and the use of medication that alters blood pressure (BP), glucose, or lipid metabolism. After initial assessments at the school or primary care center, children fulfilling the inclusion criteria were invited for a clinical examination in the appropriated, participating hospitals ("Reina Sofia" University Hospital (Córdoba), Unit of Clinical Analyses "Valle de los Pedroches" Hospital (Córdoba), Pediatric department of "Lozano Blesa" University Hospital (Zaragoza), and Pediatric Department of Clinic University Hospital of Santiago (Santiago de Compostela)). Trained pediatricians performed the clinical examinations according to standardized methods. Pubertal stage was determined in each patient according to the Tanner criteria [16] and validated by plasma sex hormone concentrations. All employed recruiting protocols have been previously published [17,18]. All children were classified for obesity using the BMI age- and sex-specific cut-off points proposed by Cole *et al.* [19] and allocated into experimental groups according to their sex, obesity and IR status. The experimental groups comprised non-IR normal weight children, non-IR children with obesity, and IR children with obesity (further stratified by sex). The minimum number of subjects per group was 40 and the maximum 43, therefore constituting a balanced design. General characteristics, anthropometry, biochemical parameters, adipokines, and cardiovascular/pro-inflammatory biomarkers of the study population 1 are presented in the Supplementary Table 1 stratified by group.

In order to perform *S100A4* gene expression analyses in WAT, two additional study populations were recruited and named as study population 2 and 3. The first one consisted of 11 prepubertal Spanish children derived from a previously published work and followed the same recruiting protocols than the study population 1. In this case, all children were enrolled from our recruiting center in the southern of Spain (Córdoba) and comprised ages 7-12 years. General characteristics of these children are presented in the Supplementary Table 2. Prepubertal stage was again determined in each patient according to the Tanner criteria [16] and validated by plasma sex hormone concentrations [20]. As an independent and validation sample, the study population 3 was composed of a total of 20 prepubertal children (10 with obesity and 9 normal-weight) [20] (Supplementary Table 3). Enrolled from our recruiting center in the southern of Spain, these children

comprised ages 6-13 years. Prepubertal stage was also determined in each patient according to the Tanner criteria [16] and validated by plasma sex hormone concentrations.

The whole study design of this work has been summarized and extensively detailed in **figure 1**.

2.2. The longitudinal study design

A longitudinal study was also conducted in 53 Spanish girls allocated into six experimental groups according to their obesity and IR status before and after the onset of puberty (study population 4). Pubertal stage was determined in each patient according to the Tanner criteria [16] and validated by plasma sex hormone concentrations. Of note, 17 girls from this subpopulation were also part of the previously mentioned cross-sectional study population 1. All these girls were first recruited as prepubertal children during the year period (2012-2015) and called again for follow-up medical consultation in 2018. At the moment of recruitment, girls were aged 4-10.7 and came from our three recruiting centers (Córdoba (southern Spain) (*N*=4), Santiago de Compostela (northwestern Spain) (*N*=27), and Zaragoza (northeastern Spain) (*N*=22)). At the second visit, girls were aged 9.7-17.4. All subjects with clinical signs of reached puberty were finally included in the longitudinal population. During all the course of the study (2012-2018), children remained under regular medical monitoring by the same pediatricians. In this longitudinal design, both S100A4 levels and *S100A4* DNA methylation were investigated in blood samples.

Further details regarding the adopted longitudinal design are illustrated in **Figure 1**.

2.3. Ethics statement

These studies were conducted in accordance with the Declaration of Helsinki (Edinburgh 2000 revised) and they followed the recommendations of the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the Spanish legislation in force, which regulates the clinical investigation of human beings (RD 223/04 about clinical trials). The Ethics Committee on Clinical Research of Aragon, the Bioethics Committee of the University of Santiago de Compostela, the Ethics Committee of the Reina Sofia University Hospital of Cordoba, and the Ethics Committee on Human Research of the University of Granada approved all experiments and procedures. All parents or guardians provided written informed consent and the children gave their assent.

2.4. HOMA-IR cut-off points

The IR status was defined by means of the HOMA-IR index. Since HOMA-IR strongly varies between ages, genders, and diseases [21] and since no reference values have been yet established in either children or adult populations [22], we extracted our own cut-off points from a previous well-described Spanish cohort composed of 1669 children and adolescents [23,24]. For the prepubertal stage, a single cut-off value of HOMA-IR=2.5 was considered for IR [23,25]. For the pubertal stage, sex

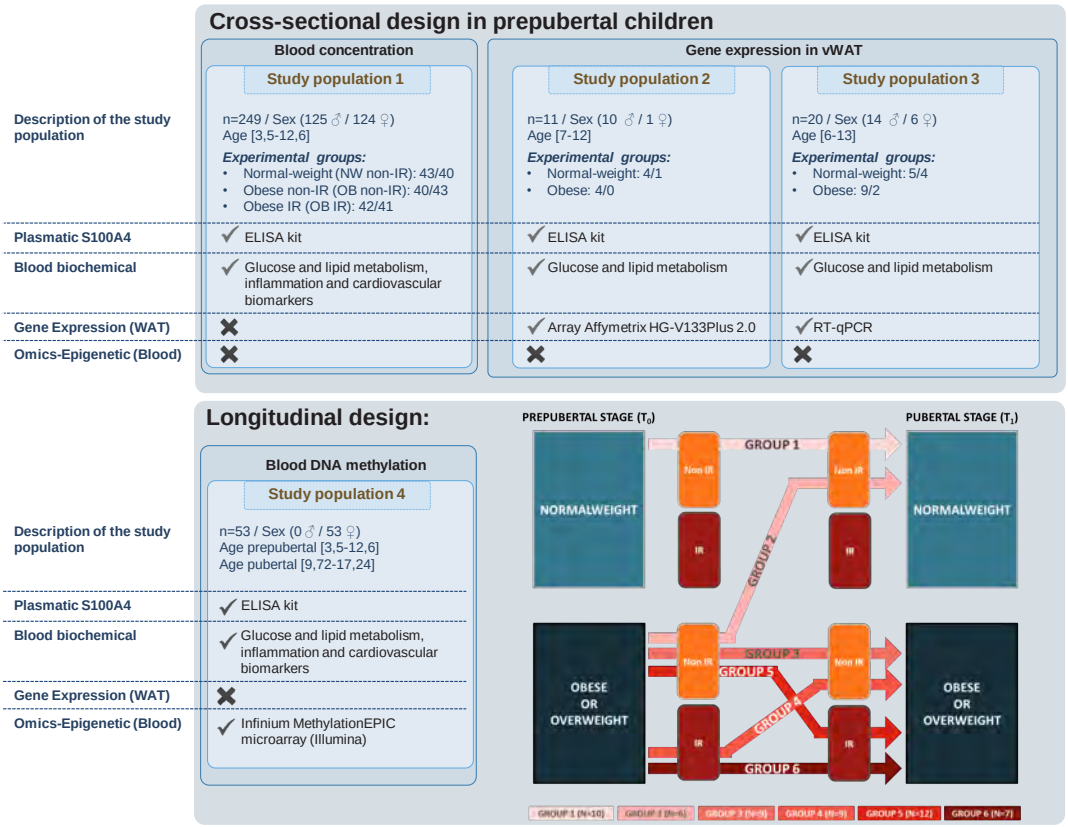


Figure 1. Study design and experimental details of the study populations. Abbreviations: NW, normal-weight; IR, insulin resistance; WAT, white adipose tissue; vWAT, visceral WAT.

information was taken into consideration and different cut-off points were adopted for IR according to the 95th HOMA-IR percentile. Extracted from a subset of 778 pubertal Spanish children, cut-off values corresponded to HOMA-IR=3.38 in boys and HOMA-IR=3.905 in girls.

2.5 Anthropometric and biochemical measurements

Body weight (kg), height (cm) and waist circumference (WC) (cm) were measured using standardized procedures and BMI was calculated. BMI z-score was calculated based on the Spanish reference standards published by Sobradillo *et al.* [26]. Blood pressure was measured three times for each individual by the same examiner using a mercury sphygmomanometer and following international recommendations [27]. Routine measures of lipid and glucose metabolism were performed at the laboratories of each participating hospital following internationally-accepted quality control protocols.

Blood samples were collected in overnight fasting conditions, centrifuged, and plasma and serum were stored at -80°C. Plasma adipokines, inflammation, and cardiovascular risk biomarkers

(adiponectin, leptin, resistin, tumour necrosis factor alpha (TNF-α), high-sensitivity CRP (hsCRP), interleukin (IL)-6, IL-8, total plasminogen activator inhibitor-1 (PAI-1), myeloperoxidase (MPO), monocyte chemoattractant protein 1 (MCP-1), matrix metalloproteinase-9 (MMP-9), soluble intercellular cell adhesion molecule-1 sICAM-1, and soluble vascular cell adhesion molecule-1 (*s*VCAM)) were analyzed in samples of prepubertal children using XMap technology (Luminex Corporation, Austin, TX) and human monoclonal antibodies (Milliplex Map Kit; Millipore, Billerica, MA) as previously reported [24].

2.6. Plasma S100A4 levels

S100A4 protein levels were determined in plasma by CSB-EL02032HU (Cusabio Biotech Co, Ltd, Wuhan, China), an enzyme-linked immuno-absorbent assay kit according to the manufacturers' instructions. The coefficient of variance was 7%.

2.7. DNA extraction and methylation analysis

Buffy coat fractions from blood samples belonging to groups 1,3,4,5, and 6 of the longitudinal design (study population 4) were selected for DNA methylation analysis (*N*=48). Genomic DNA was extracted from peripheral white blood cells as previously described [24]. High-quality DNA samples (≥ 500 ng) were treated with bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research Corporation, Irvine, CA). DNA methylation was measured with the Infinium MethylationEPIC microarray using bead chip technology (Illumina, San Diego, CA, USA). A detailed description of the methods employed for EWAS pre-processing, primary analyses, and statistical designs are available in the supplementary material.

2.8. Gene expression analysis

Genome-wide expression data from visceral WAT (vWAT) of 11 prepubertal Spanish children (study population 2) were obtained from a previously published work [20]. The dataset is freely available from the Gene Expression Omnibus repository [28] (GSE9624) and was composed of six normal-weight and five obese-derived RNA samples (Supplementary Table 2). Among all available probes within the HG-U133 Plus-2.0 array, one was found within the *S100A4* locus, with the identifier "203186_s_at". Stored plasma samples from each participant were thawed and plasma levels of S100A4 were determined in order to study their correlation with *S100A4* gene expression. The *LIMMA* R package was employed for statistical analysis. Robust Multichip Average-normalized gene expression from the "203186_s_at" probe was tested for association with experimental conditions (normal-weight vs. obese) and with each continuous outcome (BMI Z-score, HOMA-IR, insulin levels, glucose levels, and S100A4 plasma levels). These analyses were adjusted for confounders when necessary.

In order to validate microarray analyses, qPCR experiments for the *S100A4* gene were performed in vWAT samples derived from an independent sample of 20 prepubertal children (10 with obesity and 9 normal-weight) (study population 3) previously described [20] (Supplementary Table 3). Hypoxanthine-guanine phosphoribosyltransferase-1 (HPRT1) was used as a reference gene. Primers were acquired (refs. qHsaCED0048045 and qHsaCID0016375) from Bio-Rad Laboratories, California, USA) and qPCR was performed using the ABI Prism 7900 HT instrument (Applied Biosystems, Foster City, USA) and the SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad Laboratories, California, USA). Data were analyzed using the 2−ΔΔCt approach and normalized against *HPRT1* expression.

2.9. Statistical analysis

In the cross-sectional prepubertal study, continuous variables were tested for normality using the Shapiro–Wilk test and transformed when necessary by means of the natural log or the rank-based inverse normal transformation. Heteroscedasticity between experimental groups was explored by means of the Levene test. Then, the one-way ANOVA, Kruskal-Wallis and the Welch test were employed to assess group differences in S100A4 levels and other measurements according to standard statistical assumptions. The pairwise-*t*-tests, the pairwise Mann–Whitney *U*-tests, and the Dunn tests were applied conveniently as post-hoc analyses to determine which experimental groups differed from each other. Multiple linear regression (MLR) analyses were applied for all continuous variables in order to study their association with S100A4 levels. In these analyses, origin, age, gender, pubertal stage, height, BMI Z-score, and insulin were adjusted as confounders when necessary. A p-value < 0.05 was considered as significant. Given the number of analyzed outcomes, we considered false discovery rate (FDR) as in Benjamini and Hochberg to correct for multiple hypothesis testing.

In the longitudinal study, within-group changes from baseline (T_0) to puberty (T_1) were assessed by means of a paired design in all continuous variables (including S100A4 levels); employing either a paired t-test or a Wilcoxon signed-rank test. Between-group differences were assessed by the one-way ANOVA, Kruskal-Wallis, or Welch tests to the computed delta values (T_1–T_0) for each continuous measurement. Between-group differences in S100A4 levels were particularly investigated by means of a linear mixed-effects (LME) model including the covariates age, Tanner stage and time as fixed effects and a random intercept for each participant. Test significance for the LME model was evaluated on the "Time * Experimental Group" interaction term and further investigated using pairwise comparisons between groups. The existence of random intercepts was confirmed through the two-way ANOVA and the scatter plots. Homoscedasticity of within-group errors was assessed using residual *vs.* predicted plots and the Levene test (P=0.108, F-Value=1.923, df=47). The Akaike information criterion (AIC) was employed for validating the selected LME model versus model variations (AIC=1097.931, df=16). The statistical power was assessed by comparing the full model (including the Experimental Group*Time interaction) versus a model without the

interaction term. As a result of the application of 1000 simulations with the powerSim() function of the simr R package [29], we obtained that our approach presented a β = 94.90 % with a 95% confidence interval of [93.35-96.18]. Simulations were run for an alpha value of 0.05, using the 106 available records and elapsed during 0 h 1 min and 47 sec in an i7-8700K CPU 3.70GHz with 6 cores (12 threads) and 32 GB of RAM memory. MLRs were also applied for all calculated deltas in order to study their correlation with the change in S100A4 levels. On measurements showing significant results, cross-sectional MLRs were further investigated at each time point with S100A4 as independent variable. All described analyses were performed in R environment version 3.6.0 [30].

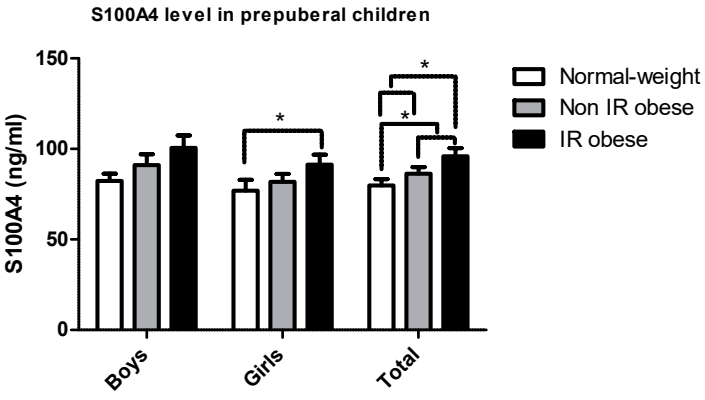


*Figure 2. Group comparisons for S100A4 plasma levels in the prepubertal population of 249 children (study population 1). The one-way ANOVA, Kruskal-Wallis and the Welch test were employed to assess group differences in S100A4 levels according to standard statistical assumptions. The pairwise-t-tests, pairwise Mann–Whitney U-tests and Dunn tests were applied conveniently as post-hoc analyses to determine which experimental groups differed from each other. * refers to comparisons yielding significant results (P < 0.05).*

3. Results

3.1. The cross-sectional study

General characteristics of the 249 children in the cross-sectional study (study population 1) are shown in the Supplementary Table 1. S100A4 plasma levels according to obesity and IR by sex are shown in **Figure 2**. Higher S100A1 plasma levels were observed in girls with IR and obesity when compared with normal-weight non-IR girls (P=0.04). In the same way, higher S100A4 plasma levels were observed for boys with obesity and IR than for normal-weight boys, without statistical significance though. When all subjects of the sample were compared together, we reported significant differences between normal-weight and children with obesity (P=0.02) and between non-IR and IR children (P=0.02), with higher values in the obesity and IR groups, respectively (**Figure 2**).

In order to clarify the relationship between S100A4, IR, and obesity, MLRs were further conducted in a wide range of metabolic outcomes (**Table 1**). The strongest association was found for glucose levels, for which each additional ng/mL of S100A4 in plasma was associated with an increase of 0.05 mg/dL (P=0.005). We also identified significant and positive associations between the S100A4 plasma levels and HOMA-IR (P=0.02) and plasma *s*ICAM1 concentrations (P=0.02), all

of them properly adjusted for confounders (please, see table 1 legend). Otherwise, no significant association was identified with BMI Z-score.

3.2 The longitudinal study

All details regarding the adopted longitudinal design are illustrated in **Figure 1**. Longitudinal within-group and between-group changes for analyzed variables are shown in **Table 2**. Changes in anthropometric variables showed a coherent behavior according to the experimental condition. In particular, for WC, which is a metabolic health indicator in obesity, we found within-group significant increases in groups 1, 4, 5 and 6. The higher change corresponded to the group 5, in which girls with obesity become IR with pubertal maturation. The metabolic health derangement observed in groups 5 and 6 was also confirmed by changes in blood pressure, glucose levels, QUICKI, HOMA-IR and triglycerides (**Table 2**).

Table 1
Association between S100A4 plasma levels and metabolic outcomes in the cross-sectional prepubertal population of 249 children (study population 1), in decreasing order of statistical significance.

Outcome	Beta	SE	CI.LOW	CI.HIGH	T-value	P-Value	FDR
Glucose (mg/dL)	0.05	0.02	0.01	0.08	2.85	0.005	0.15
T4 (ng/dL)	−0.001	4.00E-04	−0.002	−0.0003	−2.58	0.01	0.15
GGT (U/L)	−0.02	0.009	−0.04	−0.003	−2.33	0.02	0.19
HOMA-IR	0.007	0.003	0.001	0.01	2.34	0.02	0.19
sICAM-1 (mg/L)	3,00E-04	1,00E-04	2.00E-05	0.0005	2,14	0.03	0.21
Haematocrit %	−0.01	0.006	−0.023	−0.0008	−2.09	0.04	0.21
Resistin (ug/L)	−0.04	0.02	−0.09	−0.003	−2.08	0.04	0.21
Creatinine (mg/dL)	4.00E-04	2.00E-04	0.0003	9.00E-04	1.94	0.05	0.26
Insulin (mU/L)	0.03	0.02	−0.001	0.06	1.88	0.06	0.27
Leukocytes (cell n°)	71.10	39.73	−0.68	1.489.60	1.78	0.07	0.30
ALT (U/L)	−0.03	0.02	−0.06	0.004	−1.73	0.08	0.31
MPO (ug/L)	−0.15	0.09	−0.33	0.03	−1.66	0.09	0.32
AST (U/L)	−0.02	0.01	−0.05	0.004	−1.64	0.10	0.32
Alkaline phosphatase (U/L)	−0.42	0.29	−0.99	0.15	−1.44	0.15	0.41
Adiponectin (mg/L)	0.03	0.024	−0.014	0.08	1.39	0.17	0.41
MCP1 (ng/L)	0.13	0.09	−0.05	0.31	1.38	0.17	0.41
Follicle_stimulating (U/L)	0.003	0.002	−0.001	0.007	1.37	0.17	0.41
SBP (mm Hg)	−0.04	0.03	−0.09	0.02	−1.35	0.18	0.41
IL8 (ng/L)	−0.005	0.004	−0.01	0.003	−1.23	0.22	0.47
Iron (ug/dL)	0.07	0.06	−0.05	0.19	1.22	0.22	0.47
Ferritin (ng/mL)	−0.06	0.05	−0.17	0.05	−1.12	0.26	0.52
Protein (g/dL)	−0.001	9.00E-04	−0.003	0.0008	−1.05	0.29	0.54
HDLc (mg/dL)	−0.03	0.03	−0.09	0.027	−1.05	0.29	0.54
Haemoglobin (g/dL)	−0.002	0.002	−0.005	0.002	−1.00	0.31	0.55
HDL/LDL	7.00E-04	7.00E-04	−0.0006	0.002	0.99	0.32	0.55
BMI Z-score	0.004	0.004	−0.003	0.01	0.96	0.33	0.55
TSH (mU/L)	0.002	0.003	−0.003	0.007	0.82	0.41	0.65
hsCRP (mg/L)	0.006	0.008	−0.009	0.02	0.81	0.42	0.65
TAG (mg/dL)	0.05	0.064	−0.077	0.17	0.75	0.46	0.65
DBP (mm Hg)	0.02	0.02	−0.03	0.06	0.71	0.48	0.65
QUICKI	−1.00E-04	1.00E-04	−0.0003	0.0001	−0.70	0.48	0.65
Cholesterol (mg/dL)	−0.04	0.06	−0.15	0.07	−0.66	0.51	0.65
LDLc (mg/dL)	−0.03	0.05	−0.14	0.07	−0.65	0.51	0.65
ALR	−0.001	0.002	−0.005	0.003	−0.65	0.51	0.65
Leptin (ug/L)	0.01	0.02	−0.03	0.05	0.65	0.52	0.65
Urea (mg/dL)	−0.009	0.01	−0.04	0.02	−0.62	0.55	0.66
Cortisol (nmol/L)	−0.19	0.32	−0.81	0.44	−0.58	0.56	0.67
WC (cm)	0.003	0.01	−0.02	0.03	0.23	0.82	0.92
Estradiol (ng/L)	−0.004	0.02	−0.04	0.03	−0.20	0.84	0.92
WC/Height	0.00001	1.00E-04	−0.0002	0.0002	0.14	0.89	0.92
tPAI1 (ug/L)	0.004	0.03	−0.06	0.06	0.14	0.89	0.92
TNF-α (ng/L)	5.00E-04	0.004	−0.007	0.007	0.14	0.89	0.92
Uric acid (mg/dL)	−2,00E-04	0.002	−0.004	0.003	−0.12	0.90	0.92
APO A (mg/dL)	−0.002	0.07	−0.14	0.14	−0.03	0.97	0.97

Multiple regression analyses with S100A4 plasma levels as independent variable. Models were adjusted for BMI Z-Score, Sex, Age and Origin (as well as Insulin when the dependent variable was BMI Z-Score; or Height when the dependent variable was Blood pressure). Abbreviations: SE, standard error; CI, confidence interval; AST, aspartate aminotransferase; ALT, alanine aminotransferase; BMI, body mass index; WC, waist circumference; SBP, systolic blood pressure; DBP, diastolic blood pressure; HOMA-IR, homeostasis model assessment for insulin resistance; QUICKI, quantitative insulin sensitivity check index; TAG, triglycerides; HDL-c, high-density lipoproteins-cholesterol; LDL-c, low-density lipoproteins-cholesterol; Apo B, apolipoprotein B; Apo A, apolipoprotein A; ALR, adiponectin leptin ratio; hsCRP, high-sensitivity CRP; MCP-1, monocyte chemoattractant protein 1; GGT, Gamma-glutamyltransferase; TNF-α, tumour necrosis factor alpha; TSH, thyroid-stimulating hormone; IL, interleukin; PAI-1, plasminogen activator inhibitor-1; MPO, myeloperoxidase; sICAM, soluble intercellular cell adhesion molecule-1.

Table 2
Descriptive statistics for the longitudinal population (study population 4).

	NW non-IR no change GROUP 1		OB/OW non-IR to NW non-IR - GROUP 2		OB/OW non-IR no change – GROUP 3		OB/OW IR to non-IR GROUP 4		OB/OW non-IR to IR GROUP 5		OB/OW IR no change GROUP 6		FDR §
	T0	Δ (T1 – T0)	T0	Δ (T1 – T0)	T0	Δ (T1 – T0)	T0	Δ (T1 – T0)	T0	Δ (T1 – T0)	T0	Δ (T1 – T0)	
N	10		6		9		9		12		7		
Age (y)	7.93 (1.16)	6.88 [5.35, 8.42] ^{***ab}	8.92 (1.47)	5.57 [3.26, 7.88] ^{**ab}	7.88 (1.77)	5.35 [3.06, 7.64] ^{***ab}	8.29 (1.36)	5.19 [3.2, 7.18] ^{***b}	7.16 (1.77)	6.59 [5.12, 8.05] ^{***a}	8.21 (0.87)	4.01 [2.82, 5.21] ^{***b}	0.0035
S100A4 (ng/mL)	60, 78 (19, 17) ^{a}	34.77 [14.27, 55.27] ^{**a}	95, 29 (23, 11) ^{ab}	14.21 [−14.46, 42.88] ^{ab}	104, 97 (47, 73) ^{ab}	−4.39 [−40.71, 31.93] ^{ab}	133, 21 (98, 39) ^{b}	−30 [−101.48, 41.48] ^{b}	94, 82 (39, 78) ^{ab}	40.71 [11.68, 69.73] ^{*a}	78, 54 (35, 5) ^{ab}	46.15 [14.73, 77.57] **	0.03 #
Anthropometry													
WC (cm)	56.7 (4.14) ^{de}	9.02 [3.67, 14.36] ^{**b}	76.43 (7.11) ^{d}	−2.56 [−10.92, 5.81] ^{c}	75.28 (9.74) ^{a}	8.9 [−0.35, 18.16] ^{b}	77.2 (6.8) ^{a}	13.19 [2.95, 23.43] ^{*ab}	73.42 (8.51) ^{a}	18.78 [10.66, 26.91] ^{***a}	76.9 (10) ^{a}	15.29 [−1.77, 32.36] ^{**ab}	1.30-E07
HC (cm)	66.16 (3.98) ^{de}	22.48 [17.74, 27.22] ^{***a}	79.52 (4.46) ^{a}	10.87 [4.46, 17.29] ^{*b}	80.46 (5.94) ^{a}	21.58 [10.22, 32.93] ^{**ab}	84.84 (7.7) ^{a}	19.21 [8.48, 29.95] ^{**ab}	77.48 (6.32) ^{a}	29.5 [21.08, 37.92] ^{***a}	85.17 (5.6) ^{a}	19.88 [10.45, 29.32] ^{***ab}	0.0009
BMI Z-Score	−0.33 (0.33) ^{b}	0.24 [−0.17, 0.65] ^{a}	1.16 (0.87) ^{a}	−1.17 [−2.08, −0.26] ^{*b}	2.26 (1.02) ^{a}	0.14 [−0.95, 1.24] ^{a}	2.12 (0.78) ^{a}	0.23 [−0.94, 1.41] ^{a}	2.28 (1.18) ^{a}	0.73 [−0.33, 1.78] ^{a}	2.4 (0.76) ^{a}	0.38 [−0.96, 1.72] ^{a}	0.0003
Biochemistry													
DBP (mm Hg)	67.2 (5.96)	1.5 [−6.06,9.06] ^{ab}	64.5 (5.82)	−3.6 [−12.49, 5.29] ^{a}	58.44 (6.75)	16.5 [3.25, 29.75] ^{*c}	67.75 (9.85)	3.75 [−6.34, 13.84] ^{abc}	61.67 (8.72)	9 [1.44, 16.56] ^{*bc}	56.86 (4.85)	10.79 [2.32, 19.25] ^{*bc}	0.0004
SBP (mm Hg)	105 (9.33)	−0.8 [−9.92, 8.32] ^{b}	105.67 (12.23)	−1.57 [−18.23, 15.1] ^{b}	106.56 (7.37)	0.72 [−9.08, 10.52] ^{b}	111.12 (10.7)	0.59 [−11.71, 12.89] ^{ab}	102.83 (8.34)	12.54 [5.81, 19.28] ^{***a}	101.43 (15.41)	6.71 [−10.53, 23.96] ^{ab}	0.001
Glucose (mg/dL)	85.4 (5.7)	0.2 [−0.27, 0.67] ^{ab}	83.17 (6.82)	−0.07 [−0.53, 0.38] ^{bc}	76.44 (9.7)	0.59 [0.12, 1.05] ^{**a}	82.44 (8.86)	−0.27 [−0.71, 0.18] ^{*c}	81.75 (5.5)	0.34 [0.05, 0.64] ^{*ab}	86 (6.66)	0.38 [−0.05, 0.81] ^{*ab}	6.14E-07
Insulin (mU/L)	3.58 [1.99–8.5] ^{a}	5.29 [2.16, 8.41] ^{**b}	6.2 [2.9–11.3] ^{a}	2.98 [−1.88, 7.84] ^{b}	7.41 [2.2–10.4] ^{a}	4.31 [1.28, 7.34] ^{**b}	16.7 [12.8–32.4] ^{b}	−7.24 [−14.22, −0.25] ^{c}	8.77 [3.2–11.8] ^{a}	20.23 [13.7, 26.75] ^{***a}	16.8 [12.3–27.5] ^{b}	9.13 [1.32, 16.93] ^{**ab}	8.60E-09
QUICKI	0.4 [0.35–0.46] ^{a}	−0.05 [−0.08, −0.02] ^{**bc}	0.37 [0.34–0.41] ^{a}	−0.02 [−0.05, 0.02] ^{b}	0.36 [0.35–0.45] ^{a}	−0.04 [−0.07, −0.01] ^{**bc}	0.32 [0.29–0.33] ^{b}	0.02 [0, 0.05] **	0.35 [0.34–0.42] ^{a}	−0.06 [−0.08, −0.04] ^{***c}	0.31 [0.3–0.33] ^{b}	−0.02 [−0.03, −0.01] ^{**b}	1.10E-08
HOMA	0.92 (0.48) ^{a}	1.2 [0.49, 1.91] ^{**b}	1.37 (0.53) ^{a}	0.61 [−0.34, 1.55] ^{b}	1.2 (0.52) ^{a}	1.12 [0.44, 1.8] ^{**b}	4.1 (1.75) ^{b}	−1.63 [−3.06, −0.19] ^{c}	1.65 (0.57) ^{a}	4.6 [2.97, 6.24] ^{***a}	3.7 (0.97) ^{b}	2.42 [0.7, 4.13] ^{**b}	8.60E-09
Total Cholesterol (mg/dL)	172.9 (30.79)	−8.4 [−36.77, 19.97]	168.17 (38.47)	−6 [−47.13, 35.13]	154.67 (35.25)	−6.22 [−38.41, 25.97]	153.67 (15.67)	−7.44 [−22.66, 7.77]	156.83 (24.19)	−0.33 [−21.21, 20.55]	170.43 (20.57)	−9.71 [−31.48, 12.05]	0.76
TAG (mg/dL)	48.1 (6.28) ^{a}	28.5 [0.33, 56.67] ^{a}	46.67 (10.58) ^{a}	6 [−10.77, 22.77] ^{ab}	58.89 (20.03) ^{a}	14.33 [−11.52, 40.18] ^{ab}	75.11 (27.93) ^{a}	−12.78 [−39.2, 13.65] ^{b}	67.33 (40.77) ^{a}	33.92 [−4.23, 72.06] ^{**a}	84.86 (46.7) ^{a}	17.57 [−35.04, 70.18] ^{ab}	0.0009
HDLc (mg/dL)	67.56 (16.58) ^{a}	−7.06 [−22.21, 8.1]	57.83 (5.67) ^{ab}	−2.83 [−15.2, 9.54]	50.11 (12.17) ^{ab}	−3.67 [−13.9, 6.56]	50.56 (12.58) ^{ab}	1.67 [−15.82, 19.15]	47.42 (13.81) ^{b}	−2.83 [−12.6, 6.94]	51.71 (11.67) ^{ab}	−10.29 [−28.87, 8.3]	0.22
LDLc (mg/dL)	99.89 (30.91)	−11.23 [−40.96, 18.5]	92.83 (30.78)	−2.33 [−34.75, 30.08]	92.44 (26.94)	−6.8 [−32.42, 18.82]	83.78 (15.74)	−0.33 [−16.05, 15.38]	93.92 (22.2)	−3.73 [−22.68, 15.22]	96.29 (23.5)	−8.71 [−30.83, 13.4]	0.71
AST (U/L)	27 [24–33]	−7.9 [−11.07, −4.73] ^{*bc}	30 [22–34]	−7.83 [−13.51, −2.15] ^{*c}	25.5 [14–51]	−9.28 [−20.33, 1.78] **	24 [14–29]	−0.01 [−7.42, 7.4] ^{a}	25.5 [17–40]	−4.5 [−9.15, 0.15] ^{*ab}	23 [17–31]	−5.91 [−12.92, 1.09] ^{**c}	0.013
ALT (U/L)	15 [11–25]	−4.6 [−8.57, −0.63] ^{*ab}	24 [11–33]	−9.13 [−18.67, 0.4] **	15 [13–56]	−6.76 [−18.82, 5.29] ^{ab}	20 [14–38]	−3.67 [−11.84, 4.51] ^{ab}	20 [12–29]	−1.83 [−6.09, 2.42] ^{b}	17 [13–40]	−6.63 [−16.99, 3.73] ^{ab}	0.02
GGT (U/L)	10.4 (2.07)	−0.1 [−1.91, 1.71]	11.67 (3.01)	−0.5 [−5.11, 4.11]	11.67 (2.65)	−0.67 [−3.29, 1.95]	14.12 (4.12)	−0.35 [−6.15, 5.45]	13.25 (4.03)	1.33 [−1.82, 4.49]	14.71 (1.6)	−1.71 [−4.22, 0.79]	0.35
FSH (U/l)	1.03 [0.45–4.67]	3.96 [2.62, 5.29] ^{***ab}	1.2 [0.5–4.5]	3.23 [0.75, 5.72] ^{**ab}	1.9 [0.73–3.4]	2.99 [1.57, 4.4] ^{***b}	1.3 [0.4–5.16]	2.78 [0.65, 4.91] ^{*ab}	1.12 [0.1–2]	4.51 [3.3, 5.71] ^{***a}	2.13 [0.95–4.12]	2.47 [0.11, 4.82] ^{b}	0.06
Cortisol (nmol/L)	404.2 [189.8–598.7]	62.3 [−63.33, 187.93]	251.6 [220.8–480.1]	16.46 [−134.24, 167.16]	248 [176.6–626.5]	44.66 [−97.38, 186.7]	354.5 [8.5–623.8]	47.88 [−168.52, 264.28]	237.3 [107.6–676.2]	100.86 [−45.75, 247.46]	325.7 [153.5–543.7]	35.76 [−141.12, 212.65]	0.78
Estradiol (ng/L)	19 [5–40]	47.99 [14.01, 81.96] ^{**b}	5 [4–19]	66.55 [−39.79, 172.89] ^{a}	17.5 [9–44]	33.46 [−27.55, 94.46] ^{b}	23 [4–39.3]	17.79 [1.3, 34.27] ^{ab}	9 [5–19]	38.27 [21.7, 54.84] ^{***a}	28 [5–67]	52.32 [−51.52, 156.16] ^{ab}	0.002
hsCRP (mg/L)	0.3 [0.1–8.5]	−1.54 [−3.45, 0.36]	1.19 [0.3–9.74]	−3.59 [−8.26, 1.07]	0.95 [0.2–2.9]	−1.05 [−1.95, −0.14]	0.89 [0.36–6.7]	−1.59 [−3.19, 0.01]	2.5 [0.9–2.7]	−1.6 [−2.16, −1.05] **	2.05 [1.6–2.6]	−1.5 [−2.35, −0.65] *	0.15

Baseline data are expressed as mean (standard deviation) or median [min-max] if not normally distributed. For Δ (T1 – T0) changes, data are expressed as mean change accompanied by [CI low, CI high]. Distributions within the same row with unlike superscript letters were significantly different (P < 0.05). * for P ≤ 0.05 in within-group changes (Δ) from start. ** for P ≤ 0.01 in for within-group changes (Δ) from start. *** for P ≤ 0.0001 in for within-group changes (Δ) from start. Within-group changes from baseline (T_0) to puberty (T_1) were assessed by means of a paired design in all continuous variables (including S100A4 levels); employing either a paired *t*-test or a Wilcoxon signed-rank test. Between-group differences were assessed by the one-way ANOVA, Kruskal-Wallis or Welch tests to the computed delta values (T_1–T_0) for each continuous measurement. # Refers to between-group differences in S100A4 levels, which were investigated by means of a linear mixed-effects (LME) model adjusting for age and tanner stage and including time as fixed effects term and a random intercept for each patient. § refers to FDR for between group Delta comparisons. Abbreviations: ALR, adiponectin leptin ratio; Apo A, apolipoprotein A; Apo B, apolipoprotein B; BMI, body mass index; CI, confidence interval; DBP, diastolic blood pressure; FSH, follicle stimulating hormone; HC, hip circumference; HDL-c, high-density lipoproteins-cholesterol; HOMA-IR, homeostasis model assessment for insulin resistance; hsCRP, high-sensitivity C reactive protein; LDL-c, low-density lipoproteins-cholesterol; QUICKI, quantitative insulin sensitivity check index; SBP, systolic blood pressure; TAG, triglycerides; WC, waist circumference.

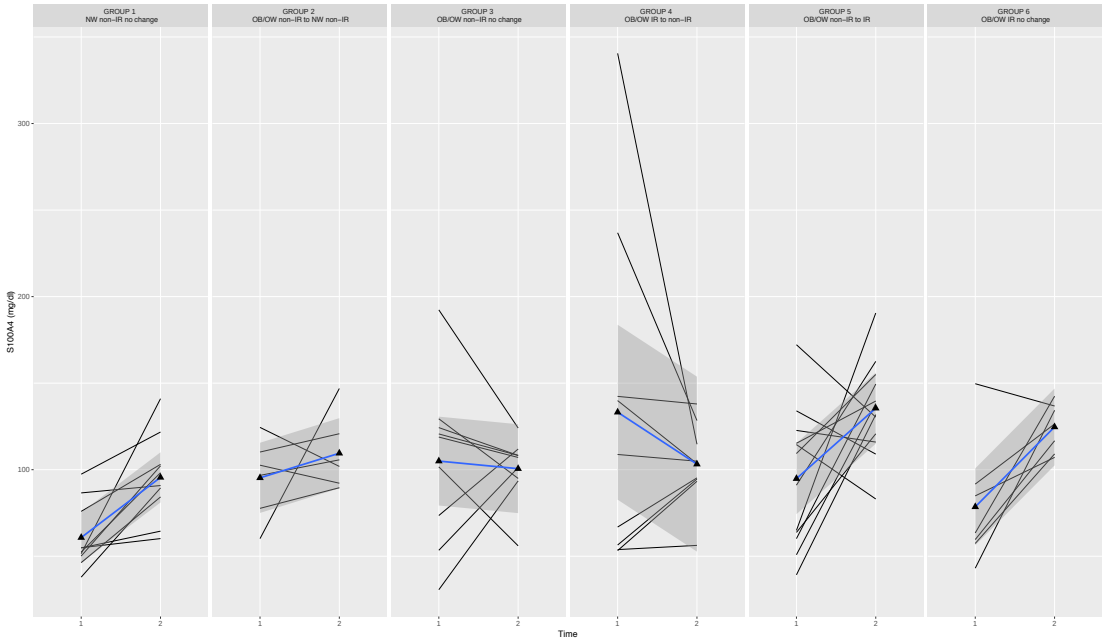


Figure 3. Longitudinal trajectories in S100A4 plasma levels by experimental condition in the longitudinal study (study population 4). Between-group differences in S100A4 levels were investigated by means of a linear mixed-effects model including the covariates age, Tanner stage and time as fixed effects and a random intercept for each participant. Test significance was evaluated on the "Time x Experimental Group" interaction term and further investigated using pairwise comparisons between groups.

Regarding S100A4 plasma levels at baseline, we found significant lower levels in the normal-weight group (group 1) than in the obese with IR group (group 4) (**Table 2** and **Figure 3**). Concerning longitudinal S100A4 changes, we identified significant within-group differences for groups 1, 5, and 6 (FDR=0.002 for group 1; FDR= 0.02 for group 5; FDR=0.01 for group 6). Interestingly, the observed increase in healthy normal-weight girls put the S100A4 levels in similar values to baseline levels of girls with obesity (Figure 3). These reached levels in healthy normal-weight pubertal girls were comparable to the mean S100A4 levels observed in healthy adult woman populations [8], suggesting that pubertal development is an important step for the stabilization/destabilization of S100A4 levels and the further appearance of its related phenotypes. Regarding between-group differences, an LME model reported a significant association (F-Value=2.72 and P=0.03) between the interaction term "Time*Experimental group" and the S100A4 levels (**Table 2**). The statistical power of the approach was evaluated obtaining a β = 94.90 % with a 95% confidence interval of [93.35-96.18]. Post-hoc pairwise comparisons between experimental groups revealed a strong association between a worsening/improvement of the IR status and the increase/decrease in S100A4 levels, yielding significant results in 4 of the 15 tested comparisons (P=0.02 for group 1-vs-group 4, P=0.03 for group 3-vs-group 6, P=0.01 for group 4-vs-group 5, and P=0.004 for group 4-vs-

group 6) (**Figure 3** and **Table 2**). In order to validate these findings, we further applied MLRs with deltas for continuous measurements as input variables (computed as T_1-T_0) (**Table 3**). We found highly significant positive correlations between the increase in S100A4 levels and the increase in HOMA-IR (P=0.0003, FDR=0.005) and insulin (P=0.0003, FDR=0.005) during the course of puberty (**Figure 4**). A significant inverse correlation was also reported with the change in HDLc levels (P=0.003, FDR=0.03). Otherwise, no significant association was identified with the change in BMI Z-score.

In the pubertal stage of the longitudinal population, we identified strong associations between S100A4 levels and HOMA-IR (P=0.003) and QUICKI (P=0.024) (**Table 4**), reporting effect sizes and significant values comparable to previous findings from adult populations [8], while a null association with BMI Z-score was reported. Details for adjusting covariates in all tested models can be found in the table legends or in the method section.

Table 3
Association between the change in S100A4 plasma levels and the change in 38 metabolic outcomes in the longitudinal PUBMEP population (study population 4).

Outcome	Beta	SE	CI.LOW	CI.HIGH	T-value	P-Value	FDR
HOMA-IR	0.02	0.006	0.01	0.03	40.23	0.0003	0.005
Insulin (mU/L)	0.09	0.02	0.05	0.15	39.56	0.0003	0.005
HDLc (mg/dL)	−0.12	0.04	−0.20	−0.05	−31.85	0.003	0.03
Glucose (mg/dL)	0.05	0.02	0.002	0.09	20.44	0.05	0.37
LH (U/L)	−0.03	0.01	−0.06	−4.00E-04	−1.99	0.06	0.39
WC/ Height	−5.00E-04	3.00E-04	−0.001	1.00E-04	−17.38	0.09	0.49
BMI Z-Score	−0.005	0.003	−0.01	8.00E-04	−16.90	0.09	0.49
Estradiol (ng/L)	0.40	0.28	−0.14	0.94	14.42	0.16	0.52
Cortisol (nmol/L)	−0.73	0.5	−17.39	0.27	−14.29	0.16	0.52
hsCRP (mg/L)	−0.01	0.008	−0.03	0.004	−14.22	0.17	0.52
FSH (U/L)	0.009	0.007	−0.004	0.02	13.95	0.17	0.52
SBP (mm Hg)	0.06	0.04	−0.02	0.14	13.58	0.18	0.52
Alkaline phosphatase (U/L)	0.39	0.32	−0.23	1.02	12.28	0.23	0.60
WC (cm)	−0.04	0.04	−0.12	0.03	−1.12	0.27	0.67
Uric acid (mg/dL)	−0.003	0.003	−0.008	0.0027	−0.94	0.35	0.77
QUICKI	−1.00E-04	1.00E-04	−4.00E-04	1.00E-04	−0.92	0.36	0.77
Hip Circumference (cm)	−0.02	0.03	−0.08	0.03	−0.91	0.37	0.77
Urea (mg/dL)	−0.04	0.05	−0.13	0.05	−0.85	0.40	0.77
TSH (mU/L)	0.006	0.007	−0.007	0.02	0.84	0.40	0.77
HDL/LDL	−89.54	114.50	−313.96	134.88	−0.78	0.44	0.78
Protein (g/dL)	8.00E-04	0.001	−0.001	0.00	0.76	0.45	0.78
Waist/Hip	−2.00E-04	3.00E-04	−8.00E-04	3.00E-04	−0.74	0.47	0.78
AST (U/L)	0.02	0.02	−0.03	0.06	0.68	0.50	0.78
Haemoglobin (g/dL)	0.001	0.002	−0.003	0.005	0.65	0.52	0.78
Haematocrit %	0.02	0.02	−0.03	0.06	0.64	0.53	0.78
DBP (mm Hg)	−0.22	0.49	−11.78	0.73	−0.45	0.65	0.90
LDLc (mg/dL)	−0.015	0.04	−0.096	0.06	−0.38	0.71	0.90
Iron (ug/dL)	0.04	0.09	−0.16	0.23	0.37	0.72	0.90
Ferritin (ng/mL)	−0.02	0.06	−0.13	0.09	−0.34	0.74	0.90
T4 (ng/dL)	2.00E-04	7.00E-04	−0.001	0.002	0.32	0.75	0.90
GGT (U/L)	−0.003	0.01	−0.02	0.02	−0.30	0.77	0.90
Leukocytes (cell n°)	31.7	106.68	−177.32	240.87	0.30	0.77	0.90
APO A (mg/dL)	−0.02	0.13	−0.28	0.23	−0.18	0.86	0.96
CHOL_ (mg/dL)	0.009	0.05	−0.09	0.11	0.16	0.87	0.96
Creatinine (mg/dL)	0	3.00E-04	−5.00E-04	6.00E-04	0.14	0.89	0.96
APO B (mg/dL)	0.01	0.17	−0.32	0.34	0.08	0.94	0.97
ALT (U/L)	−0.001	0.02	−0.04	0.04	−0.05	0.96	0.97
TAG (mg/dL)	0.003	0.09	−0.18	0.19	0.03	0.97	0.97

Multiple regression analyses with the change in S100A4 plasma levels as independent variable. Models were adjusted for the change in BMI Z-Score, the origin, the elapsed time from baseline to T1, and the pubertal stage reached. When the dependent variable was the change in the BMI Z-Score, we replaced the BMI confounder with the change in Insulin levels. The change in Height was further included in models when the dependent variable was change in Blood pressure. Abbreviations: SE, standard error; CI, confidence interval; AST, aspartate aminotransferase; ALT, alanine aminotransferase; BMI, body mass index; WC, waist circumference; SBP, systolic blood pressure; DBP, diastolic blood pressure; HOMA-IR, homeostasis model assessment for insulin resistance; QUICKI, quantitative insulin sensitivity check index; TAG, triglycerides; HDL-c, high-density lipoproteins-cholesterol; LDL-c, low-density lipoproteins-cholesterol; Apo B, apolipoprotein B; Apo A, apolipoprotein A; ALR, adiponectin leptin ratio; hsCRP, high-sensitivity CRP; MCP-1, monocyte chemoattractant protein 1; GGT, Gamma-glutamyltransferase; TNF-α, tumour necrosis factor alpha; TSH, thyroid-stimulating hormone; IL, interleukin; PAI-1, plasminogen activator inhibitor-1; MPO, myeloperoxidase; sICAM, soluble intercellular cell adhesion molecule-1.

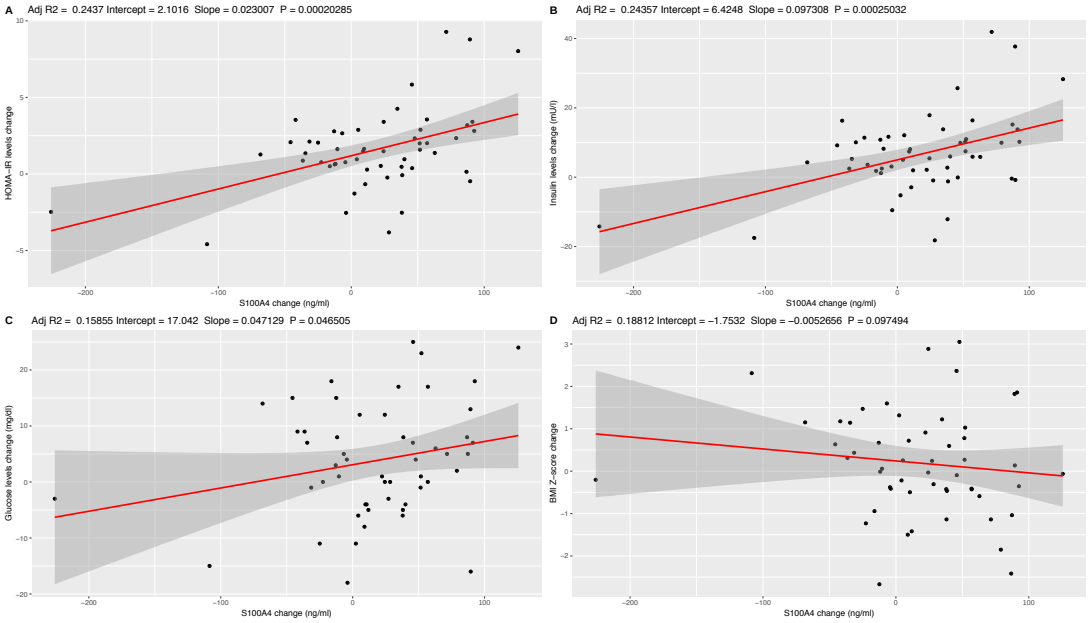


Figure 4. Multiple linear regressions analyses between the change in S100A4 plasma levels and the changes in glucose metabolism outcomes and BMI Z-Score in the longitudinal cohort (study population 4). Deltas were calculated as T1 – T0. Figure A reports the linear model with delta HOMA-IR as dependent variable, B refers to the model with delta insulin levels as dependent variable, C the model for delta glucose levels and D the one for delta BMI Z-score.

3.3 *S100A4* DNA methylation

Fourteen methylation sites were selected from the Infinium Methylation EPIC microarray of which thirteen were annotated as promoter associated CpGs (Supplementary Table 4). All the CpGs were annotated as open sea. At baseline, we found a positive significant association between the methylation status of the probe cg07245635 and HOMA-IR, insulin levels, and S100A4 plasma levels (**Figure 5**, and Supplementary Table 5). Interestingly, the association with plasma S100A4 remained statistically significant after adjusting for HOMA-IR (Supplementary Table 5). At the pubertal stage, we found a significant inverse association between the methylation status of the probe cg10447638 and HOMA-IR, insulin levels, and S100A4 plasma levels (**Figure 5** and Supplementary Table 5). These cross-sectional findings from pre- and pubertal stages were further validated at the longitudinal level by applying MLRs to delta measurements (computed as T_1-T_0) (Supplementary Figure 1 and Supplementary Figure 2).

Following the same experimental design than in the longitudinal approach, we studied within-group changes in the DNA methylation status of analyzed probes. Both cross-sectional findings were validated with this approach (Supplementary table 6). For the cg07245635, we reported a significant fold change of 1.25 more methylation in the pubertal stage than in baseline

Table 4
Association between S100A4 plasma levels and metabolic outcomes in the pubertal stage of the longitudinal PUBMEP population (study population 4).

Outcome	Beta	SE	CI.LOW	CI.HIGH	T-value	P-Value	FDR
HOMA-IR	0.0326	0.0106	0.0119	0.0533	3089	0.003	0.13
Insulin (mU/L)	0.1324	0.0449	0.0444	0.2205	2948	0.005	0.13
Haematocrit %	9.17E−02	3.40E−02	2.51E−02	1.58E-01	2.7	0.010	0.17
QUICKI	-0.0003	0.0001	-0.0006	-0.0001	−2344	0.024	0.28
Luteinizing hormone (U/L)	0.0413	0.0181	5.90E−03	0.0767	2284	0.027	0.28
Glucose (mg/dL)	0.0877	0.047	−0.0046	1.80E−01	1863	0.069	0.51
Glucose (mmol/L)	0.0049	0.0026	−0.0003	0.01	1863	0.069	0.51
Haemoglobin (g/dL)	0	0	−0.0013	0.0155	1664	0.103	0.60
GGT (U/L)	0.0282	0.018	−0.0072	0.0635	1562	0.126	0.60
APO B/APO A Ratio	-0.0022	0.0014	-0.0049	0.0006	−1563	0.135	0.60
Prolactin (ng/mL)	0.078	0.0521	−0.0241	0.18	1498	0.142	0.60
TSH (mU/L)	−0.0121	0.0082	−0.0282	0.0041	−1466	0.150	0.60
T4 (ng/dL)	−0.0121	0.0082	−0.0282	0.0041	−1466	0.150	0.60
Hip-Circumference (cm)	4.01E−02	0.0301	−0.0188	0.099	1333	0.189	0.67
Waist/Hip Ratio	−0.0005	0.0004	-0.0013	0.0002	−1321	0.193	0.67
Urea (mg/dL)	−0.0343	0.028	−0.0892	0.0206	−1226	0.227	0.71
ALT (U/L)	3.14E−02	0.0267	-0.0209	0.0837	1178	0.245	0.71
Leukocytes (cell n°)	−78.976	69.203	−21.4613	56.662	−1141	0.26	0.709
Cholesterol (mg/dL)	0.1616	0.1464	−0.1254	0.4486	1103	0.276	0.709
WC/Height Ratio	−0.0003	0.0002	−0.0008	0.0002	−1085	0.284	0.709
ASTUI_T2	0.0262	0.0243	−0.0214	0.0739	1.08	0.286	0.709
Ferritin (ng/mL)	−0.1257	0.1206	−0.3621	0.1106	−1043	0.303	0.716
TAG (mg/dL)	0.229	0.2274	−0.2167	0.6748	1007	0.319	0.718
APO B (mg/dL)	−0.1633	0.1638	−0.4843	0.1577	−0.997	0.331	0.718
HDLc (mg/dL)	−0.0638	0.072	−0.205	0.0774	−0.885	0.381	0.723
HDL/LDL Ratio	−0.0011	0.0013	−0.0037	0.0014	−0.877	0.385	0.723
APO_B/LDL Ratio	−0.0011	0.0013	−0.0037	0.0014	−0.877	0.385	0.723
Hydroxyprogesterone (ng/mL)	0.007	0.0087	−0.01	0.024	0.804	0.426	0.723
Free testosterone (pg/mL)	0.007	0.0087	−0.01	0.024	0.804	0.426	0.723
Transferrin (mg/dL)	0.1866	0.2321	−0.2683	0.6415	0.804	0.426	0.723
SBP (mm Hg)	0.0428	0.0548	−0.0647	0.1502	0.78	0.44	0.723
iron (ug/dL)	0.1274	0.1672	−0.2004	0.4552	0.762	0.45	0.723
hsCRP (mg/dL)	0.0012	0.0016	−0.002	0.0044	0.748	0.459	0.723
Follicle stimulating (U/L)	0.0077	0.0108	−0.0135	0.0288	0.712	0.48	0.734
Sum of Folds	0.1005	0.1476	−0.1887	0.3898	0.681	0.499	0.742
Uric acid (mg/dL)	0.0034	0.0058	−0.008	0.0148	0.586	0.561	0.799
Total proteins (g/dL)	0.0012	0.0024	−0.0035	0.0059	0.505	0.616	0.843
BMI Z-score	−0.004	0.01	−0.024	0.0155	−0.419	0.676	0.901
Dehydroepiandrosterone sulfate (ng/mL)	−20.677	70.321	−15.8505	11.7152	−0.294	0.772	0.923
Total Testosterone (ng/mL)	0.0002	0.0006	−0.0011	0.0014	0.268	0.79	0.923
DBP (mm Hg)	−0.0127	0.0504	−0.1114	0.0861	−0.251	0.803	0.923
Waist circumference (cm)	−0.0097	0.0416	−0.0913	0.0719	−0.233	0.817	0.923
Alkaline phosphatase (U/L)	0.1286	0.5705	−0.9896	1.2468	0.225	0.823	0.923
Creatinine (mg/dL)	−0.0001	0.0005	−0.0011	0.0009	−0.194	0.847	0.923
Cortisol (mg/dL)	0.0066	0.035	−0.0621	0.0752	0.188	0.852	0.923
Estradiol (ng/L)	−0.0445	0.3246	−0.6807	0.5917	−0.137	0.892	0.946
APO A (mg/dL)	−0.0199	0.226	−0.4629	0.4231	−0.088	0.931	0.957
LDL (mg/dL)	−0.0102	0.132	−0.2689	0.2486	−0.077	0.939	0.957
LDL/APO B Ratio	−0.0001	0.0017	−0.0034	0.0032	−0.054	0.957	0.957

Models were adjusted for BMI Z-Score, Sex, Age, Pubertal Status and Origin. The Height was further included in models when the dependent variable was Blood pressure. Instead, BMI was replaced by Insulin when the dependent variable was BMI Z-Score. Abbreviations: SE, standard error; CI, confidence interval; AST, aspartate aminotransferase; ALT, alanine aminotransferase; BMI, body mass index; WC, waist circumference; SBP, systolic blood pressure; DBP, diastolic blood pressure; HOMA-IR, homeostasis model assessment for insulin resistance; QUICKI, quantitative insulin sensitivity check index; TAG, triglycerides; HDL-c, high-density lipoproteins-cholesterol; LDL-c, low-density lipoproteins-cholesterol; Apo B, apolipoprotein B; Apo A, apolipoprotein A; ALR, adiponectin leptin ratio; hsCRP, high-sensitivity CRP; MCP-1, monocyte chemoattractant protein 1; GGT, Gamma-glutamyltransferase; TNF-α, tumour necrosis factor alpha; TSH, thyroid-stimulating hormone; IL, interleukin; PAI-1, plasminogen activator inhibitor-1; MPO, myeloperoxidase; sICAM, soluble intercellular cell adhesion molecule-1.

for the group 3, which corresponds to non-IR prepubertal girls with obesity that remained with insulin-sensitivity after puberty entrance. On the other hand, for the cg10447638, we revealed a significant fold change of 0.69 less methylation in the pubertal stage comparing with the baseline levels of the group 6, which corresponds to IR prepubertal girls with obesity for which IR remained after puberty onset. No significant results were obtained for the rest of experimental groups nor analyzed probes.

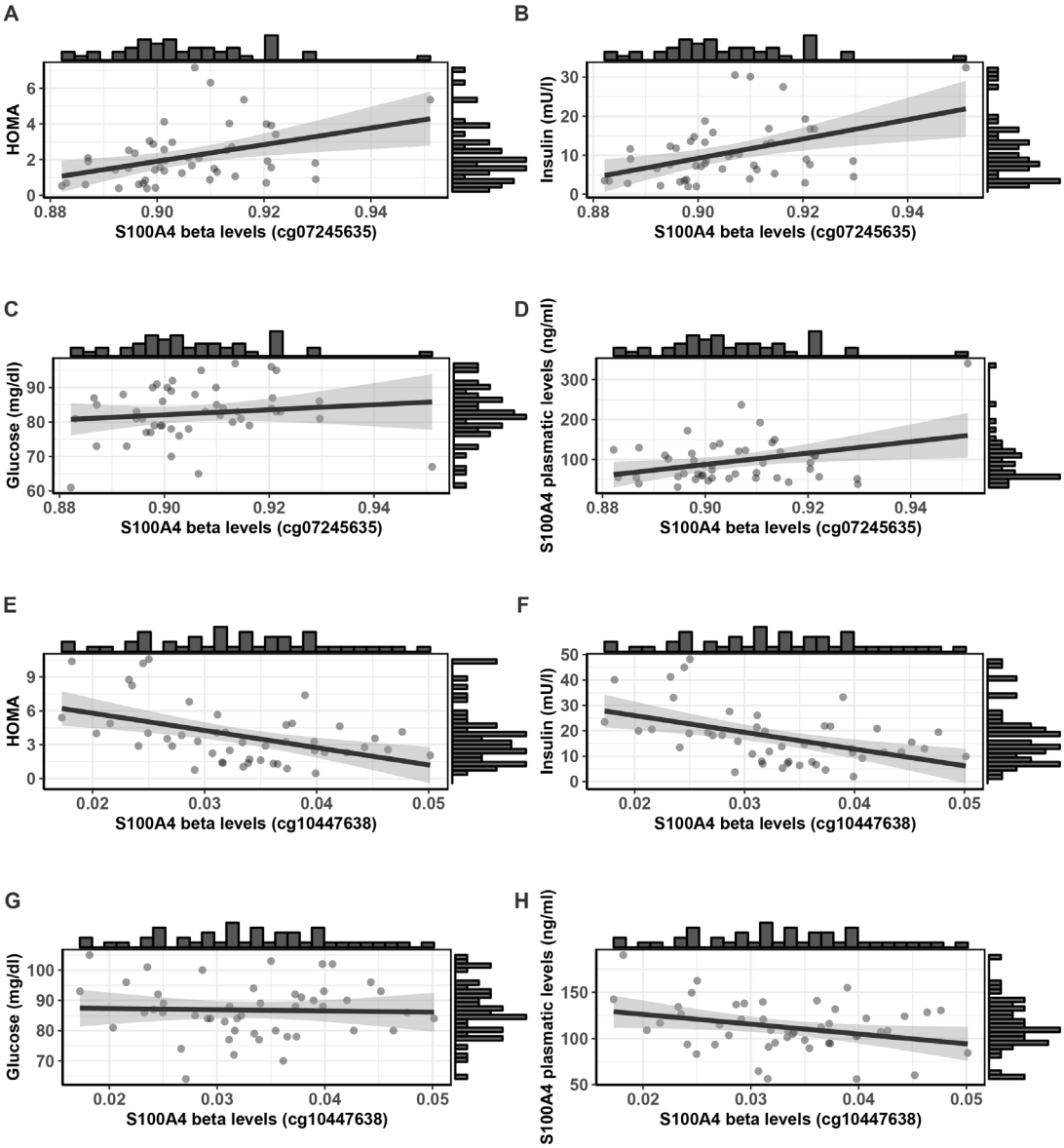


Figure 5. Cross-sectional associations between S100A4 DNA methylation status and glucose metabolism outcomes. Figures A to D refer to baseline (prepubertal) data while figures E to H refer to the pubertal stage. Multiple linear regressions were employed with M values as independent variables and each outcome as dependent variable. Models were adjusted for confounders when necessary (Supplementary Table 5). Because percentage methylation is easily interpretable, beta values were employed for graphical representation of results.

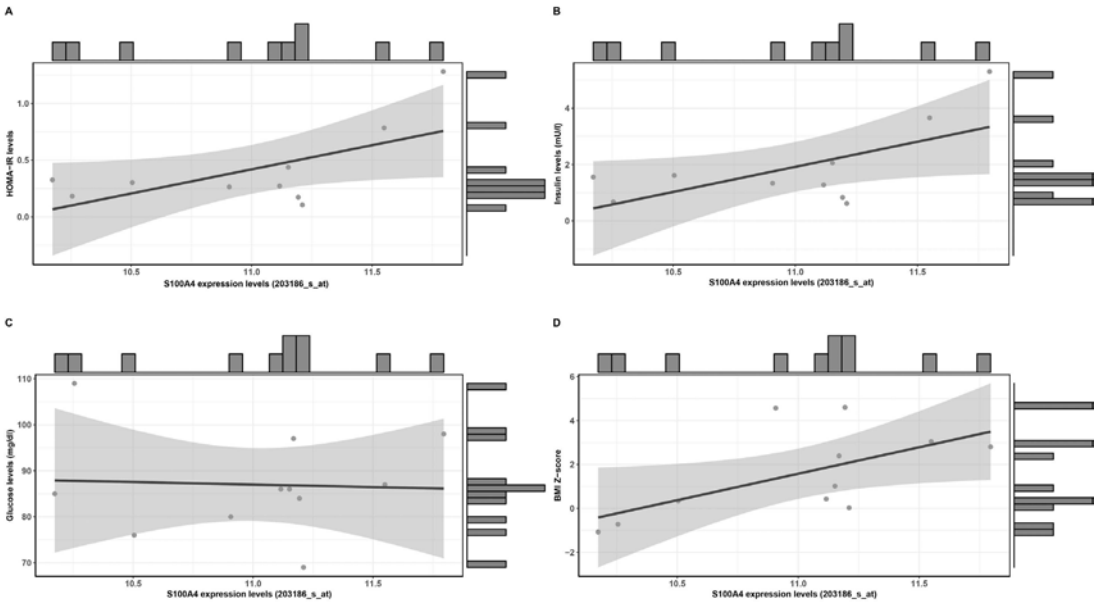


Figure 6. Visceral white adipose tissue genetic expression analyses in study population 2 for the S100A4 probe "203186_s_at". Robust Multichip Average-normalized gene expression from the "203186_s_at" probe was tested for association with each continuous outcome; A for HOMA-IR, B for insulin levels, C for glucose levels and D for BMI Z-score. These analyses were adjusted for confounders when necessary.

3.4 S100A4 gene expression

Descriptive statistics for the study population 2 employed in gene expression analyses are available in the Supplementary Table 2. The *203186_s_at* probe, localized in the *S100A4* locus, was found 1.6 times up-regulated in vWAT samples of children with obesity compared with their normal-weight counterparts (Signal Log Ratio=0.67, Average Normalized Expression=11.03, P=0.007 and $FDR_{whole\text{-}array}$=0.78). This finding was supported by additional correlations with HOMA-IR, insulin levels, and the BMI Z-score (Adj-R2=0.63, Slope [CI]=0.46 [0.13, 0.79] and P=0.03; Adj-R2=0.56, Slope [CI]=1.91 [0.39, 3.43] and P=0.04; and Adj-R2=0.30, Slope [CI]=2.41 [0.36, 4.46] and P=0.04, respectively in each confounding-adjusted model) (**Figure 6**). We did not find any correlation between *S100A4* gene expression and S100A4 plasma levels (AdjR2 = -0.11, Slope [CI]=2.16 [-52.85, 57.18] and P=0.94 in the adjusted model), which might suggest that AT is not the main contributor to the systemic S100A4 levels in humans. A descriptive statistic of the population (study population 4) used for the validation of microarray results by qPCR analyses is available in the Supplementary Table 3. Regarding the relationship between qPCR *S100A4* expression and the studied biomarkers, we observed a trend in the association with HOMA-IR (AdjR2 = 0.03, Slope [CI]=0.03 [-0.002,0.05] and P=0.09 in the adjusted model), while no association with the rest of studied markers was found (Supplementary table 7).

In the Study population 2, we chose the BMI Z-Score and age as confounders for the models with HOMA-IR and Insulin as dependent variables, the variables Insulin and Age as confounders for the model with BMI Z-Score as dependent variable, and the variable age for adjusting the model with S100A4 plasma levels as the dependent variable. For the study population 3, confounder models were the same than in the study population 2 but with the inclusion of sex (since in this population we counted on data for both sexes).

4. Discussion

In the present work, we show a strong association between S100A4 plasma levels and a bulk of metabolic and anthropometric outcomes, with special relevance of IR status in children and adolescents with obesity. Our findings illustrate how this protein can be found in high levels already in the prepubertal stage of non-diabetic children with obesity, and how the evolution in S100A4 levels is related to trends in the IR status of children during sexual maturation.

Previously, high serum levels of this protein have been associated with a greater prevalence of type 2 diabetes, obesity and IR in adult populations [9], as well as proposed as a circulating marker of hepato-visceral fat excess in adolescents [10]. Now, it is the first time that the relationship S100A4-IR is investigated in a prepubertal population and during the first stages of puberty. The motivation for focusing in the course of puberty in the present work lies in the fact that sexual maturation has been presented as a significant metabolic risk period for children with obesity [31]. As far as we know, this work is also the first to address the role of S100A4 in IR through a multi-omics approach. As a result, we provide interesting knowledge into the plausible molecular mechanisms underlying this association.

In a prepubertal sample of 249 children (study population 1), we showed a strong association between S100A4 plasma levels and glucose, and a weaker association with IR (assessed by the HOMA-IR index), independently of BMI (**Table 3**). Group comparisons for S100A4 levels also revealed significant results, although only when comparing extreme experimental conditions (normal-weight *vs.* obese with IR) (**Figure 2** and **Table 2**). In view of the results, it could be elucidated that the increase in S100A4 levels in the prepubertal stage of obese children is directly related to the pre-IR state of children with obesity rather than to the excess of adiposity itself. When analyzing the pubertal cross-data of a longitudinal cohort instead (study population 4), robustly significant associations were reported between the S100A4 levels and IR outcomes, whereas no association was detected with the BMI Z-Score **(Table 3**). MLRs for the change in obesity-related outcomes further corroborated the association with IR and discarded associations with the change in the BMI Z-Score **(Table 4**). Considering these results, we might state that S100A4 is apparently involved in the generation of an IR status in children and adolescents with obesity. Regarding literature, Arner *et al.* [9] demonstrated that S100A4 is associated with IR, type 2 diabetes and a pernicious

adipose phenotype (in a BMI-independent manner) in adult obese populations. Particularly, they showed strong associations with fat cell size in a BMI-independent manner. They also showed general differences in circulating S100A4 between non-obese and obese subjects as well as a correlation between *S100A4* sWAT expression and BMI. Although interesting, none of their analyses were adjusted for HOMA-IR and thus there is no way to discard that their reported BMI-S100A4 association actually could be a consequence of the strong overlapping between obesity and IR. The other available S100A4 study is the one of Malpique *et al.* [10], in which S100A4 was directly associated with hepato-visceral adiposity of non-obese adolescent girls with polycystic ovary syndrome (PCOS). Under an intervention design with spironolactone/pioglitazone/metformin, authors showed how the one-year changes in S100A4 correlated with a reduction of hepato-visceral fat as well as with an improvement of fasting insulin, HOMA-IR, and LDL-cholesterol levels. Although these results would point to S100A4 as a protein with a direct role in the increase of adiposity and the total fat load, again analyses were not adjusted for HOMA-IR, which hinders the drawing of firm conclusions regarding the S100A4-IR-Obesity relationship. In any case, our findings, along with Arner's previous work [9], clearly show how S100A4 could serve as an early-life predictive marker for the appearance of IR and type 2 diabetes later in life.

In order to investigate the molecular basis of the reported IR-S100A4 relationship, we also conducted a multi-omics approach based on gene expression data and methylation reads derived from three independent children samples. Under the epigenetic approach, we counted on the methylation data derived from a longitudinal study conducted in 48 girls (study population 4). We showed how a differential methylation status in several *S100A4* probes associated well with the prepubertal and pubertal IR status of analyzed children, as well as with their longitudinal trajectories for IR. The identified epigenetic associations corresponded with two-enhancer associated CpG sites (cg07245635 and cg10447638) and were mechanistically validated with the additional associations found between the methylation percentage and the plasma S100A4 levels. To date, only a few studies have investigated the relationship between the DNA methylation of *S100A4* and the S100A4 levels [14,32,33]. Restricted to cancer tissues, these studies have shown a direct causal relationship between hiper/hipo methylation of different *S100A4* domains and *S100A4* mRNA expression levels. In the context of type 2 diabetes and obesity, a previous work reported differences in the methylation percentage of several *S100A4* probes in WAT samples when comparing healthy individuals and type-2-diabetes patients [34]. Although these findings have been evidenced for different probes than the ones reported in our study, they are in concordance (regarding the direction of association) with the present results. Indeed, one of the previously associated CpGs (cg26894575) maps very near to one of our top identified sites (cg10447638). Altogether, these findings reinforce the theory that differential methylation of the *S100A4* genetic region could be one of the molecular mechanisms by which this protein is deregulated in obesity and type 2 diabetes to exert a negative effect on the IR axis.

Among the limitations of our epigenetic approach, we can highlight the fact of using blood instead of AT samples, though AT is the target tissue in obesity. As a result, we might be missing some of the key epigenetic signatures of AT cells (mainly represented by leukocytes and adipocytes) with real relevance for the obesity-associated IR. In spite of it, more and more studies are pointing a correlation between the global state of methylation in blood and AT [35]. This correlation might be explained by the abundant presence of leukocytes in both tissues and suggests that buffy coat might be a valid indicator of what happens at the methylation level in AT, especially for the case of inflammatory and immune system related questions.

On the other hand, in prepubertal children (study population 2) we found significant correlations between vWAT *S100A4* expression levels and HOMA-IR, insulin levels, and BMI Z-Score, but not with plasma levels of S100A4. These results therefore might suggest that vWAT is not the main contributor to the systemic S100A4 load in humans. The associations between *S100A4* vWAT expression and IR reported here are in accordance with the insights identified from our longitudinal and cross-sectional approaches and reinforce the fact that S100A4 might play a key role in AT dysfunction and IR. Contrarily to these results, S100A4 secretion in sWAT explants from obese and non-obese adults has previously shown a positive correlation with circulating S100A4 levels [9]. On this matter, further investigation is required for a better understanding of the contribution of sWAT and vWAT in systemic S100A4 levels, and the obesity-associated IR.

Although we have yielded interesting insights regarding the role of S100A4 in IR and obesity, the main cell types contributing to S100A4 levels in obesity and the molecular pathways through which S100A4 induces IR in WAT remain unknown. Some studies support the idea that adipocytes could not be the main contributors and targets of S100A4 in WAT. Arner *et al.* [9] observed that, although S100A4 was present in every cell type of WAT explants, its gene expression was greater in progenitor and immune cells than in mature adipocytes [9]. These authors also treated human adipocytes with S100A4 without reporting any effect [9], suggesting that fat cells could not be the main target of this protein. In the same way, *S100A4* expression has been mainly localized in stromal vascular fraction of mice WAT but seldom in mature adipocytes [36]. These facts are in accordance with our vWAT gene expression insights. S100A4 is a protein secreted by a wide range of inflammatory and immune cells, being leukocytes, fibroblasts, and macrophages the main sources [2,9,36]. Moreover, the binding of S100A4 with target proteins leads to pro-inflammatory processes including chemotaxis, cell migration, ECM remodeling and altered angiogenesis [37]. In relation to this, we observed a positive significant association between circulating S100A4 and SICAM levels in our prepubertal study population 1, a cell adhesion molecule with a role in inflammatory processes [38]. It is known that S100A4 is involved in the epithelial-mesenchymal transition (EMT) through the activation of the transcription of β-catenin, a process characterized by changes in cell morphology and inflammation effects, tissue fibrosis and cancer progression [34]. Moreover, the association between obesity and cancer is well known, along with the involvement

of S100A4 in metastasis progression, where this protein has a key role as a biomarker for poor prognosis in several tumors [5,37,39]. Although it needs to be further investigated, S100A4 could represent a link between obesity, IR, and cancer. Even though it is unclear if circulating S100A4 is a cause or a consequence of obesity and IR, this protein could exacerbate the dysfunction of WAT, promoting an inflammatory environment where leukocytes, macrophages, and other immune cell types could be attracted to the tissue, triggering WAT fibrosis and consequently, IR.

This is the first work reporting a robust association between S100A4 and IR in prebupertal children and under a longitudinal design in children undergoing pubertal development. Furthermore, it represents a complete research approach since it counts on data for S100A4 protein levels in blood, *S100A4* gene expression in vWAT and *S100A4* DNA methylation in blood. Other strengths of this work include the inherent study design with multiple and independent cohorts under both cross-sectional and longitudinal designs as well as the fact of presenting cross-replications among studied populations remaining significant after multiple-test corrections.

Concerning the limitations of the study, it could be remarked that the gene expression and methylation approaches were not performed in the same tissue, as well as the lack of a wide sample size in the longitudinal design.

Regarding the translational potential of our research, the early-life identification of high-risk individuals for IR and T2D complications is of special importance for reducing obesity-associated mortality. Nowadays, new adipokines such as S100A4 offer a great window opportunity for disease prevention and could be monitored as non-invasive biomarkers. Beyond diagnostic implications, this investigation further offers interesting opportunities for novel therapeutic approaches.

5. Conclusion

In summary, we report for the first time the implication of S100A4 in IR and WAT dysfunction in prepubertal populations as well as how the change in plasma S100A4 levels accompanies longitudinal trajectories of IR in children during sexual maturation. Moreover, we propose epigenetic changes in two methylation sites and an altered *S100A4* vWAT expression as the plausible molecular mechanisms underlying this disturbance in obesity. These results could encourage the use of circulating S100A4 as an early predictor of IR in pediatric population and lay the groundwork for future investigations and functional analyses.

Acknowledgements

The authors would like to thank the children and parents who participated in the study.

Funding source

This work was supported by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I + D + I), Instituto de Salud Carlos III-Health Research Funding (FONDOS FEDER) (PI051968, PI1102042 and PI1600871); Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015) and the Mapfre Foundation. The authors also acknowledge Instituto de Salud Carlos III for personal funding: Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016 (IFI17/00048) and the Spanish Ministry of Education (FPU 16/03653).

Conflicts of interest

The authors declare there are no conflicts of interest associated with this manuscript.

Author contributions

CAG is the guarantor of this work. AG and CAG contributed to the study concept and design. AIR, RL, GB, MG, IT, and LM participated in the child recruitment and anthropometric measures (data acquisition). CAG and AAR revised DNA extraction and methylation analyses. CAG and IT revised genome-wide expression analyses. AAR performed all data analyses (including multi-omics approaches). AMG performed all S100A4 measurements and PCR validation analyses. CGL participated in PCR validation analyses (data analysis and interpretation). AAR, AMG and CAG took part in the interpretation of data, the drafting of the manuscript. All authors took park in the critical revision of the manuscript. AAR, AMG and CAG wrote the manuscript. AG, CAG, RL and GB obtained funding. All authors approved the final version of the manuscript.
This paper will be part of Augusto Anguita-Ruiz's doctorate thesis, which is being performed under the "Nutrition and Food Sciences Program" at the University of Granada.

Supplementary data

Supplementary Data are available online at https://doi.org/10.1016/j.metabol.2020.154187?.

References

[1] Vallely KM, Rustandi RR, Ellis KC, Varlamova O, Bresnick AR, Weber DJ. Solution structure of human Mts1 (S100A4) as determined by NMR spectroscopy. Biochemistry 2002;41:12670–80. doi:10.1021/bi020365r.

[2] Fei F, Qu J, Li C, Wang X, Li Y, Zhang S. Role of metastasis-induced protein S100A4 in human non-tumor pathophysiologies. Cell Biosci 2017;25:64. doi:10.1186/s13578-017-0191-1.

[3] Grigorian M, Tulchinsky E, Burrone O, Tarabykina S, Georgiev G, Lukanidin E. Modulation of mts1 expression in mouse and human normal and tumor cells. Electrophoresis 1994;15:463–8. doi:10.1002/elps.1150150163.

[4] Takenaga K, Nakamura Y, Sakiyama S. Expression of a calcium binding protein pEL98 (mts1) during differentiation of human promyelocytic Leukemia HL-60 Cells. Biochem Biophys Res Commun 1994;202:94–101. doi:10.1006/bbrc.1994.1898.

[5] Fei F, Qu J, Zhang M, Li Y, Zhang S. S100A4 in cancer progression and metastasis: A systematic review. Oncotarget 2017;8:73219–39. doi:10.18632/oncotarget.18016.

[6] Michetti F, Dell'Anna E, Tiberio G, Cocchia D. Immunochemical and immunocytochemical study of S-100 protein in rat adipocytes. Brain Res 1983;262:325–6. doi:10.1016/0006-8993(83)91032-6.

[7] Takenaga K, Nakamura Y, Sakiyama S. Cellular localization of pEL98 protein, an S100-related calcium binding protein, in fibroblasts and its tissue distribution analyzed by monoclonal antibodies. Cell StructFunct 1994;19:133–41. doi:10.1247/csf.19.133.

[8] Riuzzi F, Chiappalupi S, Arcuri C, Giambanco I, Sorci G, Donato R. S100 proteins in obesity: liaisons dangereuses. Cell Mol Life Sci 2019:1–19. doi:10.1007/s00018-019-03257-4.

[9] Arner P, Petrus P, Esteve D, Boulomié A, Näslund E, Thorell A, et al. Screening of potential adipokines identifies S100A4 as a marker of pernicious adipose tissue and insulin resistance. Int J Obes 2018;42:2047–56. doi:10.1038/s41366-018-0018-0.

[10] Malpique R, Sánchez-Infantes D, Garcia-Beltran C, Taxerås SD, López-Bermejo A, de Zegher F, et al. Towards a circulating marker of hepato-visceral fat excess: S100A4 in adolescent girls with polycystic ovary syndrome — Evidence from randomized clinical trials. Pediatr Obes 2019;14:e21500. doi:10.1111/ijpo.12500.

[11] Arner P, Petrus P, Esteve D, Boulomié A, Näslund E, Thorell A, et al. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Cell StructFunct 2018. doi:10.3389/fpubh.2017.00058.

[12] Tirado-Magallanes R, Rebbani K, Lim R, Pradhan S, Benoukraf T. Whole genome DNA methylation: beyond genes silencing. Oncotarget 2017;8:5629–37. doi:10.18632/oncotarget.13562.

[13] Rehman I, Cross SS, Catto JWF, Leiblich A, Mukherjee A, Azzouzi AR, et al. Promoter hyper-methylation of calcium binding proteins S100A6 and S100A2 in human prostate cancer. Prostate 2005;65:322–30. doi:10.1002/pros.20302.

[14] Lindsey JC, Lusher ME, Anderton JA, Gilbertson RJ, Ellison DW, Clifford SC. Epigenetic deregulation of multiple S100 gene family members by differential hypomethylation and hypermethylation events in medulloblastoma. Br J Cancer 2007;97:267–74. doi:10.1038/sj.bjc.6603852.

[15] Bell CG. The Epigenomic Analysis of Human Obesity. Obesity 2017;25:1471–81. doi:10.1002/oby.21909.

[16] Tanner JM, Whitehouse RH. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. Arch Dis Child 1976;51:170–9. doi:10.1136/adc.51.3.170.

[17] Olza J, Aguilera CM, Gil-Campos M, Leis R, Bueno G, Martínez-Jiménez MD, et al. Myeloperoxidase is an early biomarker of inflammation and cardiovascular risk in prepubertal obese children. Diabetes Care 2012;35:2373–6. doi:10.2337/dc12-0614.

[18] Anguita-Ruiz A, Pastor-Villaescusa B, Leis R, Bueno G, Hoyos R, Vázquez-Cobela R, et al. Common Variants in 22 Genes Regulate Response to Metformin Intervention in Children with Obesity: A Pharmacogenetic Study of a Randomized Controlled Trial. J Clin Med 2019;8:1471. doi:10.3390/jcm8091471.

[19] Cole TJ. Establishing a standard definition for child overweight and obesity worldwide: international survey. BMJ 2000;320:1240–3. doi:10.1136/bmj.320.7244.1240.

[20] Aguilera CM, Gomez-Llorente C, Tofe I, Gil-Campos M, Cañete R, Gil Á. Genome-wide expression in visceral adipose tissue from obese prepubertal children. Int J Mol Sci 2015;16:7723–37. doi:10.3390/ijms16047723.

[21] Tang Q, Li X, Song P, Xu L. Optimal cut-off values for the homeostasis model assessment of insulin resistance (HOMA-IR) and pre-diabetes screening: Developments in research and prospects for the future. Drug Discov Ther 2015;9:380–5. doi:10.5582/ddt.2015.01207.

[22] Ziaee A, Esmailzadehha N, Oveisi S, Ghorbani A, Ghanei L. The threshold value of homeostasis model assessment for insulin resistance in Qazvin Metabolic Diseases Study (QMDS): Assessment of metabolic syndrome. J Res Health Sci 2015;15:94–100.

[23] Rupérez AI, Olza J, Gil-Campos M, Leis R, Bueno G, Aguilera CM, et al. Cardiovascular risk biomarkers and metabolically unhealthy status in prepubertal children: Comparison of definitions. Nutr Metab Cardiovasc Dis 2018;25:524–30. doi:10.1016/j.numecd.2018.02.006.

[24] Ruiz-Ojeda FJ, Anguita-Ruiz A, Rupérez AI, Gomez-Llorente C, Olza J, Vázquez-Cobela R, et al. Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism. Sci Rep 2019;9:3979. doi:10.1038/s41598-019-40482-0.

[25] Olza J, Aguilera CM, Gil-Campos M, Leis R, Bueno G, Valle M, et al. A continuous metabolic syndrome score is associated with specific biomarkers of inflammation and CVD risk in prepubertal children. Ann Nutr Metab 2015;66:72–9. doi:10.1159/000369981.

[26] Sobradillo B, Aguirre A, Aresti U. Curvas y tablas de crecimiento (estudios longitudinal y transversal). Fundación Faustino Orbegozo Eizaguierre. Inst Investig Sobre Crecim y Desarro Fund Faustino Obegozo Eizaguirre 2009.

[27] McCrindle BW. Assessment and management of hypertension in children and adolescents. Nat Rev Cardiol 2010;7:155–63. doi:10.1038/nrcardio.2009.231.

[28] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res 2013;41:D991-5. doi:10.1093/nar/gks1193.

[29] Green P, Macleod CJ. SIMR: An R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol 2016;7:493–8. doi:10.1111/2041-210X.12504.

[30] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria 2018. doi:10.1108/eb003648.

[31] Reinehr T, Roth CL. Is there a causal relationship between obesity and puberty? Lancet Child Adolesc Heal 2019;3:44–54. doi:10.1016/S2352-4642(18)30306-7.

[32] Lin Z, Deng L, Ji J, Cheng C, Wan X, Jiang R, et al. S100A4 hypomethylation affects epithelial–mesenchymal transition partially induced by LMP2A in nasopharyngeal carcinoma. Mol Carcinog 2016;55:1467–76. doi:10.1002/mc.22389.

[33] Li Y, Liu ZL, Zhang KL, Chen XY, Kong QY, Wu ML, et al. Methylation-associated silencing of S100A4 expression in human epidermal cancers. Exp Dermatol 2009;18:842–8. doi:10.1111/j.1600-0625.2009.00922.x.

[34] Nilsson E, Jansson PA, Perfilyev A, Volkov P, Pedersen M, Svensson MK, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. Diabetes 2014;63:2962–76. doi:10.2337/db13-1459.

[35] Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature 2017;5:81–6. doi:10.1038/nature20784.

[36] Hou S, Jiao Y, Yuan Q, Zhai J, Tian T, Sun K, et al. S100A4 protects mice from high-fat diet-induced obesity and inflammation. Lab Investig 2018;98:1025–38. doi:10.1038/s41374-018-0067-y.

[37] Ambartsumian N, Klingelhöfer J, Grigorian M. The multifaceted S100A4 protein in cancer and inflammation. Methods Mol. Biol., 2019. doi:10.1007/978-1-4939-9030-6_22.

[38] Illán Gómez F, Gonzálvez Ortega M, Aragón Alonso A, Orea Soler S, Alcaraz Tafalla M. S, Pérez Paredes M, et al. Obesity, endothelial function and inflammation: the effects of weight loss after bariatric surgery. Nutr Hosp 2016;33:1340–6. doi:10.20960/nh.793.

[39] Avgerinos KI, Spyrou N, Mantzoros CS, Dalamaga M. Obesity and cancer risk: Emerging biological mechanisms and perspectives. Metabolism 2019; 92:121-135. doi: 10.1016/j.metabol.2018.11.001

Publication in preparation process.

Study 6 Integrative analysis of blood cells DNA methylation, transcriptomics and genomics identifies novel epigenetic regulatory mechanisms of insulin resistance during puberty in children with obesity: a longitudinal study

Anguita-Ruiz A[1,2,3], Ruiz-Ojeda FJ[1,2,3], Alcalá-Fdez J[2,4], Bueno G[3,5,6,7], Gil-Campos M[3,8], Roa-Rivas J[9], Moreno LA[3,5,6,7], Gil A[1,2,3], Leis R[3,10,11,12†*], Aguilera CM[1,2,3†*]

Abstract Background: Puberty is a time of considerable metabolic and hormonal changes associated with a physiological increase in peripheral tissue insulin resistance (IR). There is evidence that physiological IR does not resolve in youth who are obese, which may result in increased cardio-metabolic risk. Understanding the molecular and biological processes underlying the development of IR in puberty and the additional impact of obesity on these processes is crucial to prevent type 2 diabetes.

Methods: This is a longitudinal study based on the follow-up until puberty of a cohort of prepubertal Spanish boys and girls. The study population was composed of 139 children

Affiliations 1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Armilla, 18016 Granada, Spain. (A.A.-R., F. R.-O., A.G., C.M.A.) / 2. Instituto de Investigación Biosanitaria ibs. GRANADA 18014 Granada, Spain. (A.A.-R., F. R.-O., A.G., C.M.A.) / 3. CIBEROBN, (Physiopathology of Obesity and Nutrition) Institute of Health Carlos III (ISCIII), 28029 Madrid, Spain. (A.A.-R., F. R.-O., R.L., G.B., M.G.-C., J.R.-R, L.A.M., A.G., C.M.A.) / 4. Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain. (J.A.-F.) / 5. GENUD Research group, University of Zaragoza, Institute of Sanitary Research of Aragón (IIS Aragón), Zaragoza, Spain. (L.A.M., G.B.) / 6. Agri-food Institute of Aragon (IA2), Zaragoza, Spain (L.A.M., G.B.) / 7. Unit of Pediatric Endocrinology, University Clinical Hospital Lozano Blesa, 50009 Zaragoza, Spain. (G.B.) / 8. Metabolism and Investigation Unit, Reina Sofia University Hospital, Maimónides Institute of Biomedicine Research of Córdoba (IMIBIC), University of Córdoba, 14071 Córdoba, Spain. (M.G.-C.) / 9. Instituto Maimónides de Investigación Biomédica de Córdoba; Department of Cell Biology, Physiology and Immunology, University of Córdoba; Hospital Universitario Reina Sofia (IMIBIC/ HURS) 14004 Córdoba, Spain. (J.R.-R.) / 10. Unit of Investigation in Human Nutrition, Growth and Development of Galicia (GALINUT), University of Santiago de Compostela (USC), Santiago de Compostela, Spain. (R.L.) / 11. Pediatric Nutrition Research Group. Institute of Sanitary Research of Santiago de Compostela (IDIS). CHUS–USC. 15706 Santiago de Compostela, Spain. (R.L.) / 12. Unit of Pediatric Gastroenterology, Hepatology and Nutrition. Pediatric Service. University Clinical Hospital of Santiago (CHUS). 15706 Santiago de Compostela, Spain. (R.L.)
*†Concepción M. Aguilera and Rosaura Leis are the senior authors.
*Corresponding Authors

organized in a longitudinal approach of 90 subjects (47 females) and two cross-sectional approaches of 99 (52 females) and 130 (71 females) subjects for prepubertal and pubertal stages, respectively. Children were allocated into experimental groups according to their obesity and IR status before and after the onset of puberty. All participants presented blood DNA samples for GWAS and EWAS analyses. In 44 children of the pubertal stage, we counted on blood RNA samples for RNA-seq analysis.

Results: Our large-scale integrative molecular analysis identified novel blood multi-omics signatures (mapping the loci *ABCG1*, *ESR1* and *VASN*, among others) significantly associated with IR longitudinal trajectories in children with obesity during pubertal maturation. Functional enrichment analysis revealed that identified loci participate in systemic metabolic pathways and sexual maturation processes relevant to the pathogenesis of IR in the context of puberty. Additional analyses on cardiometabolic and inflammatory phenotypes showed that blood DNAm patterns of some of the identified loci are further associated, beyond IR, with an overall risky-cardiometabolic profile in children. Serum protein levels of vasorin (VASN), one of the most promising novel biomarkers identified in this study, were further associated with IR in the pubertal stage.

Conclusions: To our knowledge, this is the first longitudinal multi-omics approach characterizing molecular blood alterations for IR and obesity during the metabolically critical period of puberty. Our results shed light on the molecular mechanisms underlying epigenetic alterations in obesity and propose novel and promising biomarkers for IR and metabolic alterations in children.

Keywords: Adolescent; Child; DNA Methylation; Epigenetics; Epigenome-Wide Association Study, EWAS; Gene expression; Genetics; Genome-Wide Association Study, GWAS; Insulin resistance; Multi-omics; Pediatric obesity; Puberty; Vasorin; *VASN*

Abbreviations: Cardiovascular disease (CVD); Differentially methylated site (DMS); DNA methylation (DNAm); Epigenome-Wide Association Studies (EWAS); Expression quantitative trait methylations (eQTMs); False discovery rate (FDR); Genome-Wide Association Studies (GWAS); Hardy-Weinberg equilibrium (HWE); Insulin resistance (IR); Methylation quantitative trait loci (mQTLs); Minor allele frequency (MAF); Multiple linear regression (MLR); Single nucleotide polymorphisms (SNPs).

Background

Insulin resistance (IR) is a pathological condition of glucometabolic sufferance contributing to type 2 diabetes and cardiovascular disease (CVD) in both adults and children. Of note, obesity is the main driver of IR in children [1, 2]. Many children who are overweight or suffer from obesity before puberty maintain obesity in early adulthood, which is associated with increased morbidity and mortality [3–5]. The high mortality rates among people with obesity are mainly due to the development of type 2 diabetes and the increased risk of CVD [6]. Therefore, it is crucial to prevent and treat obesity and IR from the early periods of life [7].

Puberty is a period characterized by dynamic physiological changes, including activation of the reproductive axis and subsequent increase in sex steroids secretion, acceleration in growth, and accumulation of both lean and fat mass [8]. Besides physiological events, puberty has also been associated with differential disease prognosis for conditions such as IR, reinforcing the relevance of this development period to life-long health. Nevertheless, pubertal changes seem not to affect all individuals equally [9–11]. In healthy normal-weight youths, there is a drop in insulin sensitivity in mid-puberty, which recovers at puberty completion. In youth who are obese going into puberty, otherwise, there is evidence that such IR does not resolve, which may result in increased cardio-metabolic risk. Accordingly, youth-onset type 2 diabetes incidence is also tightly linked with pubertal development [12]. Understanding the molecular and biological processes underlying metabolic changes during puberty and the additional impact of obesity on these changes is crucial for preventing type 2 diabetes. Thanks to that, novel non-invasive early diagnostic markers could arise with a great utility for reducing obesity-associated mortality.

DNA methylation (DNAm) is a heritable epigenetic mark consisting of the covalent addition of a methyl group to a cytosine followed by a guanine (CpG). DNAm is potentially reversible and can be altered by environmental factors, resulting in gene expression alterations and providing an interactive connection between genetics, specific diseases and the environment. Indeed, differential DNAm in certain loci has been related to obesity [13], systemic IR [14–22], and type 2 diabetes [13, 15, 16, 23–27] in adults, either in blood or in other metabolically relevant tissues. During puberty, the dynamics of DNAm have also been investigated in one or both genders, emphasizing how DNAm is stable at some CpG sites and varies at others [28, 29]. On the other hand, transcriptional dysregulation of genes has been reported as a key molecular mechanism associated with IR and obesity, possibly connected to DNAm alterations [30, 31]. In this regard, there is evidence of the interacting effects between DNAm and gene expression and the risk for glucometabolic alterations in adult women with obesity (phenomena known as expression quantitative trait methylations (eQTMs)) [17].

Although recent genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs) associated with type 2 diabetes and its related traits [32–38],

these variants only explain a small proportion of the estimated heritability (15–18 %), proposing that there are additional genetic factors left to be discovered. Among the best explanations, it highlights the existence of interacting phenomena between SNPs and DNAm epigenetics marks, known as methylation quantitative trait locis (mQTLs). Interestingly, a previous study demonstrated how interactions between SNPs and DNAm influence mRNA expression and insulin secretion in adult human pancreatic islets [39].

Although the understanding of the molecular and biological processes underlying IR in obesity is growing (especially in adults), none is known about the omics alterations characterizing IR in obesity during the metabolically critical period of puberty, and how it might contribute to the increased disease risk (e.g., for type 2 diabetes). For this purpose, multi-omics approaches present as a promising resource in which systems biology can be applied to mine the complex interactions between genetics, epigenetics, and transcriptomics. The identification of multi-omics signatures and how they relate with the progression of obesity and IR during puberty will allow us to provide pediatricians with robust non-invasive early biomarkers for type 2 diabetes risk. For this task, longitudinal designs are also encouraged.

Considering all this, in the present study, we have identified the multi-omics signatures (DNAm in CpGs, eQTMs and mQTLs) associated with IR in children with obesity, before, during, and after the onset of puberty. This research is a continuation of the PUBMEP study, which evaluates the prevalence of metabolic syndrome and the progression of the cardio metabolic risk factors related to it, from pre-puberty to puberty, in a longitudinal cohort of Spanish children [40].

Research Design and Methods

2.1. Study population

This analysis was conducted within the context of the PUBMEP study. The main clinical findings derived from the PUBMEP study and additional details on the whole study cohort have been already published and are available elsewhere. In the PUBMEP study, all children were first recruited as prepubertal children during 2012–2015 and called again for follow-up medical consultation in 2018. At the moment of recruitment, children were aged 4–12.1 years and came from three Spanish recruiting centers (cities). At the second visit, children were aged 9.72–18.07. All subjects with clinical signs of reached puberty were finally included in the longitudinal population. During the course of the study (2012–2018), children remained under regular medical monitoring by the same pediatricians. The assessment of the pubertal stage was carried out following the Tanner classification [41] and confirmed with a hormonal study. Here, a sub-population of 139 children (76 females) from the whole PUBMEP cohort was selected for omics analyses. The main inclusion criterion for this sub-project was presenting a good-quality DNA sample in the prepubertal stage. The following characteristics were considered as exclusion criteria: birth weight <2500 g; intake

of any drug that could alter blood glucose, blood pressure or lipid metabolism; not being able to comply with the study procedures and being participating or having participated in the last three months in an investigation project. The 139 participating children were organized in a longitudinal approach of 90 children (47 females) and two cross-sectional approaches of 99 (52 females) and 130 (71 females) children for prepubertal and pubertal stages, respectively. A general overview of the study design, populations and statistical analyses performed is well-described in **Figure 1**. The longitudinal approach consisted of 90 Spanish children (47 females) allocated into five experimental groups according to their obesity and IR status before and after the onset of puberty **(Figure 2A**).

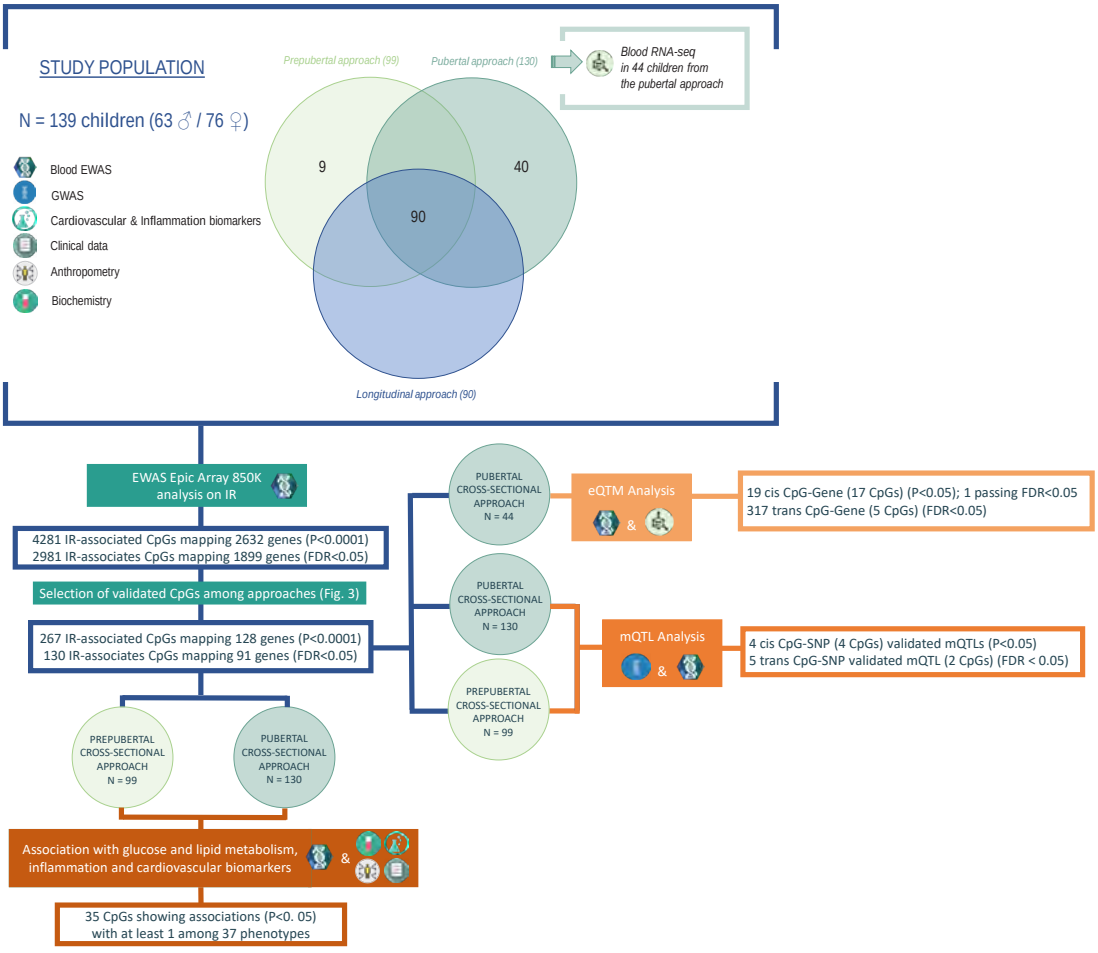


Figure 1. Overview of the study design, populations under study and statistical analyses performed.

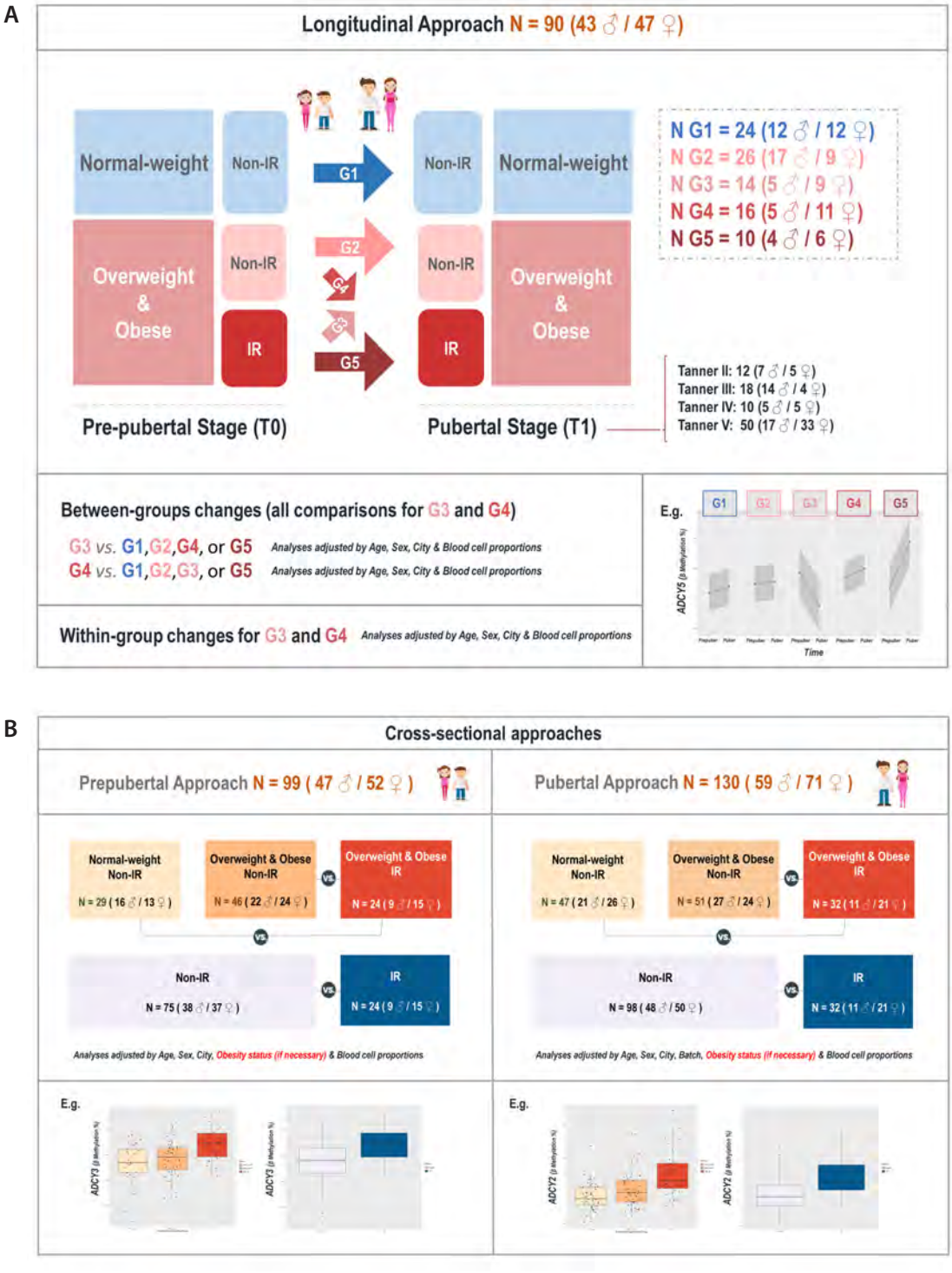


Figure 2. The statistical design adopted for DNA methylation bioinformatics analysis of IR.

All participants (N=139) presented DNA samples with enough quality for genomics (GWAS) and epigenomics (EWAS) analyses. Moreover, we also collected blood samples in 44 children of the pubertal stage using PAXGEN-RNA tubes for posterior RNA-seq analysis. Descriptive statistics for these longitudinal and cross-sectional approaches are available in the Additional files 1, 2 and 3. The main blocks of analyses performed in this work consisted of 1) EWAS analysis of IR and 2) multi-omics integration of EWAS results along with GWAS and RNAseq.

2.2. Ethics statement

These studies were conducted following the Declaration of Helsinki (Edinburgh 2000 revised), and they followed the recommendations of the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legal in-forced Spanish regulation, which regulates the clinical investigation of human beings (RD 223/04 about clinical trials). Accordingly, the corresponding ethics committees approved the study at each of the participating centers (Code IDs GENOBOX: Córdoba01/2017, Santiago 2011/198, Zaragoza 12/2010; and PUBMEP: Córdoba 260/3408, Santiago 2016/522, Zaragoza 22/2016, Granada 01/2017).

2.3. Anthropometry, biochemical measurements, and inflammation and cardiovascular risk biomarkers

Anthropometric measurements such as body weight (kg), height (cm), hip circumference (cm) and waist circumference (WC) (cm) were measured at each time point using standardized procedures, and BMI (kg/m^2) was calculated. BMI z-score was estimated based on the Spanish reference standards published by Sobradillo *et al.* [42]. Blood pressure was measured three times for each individual by the same examiner using a mercury sphygmomanometer and following international recommendations [43]. Measures of lipid and glucose metabolism, hormones and classical biochemical parameters were performed at the laboratories of each participating hospital following internationally accepted quality control protocols.

Blood samples from both time points were collected in overnight fasting conditions, centrifuged, and plasma and serum were stored at -80°C. Plasma adipokines, inflammation, and cardiovascular risk biomarkers (adiponectin, leptin, resistin, tumor necrosis factor alpha (TNF-α), high-sensitivity CRP (hsCRP), interleukin (IL)-6, IL-8, total plasminogen activator inhibitor-1 (PAI-1), P-Selectin, myeloperoxidase (MPO), monocyte chemoattractant protein 1 (MCP-1), matrix metalloproteinase-9 (MMP-9), soluble intercellular cell adhesion molecule-1 sICAM-1, and soluble vascular cell adhesion molecule-1 (sVCAM)) were analyzed in all samples and time points using XMap technology (Luminex Corporation, Austin, TX) and human monoclonal antibodies (Milliplex Map Kit; Millipore, Billerica, MA) as previously reported [44, 45]. S100A4 protein levels were determined in plasma using the CSBEL02032HU (Cusabio Biotech Co, Ltd., Wuhan, China), an enzyme-linked immune-absorbent assay kit according to the manufacturers' instructions. The coefficient of variance was 7%.

2.4. HOMA-IR cut off points

The IR status was here defined by means of the HOMA-IR index. Since HOMA-IR strongly varies between ages, genders and diseases, and since no reference values have been yet established in either children or adult populations [46, 47], own cut-off points were extracted from a previous well-described Spanish cohort composed of 1669 children and adolescents [44, 48]. For the prepubertal stage, a single cut-off value of HOMA-IR=2.5 was considered for IR [44, 48]. For the pubertal stage instead, gender information was taken into consideration and different cut-off points were adopted for IR according to the 95th HOMA-IR percentile. Extracted from a subset of 778 pubertal Spanish children, cut-off values corresponded to HOMA-IR=3.38 in boys and HOMA-IR=3.905 in girls. These cut-off points have already been tested and validated as good metabolic risk classifiers in our population according to the results from a previous PUBMEP paper [40].

2.5. Epigenome-wide association study (EWAS)

EWAS analysis was performed in all children (N=139) and time points, including longitudinal and cross-sectional approaches (**Figure 2**). Buffy coat fractions from blood samples in all children and time points were selected for DNA methylation analysis. Genomic DNA was extracted from peripheral white blood cells using two automated kits, the Qiamp DNA Investigator Kit for coagulated samples and the Qiamp DNA Mini & Blood Mini Kit for non-coagulated samples (QIAgen Systems, Inc., Valencia, CA, USA). All extractions were purified using the DNA Clean and Concentrator kit from Zymo Research (Zymo Research, Irvine, CA, USA). High-quality DNA samples (≥ 500 ng) were treated with bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research Corporation, Irvine, CA). DNA methylation was measured with the Infinium Methylation EPIC array using bead chip technology (Illumina, San Diego, CA, USA).

Raw intensity signals from IDAT files were loaded into the R environment using the MINFI R package. As a result, we obtained an RGChannelSet object containing all the raw intensity data, from both the red and green color channels, for each of the samples and time records. We generated a detection p-value for every CpG in every sample by comparing the total signal for each probe to the background signal level, which was estimated from the negative control probes. To minimize the unwanted variation within and between samples, we applied Beta-Mixture Quantile (BMIQ) intra-array normalization, including all individuals and time records. Poor performing probes were filtered out according to different criteria: probes with a detection p-value above 0.01 in more than 10 % of the samples (number of probes= 230), probes with SNPs (number of probes= 30,432), cross-reactive probes aligning to multiple locations (number of probes= 25,570) and probes located on the Y chromosome (number of probes= 246). After applying all these filters, 809,381 probes remained in the dataset.

As methylation is cell type-specific and methylation arrays provide CpG methylation values for a population of cells, biological findings from samples comprised of a mixture of cell types,

such as the case of peripheral blood, can be confounded due to variable cell-type composition. To correct analyses for the variable proportion of each white cell type in our subjects, we employed the reference EPIC 850k dataset published by Salas *et al.* (2012) and the Houseman procedure [49]. The influence of each confounding variable on the global state of methylation in our population was assessed by means of correlation studies and heatmap plots using the SWAMP R package v1.4.1.

Intensity values were used to determine the proportion of methylation at each CpG site. Methylation levels were reported as either beta values (β-Values = M/(M + U)) or M-values (M value = $\log_2$(M/U)), where M and U correspond to the Methylated and Unmethylated signals, respectively. Beta values and M-values are related through a logit transformation (M-value = $\log_2$(β-value /1-β-value)). Because percentage methylation is easily interpretable, beta values in the present paper were employed for describing the level of methylation at each locus and for graphical presentation of results. On the other hand, due to their distributional properties, M-values were selected for statistical testing. All described analyses were performed in R environment version 4.0.3.

2.6. Genome-wide association study (GWAS)

GWAS analysis was performed in all children (N=139) (**Figure 1**). When two samples of DNA were available for the same individual (e.g., in the longitudinal approach), the most recently extracted DNA sample was selected for genomics analysis. Whole-genome genotyping analysis was performed on the i-SCan platform using the Infinium HTS Assay (Illumina, San Diego, CA, USA). The Bead Chip selected for the project was the Infinium Global Screening Array-24 v3.0 Kit, which includes ~ 654,000 genetic markers associated with complex diseases. After quantification of DNA samples by fluorimetry, they were normalized to 200-400 ng of DNA per sample in deep well plates, as established in the Infinium HTS Assay Protocol.

The first step of the primary data analysis consisted of the extraction of genotype calls from fluorescence data and the construction of work data files for data manipulation and analysis. Using the *GenomeStudio* software, we obtained genotype calls for all individuals and generated the standard format files *(.ped and .map)*. Data were then imported into *PLINK* 1.9 software [50], and converted into binary format files using the *--make-bed* flag. These binary formats *(.bed, .bim and .fam)* are a more compact representation of the data that saves space and speeds up subsequent analyses. We implemented a quality control (QC) process in PLINK 1.9 software before high-level statistical analyses. According to literature, we applied standard QC filters including: 1) Exclusion of SNPs and individuals with a missing data rate >= 10%, and 2) Exclusion of SNPs with a minor allele frequency (MAF) < 1% or a Hardy-Weinberg Equilibrium (HWE) P-value < 1×10−5 in controls. As a result, 471,192 SNPs remained in the dataset. These filters were selected in accordance with procedures described elsewhere [51] to minimize the influence of genotype-calling artefacts in a GWAS.

2.7. Next-generation transcriptome sequencing (RNA-Seq)

RNA-seq analysis was performed in a subset of 44 children from the pubertal cross-sectional approach (**Figure 1**), which had also been included in the EWAS and GWAS. RNA was extracted from peripheral blood using the PAXgene® Blood RNA Kit (PreAnalytiX/QIACUBE) according to the manufacturer's instructions. The concentration and quality of extracted RNA were measured using the Qubit 4 Fluorometer (Thermo Fisher Scientific, MA, USA) and the 2100 Bioanalyzer Instrument (Agilent Technologies, CA, USA). Libraries from mRNA were prepared using 1µg of RNA starting material and the TruSeq Stranded mRNA Library Prep Kit (Illumina, CA, USA) according to the manufacturer's protocol. This protocol captures poly-adenylated RNA by transcription by oligo-dT primer, after which the RNA is fragmented. The sample is back transcribed to generate the cDNA, both in the first and second strands. The 3' ends are adenylated, the adapters and barcodes are ligated, and finally, the sample is enriched by PCR. Adapters and sample codes (index-barcodes) are added to the libraries to be simultaneously sequenced. mRNA libraries were sequenced on the Next-Seq 500 system (Illumina, CA, USA) using the highest output mode and paired-end 75 bp read lengths with a depth of 20 million reads for each sample. To get a depth of 20 million reads per sample 2 runs with 4 lanes for each run were conducted.

Primary RNA-seq bioinformatics analyses were implemented in R environment separately for each run following standardized published recommendations. Primary analyses included processing raw sequencing reads, aligning to the reference genome, and quantitating the expression levels. The aligning of RNA-seq reads to the genome was conducted using *HISAT* software 2.2.1 release [52]. We sorted and converted the generated SAM files into BAM using *SAMTOOLS* [53]. Then, we used the *FEATURECOUNTS* R package [54] to generate count matrices from reads aligned to the genome. As reference genome, we used the *hg38* version from the Ensembl [55]. From it, we created a transcript database, using the function *makeTxDbFromGFF* from the *GENOMICFEATURES* R package. Finally, we obtained two datasets of 60,058 quantified ENSG ids (one per run). After confirming the grouping of technical replicates (samples) among runs by PCA plot (Additional file 4), we merged the two counts datasets by applying a sum.

2.8. Descriptive statistics

At each cross-sectional approach (**Figure 2B**), continuous non-omics variables were tested for normality using the Shapiro–Wilk test. Heteroscedasticity between experimental groups was explored through the Levene test. T-tests and Mann–Whitney U-tests were applied conveniently to determine group differences at each cross-sectional stage (prepubertal and pubertal). The resulting descriptive statistics are available in Additional files 2 and 3. In the longitudinal approach (**Figure 2A**), within-group changes from prepuberty (T_0) to puberty (T_1) were assessed using a paired design, employing either a paired t-test or a Wilcoxon signed-rank test. Between-group differences were assessed by one-way ANOVA, Kruskal-Wallis or Welch tests to the computed delta values (T_1–

T_0) for each continuous measurement according to standard statistical assumptions. The one-way ANOVA, Kruskal-Wallis and the Welch test were also employed to assess group differences (among the 5 experimental groups) at each stage (time point) of the longitudinal approach. Pairwise t-tests, pairwise Mann–Whitney U-tests and Dunn tests were applied conveniently as post-hoc analyses to determine which experimental groups differed from each other in these analyses. The resulting descriptive statistics are available in Additional file 1. All described analyses were performed in R environment version 4.0.3 [56].

2.9. DNA methylation bioinformatics analysis on insulin resistance

After pre-processing, 809,381 CpGs probes passing quality filters in the EWAS were selected for high-level statistical analyses. The statistical design adopted for DNA methylation bioinformatics analysis of IR is presented in **Figure 2**. The main objective of this analysis was identifying the DNA methylation patterns associated with IR development, amelioration or worsening in children with obesity during the onset of puberty. For avoiding confounding with non-pathological aging epigenetics marks, derived findings were contrasted to the DNA methylation patterns associated with the onset of puberty in the normal-weight group *G1*. For the longitudinal approach (N=90), we investigated the changes in DNA methylation from prepuberty to puberty of each included CpG site performing both within- and between- groups comparisons. Within-group changes in DNA methylation from prepuberty (T_0) to puberty (T_1) were studied exclusively for the groups *G3* and *G4*, since these were the only groups presenting changing trajectories for IR. Between-groups changes in DNA methylation from prepuberty (T_0) to puberty (T_1) were otherwise investigated for all pairwise group combinations that involved either the *G3* or *G4* group (for more details see **Figure 2A**). These analyses were implemented in the R environment using linear models from the *LIMMA* R package and considering the M-values of DNA methylation for each CpG as the outcome or dependent variable. For that purpose, a multi-level experiment was considered, treating the patient as a random effect, and the experimental group and time as a combined fixed factor. The inter-subject correlation was the input for the linear model fit. Contrasts of interest over the constructed linear model were then applied using a moderated t-test. All this was implemented in *LIMMA* using the functions *duplicateCorrelation*, *lmFit, contrasts.fit* and *eBayes*. Analyses were conveniently adjusted for confounders, including gender, city of origin, blood cell proportions and age. Raw p-values were corrected using the false discovery rate (FDR) according to the Benjamin-Hochberg procedure for multiple comparisons.

In cross-sectional approaches (N=99 and N=130) **(Figure 2B**), the objective was also focused in identifying the DNA methylation patterns associated with IR, but including now those marks associated with IR before the onset of puberty. Since cross-sectional approaches included extra children with regard to the longitudinal cohort of 90 children (**Figure 1**), these analyses were also intended as validation approaches for the longitudinal findings. In these analyses, *LIMMA*

linear models were conducted with DNA methylation levels as outcome, experimental group as a categorical predictor and gender, city of origin, blood cells proportions, batch, obesity status (when necessary) and age as covariates. Raw p-values were corrected using the false discovery rate (FDR) according to the Benjamin-Hochberg procedure for multiple comparisons.

We selected exclusively overlapping findings from these analyses and different approaches (**Figure 3**), obtaining a curated list of loci whose methylation is robustly and repeatedly associated with IR in our design. This list of IR-associated CpGs (and their mapped genes) (thereinafter known as 'validation-list') was the input for the next blocks of analysis. For this selection, we first created a Venn diagram including the genes reported as significant in each statistical approach. Once overlapping IR-associated genes were identified, all significant CpGs mapping these genes were selected. The resulting list was composed of 267 IR-associated CpGs mapping 128 genes.

In cross-sectional approaches, we further applied multiple linear regressions (MLR) between DNA methylation M-values of the CpGs in the 'validation-list' and continuous non-omics outcomes (including all collected anthropometric and biochemical measurements, and inflammation and cardiovascular biomarkers) at each temporal record (T_0 and T_1). The main purpose of these additional analyses was to identify those CpGs, which instead or besides IR, are associated with other obesity-related metabolic alterations or parameters. In these analyses, IR, BMI Z-score, gender, age, city of origin and height were included as covariates conveniently. Given the number of analyzed outcomes, we again considered FDR as in Benjamin-Hochberg to correct for multiple hypothesis testing.

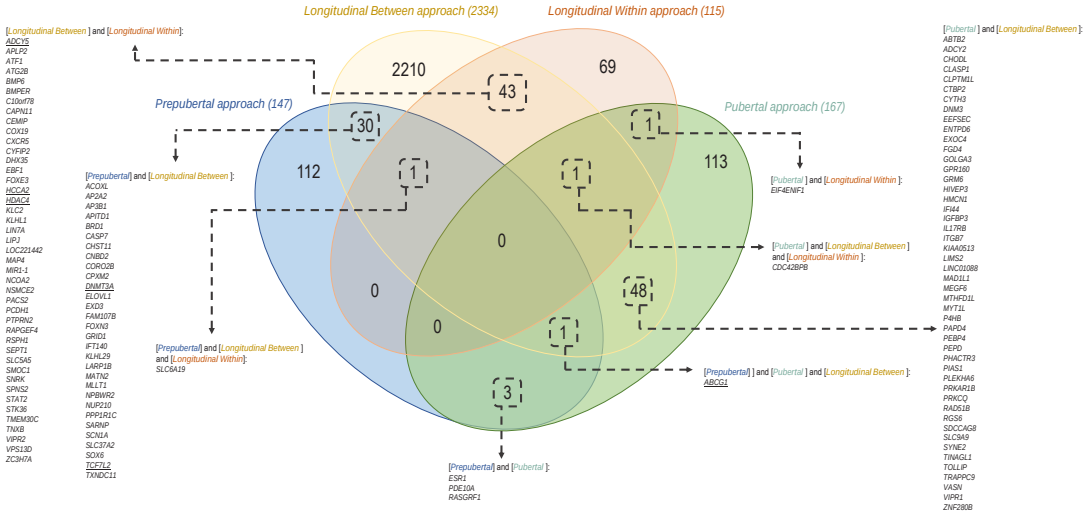


Figure 3. Overlapping findings (at loci level) between the statistical approaches of the EWAS on insulin resistance. These findings constitute the so called 'validation-list'.

2.10. Expression quantitative trait methylation (eQTM) analyses

The expression quantitative trait methylation (eQTM) analysis is intended for identifying gene expression regulatory phenomena, in which the methylation of a CpG influence (up- or down-regulating) the expression of a target transcript. For these analyses, we used linear regression, as implemented in the *MatrixEQTL* R package, to test whether DNAm levels of the CpG sites from the 'validation-list' are associated with transcript expression levels (considering all the transcripts mapped by our RNA-seq analysis). For these analyses, gene expression data for 60,058 transcripts were normalized using the quantile normalization method. Here, the transcript expression level (normalized) was the outcome and the methylation level (M-value) of each CpG site was the predictor, with gender, age and city of origin as adjusting covariates. We searched for cis-eQTMs in and around 10,000 bp of each transcript and trans-eQTMs if the distance between the CpG and the transcript was higher than 10,000 bp. As a measure of eQTMs effect size, we reported the beta regressors estimated by the linear model. The p-values from the linear regression analysis were adjusted for multiple comparisons using the Benjamini-Hochberg FDR procedure. These analyses focused on 267 CpGs and 60,058 high-quality transcripts as input. In the cis-analysis, no correction value was used for correcting multiple testing. The correction value for the trans-analysis was calculated as the total number of analyzed CpG sites multiplied by the number of transcripts in the whole dataset.

2.11. Methylation quantitative trait loci (mQTL) analysis

The methylation quantitative trait loci (mQTL) analysis is intended for identifying epigenetics regulating phenomena, in which an SNP regulates the methylation levels of a CpG. These phenomena could be, therefore the molecular explanation for some epigenetics IR-associated marks for which the environment is not the causal mediator. For these analyses, we used linear regression, as implemented in *MatrixEQTL* R package. In the linear model, DNA methylation values were modelled as the outcome, SNP genotypes from GWAS were encoded as 0, 1 or 2 according to the number of minor alleles (additive genetic model), and gender, age and city of origin were included as covariates. To distinguish between local (cis-mQTLs) and distant (trans-mQTLs), an arbitrary boundary with the maximum distance of 500 bp between SNPs and CpG sites was used to define cis-mQTLs. All other SNP-CpG pairs were considered as trans-mQTLs. P-values were adjusted with a correction value for multiple testing, which considers the dependency of linkage disequilibrium (LD) between SNPs by LD-based pruning and thereby uses the number of independent tests. In the cis-analysis, no correction value was used for correcting multiple testing. The correction value for the trans-analysis was calculated as the total number of analysed CpG sites multiplied by the number of SNPs in the whole dataset.

2.13. Reference genome assembly

EWAS, GWAS and RNA-seq datasets were functionally annotated based on GO and KEGG ontologies using entrez gene identifiers and the database *org.Hs.eg.db* [57]. For eQTM and mQTL analyses, the three employed omics datasets were re-annotated or flipped (in terms of chromosome number and genomic locations) to the same assembly (*hg38*) as reference.

2.14. Functional annotation of CpG sites and regions

Selected CpG sites and regions were further annotated using the *ILLUMINAHUMANMETHYLATIONEPICANNO.ILM10B4.HG19* and *MISSMETHYL* R packages. Genes associated with each CpG site were obtained using the *getMappedEntrezIDs* function. The annotation in terms of genomics regulatory elements consisted of two categories: 1) distance to a CpG island and 2) annotation to gene region. The distance related annotations identify whether CpG sites overlap a known CpG island, 2000bp of the flanking regions of the CpG islands (shores), 2000bp of the flanking regions of the shores (shelves), or outside these regions (open sea). CpGs overlapping gene bodies were annotated as (Body). The gene region analysis classified CpGs in the context of genes, namely, exons, UTRs, introns, promoters, and intergenic regions. Additional annotation of CpG sites for nearby SNPs was determined using the UCSC database. To identify the regulatory potential of CpG sites, each site was categorized based on its predicted chromatin state. These data and additional information have been gathered for significant CpGs and are available in each table result.

2.15. Gene ontology and biological pathway enrichment analysis

The *gometh* function from the R package *MANIFEST* was used to determine enrichment of CpG-annotated genes in KEGG terms, biological pathways, and cellular and molecular functions. This function takes a character vector of significant CpG sites, maps the CpG sites to Entrez Gene IDs, and tests for GO term or KEGG pathway enrichment using a hypergeometric test [58], taking into account the number of CpG sites per gene on the EPIC array. The *gometh* is based on the *goseq* method [59] and calls the goana function from the *LIMMA* R package [60]. The *gometh* tests all GO or KEGG terms, and FDR are calculated using the method of Benjamini and Hochberg (1995). The *LIMMA* functions *topGO* and *topKEGG* were used to display the top most enriched pathways.

Results

1. Clinical characteristics of participants

A general overview of the study design, populations and statistical analyses conducted is well-described in **Figure 1** and **Figure 2**. Descriptive statistics for longitudinal and cross-sectional approaches are available in the Additional files 1, 2 and 3. **Figure 2A** describes all details on the

longitudinal approach. At baseline, there was no age difference between experimental groups, with all but the G5, which maintains IR with pubertal maturation, presenting not significant differences in the elapsed time between visits. As expected from the study design, groups maintaining or developing IR with the onset of puberty, G4 and G5, presented the highest increases in HOMA-IR and fasting serum insulin as compared with the insulin-sensitive groups (Additional file 1). Among them, the G4 group, which develops IR with the onset of puberty, also displayed the highest increase in waist circumference, hip circumference and SBP, and inflammatory and cardiovascular biomarkers (e.g., P-selectin and t-PAI) (Additional file 1). Interestingly, the G3 group, presenting the opposite behavior for IR (amelioration) than G4, showed a decrease in plasma glucose, triacylglycerol concentrations and MCP1 cytokine levels, and these changes were significantly different from the patterns observed in the rest of the groups. Thus, the experimental groups of our design were representative of the longitudinal IR and insulin-sensitive trajectories during the onset of puberty (Additional file 1). As expected from the experimental design, an overall decrease in total and low-density lipoprotein cholesterol levels during puberty was observed for all groups, as well as a decrease in adiponectin concentrations. Likewise, cross-sectional experimental groups showed coherent behaviors in anthropometry, glucose, lipid, inflammatory, and cardiovascular biomarkers (Additional files 2 and 3).

2. Blood cells CpG methylation is associated with obesity insulin resistance

After pre-processing, 809,381 CpGs EWAS probes passing quality filters were selected for high-level statistical analyses. The statistical design adopted for DNA methylation analysis of IR is presented in **Figure 2**. The main objective of this design was to identify the DNA methylation patterns associated with IR development, amelioration or worsening in children with obesity before, during, and after the onset of puberty. Our analysis identified 4,281 IR-associated unique CpG sites (P-value < 0.0001), from which 2,981 further presented an FDR < 0.05 (Additional file 5). Annotation of differentially methylated sites (DMS) informed that reported CpGs were linked to 2,632 and 1,899 genes, respectively for raw P-value and FDR thresholds. We assessed the robustness of our findings by comparing our list of 2,632 IR-associated genes to a curated list of genes whose methylation degree is strongly associated with type 2 diabetes according to a recent large EWAS meta-analysis conducted in Europeans adults [23]. Our list contained 8 of the 38 type 2 diabetes-associated EWAS genes validated by Juvinao *et al.* 2021 [23] (*ABCG1, CDH23, CPT1A, HCCA2, HDAC4, KRT4, PBX1* and *SGK2*), highlighting three of them among the top 6 associated genes in the meta-analysis. Other 24 well-known diabetes and obesity epigenetics loci recently reviewed by Ling C *et al*. 2019 [13], were also present among our associations *(ABCC3, ADCY5, ATP10A, CDKN1A, CXCL14, DNMT3A, FADS2, FTO, GLP1R, GRB10, HIF3A, KCNQ1, MALT1, MOGAT1, NCOR2, NFAM1, PLCB1, PPARG, PRDM16, PRKCE, SEPT9, TCF7L2, THADA* and *VAC14)*. Genes encoding proteins with a key role in the IR-puberty axis and strongly related to the growth hormone, such is the *IGF-1*, were also highlighted in our analysis [61]. Among associations, there were also loci whose DNA methylation,

beyond type 2 diabetes, have been specifically associated with IR according to literature *(COL18A1, CTNND2, CXCL1, DNMT3A, GRB10, HDAC4, LAT, PAX6, SH3RF3* and *SIRT2)*.

KEGG-pathways enrichment analysis showed that the 2,632 IR-associated genes (passing the raw P-value threshold) were over-represented for pathways with relevance in inflammation and human metabolism in the context of our research including; 'Ovarian steroidogenesis', 'Cortisol synthesis and secretion', 'Extracellular Matrix-receptor interaction', 'Glucagon signalling pathway', 'cAMP signalling pathway', 'Insulin secretion', 'Phospholipase D signalling pathway' or 'PPAR signalling pathway' (P-value < 0.05) (Additional file 6).

As expected from the study design, our analysis also revealed genes previously described as markers of dynamic DNA methylation changes during the course of puberty (e.g., *ADCY9, ATK3, GRIK5, GNG7, PDE10A,* or *TRAF3IP2)* [28].

From the initial list of 2,632 IR-associated genes, we exclusively selected those overlapping findings among statistical approaches (**Figure 3**), obtaining a reduced list of genes whose methylation is robustly and repeatedly associated with IR in our design. The resulting list of IR-associated genes and their mapping CpGs (thereinafter known as 'validation-list') was the input for the next blocks of analysis. This list was composed of 267 IR-associated unique CpGs mapping 128 genes (P-value < 0.0001), from which 130 IR-associated CpGs, mapping 91 genes, presented an FDR < 0.05 (**Figure 4** and Additional file 7). Association results for the top 25 CpGs from the list are shown in **table 1**. Interestingly, functional enrichment analyses on KEGG terms for these CpGs, conserved interesting biological pathways such as 'Ovarian steroidogenesis', 'cAMP signalling pathway', 'Insulin secretion' and 'Phospholipase D signaling pathway', or reported new ones as 'estrogen signaling pathway' (P-value < 0.1) (Additional file 8). This list also maintained top adult type 2 diabetes loci previously described in the literature *(ABCG1, ADCY5, DNMT3, HDAC4, TCF7L2* and *HCCA2)*, and revealed new promising loci never reported as epigenetic marks of IR (e.g. *CDC42BPB, ESR1, HMCN1, PRKAR1B, SNRK* and *VASN*, among others). GO-terms enrichment analysis further revealed genes from the 'validation-list' mapping important pathways such is the 'G protein-coupled receptor signalling pathway' (P-value = 0.002) (Additional file 9).

3. Association between DNAm and other phenotypic traits at key CpGs identified by our EWAS

At each cross-sectional approach, we applied multiple linear regressions to the CpGs from the 'validation-list' and continuous non-omics outcomes (including all collected anthropometric and biochemical measurements, and inflammation and cardiovascular biomarkers). The main purpose of these additional analyses was to identify those CpGs, which instead or besides IR, are associated with other obesity-related metabolic alterations or parameters. In additional file 10, we present all significant associations showing a P-value < 0.05 with at least one trait after confounding

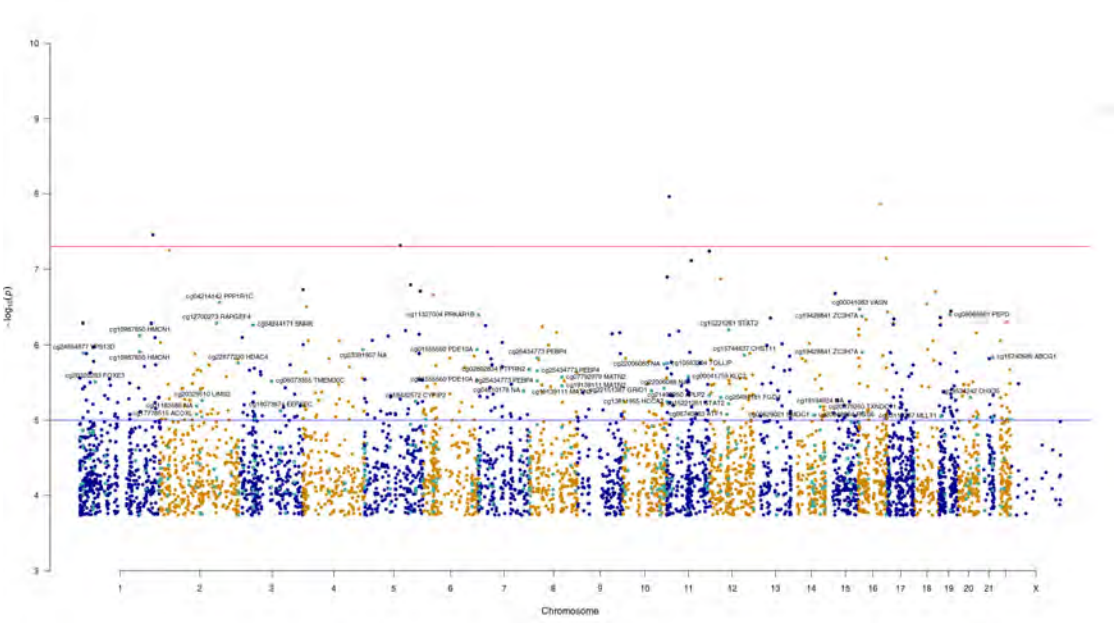


Figure 4. Manhattan plot showing the full list of associations derived from the EWAS on insulin resistance. Loci from the 'validation-list' are highlighted in green and labelled.

Table 1. Association results for the top 25 CpGs from the 'validation-list'.

Chromosome	Position	CpG	Gene Symbol	logFC	AveExpr	t	P-Value	FDR	Beta	Approach or Group Comparison
2	182966420	cg04214142	*PPP1R1C*	0.49	2.09	5.35	2.77E-07	0.01	6.31	Longitudinal G5 vs. G3
16	4424148	cg00041083	*VASN*	-0.62	3.87	-5.40	3.39E-07	0.16	3.80	Pubertal non-IR NW vs. IR Obese & Overweight
19	34013539	cg08085561	*PEPD*	-0.35	2.84	-5.36	4.06E-07	0.16	3.69	Pubertal non-IR NW vs. IR Obese & Overweight
7	645500	cg11327004	*PRKAR1B*	1.27	5.46	5.27	4.09E-07	0.02	5.97	Longitudinal G5 vs. G3
16	11891221	cg19428841	*ZC3H7A*	0.93	-5.39	5.26	4.21E-07	0.15	4.57	Longitudinal G3 vs. G2
2	173784502	cg12700273	*RAPGEF4*	0.78	3.32	5.22	5.20E-07	0.02	5.76	Longitudinal G5 vs. G3
3	43331949	cg04244171	*SNRK*	-0.44	1.26	-5.21	5.40E-07	0.15	4.39	Longitudinal G3 vs. G2
12	56747353	cg15221261	*STAT2*	5.25	5.74	5.18	6.28E-07	0.34	2.21	Longitudinal G3 (within)
1	186051470	cg10987850	*HMCN1*	-0.67	-1.00	-5.22	7.57E-07	0.47	3.55	Pubertal non-IR Obese & Overweight vs. IR Obese & Overweight
4	184693663	cg23391907	-	1.97	5.39	5.05	1.15E-06	0.02	5.06	Longitudinal G5 vs. G3
6	165958729	cg01555560	*PDE10A*	-0.53	3.78	-5.13	1.15E-06	0.47	3.26	Pubertal non-IR Obese & Overweight vs. IR Obese & Overweight
1	186051470	cg10987850	*HMCN1*	0.67	-1.01	5.12	1.23E-06	0.61	2.90	Pubertal IR vs. non-IR
16	11891221	cg19428841	*ZC3H7A*	0.85	-5.39	5.03	1.25E-06	0.34	1.84	Longitudinal G3 (within)
1	12336998	cg24654877	*VPS13D*	1.66	4.96	5.02	1.28E-06	0.55	-1.04	Longitudinal G4 (within)
12	105073955	cg15744837	*CHST11*	0.28	0.13	5.01	1.36E-06	0.20	4.29	Longitudinal G1 vs. G3
21	43655919	cg16740586	*ABCG1*	-0.34	0.88	-5.07	1.45E-06	0.24	2.86	Pubertal non-IR NW vs. IR Obese & Overweight
8	22755479	cg25434773	*PEBP4*	-0.38	3.67	-5.07	1.47E-06	0.24	2.86	Pubertal non-IR NW vs. IR Obese & Overweight
2	240115678	cg22877230	*HDAC4*	0.91	4.31	4.95	1.74E-06	0.02	4.69	Longitudinal G5 vs. G3
11	1299477	cg10583204	*TOLLIP*	-0.55	2.96	-5.03	1.76E-06	0.24	2.74	Pubertal non-IR NW vs. IR Obese & Overweight
10	126437015	cg22006088	-	-0.52	4.95	-5.02	1.80E-06	0.48	2.97	Pubertal non-IR Obese & Overweight vs. IR Obese & Overweight
7	157650205	cg02802834	*PTPRN2*	1.08	3.05	4.91	2.11E-06	0.02	4.52	Longitudinal G5 vs. G3
8	22755479	cg25434773	*PEBP4*	0.33	3.68	4.98	2.20E-06	0.61	2.53	Pubertal IR vs. non-IR
11	66024941	cg00041759	*KLC2*	0.94	-5.28	4.86	2.65E-06	0.54	1.42	Longitudinal G3 (within)
8	98900139	cg07792979	*MATN2*	0.93	3.84	4.86	2.67E-06	0.02	4.31	Longitudinal G5 vs. G3
6	165958729	cg01555560	*PDE10A*	0.51	3.78	4.91	2.92E-06	0.61	2.35	Pubertal IR vs. non-IR

The logFC field represents the change in the average M-value between conditions (a change in the log-odds of methylation; larger logFC refers to stronger differential methylation). The AveExpr field represents the average M-value across all samples, which gives a measure of the overall amount of methylation for each probe. The B-statistic is the log-odds of differential methylation to constant methylation (note, not the log-odds of methylation to nonmethylation, which is the M-value itself).

adjustment. A sub-selection of the top significant associations from these analyses is available in **figure 5** (P-value threshold < 0.005). Interestingly, the methylation degree of three genes showed significant associations with assessed phenotypic traits at both cross-sectional stages (prepubertal and pubertal) *(CNBD2, FGD4*, and *VASN)*. Among them, the loci *CNBD2* and *FGD4* reported an association with anthropometry (BMI Z-Score and WC). At the same time, *VASN* reinforced its association with glucose metabolism (glucose, insulin and HOMA-IR) in the pubertal stage. *CNBD2* and *FGD4* also presented associations with metabolic traits in the pubertal stage including, leptin concentrations and QUICKI, and triacylglycerol levels, respectively. Previously literature described genes such as *ABCG1*, or other novels such as *VASN, CEMIP* and *HMCN1* reported strongly significant associations with glucose metabolism at the pubertal stage (glucose, insulin and HOMA-IR) in this study. On the other hand, DNAm in the *ESR1* and *VASN* genes was associated with kidney-function markers (creatinine and uric acid levels).

4. Association between gene expression and DNA methylation – A genome-wide eQTM analysis in human blood cells

With the intention of investigating the mechanistic relevance of identified IR epigenetics marks, RNA-seq analysis was performed in a subset of 44 children from the pubertal cross-sectional approach (**Figure 1**), which had also been included in the EWAS analysis. In this sub-population, we searched for cis-eQTMs in and around 10.000 bp of each transcript, and trans-eQTMs if the distance between the CpG and the transcript was higher than 10.000 bp. These analyses focused on the 267 CpGs from the 'validation-list' and 60,058 high-quality transcripts (whole-genome distributed) as input.

The cis-eQTM analysis identified 19 CpG-transcript pairs that met a P-value < 0.05, comprising 19 transcripts and 17 CpG sites (**Table 2A**). Methylation levels of CpG sites were both positively (45%) and negatively (55%) correlated with expression levels. Some genes reported in our regression analysis on obesity phenotypes were also revealed here as cis-eQTMs (highlighting *ABCG1, CEMIP, CNBD2, ESR1, FGD4, HMCN1* and *VASN*). Among them, only the *FGD4*, showed an FDR < 0.05. Identifed cis-eQTM CpGs for the genes *HMCN1, CASP7* and *VASN*, were annotated within enhancer regions according to the list of 450k enhancer predicted elements.

The trans-eQTM analysis identified 317 CpG-transcript pairs that met an FDR < 1x10-5, comprising 317 transcripts, and 5 CpG sites mapping the genes *CDC42BPB, CEMIP, LIN7A* and *RASGRF1* (Additional file 11). Methylation levels of CpG sites were positively correlated with expression levels in the majority of trans-eQTM pairs (98.74%).

Identified cis- and trans-eQTMs were annotated according to the genomics regulatory elements they map (Additional file 12). The annotation consisted of two categories: 1) distance to a CpG island and 2) annotation to gene region. In terms of distance to CpG islands, we did not

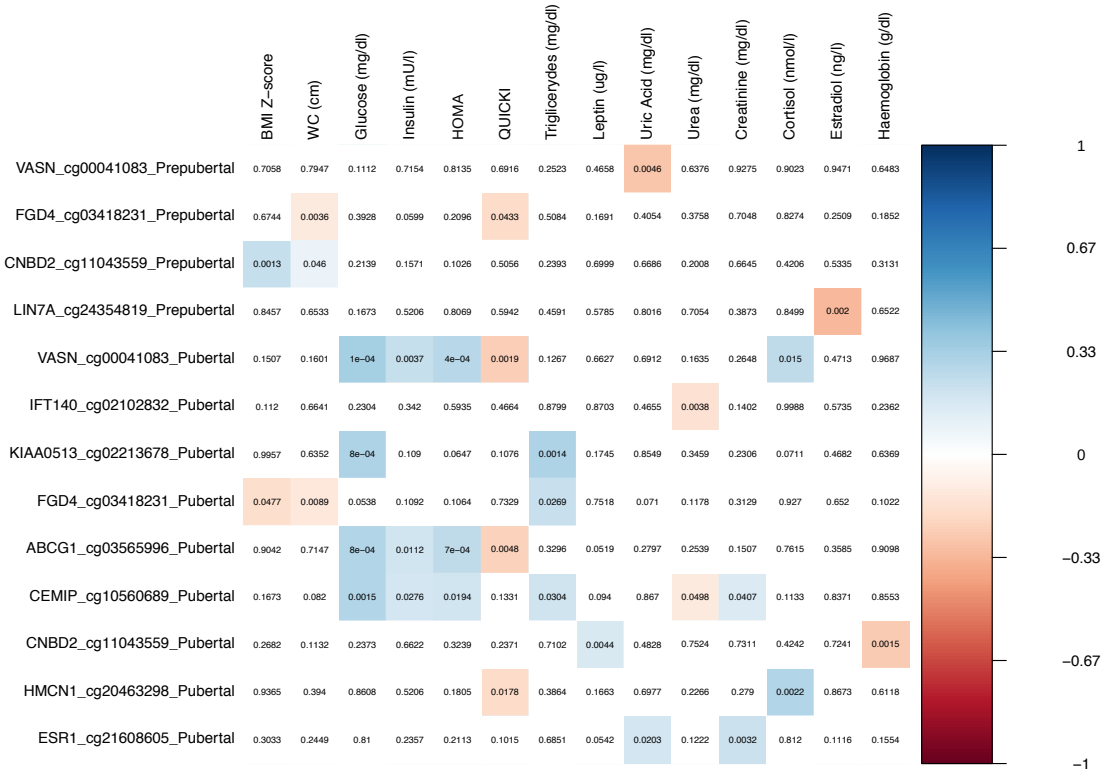


Figure 5. All significant associations showing a P-value < 0.005 with at least one trait in our continuous outcome DNA methylation analyses.

find differences in the annotation of eQTMs compared to the annotation derived from the whole list of CpGs in the EPIC array. On the contrary, in terms of annotation to gene regions, we found a higher proportion of CpGs mapping promoter regions (e.g., *TSS200* and untranslated regions [UTRs]) among the eQTMs in comparison to the whole list of CpGs in the EWAS EPIC array. Using the BIOS QTL browser [62, 63], we validated *in silico* some of the identified eQTMs. Particularly, we found evidences of blood cis-eQTMs for the same transcripts but distinct CpGs in the genes *ABCG1, CDC42BPB, ESR1, IFT140* and *LIN7A*.

Some identified cis- and trans-eQTM loci, like the *ABCG1, ESR1, FGD4, VIPR2, RGS6,* and *CEMIP*, mapped biological process GO-terms with relevance in obesity, puberty and metabolism; 'activation of GTPase activity', 'antral ovarian follicle growth', 'cellular response to estrogen stimulus', 'G protein-coupled receptor signaling pathway', 'positive regulation of protein kinase C activity', or 'regulation of cholesterol esterification'.

Table 2. *Cis* CpG-transcript pairs (distance of 10,000 bp) identified in the eQTM analysis.

Cis-eQTMs					
CpG_Gene	**Transcript_Gene**	**Statistic**	**P-Value**	**FDR**	**Beta**
cg03418231_FGD4	ENSG00000139132_FGD4	-5.059	1.04E-05	0.004	-0.854
cg10956605_VIPR2	ENSG00000106018_VIPR2	3.796	0.001	0.108	1.759
cg25069618_CEMIP	ENSG00000103888_CEMIP	3.487	0.001	0.112	1.974
cg26675212_KLHL29	ENSG00000119771_KLHL29	3.483	0.001	0.112	2.419
cg02102832_IFT140	ENSG00000131634_TMEM204	3.467	0.001	0.112	1.452
cg00041083_VASN	ENSG00000262246_CORO7	-2.817	0.008	0.497	-0.882
cg02102832_IFT140	ENSG00000187535_IFT140	2.793	0.008	0.497	1.166
cg03565996_ABCG1	ENSG00000160179_ABCG1	-2.739	0.009	0.500	-0.823
cg21608605_ESR1	ENSG00000091831_ESR1	2.604	0.013	0.554	0.733
cg02213678_KIAA0513	ENSG00000135709_KIAA0513	-2.573	0.014	0.554	-0.794
cg09050582_LIPJ	ENSG00000204022_LIPJ	-2.570	0.014	0.554	-0.821
cg20463298_HMCN1	ENSG00000143341_HMCN1	2.438	0.019	0.699	0.595
cg11078674_RGS6	ENSG00000182732_RGS6	-2.368	0.023	0.762	-1.411
cg23202420_NPBWR2	ENSG00000286999_NA	2.280	0.028	0.870	1.599
cg00041083_VASN	ENSG00000168140_VASN	-2.168	0.036	1.000	-0.712
cg12363898_IL17RB	ENSG00000271976_NA	-2.137	0.039	1.000	-0.405
cg21891499_CLPTM1L	ENSG00000286388_NA	-2.126	0.040	1.000	-0.696
cg11043559_CNBD2	ENSG00000149646_CNBD2	2.066	0.046	1.000	0.998
cg17580480_CASP7	ENSG00000165806_CASP7	-2.053	0.047	1.000	-0.935

As a measure of eQTMs effect size, we reported the beta regressors estimated by the linear model. The P-Values from the linear regression analysis were adjusted for multiple comparisons using the Benjamini-Hochberg FDR procedure. Theses analyses focused on 267 CpGs and 60,058 high quality transcripts as input.

5. Association between genetic variation and DNA methylation – A genome-wide mQTL analysis in human blood cells

At each cross-sectional approach (prepubertal and pubertal), mQTL analyses were conducted, revealing epigenetics regulating phenomena by which SNPs affect the methylation levels of a CpG. These analyses focused on the 267 CpGs from the 'validation-list' and 471,192 SNPs, whole-genome distributed, as input.

At the prepubertal stage, a total of 7 SNP-CpG pairs were found to be located in cis and 5 SNP-CpG pairs were located in trans (P-value < 0.05 and FDR < 0.005 respectively) (**Table 3**). Cis-mQTLs involved genes with special relevance to type 2 diabetes, such as the *ADCY5* or others previously highlighted in our pipeline, like *ESR1*. All but one SNP-CpG pair located in trans involved the gene *BRD1*. At the pubertal stage, a total of 10 SNP-CpG pairs were found to be located in cis and 10 SNP-

CpG pairs were located in trans (P-value < 0.05 and FDR < 0.005 respectively) (**Table 4**). Cis-mQTLs for the loci *ADCY5, TINAGL1*, *GOLGA3* and *GRM6*, identified in the prepubertal approach, were also validated in the pubertal stage. The CpG mapping the *TINAGL1* was further annotated within an enhancer region. Two cis-mQTLs from the pubertal approach presented FDR < 0.05 (*MEGF6* and *SCN1A*). Another interesting gene, according to literature, and highlighted as cis-mQTL in the pubertal approach, is the *TNXB*. DNAm levels in this gene have been associated with an under nutrition status in adults [64] and previously reported as a mQTL in human pancreatic islets [39]. Using the BIOS QTL browser [62, 63], we validated in silico some of identified mQTLs. Particularly, we found evidence of blood cells cis-mQTLs for the genes (*ESR1*, *MEGF6* and *TNXB*), although involving different CpGs and SNPs.

Table 3a. *Cis* SNP-CpG pairs (distance of 500 bp) identified in the mQTL analysis of the prepubertal stage.

Cis-mQTLs					
SNP	**CpG_Gene**	**Statistic**	**P-Value**	**FDR**	**Beta**
rs73186452	cg00978808_ADCY5	3.391	9.99E-04	0.074	0.939
GSA.rs114659838	cg07504762_TINAGL1	3.074	0.003	0.089	1.672
rs12282	cg21561989_GOLGA3	-2.980	0.004	0.089	-0.592
GSA.rs117733790	cg23792592_MIR1-1	2.434	0.017	0.306	1.053
GSA.rs35882398	cg12001846_ESR1	-2.351	0.021	0.306	-1.302
rs2067011	cg10648542_GRM6	2.171	0.032	0.398	0.291
GSA.rs114827188	cg20329510_LIMS2	-2.107	0.038	0.398	-0.576

As a measure of mQTLs effect size, we reported the beta regressors estimated by the linear model. The p-values from the linear regression analysis were adjusted for multiple comparisons using the Benjamini-Hochberg FDR procedure.

Table 3b. *Trans* SNP-CpG pairs (distance higher than 500 bp) identified in the mQTL analysis of the prepubertal stage.

Trans-mQTLs					
SNP	**CpG_Gene**	**Statistic**	**P-Value**	**FDR**	**Beta**
rs28372042	cg16053902_BRD1	-8.873	2.91E-14	2.53E-06	-1.052
rs14065	cg20872261_SLC37A2	-8.808	4.03E-14	2.53E-06	-1.050
rs138843	cg16053902_BRD1	-8.622	1.03E-13	4.30E-06	-1.069
GSA.rs11912619	cg16053902_BRD1	-7.669	1.16E-11	3.65E-04	-1.039
GSA.rs7287579	cg16053902_BRD1	-7.055	2.29E-10	0.006	-1.010

As a measure of mQTLs effect size, we reported the beta regressors estimated by the linear model. P-values were adjusted with a correction value for multiple testing, which takes into consideration the dependency of linkage disequilibrium (LD) between SNPs by LD based pruning and thereby uses the number of independent tests. These analyses focused on the 267 CpGs from the 'validation-list' and 471,192 SNPs, whole-genome distributed, as input. The correction value for the trans-analysis was calculated as the total number of analysed CpG sites multiplied by the number of SNPs in the whole dataset.

Table 4a. *Cis* SNP-CpG pairs (distance of 500 bp) identified in the mQTL analysis of the pubertal stage.

Cis-mQTLs					
SNP	**CpG_Gene**	**statistic**	**P-Value**	**FDR**	**Beta**
GSA.rs78267041	cg00418943_MEGF6	3.750	0.00026329	0.019	1.583
rs16851382	cg15434576_SCN1A	-3.563	0.00051099	0.019	-0.533
rs12282	cg21561989_GOLGA3	-3.088	0.002	0.051	-0.505
rs7774197	cg24252708_TNXB	-2.968	0.004	0.051	-0.952
GSA.rs114888185	cg20320283_FOXE3	-2.959	0.004	0.051	-1.121
GSA.rs114659838	cg07504762_TINAGL1	2.918	0.004	0.051	1.258
rs73186452	cg00978808_ADCY5	2.447	0.016	0.166	0.679
rs2067011	cg10648542_GRM6	2.189	0.030	0.281	0.250
rs3742476	cg12913090_ATG2B	-2.131	0.035	0.287	-0.254
GSA.rs33932952	cg25711726_SLC37A2	2.019	0.046	0.337	0.296

As a measure of mQTLs effect size, we reported the beta regressors estimated by the linear model. The p-values from the linear regression analysis were adjusted for multiple comparisons using the Benjamini-Hochberg FDR procedure.

Table 4b. *Trans* SNP-CpG pairs (distance higher than 500 bp) identified in the mQTL analysis of the pubertal stage.

Trans-mQTLs					
SNP	**CpG_Gene**	**statistic**	**P-Value**	**FDR**	**Beta**
GSA.rs11912619	cg16053902_BRD1	-11.263	4.87E-21	6.13E-13	-1.189
rs138843	cg16053902_BRD1	-11.116	1.14E-20	7.19E-13	-1.148
rs14065	cg20872261_SLC37A2	-10.820	6.34E-20	2.66E-12	-0.953
GSA.rs7287579	cg16053902_BRD1	-10.723	1.11E-19	3.50E-12	-1.179
rs7410612	cg16053902_BRD1	-10.361	8.97E-19	2.26E-11	-1.121
rs28372042	cg16053902_BRD1	-10.282	1.41E-18	2.97E-11	-1.069
rs1009321	cg21561989_GOLGA3	-7.104	6.79E-11	0.001	-0.678
GSA.rs761878	cg16053902_BRD1	-7.021	1.05E-10	0.002	-1.077
rs4477450	cg20872261_SLC37A2	-6.936	1.63E-10	0.002	-0.732
rs2824560	cg20401955_CHODL	6.807	3.17E-10	0.004	0.678

As a measure of mQTLs effect size, we reported the beta regressors estimated by the linear model. P-values were adjusted with a correction value for multiple testing, which takes into consideration the dependency of linkage disequilibrium (LD) between SNPs by LD based pruning and thereby uses the number of independent tests. These analyses focused on the 267 CpGs from the 'validation-list' and 471,192 SNPs, whole-genome distributed, as input. The correction value for the trans-analysis was calculated as the total number of analysed CpG sites multiplied by the number of SNPs in the whole dataset.

6. Serum protein levels of vasorin are associated with IR and obesity in the pubertal stage

To further investigate the role of one of the most promising biomarkers identified for IR, we measured VASN serum protein levels in the cohort. Descriptive statistics for experimental groups reported lower levels of VASN protein significantly associated with IR and obesity in the pubertal stage of the children (P=0.007) (Additional file 3). Moreover, in the longitudinal approach (N = 90), groups maintaining or developing IR with the onset of puberty, G4 and G5, presented the lowest increases in VASN levels (P = 0.06) (Additional file 13). At the opposite, insulin-sensitive groups such as G1 and G3 showed a pronounced increase in VASN levels.

With the aim of functionally validate our EWAS and eQTM findings for *VASN* in the pubertal stage, we also studied the correlation between *VASN* DNAm and VASN serum protein levels, as well as between *VASN* mRNA and VASN serum protein levels. Interestingly, in the 130 pubertal children, a suggestive trend was reported for the correlation between DNAm at the cg00041083 and VASN protein levels after adjusting for confounders such as age, sex, origin and BMI Z-Score (P = 0.09) (Additional file 14); higher DNAm levels related to lower protein levels. Contrarily, in the 44 pubertal children with available RNAseq data, we did not find a significant correlation between mRNA levels at the *ENSG00000168140* and VASN serum levels (P = 0.34).

Discussion

The current large-scale integrative molecular analysis identifies novel blood multi-omics signatures such as DNAm marks, eQTMs and mQTLs, underlying the development, amelioration and worsening of IR in children with obesity during puberty. Functional enrichment analysis revealed that identified loci participate in systemic metabolic pathways and sexual maturation processes with relevance to the pathogenesis of IR. Additional analyses on cardiometabolic and inflammatory phenotypes show that blood DNAm patterns of some identified loci are further associated, beyond IR, with an overall risky-cardiometabolic profile in children. To our knowledge, this is the first longitudinal multi-omics approach characterizing molecular blood alterations for IR and obesity during the metabolically critical period of puberty. With our results, we propose novel and promising biomarkers with predictive utility for the identification of children with obesity at high risk of developing IR and metabolic alterations. Likewise, we also aid insights into the molecular and functional mechanisms linking epigenetics alterations and the IR phenotype in obesity.

Our EWAS analysis on IR identified 4,281 associated unique CpG sites, from which 2,981 further presented an FDR < 0.05 (linked to 2,632 and 1,899 genes, respectively) (**Figure 1**). Among them, we selected only those loci presenting significant associations in at least two of our statistical approaches (**Figure 3**). The resulting list was composed of 267 IR-associated unique CpGs mapping 128 genes, from which 130 CpGs (mapping 91 genes) presented an FDR < 0.05 (**Figure 4** and

Additional file 7). Among our top significant results (**Table 1** and Additional file 7), there were new and promising regions never reported as epigenetics marks of IR (e.g., *CDC42BPB, ESR1, HMCN1, PRKAR1B, SNRK* and *VASN*, among others). From them, DNAm levels of the *ESR1* showed association with IR not only in the pubertal but also in the prepubertal stage (**Figure 3**), indicating that they might accompany IR already from early childhood. The rest of them otherwise showed association with IR in our longitudinal and pubertal approaches, resembling marks associated with IR in the context of puberty. On these and the rest loci from our 'validation-list', associations with a bulk of cardiometabolic phenotypes other than IR were also investigated (**Figure 5**). As a novelty, these confounding-adjusted analyses allowed us to distinguish between regions in which the initial IR-association is direct (e.g., *ABCG1* and *VASN*), or rather derives from an indirect or secondary association with anthropometry and obesity traits (e.g., *CNBD2* and *FGD4*), or with inflammation (e.g., *CDC42BPB*). Interestingly, the functional enrichment analysis of reported CpGs indicated that identified loci participate in systemic metabolic pathways, inflammatory and sexual maturation processes with relevance to the pathogenesis of IR. Among them, the terms related to the synthesis and secretion of sexual hormones outline the importance of puberty and its hormonal and biochemical changes as plausible contributors to the development and worsening of obesity IR.

Beyond the new targets identified, we also found genes whose methylation levels have been previously and repeatedly associated with adult type 2 diabetes and obesity in the literature (e.g., *ABCG1, ADCY5, CPT1A, FTO, HCCA2, HDAC4, HIF3A, IGF-1, KCNQ1, PPARG,* and *TCF7L2*, among others) [13, 23]. Therefore, our results remark the role of IR as an important pathophysiological mechanism linking obesity and cardiometabolic comorbidities, and reinforce the fact that epigenetics marks of IR may have utility as predictive markers of future disease outcomes. In addition to type 2 diabetes, our associations also highlighted loci specifically associated with adult IR in the literature (e.g., *COL18A1, CTNND2, CXCL1, DNMT3A, GRB10, HDAC4, LAT, PAX6, SH3RF3* and *SIRT2*) [18–22]. This is important since literature IR studies had mostly focused on studying the relationship between the methylation levels of candidate genes and the HOMA-IR [22], and EWAS on IR are still scarce [14–16, 22, 65] (with barely one study conducted in children) [66]. The fact of validating here previously known adult epigenetics marks of IR may indicate that the DNA methylation patterns of IR are established early, under the influence of childhood or puberty obesogenic environments, and remain stable throughout adulthood.

In order to elucidate the molecular mechanisms behind identified IR epigenetics marks, we integrated our EWAS data along with other omics sources in the same cohort (GWAS and RNAseq data). As a result, we reported that some of the identified loci might be participating in phenomena that alter gene expression levels (eQTMs) while others could be explained by the existence of SNPs (mQTLs) (**Figure 1**).

Regarding eQTMs phenomena, our analysis reported that some of the most promising regions identified could exert their effects of IR through a modification, either up- or down-regulating, the expression of target transcripts *in situ* (*cis*-eQTMs), or at long distances from their occurrence (trans-eQTMs) (**Table 2**). Among identified cis-eQTM phenomena, we can highlight previously well-known eQTM loci (like the *ABCG1*) [23] but also some of our promising new markers (*CDC42BPB*, *ESR1*, *HMCN1* and *VASN*). Interestingly, most identified CpGs mapped into genomics regulatory elements (e.g., enhancers, transcription start sites or UTR), reinforcing their role as plausible gene expression controllers. To date, this is the first study integrating RNAseq and EWAS data in the blood of pubertal insulin-resistant children with obesity. Previously, a recent study investigated the existence of eQTMs phenomena in the adipose tissue of African American adult women with IR [17]. Although with no overlapping regions between their and our approach, our findings reinforce the idea that DNAm-mediated regulation of gene expression could be implicated into the pathogenesis of IR.

Previous studies have shown that DNAm alterations at the cg06500161 of *ABCG1* strongly correlate with glucose metabolism dysfunction and diabetes in adults [14, 65, 67]. Many of these studies have also evidenced that these DNAm alterations elicit effects on *ABCG1* gene expression through eQTM interactions. Moreover, the connection between *ABCG1* and diabetes-related traits and dyslipidemia has been supported by animal and human studies [14, 65, 67]. The protein encoded by *ABCG1* is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes. More specifically, ABCG1 is involved in macrophage cholesterol and phospholipid transport and may regulate cellular lipid homeostasis in other cell types. Although a bulk of insights had been reported in adults, our results are the first evidence of such relationships in children with obesity and IR. Thus, we contribute to the body of evidence supporting the role of lipid metabolism, specifically of ABC transporters, in IR and highlight the interest of *ABCG1* as a potential non-invasive biomarker for future glucose metabolism complications.

Here, mQTL analyses were also conducted revealing epigenetics regulating phenomena, by which SNPs affect the methylation levels of CpGs (**Figure 1**). For this approach, we counted on the EWAS data from both the prepubertal and pubertal stages, which allowed us to look for replicated mQTL phenomena across time points. Among results, we again reported some previously known diabetes loci (such is the case of *ADCY5*) [68, 69] but also interesting new IR epigenetics marks (*ESR1*), for which previous mQTL evidence had been reported in the literature [13]. These phenomena could be, therefore, the molecular explanation for some epigenetics IR early-life marks for which the environment is not the causal mediator.

ESR1 is a non-imprinted gene that encodes the estrogen receptor-α (ER-α), a transcription factor involved in the regulation of energy homeostasis [70]. In females, estrogens maintain energy

homeostasis via ERα by suppressing energy intake and lipogenesis, enhancing energy expenditure [71] and ameliorating insulin secretion and sensitivity [72]. In males, however, testosterone is converted to estrogen and maintains fuel homeostasis via ERα and androgen receptors, which share related functions to suppress adipose tissue accumulation and improve insulin sensitivity. Conversely, the lowering in estrogens levels observed in postmenopausal women provoke IR and increase the risk of type 2 diabetes [70]. Although no previous evidence of an association between DNAm levels and IR has been reported for the *ESR1* in the literature, dynamic changes in the DNAm of this region have been previously associated with aging [73]. It is noteworthy to see how the *ESR1*, which had not been previously evidenced as an epigenetic marker of IR, appears as a significant locus in all the approaches of our study (EWAS on IR, association with cardiometabolic traits, *cis*-eQTMs and *cis*-mQTLs). From this, we can conclude several things; 1) DNAm alterations in the *ESR1* locus during puberty could be an important contributor to the appearance and worsening of IR in children with obesity, 2) these alterations could exert their effects on the phenotype through the alteration of *ESR1* gene expression levels, and 3) there could also be some from-birth predisposing SNPs favouring the alteration of DNAm levels in the region. The evidenced mQTL phenomenon of the region agrees with the fact of *ESR1* appearing as a significant epigenetic marker of IR in both prepubertal and pubertal stages in our study. Considering our results and the implication of the estrogen axis into the context of IR and puberty, we propose that *ESR1* could be a promising epigenetic target to prevent age-related metabolic disorders associated with obesity.

Besides the *ESR1*, the most promising and novel biomarker identified from our approach is the *VASN*, which was reported as a top association from the 'validation-list' in the EWAS on IR (**Figure 4** and **table 1**), as well as a participant of a *cis*-eQTM phenomenon. Particularly, we report for the first time that both higher blood *VASN* DNAm levels and lower serum protein concentrations are strongly associated with IR in the pubertal stage in children with obesity. Although our eQTM analysis showed that the higher DNAm of *VASN* is associated with lower mRNA *VASN* levels in our children, we could not validate the results with an association between mRNA *VASN* levels and VASN serum levels. This is not surprising otherwise since the elevated VASN serum levels associated with IR are a systemic finding that could derive from many other tissues than blood cells. Moreover, the population sample size with RNAseq data and VASN serum levels measured was small (barely 40 subjects). VASN is a type I transmembrane protein (SLIT-like 2), highly expressed in smooth muscle cells and with reported expression in adipocytes [74, 75]. VASN was originally found to play a role in vascular injury repair and angiogenesis, and is a potential biomarker for hepatocarcinoma [75]. Mechanistically, VASN directly binds to the transforming growth factor (TGF-B) and attenuates TGF-B signaling *in vitro*. A recent study has also shown that hypoxia increases Notch signaling in glioma-like cells through the induction of VASN and the *hypoxia-inducible factor-1* (HIF1)/STAT3), thus describing a possible action mechanism of VASN. However, the relationship between VASN, obesity and IR remains unknown. Our main hypothesis is that VASN could play an important role in

obesity as a potential biomarker of IR and/or a predictor of future development of type 2 diabetes in children.

The main limitations of our study are, on the one hand, the low number of participants included in our populations. On the other hand, it is the fact that our findings are based on data from blood, which was the only accessible tissue, and may not be representative of other metabolically relevant organs such as live and adipose and muscle tissues. In this regard, there is a trend pointing to a correlation between the global state of methylation in blood and adipose tissue [76]. This correlation might be explained by the abundant presence of white cells in both tissues and suggests that buffy coat might be a valid indicator of what happens at the methylation level in adipose tissue, especially for the case of inflammatory and immune system-related aspects. Another possible source of bias would be the difference in time elapsed between the two measurements (prepubertal and pubertal times) between the different participants.

The main strengths of our study are the high significance of our associations (many of them passing multiple-test correction thresholds) as well as the multi-omics design, from which we validate our top associations in a multi-omics dimensional space. Likewise, another positive point is the future pubertal study design, which strengthens the statistical robustness of our reports. Finally, it is a fact of being the first longitudinal multi-omics approach characterizing molecular blood alterations for IR and obesity during the metabolically critical period of puberty.

Conclusions

With our results, we propose novel and promising biomarkers of IR and metabolic alterations in children with obesity (*ABCG1, CDC42BPB, ESR1, HMCN1, PRKAR1B, SNRK* and *VASN*, among others). Thanks to our multi-omics design, we also aid insights into the molecular and functional mechanisms linking epigenetics alterations and the IR phenotype in obesity (mQTLs and eQTMs). If validated in other cohorts and longitudinal designs, our identified loci could serve as predictive non-invasive biomarkers for reducing the high rates of mortality and morbidity associated with obesity. Especially for genes with a promising but unknown role in the development of IR in the adipose tissue, such is the case of *VASN*, additional *in vitro* and *in vivo* functional analyses should be conducted in the near future.

Acknowledgments

The authors would like to thank the children and parents who participated in the study. This work was supported by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I + D + I), Instituto de Salud Carlos III-Health Research Funding (FONDOS FEDER) (PI05/1968, PI11/01425, PI11/02042, PI11/02059, PI16/01301, PI16/01205 and PI16/00871); CIBEROBN Network (CB15/00131, CB15/00043); the Regional Government of Andalusia ("Plan Andaluz de investigación, desarrollo e innovación (2018), P18-RT-2248); Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015) and the Mapfre Foundation. The authors also acknowledge Instituto de Salud Carlos III for personal funding: Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016 (IFI17/00048).

Supplementary data (additional files)

Additional files are available online at
https://drive.google.com/drive/folders/19SrdQjUZSXxPKi-dn4GMFjrzKZ4DR9mN?usp=sharing.

References

1. Collaborators TG 2015 O. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N Engl J Med. 2017;377:13–27. doi:10.1056/NEJMoa1614362.
2. Abarca-Gómez L, Abdeen ZA, Hamid ZA, Abu-Rmeileh NM, Acosta-Cazares B, Acuin C, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet. 2017;390:2627–42. doi:10.1016/S0140-6736(17)32129-3.
3. Lloyd LJ, Langley-Evans SC, McMullen S. Childhood obesity and risk of the adult metabolic syndrome: A systematic review. International Journal of Obesity. 2012;36:1–11. doi:10.1038/ijo.2011.186.
4. Gepstein V, Weiss R. Obesity as the Main Risk Factor for Metabolic Syndrome in Children. Frontiers in Endocrinology. 2019;10. doi:10.3389/fendo.2019.00568.
5. Jones RE, Jewell J, Saksena R, Ramos Salas X, Breda J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. Front Public Heal. 2017;5:58. doi:10.3389/fpubh.2017.00058.
6. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346:393–403. doi:10.1056/NEJMoa012512.
7. Langenberg C, Sharp SJ, Schulze MB, Rolandsson O, Overvad K, Forouhi NG, et al. Long-term risk of incident type 2 diabetes and measures of overall and regional obesity: The epic-interact case-cohort study. PLoS Med. 2012;9:17. doi:10.1371/journal.pmed.1001230.
8. Abbassi V. Growth and Normal Puberty. 1998 Aug;102(2 Pt 3):507-11.
9. Kelsey MM, Pyle L, Hilkin A, Severn CD, Utzschneider K, van Pelt RE, et al. The impact of obesity on insulin sensitivity and secretion during pubertal progression: A longitudinal study. J Clin Endocrinol Metab. 2020;105. doi:10.1210/clinem/dgaa043.
10. Reinehr T, Wolters B, Knop C, Lass N, Holl RW. Strong effect of pubertal status on metabolic health in obese children: A longitudinal study. J Clin Endocrinol Metab. 2015;100:301–8. doi:10.1210/jc.2014-2674.
11. Reinehr T, Roth CL. Is there a causal relationship between obesity and puberty? Lancet Child Adolesc Heal. 2019;3:44–54. doi: 10.1016/S2352-4642(18)30306-7.
12. Dabelea D, Bell RA, D'Agostino RB, Imperatore G, Johansen JM, Linder B, et al. Incidence of diabetes in youth in the United States. J Am Med Assoc. 2007;297:2716–24. doi:10.1001/jama.297.24.2716.
13. Ling C, Rönn T. Epigenetics in Human Obesity and Type 2 Diabetes. Cell Metabolism. 2019;29:1028–44. doi:10.1016/j.cmet.2019.03.009.

14. Hidalgo B, Irvin MR, Sha J, Zhi D, Aslibekyan S, Absher D, et al. Epigenome-wide association study of fasting measures of glucose, insulin, and homa-ir in the genetics of lipid lowering drugs and diet network study. Diabetes. 2014;63:801–7. doi:10.2337/db13-1100.

15. Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: A nested case-control study. Lancet Diabetes Endocrinol. 2015;3:526–34. doi:10.1016/S2213-8587(15)00127-8.

16. Kulkarni H, Kos MZ, Neary J, Dyer TD, Kent JW, Göring HHH, et al. Novel epigenetic determinants of type 2 diabetes in Mexican-American families. Hum Mol Genet. 2015;24:5330–44. doi:10.1093/hmg/ddv232.

17. Sharma NK, Comeau ME, Montoya D, Pellegrini M, Howard TD, Langefeld CD, et al. Integrative analysis of glucometabolic traits, adipose tissue dna methylation, and gene expression identifies epigenetic regulatory mechanisms of insulin resistance and obesity in African Americans. Diabetes. 2020;69:2779–93. doi:10.2337/db20-0117.

18. Ling C, Rönn T. Epigenetic markers to further understand insulin resistance. Diabetologia. 2016;59:2295–7. doi:10.1007/s00125-016-4109-y.

19. Arner P, Sahlqvist A-S, Sinha I, Xu H, Yao X, Waterworth D, et al. The epigenetic signature of systemic insulin resistance in obese women. Diabetologia. 2016;59:2393–405. doi:10.1007/s00125-016-4074-5.

20. Małodobra-Mazur M, Alama A, Bednarska-Chabowska D, Pawelka D, Myszczyszyn A, Dobosz T. Obesity-induced insulin resistance via changes in the DNA methylation profile of insulin pathway genes. Adv Clin Exp Med. 2019;28:1599–607. doi:10.17219/acem/110321.

21. Arpón A, Santos JL, Milagro FI, Cataldo LR, Bravo C, Riezu-Boj JI, et al. Insulin sensitivity is associated with lipoprotein lipase (Lpl) and catenin delta 2 (ctnnd2) dna methylation in peripheral white blood cells in non-diabetic young women. Int J Mol Sci. 2019;20. doi:10.3390/ijms20122928.

22. Arpón A, Milagro FI, Ramos-Lopez O, Mansego ML, Santos JL, Riezu-Boj JI, et al. Epigenome-wide association study in peripheral white blood cells involving insulin resistance. Sci Rep. 2019;9. doi:10.1038/s41598-019-38980-2.

23. Juvinao-Quintero DL, Marioni RE, Ochoa-Rosales C, Russ TC, Deary IJ, van Meurs JBJ, et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin Epigenetics. 2021;13:40. doi:10.1186/s13148-021-01027-3.

24. Florath I, Butterbach K, Heiss J, Bewerunge-Hudler M, Zhang Y, Schöttker B, et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia. 2016;59:130–8. doi:10.1007/s00125-015-3773-7.

25. Soriano-Tárraga C, Jiménez-Conde J, Giralt-Steinhauer E, Mola-Caminal M, Vivanco-Hidalgo RM, Ois A, et al. Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. Hum Mol Genet. 2016;25:609–19. doi:10.1093/hmg/ddv493.

26. Al Muftah WA, Al-Shafai M, Zaghlool SB, Visconti A, Tsai P-C, Kumar P, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clin Epigenetics. 2016;8:13. doi:10.1186/s13148-016-0177-6.

27. Meeks KAC, Henneman P, Venema A, Addo J, Bahendeka S, Burr T, et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: Findings from the RODAM study. Int J Epidemiol. 2019;48:58–70. doi:10.1093/ije/dyy171.

28. Han L, Zhang H, Kaushal A, Rezwan FI, Kadalayil L, Karmaus W, et al. Changes in DNA methylation from pre-to post-adolescence are associated with pubertal exposures. Clin Epigenetics. 2019;11:1–14. doi:10.1186/s13148-019-0780-4.

29. Suzuki MM, Bird A. DNA methylation landscapes: Provocative insights from epigenomics. Nature Reviews Genetics. 2008;9:465–76. doi:10.1038/nrg2341.

30. Sales V, Patti ME. The Ups and Downs of Insulin Resistance and Type 2 Diabetes: Lessons from Genomic Analyses in Humans. Current Cardiovascular Risk Reports. 2013;7:46–59. doi:10.1007/s12170-012-0283-8.

31. Sharma NK, Sajuthi SP, Chou JW, Calles-Escandon J, Demons J, Rogers S, et al. Tissue-specific and genetic regulation of insulin sensitivity-associated transcripts in African Americans. J Clin Endocrinol Metab. 2016;101:1455–68. doi:10.1210/jc.2015-3336.

32. Goodarzi MO. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. The Lancet Diabetes and Endocrinology. 2018;6:223–36. doi:10.1016/S2213-8587(17)30200-0.

33. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66:2888–902. doi:10.2337/db16-1253.

34. Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat Genet. 2017;49:1450–7. doi:10.1038/ng.3943.

35. Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, Chu AY, et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article. Nat Genet. 2018;50:559–71. doi:10.1038/s41588-018-0084-1.

36. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316:1331–6. doi:10.1126/science.1142358.

37. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445:881–5. doi:10.1038/nature05616.

38. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science (80-). 2007;316:1341–5. doi:10.1126/science.1142382.

39. Olsson AH, Volkov P, Bacos K, Dayeh T, Hall E, Nilsson EA, et al. Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets. PLoS Genet. 2014;10:1004735. doi:10.1371/journal.pgen.1004735.

40. Anguita-Ruiz A, Mendez-Gutierrez A, Ruperez AI, Leis R, Bueno G, Gil-Campos M, et al. The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity. Metabolism. 2020; 105:154187. doi: 10.1016/j.metabol.2020.154187.

41. Tanner JM, Whitehouse RH. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. Arch Dis Child. 1976;51:170–9. http://www.ncbi.nlm.nih.gov/pubmed/952550. Accessed 31 Aug 2018.

42. Sobradillo B, Aguirre A, Aresti U, Bilbao A, Ramos C, Lorenzo H, et al. Fernández Lizárraga A, R. I. Fernández Lizárraga A, . Curvas y tablas de crecimiento (estudios longitudinal y transversal). Fundación Faustino Orbegozo Eizaguirre Madrid, Spain. 2004.

43. McCrindle BW. Assessment and management of hypertension in children and adolescents. Nat Rev Cardiol. ;7(3):155-63. doi: 10.1038/nrcardio.2009.231.

44. Anguita-Ruiz A, Plaza-Diaz J, Ruiz-Ojeda FJ, Ruperez AI, Leis R, Bueno G, et al. X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline. Sci data. 2019;6:130. doi: 10.1038/s41597-019-0109-3.

45. Ruiz-Ojeda FJ, Anguita-Ruiz A, Rupérez AI, Gomez-Llorente C, Olza J, Vázquez-Cobela R, et al. Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism. Sci Rep. 2019;9:3979. doi:10.1038/s41598-019-40482-0.

46. Tang Q, Li X, Song P, Xu L. Optimal cut-off values for the homeostasis model assessment of insulin resistance (HOMA-IR) and pre-diabetes screening: Developments in research and prospects for the future. Drug Discov Ther. 2015;9:380–5. doi:10.5582/ddt.2015.01207.

47. Ziaee A, Esmailzadehha N, Oveisi S, Ghorbani A, Ghanei L. The threshold value of homeostasis model assessment for insulin resistance in Qazvin Metabolic Diseases Study (QMDS): assessment of metabolic syndrome. J Res Health Sci. 2015;15:94–100. http://www.ncbi.nlm.nih.gov/pubmed/26175291. Accessed 28 Jun 2019.

48. Rupérez AI, Olza J, Gil-Campos M, Leis R, Bueno G, Aguilera CM, et al. Cardiovascular risk biomarkers and metabolically unhealthy status in prepubertal children: Comparison of definitions. Nutr Metab Cardiovasc Dis. 2018;28:524–30. doi:10.1016/j.numecd.2018.02.006.

49. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:1–16. doi:10.1186/1471-2105-13-86.

50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75. doi:10.1086/519795.

51. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564–73. doi:10.1038/nprot.2010.116.

52. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60. doi:10.1038/nmeth.3317.

53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.

54. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30. doi:10.1093/bioinformatics/btt656.

55. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42:D749–55. doi:10.1093/nar/gkt1196.

56. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. 2018.

57. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2. 2019. https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

58. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics. 2013;29:1851–7. doi:10.1093/bioinformatics/btt311.

59. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11:14. doi:10.1186/gb-2010-11-2-r14.

60. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. doi:10.1093/nar/gkv007.

61. Kelsey MM, Zeitler PS. Insulin Resistance of Puberty. Current Diabetes Reports. 2016;16. doi:10.1007/s11892-016-0751-5.

62. Zhernakova D V., Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49:139–45. doi:10.1038/ng.3737.

63. Bonder MJ, Luijk R, Zhernakova D V., Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49:131–8. doi:10.1038/ng.3721.

64. Kesselmeier M, Pütter C, Volckmar AL, Baurecht H, Grallert H, Illig T, et al. High-throughput DNA methylation analysis in anorexia nervosa confirms TNXB hypermethylation. World J Biol Psychiatry. 2018;19:187–99. doi:10.1080/15622975.2016.1190033.

65. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, et al. Association between DNA Methylation in whole blood and measures of glucose metabolism: Kora F4 study. PLoS One. 2016;11. doi:10.1371/journal.pone.0152314.

66. Van Dijk SJ, Peters TJ, Buckley M, Zhou J, Jones PA, Gibson RA, et al. DNA methylation in blood from neonatal screening cards and the association with BMI and insulin sensitivity in early childhood. Int J Obes. 2018;42:28–35. doi:10.1038/ijo.2017.228.

67. Cardona A, Day FR, Perry JRB, Loh M, Chu AY, Lehne B, et al. Epigenome-wide association study of incident type 2 diabetes in a British population: EPIC-Norfolk study. Diabetes. 2019;68:2315–26. doi:10.2337/db18-0290.

68. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42:105–16. doi:10.1038/ng.520.

69. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44:659–69. doi:10.1038/ng.2274.

70. Mauvais-Jarvis F. Estrogen and androgen receptors: Regulators of fuel homeostasis and emerging targets for diabetes and obesity. Trends in Endocrinology and Metabolism. 2011;22:24–33. doi:10.1016/j.tem.2010.10.002.

71. Martínez De Morentin PB, González-García I, Martins L, Lage R, Fernández-Mallo D, Martínez-Sánchez N, et al. Estradiol regulates brown adipose tissue thermogenesis via hypothalamic AMPK. Cell Metab. 2014;20:41–53. doi:10.1016/j.cmet.2014.03.031.

72. Yan H, Yang W, Zhou F, Li X, Pan Q, Shen Z, et al. Estrogen improves insulin sensitivity and suppresses gluconeogenesis via the transcription factor Foxo1. Diabetes. 2019;68:291–304. doi:10.2337/db18-0638.

73. Kochmanski J, Goodrich JM, Peterson KE, Lumeng JC, Dolinoy DC. Neonatal bloodspot DNA methylation patterns are associated with childhood weight status in the Healthy Families Project. Pediatr Res. 2019;85:848–55. doi:10.1038/s41390-018-0227-1.

74. Ikeda Y, Imai Y, Kumagai H, Nosaka T, Morikawa Y, Hisaoka T, et al. Vasorin, a transforming growth factor β-binding protein expressed in vascular smooth muscle cells, modulates the arterial response to injury in vivo. Proc Natl Acad Sci U S A. 2004;101:10732–7. doi:10.1073/pnas.0404117101.

75. Li S, Li H, Yang X, Wang W, Huang A, Li J, et al. Vasorin is a potential serum biomarker and drug target of hepatocarcinoma screened by subtractive-EMSA-SELEX to clinic patient serum. Oncotarget. 2015;6:10045–59. doi:10.18632/oncotarget.3541.

76. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6. doi:10.1038/nature20784.

Section III

Section III

IMPLEMENTATION OF UNSUPERVISED MACHINE LEARNING (ML) MODELS FOR THE ANALYSIS OF LONGITUDINAL OMICS DATA IN OBESITY

PLoS Comput Biol. 2020;16(4):e1007792. doi:10.1371/journal.pcbi.1007792. IF: 4.475, Q1 at MATHEMATICAL & COMPUTATIONAL BIOLOGY.

Study 7

eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, Insights from obesity research

Augusto Anguita-Ruiz [1,2,3] *, Alberto Segura-Delgado[4], Rafael Alcalá[4], Concepción M. Aguilera[1,2,3], Jesús Alcalá-Fdez.[4]

Abstract Until date, several machine learning approaches have been proposed for the dynamic modeling of temporal omics data. Although they have yielded impressive results in terms of model accuracy and predictive ability, most of these applications are based on "Black-box" algorithms and more interpretable models have been claimed by the research community. The recent eXplainable Artificial Intelligence (XAI) revolution offers a solution for this issue, were rule-based approaches are highly suitable for explanatory purposes. The further integration of the data mining process along with functional-annotation and pathway analyses is an additional way towards more explanatory and biologically soundness models. In this paper, we present a novel rule-based XAI strategy (including pre-processing, knowledge-extraction and functional validation) for finding biologically relevant sequential patterns from longitudinal human gene expression data (GED). To illustrate the performance of our pipeline, we work on *in vivo* temporal GED collected within the course of a long-term dietary intervention in 57 subjects with obesity (GSE77962). As validation populations, we employ three independent datasets following the same experimental design. As a result, we validate primarily extracted gene patterns and prove the goodness of our strategy for the mining of biologically relevant gene-gene

Affiliations 1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Granada, Spain. / 2. Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain. / 3. CIBEROBN (Physiopathology of Obesity and Nutrition), Instituto de Salud Carlos III (ISCIII), Madrid, Spain. / 4. Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain.
*Corresponding author

temporal relations. Our whole pipeline has been gathered under open-source software and could be easily extended to other human temporal GED applications.

AUTHOR SUMMARY

Biological processes in humans are not single-gene based mechanisms, but complex systems controlled by regulatory interactions between thousands of genes. Within these gene regulatory networks, time-delay is a common phenomenon and genes interact each other within a four-dimension space. Hence, to fully understand or to control biological processes we need to unravel the principles of gene-gene temporal interactions. Until date, several approaches based on Artificial Intelligence methods have tried to address this issue. Nevertheless, the research community has claimed for more interpretable and biologically meaningful models. Particularly, scientists claim for methods able to infer gene-gene temporal interactions that could be later validated with real-life experiments at the lab. The recent revolution known as "eXplainable Artificial Intelligence" offers a solution for this issue, where a range of highly interpretable and explicable models has become available. Many of these methods could be applied to temporal gene expression data in order to decipher mentioned temporal gene-gene relationships in humans. Here, we propose and validate a new pipeline analysis including an eXplainable artificial intelligence method for the identification of comprehensible gene-gene temporal relationships from human intervention studies. Our method has been validated in six datasets from obesity research (consisting of low calorie diets interventions), where it was able to extract meaningful gene-gene temporal interactions with relevance to the etiology of the disease. The application of our pipeline to other type of human temporal gene profiles would greatly expand our knowledge for complex biological processes, with a special interest for drug clinical trials, in which identified gene-gene regulatory interactions could reveal new therapeutic targets.

Introduction

Biological processes in humans are not single-gene based mechanisms, but complex systems controlled by regulatory interactions between thousands of genes. Within these gene regulatory networks, time-delay is a common phenomenon and genes interact each other within a four-dimension space [1]. That is to say, it may take a time since the product of a gene is generated, until it finally causes an effect on its target molecule. Some of the main sources of time-delay in gene regulation may include; 1) the action of gene expression co-activators or co-repressors, 2) the influence of external environmental factors, and 3) the natural self-degradation of messenger RNA and proteins in cells. Time-delayed gene regulation is especially present in long-term interventions, in which changes in gene expression reflect the response of genes to external factors and may cause subsequent changes on the expression of other genes.

DNA microarray technology has provided a powerful vehicle for exploring biological processes on the genomic scale. In spite of it, in most of the genome scans performed to date, the effects of each gene on the trait of interest have been interrogated one at a time; thus presenting a limited throughput to get the overall picture of gene networks and their temporal relations. Unsupervised methods implemented in conventional microarray software (such as clustering solutions) have also failed in the discovery of network phenomena, since genes can participate in more than one network all at once. As a result, there is not a clear picture of the dynamic trends in gene-gene interactions and much of the heritability of complex human traits remains unexplained, a phenomenon termed as the "missing heritability" problem [2].

The creation of public functional genomics data repositories has enabled a huge amount of genome-wide expression profiles become available to the scientific community [3]. Among available datasets, the recent increase of massive temporal microarray experiments (such as clinical and dietary long-term interventions) open up new opportunities to uncover time-delayed gene-gene relationships. Several machine learning (ML) approaches have been proven very effective for extracting associations between different genes, highlighting Boolean models, Bayesian networks and Neural networks [4–7]. Due to their great predictive ability, these ML methods have been widely used in this and other field applications. Nevertheless, despite yielding impressive results, most of these techniques output unintelligible and complex gene networks, and can not explain how they arrive at specific decisions (which is known as the "black box" problem) [8–10]. For scientists to trust they must first understand what machines do, since in many cases it is not so much what an algorithm predicts but the relationships it establishes and how it predicts it. This is especially important in gene networking, where one of the main concerns of biologists is how to translate inferred networks into particular hypotheses that can be tested with real-life experiments. On this sense, there is a recent increased need to provide ML models with more interpretability and explicability, giving rise to what it is known as eXplainable Artificial Intelligence (XAI) [8,10,11]. As one of the most naturally interpretable and popular knowledge discovery techniques, association rule mining has become a highly relevant technique within the XAI revolution, being able to generate practical knowledge understandable from the point of view of human experts [12,13]. Practical knowledge in association rule mining is extracted in the form of association rules and it refers to concrete relationships between the elements of a database. Association rules constitute representations with the form of $X \rightarrow Y$, which means that when X occurs it is likely that Y also occurs. Due to its natural explicability, association rule mining methods have emerged as an excellent choice for the data mining of complex biological datasets in humans [14]. On this sense, they have been successfully applied to gene expression data (GED) in order to represent how the expression of one (several) gene(s) may be linked or associated with the expression of a different set of genes (gene-gene interactions discovery) [14,15].

Although interesting insights have been derived from the application of association rule mining to GED, previously mentioned time dependencies between associated genes cannot be

modelled making use of conventional association rule mining methods. To face this problem, sequential rule mining (SRM) algorithms could be used instead. SRM algorithms are intended to discover interesting sequential relationships between the elements of a sequence database, in which the data are represented sequentially (e.g. time ordered or spatially localized). The concept of a sequential rule is similar to that of association rule but, in this case, X must appear before Y according to the sequential ordering criterion of the database [13,16]. By way of example, sequential rules that can be extracted from the application of SRM to temporal microarray designs has the next form; [gene A↑, gene B↓] → (time delay) [gene C↑, gene D↑, gene E↑], which represents that the upregulation of gene A and the significant repression of gene B are followed by (or cause) a significant upregulation of genes C, D and E after a given time delay.

Until date, there has been only one adaptation of SRM methods to temporal microarray data[15], consisting of an *Ad-hoc* application for *in vitro* time-series GED in *Saccharomyces cerevisiae*. Referred to as temporal ARM (TARM), the employed method is based on the conventional association rule mining algorithm "*Apriori*" and has been exclusively designed to work with GED derived from yeast (composed of 799 genes evaluated during five transcriptional time points in the same culture). This method builds a sequence database conformed by a single sequence of events (i.e. same culture) by converting each continuous gene expression value into a discrete item by time interval (upregulation, downregulation, or none). Then, TARM is able to identify concrete and understandable temporal causal relations among genes with relevance in the yeast cell cycle.

Besides previous approach, it is also remarkable a more recent work published by Liu *et al.* (2013) [17], in which a sequential pattern mining algorithm is proposed for the identification of temporal co-expression networks from *in vitro* human data. Sequential pattern mining algorithms belong to the ML branch of frequent pattern mining and could be considered as a simpler version of SRM (sequential rule mining). The main drawback of sequential pattern mining algorithms, in comparison to SRM methods, is that they find sequential patterns that appear frequently in a sequence database but without generating sequential rules from them. Thus, they are unable of establishing causal relationships between items, and their resulting sequential patterns could be misleading in certain occasions, especially in the presence of very frequent elements in a sequence database [16]. For these reasons, although the Liu *et al.* (2013) [17] approach interestingly moves forward from yeast models into *in vitro* human data, and is based on a highly interpretable ML method, it still presents some drawbacks for its fully application in the modeling of temporal gene networks.

As Liu *et al.* (2013) [17], many other researchers have also explored co-expression for the reconstruction of gene networks in humans (although not from a pattern mining approach). These methods have become thereby the gold-standard when studying a microarray experiment from a systems biology perspective [4, 5, 18, 19]. In the case of time course genomics experiments, most of conducted co-expression approaches have been based in clustering analyses or unsupervised

learning without class labels [20]. These temporal co-expression approaches are generally based on similarity or correlation or distance measures for the identification of groups of genes with 'similar' temporal patterns of expression, and reveal hidden patterns in the original data by transforming raw temporal data into logically structured, clustered, and interconnected graphs [18,19]. Co-expression graphs can be visualized with nodes representing genes, and with edges indicating interactions, and have helped to understand how genes interact each other within the context of an integrated and global biological network. Nevertheless, despite the widespread use of these approaches [4], there are some drawbacks and limitations remaining for their application in the inference of causal gene-gene relationships:

- Co-expression networks are good to study the general interactome of an organism (free-scale network topology), but their results make hard to infer particular details such as the causal direction or the importance of each individual interaction within the whole network [20]. Thus, they may hinder the translation of inferred networks into particular hypotheses that can be tested in wet-lab experiments. On the contrary, SRM results (in the form of individual rules for each interaction) allow a concrete quality evaluation for each relationship and an easy biological interpretability of findings, which is crucial to demonstrate that a gene network is functionally meaningful, and not just biostatistical fluke [4].

- Co-expression is a very strict assumption for the extraction of gene-gene interactions from time course data [21]. That is to say, co-expression networks with temporal GED generally do not include the time order information in graphs, and they are not capable of detecting positive, negative and time-lagged gene correlations at the same time [20]. However, in living systems, gene regulations can be positive or negative possibly with time lags, and may also not span all conditions or time points. For example, some of the target genes could have a negative feedback loop and could block their own expression, which could explain fast transient dynamic changes, while other target genes could have a positive feedback loop and therefore maintain gene expression longer. Additional regulation could happen after a longer time or very fast without protein translation, i.e. by the action of functional large non-coding RNAs. All these kind of phenomena, which are missed by most of co-expression approaches, could be captured by an appropriate SRM approach.

Given all these concerns and the interesting properties of SRM methods (e.g. existence of statistical quality measures by rule, possibility of functional validation by interaction, inclusion of causality or sequential order information, discovery of complex temporal regulatory phenomena...), SRM approaches are presented as an alternative of great interest and interpretability against temporal co-expression clustering methods when inferring gene-gene temporal relations.

Unfortunately, as far as we concern, the work of Nam *et al.* (2009) [15] in *Sacharomyces cerevisiae* is the only SRM approach developed to the moment and it constitutes no more than an *Ad-hoc* application for *in vitro* experiments (whose extension to *in vivo* GED would elicits challenges that

could not be solved with simple algorithm modifications). The application of these methods to human temporal gene profiles otherwise would greatly expand our knowledge for complex biological processes, with a special interest for long-term interventions (such as clinical trials), in which identified gene-gene regulatory interactions could reveal promising and new therapeutic targets [22,23]. Among the main issues that may have prevented the adaptation of SRM for temporal gene networking in *in vivo* human data we can highlight:

1) The high dimensionality of human gene expression microarrays. With more than 30,000 probes under study in conventional human microarray platforms, the volume of the search space is so big that any available data will become sparse (especially in the case of clinical trials where sample sizes are barely composed of a few tens). The low number of temporal records that are normally assessed in this kind of interventions (rarely more than four) further worsens this sparsity. Within this context, most of ML approaches will thus present a detrimental performance and reliable results can be obtained only if the study sample size is exponentially increased or if effective 'feature-selection' methods are employed prior to analysis for dimensionality reduction.

2) The lack of gene expression discretization methods for *in vivo* datasets. Most of available SRM algorithms require categorical data as input to perform inference. Thus, the selection of an appropriate discretization strategy is a key step for a successful performance. A wide range of in vitro discretization methods has been recently revised and gathered under open-source software [24]. Of note, performance of these methods have shown strong dependence on the particularities of each biological problem. Regarding human GED, there are few issues to take into account before performing discretization, including not only the fact of having multiple sequences but also the great variability between (and within) subjects, tissues and conditions. Considering all this, the extension of existing *in vitro* discretization strategies to humans is not a trivial issue and new approaches should be proposed.

3) The problem of mining sequential rules common to multiple sequences. Contrary to what happen *in vitro*, in human or animal experiments the "subject variability" is the main issue to address, so that databases include multiple sequences instead of a single one. That is to say, we pass from single-sequence experiments where only one microarray is conducted in the culture by time record, to an experimental framework with $N > 1$ gene expression profiles evaluated at each data point (being N the number of subjects under study). Since classical SRM methods have been originally intended for mining sequential rules in a single sequence of events, gene networking in human temporal microarrays will require the adaptation of more modern and specific SRM algorithms.

4) The need for functional validation of results. As it is well known, when facing high-dimensionality data with low sample sizes, data mining methods may yield results which

seem to be significant; but which do not actually represent real behaviors of the dataset. Regarding the case of SRM and its application to GED, this problem is presented in the form of too many output rules even after pruning quality application, which can reach the order of thousands. In such cases, the extracted SRM-based gene networks will represent a chaotic set of "potential" interactions whose biological interpretation will become a serious challenge for biologists.

In this paper, we present a three-stage and rule-based XAI strategy (including pre-processing, knowledge-extraction and functional validation) for finding biologically relevant sequential rules from longitudinal human GED. Particularly, our strategy involves the proposal of an improved version of the well-known SRM algorithm CMRules [25] in order to mine time-delayed gene relationships from *in vivo* human temporal microarray data. Furthermore, we not only adapt the CMRules algorithm to the specific GED problem but also propose a full-detailed data pre- and post-processing pipeline that solve previously mentioned human data limitations and increase model explicability. As a result, our methodology is able to generate temporal gene expression networks in long-term human interventions. The proposed pre- and post-processing pipeline could be briefly summarized in the following key aspects:

- First, the initial number of probes is reduced to those differentially expressed by time interval and experimental condition. This way, we simplify the experimental problem and reduce the search space, further favoring a better performance of the algorithm.
- Secondly, we propose a new discretization approach for the conversion of continuous gene expression values into discrete categories representing temporal changes in gene expression. Based on signal log ratios by gene and time interval, this discretization strategy maps data from a vast spectrum of numeric gene expression values into three discrete categories. Therefore, it can be viewed as a secondary data dimensionality reduction technique in favor of model explicability.
- Third, we apply the SRM algorithm CMRules to the discretized dataset and generate sequential rules from it with the form of [gene A↑, gene B↓] → (time delay) [gene C↑, gene D↑, gene E↑]. Each rule is assessed in terms of quality and robustness by means of five interestingness or quality metrics as described in the *method section*.
- After the knowledge-extraction stage, we propose the integration of output gene rules along with external biological resources such as functional annotation and gene regulation databases. Three well-known and reliable biological resources (GO, KEGG and TRRUST databases) are consulted in order to compute five new biological quality measures by-rule. Through this strategy, each interaction result is biologically pruned and placed within the context of those molecular systems that commonly underlie gene-gene interactions in humans (e.g. Transcription factor (TF)-target gene regulatory relationships [26]).

- Finally, we propose data visualization for the joint representation of gene patterns and all accessed biological information. By means of hierarchical edge bundling visualization methods, we concentrate a lot of information in a single shot and facilitate the identification of a finite set of genes composing a good quality network.

The whole pipeline of our proposal is illustrated in the Supplementary Fig 1 and has been fully detailed in the *method section*. The strategy presented in this work (with special relevance of the adopted ML algorithm, and the functional validation and graphical representation of results) constitute a value proposal whose main objective is to increase model *eXplainability* and to help biologists understanding extracted gene interactions. In this way, we pretend to move away from the black box concept that is usually adopted in most of the current artificial intelligence (AI) omics applications [27], and to provide researchers with a great power to discern between random and causal gene relationships. Our whole pipeline and the SRM adaptation have been gathered as open-source software in the public hosting GitHub (https://github.com/AugustoAnguita/GeneSeqRules) and could be easily extended to other temporal GED applications. At the end, we hope this proposal becomes a helpful strategy for the identification of comprehensible genetic interactions in long-term human interventions, with special interest for the discovery of novel therapeutic targets in clinical datasets.

Since our method is the first application of a rule-based SRM strategy for the extraction of gene interactions in longitudinal human in vivo experiments, there is not currently a SRM benchmark tool that we could use to compare the performance of our pipeline with. At least, not without implementing algorithm modifications in such comparison methods. In spite of it, from the biomedical perspective the real challenge issue when inferring gene networks is their reliability for avoiding false discovery as well as their reproducibility across different patient cohorts. For this reason, we decided to validate our approach in two alternative ways: 1) First, we applied our methodology to an example dataset and give the derived results to a group of field-experts in order them to evaluate the usability of inferred networks for the generation of particular gene-gene interaction hypotheses; and 2) We repeated the application of the full pipeline to three additional datasets, following the same experimental design than the discovery sample, and mined results looking for replication patterns across studies. As a result, we validated some of the primarily extracted gene patterns and thus proved the goodness of our strategy for the mining of biologically relevant gene-gene temporal relations (see results *section*). Full details for the evaluation guidelines committed by field-experts during the interpretation of model results have been addressed in the *method section*.

Main topics covered in this paper include: 1) Preliminary concepts in ARM. 2) Methodological description of the proposed pipeline (including pre-processing, Knowledge-extraction stage and Functional validation of results), 3) Description of the research problem and employed datasets, 4) Results description, where we evaluate the performance of our pipeline in terms of the insights

extracted from a discovery sample and their validation in independent cohorts and 5) Discussion section, where we deepen the goodness of our proposal and list some drawbacks and challenges to be faced in future applications.

Methods

Preliminary: Association Rules and Sequential Rules

The concept of association rules was first proposed by Agrawal *et al* (1993) [12] as a market basket analysis tool in order to discover what items are bought together during a supermarket purchase. Many algorithms for mining association rules and other that extend the concept of association rule mining have been proposed so far to extract useful knowledge from different types of transactional datasets (T). As previously mentioned, association rules have the form of Left Hand Side (LHS) → Right Hand Side (RHS), where LHS and RHS are sets of items, and it represents that the RHS set being likely to occur whenever the LHS set occurs. Interestingly, association rules move forward from the simpler concept of frequent patterns and allows the opportunity to uncover true causal relationships between items [13]. In the field of gene networking, an example of transactional dataset would be a subset of individuals belonging to an experimental condition; where each individual from the subset would be considered as a transaction of the database, and each gene expression event for that particular individual (e.g. gene A↑, gene B↓, gene C↑, gene D↑, gene E↑...) would be considered as an item composing that particular transaction. Support and confidence are the most common measures used to assess association rules' quality, both of them based on the support of an itemset. In the previously introduced example, an itemset would refer to any combination of items from the database (e.g. gene B↓ & gene C↑), being also possible the fact of an itemset composed by only one item. In association rule mining, the support for an itemset *I* is defined as:

$$\mathrm{SUP}(I) = \frac{|\{\mathrm{e} \in \mathrm{T} | \mathrm{I} \in \mathrm{e}\}|}{|T|} \quad (1)$$

where the numerator is the number of examples (*t*) in the dataset *T* covered by the itemset *I*, and | *T* | is the total number of examples in the dataset. Thus, the support and confidence for a rule LHS → RHS are defined as

$$support(LHS \rightarrow RHS) = SUP(LHS\ U\ RHS) \quad (2)$$

$$confidence(LHS \rightarrow RHS) = \frac{SUP(LHS\ U\ RHS)}{SUP(LHS)} \quad (3)$$

In other words, support could be viewed as the percentage of transactions where the rule holds, and confidence as the conditional probability of RHS with respect to LHS (i.e. the relative cardinality of RHS with respect to LHS). The classic techniques for mining association rules attempt to discover rules whose support and confidence are greater than certain user-defined thresholds called minimum support (minSup) and minimum confidence (minConf). However, several authors have pointed out some drawbacks of this framework that lead to find many misleading rules [28]: 1) First, the confidence measure is not able to identify statistical independence or negative dependence between LHS and RHS, mainly due to the fact that the RHS support is not taken into account during the computing process, and 2) Second, itemsets with very high support will be a source of misleading rules because they exist in most of the examples (transactions) and therefore any itemset could seem to be a good predictor of the presence of the high-support itemset. The following example is from [29] and it illustrated very well previous misleading behaviors: in the CENSUS database of 1990, the rule "past active duty in military ⇨ no service in Vietnam" has a very high confidence of 0.9. This rule suggests that knowing that a person served in military we should believe that he/she did not serve in Vietnam. However, the itemset "no service in Vietnam" has a support over 95%, so in fact the probability that a person did not serve in Vietnam decreases (from 95% to 90%) when we know he/she served in military, and hence the association is negative. Clearly, this rule is misleading.

To face these problems, researchers have proposed additional quality metrics by rule and have introduced the concept of "very strong association rules" [30], which are of great aid for the selection and ranking of rules according to their potential causality and coherence. Next, we briefly describe some of the additional metrics that have been used in this paper as well as introduce the framework of "very strong association rules".

The conviction [29] measure analyzes the dependence between LHS and ¬RHS, where ¬RHS means the absence of RHS. Its domain is [0,∞), where values less than one represent negative dependence, a value of one represents independence, and values higher than one represent positive dependence. Conviction for a rule LHS →RHS is defined as

$$conviction(LHS \rightarrow RHS) = \frac{SUP(LHS)SUP(\neg RHS)}{SUP(LHS\neg RHS)} \quad (4)$$

The lift [31] measure represents the ratio between the confidence of the rule and the expected confidence of the rule. As with conviction, its domain is [0,∞), where values less than one imply negative dependence, one implies independence, and values higher than one imply positive dependence. Lift for a rule LHS → RHS is defined as

$$lift(LHS \rightarrow RHS) = \frac{SUP(LHS\ U\ RHS)}{SUP(LHS) * SUP(RHS)} \quad (5)$$

The certainty factor (CF) [32] is interpreted as a measure of variation of the probability that RHS is in a transaction when we consider only those transactions where the LHS is present. Its domain is [-1,1], where values less than zero represent negative dependence, zero represents independence, and values higher than zero represent positive dependence. CF for a rule LHS → RHS is defined in three ways depending on whether the confidence is less than, greater or equal to SUP(RHS):

if confidence(LHS → RHS) > SUP(RHS)

$$\frac{\text{confidence}(\text{LHS} \rightarrow \text{RHS}) - \text{SUP}(\text{RHS})}{1 - \text{SUP}(\text{RHS})} \quad (6)$$

if confidence(LHS → RHS) < SUP(RHS)

$$\frac{\text{confidence}(\text{LHS} \rightarrow \text{RHS}) - \text{SUP}(\text{RHS})}{\text{SUP}(\text{RHS})} \quad (7)$$

Otherwise is 0.

Some of presented metrics, such as the CF, have been further employed to create a framework intended to make easier the discovery of those patterns known as "very strong association rules" [30]. Particularly, a rule will be considered as very strong (and thus not a misleading relationship) if it fulfills the following conditions (Support > minSup, Not(Support) > (1- minSup) and CF > 0). The concept of very strong rule is very intuitive, since it is based on the logical equivalence between a rule and its counter-reciprocal, and it captures the idea that, since both rules are equivalent, finding evidence of both in data enforces our belief that the rule is important.

Although association rule mining methods and all presented metrics have shown a good ability to mine hidden relationships in many different domains (such as genetics [14], biomedicine [33], and so on), these methods are aimed at analyzing data where the sequential ordering of events is not taken into account. Consequently, when such techniques are applied on data following a specific time or sequential ordering criterion, this information will be ignored. This situation may result in the failure of association rule mining methods to extract interesting knowledge from the data, or in the extraction of knowledge that may not be useful for the experts. Otherwise, in many domains, the ordering of events or elements is important and, particularly in genetics, the temporal information is especially critical for the understanding of the regulatory mechanisms of biological processes. SRM algorithms [16] have proven to be an interesting method for discovering sequential relationships between the elements of a sequence database (in which the data are represented sequentially or time ordered). Whenever the time dimension appears, SRM approaches will present a greater predictive and descriptive power than conventional association rule mining algorithms and will provide an additional degree of interestingness. Furthermore, SRM resolve an important limitation of the previously introduced simpler technique sequential pattern mining, since a sequence pattern may appear frequently in a sequence database but may have a very low

confidence (which makes it therefore not useful for the identification of causal relationships). The concept of a sequential rule that can be extracted from SRM is similar to that of association rule except that it is required that LHS must appear before RHS. Previously mentioned quality measures (support and confidence) are also employed in SRM in order to evaluate the interestingness of each mined rule. In SRM, rules are extracted from a sequence database. Recovering the same previous example of gene networking, an example of sequence database could be a subset of individuals belonging to a long-term intervention (with more than two time point records available) in which each individual from the subset would be considered as a sequence of the database, and each gene expression change event for a particular individual and a particular time interval (e.g. gene A↑ from T1 to T2, gene B↓ from T1 to T2, gene C↑ from T2 to T3…) would be considered as items composing that particular sequence. In SRM, the introduced basic quality metrics by rule are defined as sequential support (seqSup) and sequential confidence (seqConf)

$$\text{seqSup}(\text{LHS} \rightarrow \text{RHS}) = \frac{sup(LHS \rightarrow RHS)}{|SD|} \tag{8}$$

$$\text{seqConf}(\text{LHS} \rightarrow \text{RHS}) = \frac{sup(LHS \rightarrow RHS)}{sup(LHS)} \tag{9}$$

Here, the |SD| refers to the total number of sequences in the sequence database. The element *sup(LHS→RHS)* refers to the number of sequences from the sequence database in which all the items of LHS appear before all the items of RHS (note that items within LHS (or RHS) do not need to be in the same sequence nor temporal order within each sequence). The notation *sup(LHS)* refers to the number of sequences that contains LHS. In addition to seqConf and seqSup, the rest of previously introduced association rule mining quality metrics (such as conviction, lift and CF) have also their extension in SRM based in the definition of sequential support, keeping their original meaning and domains. All of them have been incorporated by our methodology for the mining of temporal sequential patterns in GED and allow practitioners a quick identification of the robustness of each extracted pattern from a frequentist perspective.

Pre-processing Stage: Feature Selection and Data Discretization

As input files, our methodology receives raw fluorescence intensity signals (one *.cel* file per subject and time point), and perform a transformation into the form of an *N x M* matrix of gene expression values, where the *N* rows correspond to subjects under study and the *M* columns correspond to evaluated gene probes. All available time records are then merged into a single primary database so that each subject under study will present as many consecutive entries in the database (Long-format) as temporal points exist in the experiment (corresponding to the subject's gene expression profile at each time point). A primary quality control process is then conducted following straightforward pre-processing analyses in transcriptomics (generating chip pseudo-

images, histograms of log_2(intensities) and MA-plots). Finally, all microarray fluorescence signals are normalized together by means of the robust multichip average (RMA) method and probes are annotated according to the latest released version of the *"org.Hs.eg.db"* database [34].

When dealing with Affymetrix microarray technologies, the huge number of probes available in platforms (often around 33,000) may induce an exponential growth of the search space, so that the knowledge-extraction process (independently of the ML method used) will become a difficult and complex task exceeding the processing capability of conventional systems. In order to solve this problem, prior to knowledge-extraction, our methodology includes a feature-selection step in which the number of probes is reduced according to the differentially expressed (DE) genes by time interval and experimental condition. Given a longitudinal GED experiment, our method identifies DE probes by assessing the changes in gene expression during each period of intervention. A probe will be selected for downstream analyses when its Bonferroni-adjusted P-value is < 0.05 and the associated Log_2(FoldChange) (which is also known as the signal log ratio) is $>= 1$ or $<= -1$ in a paired t-test with Bayesian correction.

After feature-selection, as the second main-step of the data pre-processing process, gene expression data discretization is also incorporated in our pipeline. Data discretization is a technique commonly employed in computer science that has been successfully applied to GED applications [24]. Here, the main motivation behind the application of GED discretization is allowing the use of ML algorithms, such as SRM, that requires discrete data as an input for the inference of biological knowledge. Nevertheless, there are many other advantages that arise from data discretization in genetics; 1) Discrete states favor the inference of qualitative models [35], which are of special importance in terms of model explicability. The explicability improvement in qualitative models is achieved due to the fact that for scientists, discrete values are easier to understand, use and explain than continuous values [35,36]. On this matter, GED discretization can be viewed as a secondary data pre-processing technique that move ML approaches closer to the XAI trend. 2) Another advantage emerging from GED discretization is the homogenization of different datasets in terms of interpretability. If the same semantics is used for the discretization of heterogeneous datasets, their results will be more easily comparable and the application of the same ML method to all of them will be a more straightforward and affordable task [37]. 3) Finally, the learning process from discrete data is more efficient and effective (requiring a reduced amount of data and yielding more compact and shorter results) [36]. Therefore, this step not only allows the inference of large-size models with a higher speed of analysis but also facilitate a significant portion of the biological and technical noise presented in the raw data to be absorbed, which may indirectly lead to a better model accuracy.

A wide range of *in vitro* GED discretization methods have been recently revised [24]. In our method, we adopted an unsupervised discretization approach based on expression variations between time points and adapted it to the problem of *in vivo* human data. For that purpose,

continuous gene expression values from the filtered gene expression primary matrix *(N x M)* are transformed into three discrete categories (items) representing changes in gene expression. This approach gives a discretized matrix *A* of *M* probes and *N - 1* conditions, in which each probe by time interval may have one of three discrete states: 2, 1 and 0, meaning 'increase', 'decrease' and 'nochange' respectively. For the assignation of these discrete states by probe and time interval, each experimental condition of a dataset is considered separately as the 'data scope' framework and the next criteria are considered:

- For probes ik showing a positive signal log ratio in the previous DE analyses (feature-selection step):
  If $log_2(FoldChange)_{ikj} > log_2(FoldChange)_{ikJ}$ Then assign the label 'Upregulation'.
  Otherwise type 'No change'
- For probes ik showing a negative signal log ratio in previous DE analyses (feature-selection step):
  If $log_2(FoldChange)_{ikj} < log_2(FoldChange)_{ikJ}$ Then assign the label 'Downregulation'.
  Otherwise type 'No change'

Where the term $log_2(FoldChange)_{ikj}$ refers to the signal log ratio (i.e. change in the gene expression) for the *i* probe, in the *k* time interval and the *j* subject from the group or dataset under study, and the term $log_2(FoldChange)_{ikJ}$ otherwise refers to the mean signal log ratio for that particular probe *i*, in the *k* time interval in all subjects from the group or dataset under study J.

The discretization approach adopted in this work can be viewed as an extension of two previously described in vitro temporal methods that have been successfully applied for the reconstruction of gene regulatory networks [24,35,38]. The main motivation behind its choice is the fact that, when facing temporal GED, discretization methods based on transitions between time-points have been shown to obtain better results than those using absolute values [39]. Moreover, in our case, the use of a standardized measure of gene expression change (such is the Signal Log Ratio (SLR)) is a more sophisticated approach than the employment of simple differences between values. For example, the use of logs in the analysis eliminates difficulties caused by one very high data point in the set masking information from lower valued data points. On the other hand, base 2 is further used as the log scale, therefore a SLR of 1 represents a two-fold increase in abundance of an mRNA and a value of –1 represents a two-fold reduction in transcript expression. Finally, the use of the mean SLR $(log_2(FoldChange)_{ikJ})$ as the threshold for discretization allowed us to exclusively focus on general change trends in the dataset, which most likely will be consequence of the intervention conducted in the cohort and not particular gene expression changes due to individual's idiosyncrasy. Mapping data from a vast spectrum of numeric gene expression values to a reduced subset of three discrete states, this type of discretization could further be viewed as a secondary data reduction technique in favor of algorithm efficiency and *eXplainability*.

Knowledge-Extraction Stage: Extension of the Sequential Rule Mining Method CMRules

Once the dataset is properly formatted, for performing knowledge-extraction, our method includes an adaptation of the well-known SRM algorithm CMRules [25]. Contrarily to other SRM methods that can only discover sequential rules in a single sequence of events, CMRules is able to mine sequential rules in several sequences, which makes it an excellent choice for dealing with human microarray temporal data. Furthermore, CMRules proposes a more relaxed definition of sequential rule with unordered events within each (LHS/RHS) part of the rule. Thanks to that, it presents a great ability to recognize the fact that similar rules can describe a same phenomenon; thus avoiding undesirable losses of information. Moreover, this characteristic also allows the method to detect some particularities of gene temporal interactions in human, such is the fact that gene regulations may not span all conditions or time points or that they could not occur at the same time-delay interval in all subjects from the intervention. Generally, CMRules starts applying the classic association rule mining method *Apriori* for extracting association rules without taking into account the temporal information. Next, the sequential support of each extracted rule is calculated in order to generate sequential rules from them. A detailed description of CMRules algorithm is presented in Supplementary Fig 2 and has been reported elsewhere [25]. Besides the classical presented metrics (sequential confidence and support), we further computed more sophisticated sequential quality measures by rule as previously introduced (sequential lift, CF and conviction). Altogether included quality metrics allow practitioners a quick evaluation of the robustness of each extracted rule.

In order to deal with a common particularity of long-term interventions in humans (which is the fact of having two different intervention groups usually consisting on a placebo and a treatment group), a particular extension was implemented in the final step of the algorithm. This modification allows the user to choose that CMRules only show those sequential rules which besides fulfilling the condition (seqSup(r) > minimum sequential support (MinSeqSup) & seqConf(r) > minimum sequential confidence (MinSeqConf)), are further exclusive of each experimental group. These would be thereby all rules that assert the condition (seqSup(r) > MinSeqSup & seqConf(r) > MinSeqConf) in one experimental group but not in the other, and viceversa. Thanks to that, this extension of our method is presented as an excellent choice for the study of human clinical trials in which researchers are commonly interested in the discovery of gene-gene signatures activated by a specific treatment but not by a placebo.

Functional validation stage: New Biological Quality Measures by rule and Visualization Tool

Previous works have demonstrated that the integration of external biological resources within the gene networking process is a helpful strategy that improves model *eXplainability* and helps biologists to better understand genes and their complex relationships [14]. In recent years, a great variety of external databases containing biological knowledge has become available.

Among the most robust and reliable ones, it highlights those containing information relative to gene function, protein localization and molecular interaction (e.g. the gene ontology (GO) project [40] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [41]). The GO project is an annotation database that provides a structured and controlled vocabulary to describe gene and gene product attributes in multiple organisms according to three different categories or ontologies (cellular component, molecular function and biological process). The KEGG database, on the other hand, is a bioinformatics resource that integrates current knowledge on molecular interaction networks, cellular pathways and functional information of genes and their products. Both GO and KEGG resources have been successfully employed in previous microarray ARM analyses aiding biological explanation to the extracted associated gene sets [14].

One of the main mechanisms controlling gene expression changes in living organisms is the action of gene-specific TFs. By binding to a particular DNA sequence, TFs regulate the —turn on and off— of target genes in order to make sure that they are expressed in the right cell and at the right time. Understanding the basis of genetic interactions between TFs and their targets is therefore likely to be important for the understanding of time-delayed gene regulatory relations in humans. For this reason, we propose the use of an additional biological database, known as TRRUST [26], which includes information relative to transcriptional regulatory relationships between hundreds of genes. The current version of the TRRUST database (version 2) contains 8,444 and 6.552 TF-target regulatory relationships for 800 human and 828 mouse TFs respectively. Especially for the application of SRM to temporal microarray data, the integration of TFs information is indispensable if we want to understand the complex gene relationships that are illustrated in the form of sequential rules.

In this paper, we propose the incorporation of these three well-known biological data resources (TRRUST, KEGG and GO) in order to evaluate extracted rules within a biological framework and to aid explicability to output models. For that purpose, we compute five new by-rule quality measures named "Biological Process (BP)", "Molecular Function (MF)", "Cellular Compartment (CC)", "Signaling Pathway (SP)" and "Transcription Factor (TF)"; that respectively integrate annotation terms from the three categories of the GO project, metabolic pathway annotations from the KEGG resource and transcriptional relationships from the TRRUST database. The computing process for each measure differs according to their biological meaning and the external resource in which their are based on. The first four measures (BP, MF, CC and SP) constitute rankings computed on the identical matches (between LHS and RHS items) that each rule presents for pathway identifiers and GO-terms annotated at the gene level in the previously mentioned GO and KEGG resources. Therefore, a lower resulting value in these biological metrics for a certain rule will indicate that the rule is a good candidate for representing a potentially causal and biologically relevant gene interaction. For each of these four quality measures, the final ranking-score by rule is computed based on the type and number of reported matches between LHS and RHS. According to the

encountered types of matches, rules will be allocated into five different categories and will receive different number of points (**Figure 1**). Based on these definitions, the final ranking score for a given rule is computed as follows:

$$MEASURE\ RS(r) = CAT(r) + \left(1 - \frac{NP(r)}{(\max(NP(i)), \forall i \in Cat(r)) + 1}\right) \qquad (10)$$

Where *MEASURE* refers either to "BP", "MF", "CC" or "SP", *RS(r)* refers to the ranking score obtained for the rule under study, *CAT(r)* refers to the designated rule category, *NP(r)* refers to the number of points assigned to the rule under study and $\max(NP(i)), \forall_{(i)} \in Cat(r)$ corresponds to the maximum number of points that have been assigned to a rule from the same category. Specific details for the calculation of *NP(r)* and for the designation of a rule category *CAT(r)* are illustrated in **Figure 1**.

On the other hand, the biological quality measure TF constitutes a range of four possible values by rules (0, 1, 2 or 3) which are assigned according to the TF-target gene regulatory information hosted in the TRRUST database. The computing process for the TF measure is slightly different from the previous ones and is performed as follows; if at least one LHS gene from the evaluated rule has been reported as a validated TF in the TRRUST resource, then assign *1* point to the rule. Otherwise, assign *0* points. If the first condition has been fulfilled, then check if any of the RHS items (genes) from the rule has been presented in the TRRUST database as one of the previously-identified TF confirmed targets. In such case, assign *2* points to the rule. Otherwise, assign *0* points. If the previous condition has been further fulfilled, then check for the type of relationship that is reported in the TRRUST database for both TF and target genes (upregulation, downregulation or unknown). In the case of match between the relationship illustrated by the sequential rule (upregulation or downregulation of the target) and the information hosted in the TRRUST database, then *3* points are assigned to the rule.

The choice of that particular procedure, against other available standard GO-similarity measurements [42], was argued in the fact that we needed an evaluation method, based on categories, with the ability to discern the quality of a rule regardless of the items conforming it. If this were not the case, the rules with the highest number of antecedent/consequent elements would always have a higher score by the simple chance of coinciding in GO terms due to the greater number of genes composing them. On the contrary, with our heuristic approach, the score is adjusted by the number of items (from the total elements that constitute the rule) that share a specific GO term. Thanks to that, the rules in which all its elements share the same GO term will be identified as more robust than others, which although perhaps share a greater number of GO terms among all its antecedent elements, do not share any GO between all LHS and RHS items.

The computing process of these new five biological quality measures by rule has been implemented in python environment and can be directly applied to any output generated from the application of our ML method to temporal GED. The software requires a *.pmml* file as input

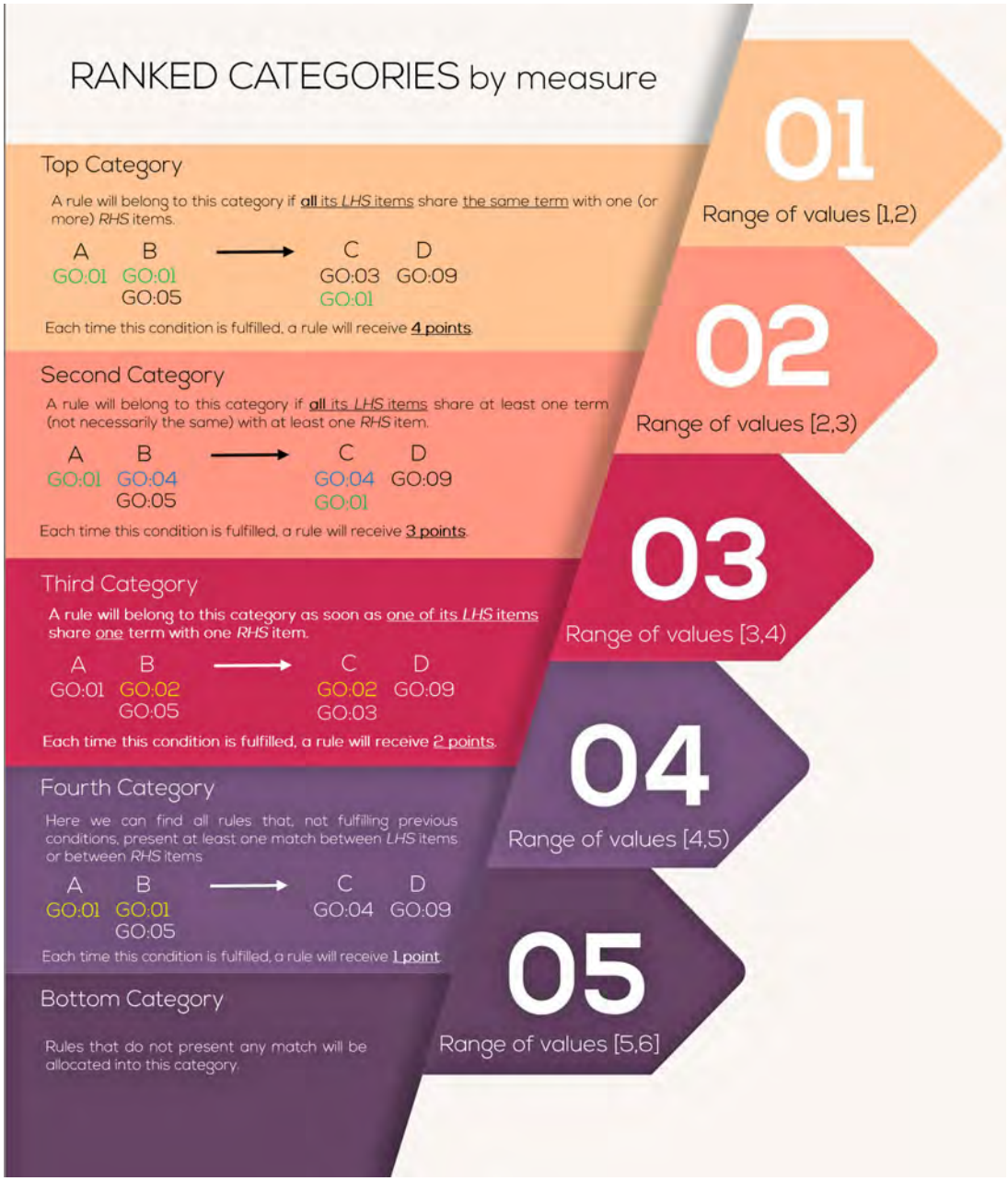
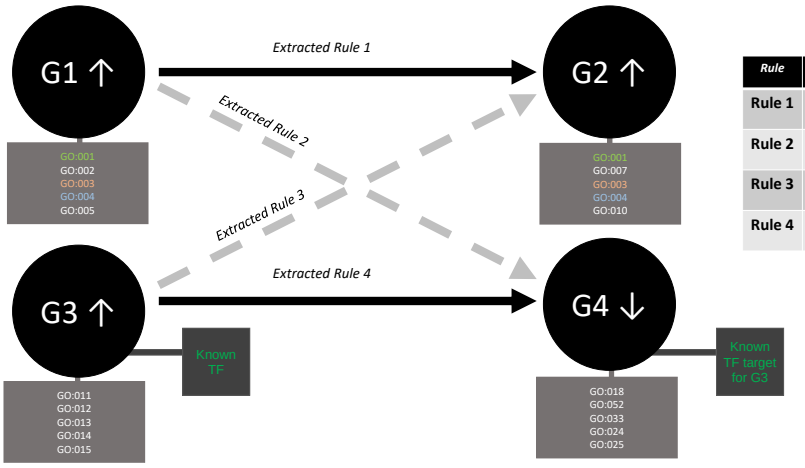


Figure 1. Assignable categories and ranking scores for a rule in the biological measures "BP", "MF", "CC" and "SP". First, a rule will be assigned to a particular category from the bottom category 5 to the top category 1. This assignation will be conducted according to the type of matches encountered for a rule in its annotated terms as is described in the figure. Once a rule is designated to a particular category, a score will be computed for the rule taking into consideration all type of matches encountered for the rule. Each match is weighted with a number of points as illustrated in the figure. The final score for a rule is computed as detailed in the method section.

(containing extracted rules with CMRules) and will output a same format file but with the new biological quality measures computed by rule. The software will also output a secondary file with *.xml* extension in which all the matches between the items of a rule are fully detailed (including information for genes and GO or KEGG terms composing each match of the rule). This software constitutes therefore a useful tool for the evaluation of extracted rules within the context of different human molecular systems and effectively complements the action of previously introduced classical SRM quality metrics. The main utility of the introduced biological quality measures has been illustrated in **Figure 2**, where we have tried to show their capacity to distinguish between true potential causal gene-gene interactions and those representing spurious biostatistical fluke. This figure represents how, although the statistical quality metrics by rule have been shown effective for detecting robust associations between items, not always a gene in the LHS will be the cause of the change in the gene expression of a RHS gene (from a biological point of view). In many cases, rules could be just referring to a range of parallel phenomena that occur simultaneously at the gene level because of co-expression. Contrarily, in other cases, the method could be effectively representing true causal relationships between genes (e.g. rules 1 or 4). Precisely to discern these spurious associations from those true phenomena of interaction is why we propose the functional validation of the results and why some of the biological quality measures presented, such as the TF measure, become especially relevant.



Rule	Support	Confidence	BP	TF
Rule 1	0.9	0.85	1.001	0
Rule 2	0.9	0.85	5	0
Rule 3	0.9	0.85	5	0
Rule 4	0.9	0.85	5	2

Figure 2. Role of biological quality measurements for the functional assessment of each discovered gene-gene relationship. While all extracted rules present acceptable and identical quality metrics (support=90% and confidence=85%), only the rule 1 presents a good BP measure value (remember that the range of values available for the BP measure was from 5 to 1, being the values near to 1 the ones corresponding to a higher number of GO matches between LHS and RHS genes). On the other hand, it is only the rule 4 the one presenting a good value for the TF measure (whose range of values was from 0 to 3, being 3 the maximal score for indicating a true TF-target gene relationship). The figure illustrate how the functional validation of results is critical to discern between spurious associations and true phenomena of interaction.

Data visualization techniques have been widely employed in data science applications given their ability to transform model results into useful knowledge [43]. Beside their high ability to simplify big amounts of extracted information, graphs further allow information to be transferred in a very intuitive way to the user. In the context of high-dimensional data, such is the case of biological data, the problem of knowledge-extraction is much greater due to the large number of rules derived from the application of ML techniques [44]. Under these conditions, visualization techniques are presented as an attractive solution that serve as an interface between scientists and the extracted knowledge [45]. In this paper, we propose a new visualization tool that integrates output gene networks along with all accessed biological information. Based on hierarchical edge bundling methods [46], our tool generates circular plots illustrating the full picture of sequential rules discovered by our algorithm. In order to be extended to other temporal GED applications, this visualization method has been implemented as open-source software in R environment. By modifying circular diameters, edge width, edge type and the color intensity of each connection, the software generates plots comprising both the new biological quality measures and the classical SRM metrics computed by rule . Of note, our visualization tool is not restricted to the information presented in figures and can be easily adapted to each user demands. The greatest virtue of generated plots lays in their ability to concentrate different types of information in single shot, which further allows an easy identification of the top and more coherent rules from the both the technical and biological points of view. On summary, our visualization tool constitutes an additional value proposal in favor of model explicability and interpretability.

Problem and Datasets Description: Long-term Interventions in Obesity

From the clinical or biomedical perspective, the real challenge issue when inferring gene networks is their reliability for avoiding false discovery as well as their reproducibility across different patient cohorts. For this reason, and given the lack of benchmark SRM methods, we decided to validate of our proposal in two alternative ways: 1) First, we applied our methodology to an example dataset (discovery sample) and give the derived results to a group of field-experts in order them to evaluate the usability of inferred networks for the generation of particular gene-gene interaction hypotheses; and 2) Second, we repeated the application of the full pipeline to three additional datasets, following the same experimental design than the discovery sample, and mined results looking for replication patterns across studies.

As an example of long-term human interventions, we chose a discovery dataset consisting of *in vivo* temporal GED derived from human adipose tissue (AT) samples collected in different time points during the course of a dietary intervention. With up to three time records available in the dataset, this study constitutes a perfect example of the *in vivo* temporal microarray experiments in which our method could extract biologically relevant gene-gene temporal relationships. Published by Vink *et al.* (2016) [47], the original clinical trial investigated the effects on weight loss (WL) of two

different dietary interventions in 57 adults with obesity. Subjects were randomly assigned to each experimental group: a low-calorie diet (LCD; 1250 kcal/day) for 12 weeks (slow weight loss) or a very-low-calorie diet (VLCD; 500 kcal/day) for 5 weeks (rapid weight loss). In both experimental conditions, the WL period was followed by a 4-week additional phase of weight stabilization (WS). Abdominal subcutaneous AT biopsies were collected from each subject at each time point (baseline, after WL and after the WS period) and submitted to microarray analysis using the *Human Gene 1.1 ST* Affymetrix platform (one array per subject and time). A more detailed description of the study design can be found in the original publication of the dataset [48].

The full dataset was downloaded from the public repository GEO with identifier GSE77962. Fluorescence data were transformed into the form of an N x M matrix of gene expression values, where the N rows correspond to subjects under study and the M columns correspond to evaluated gene probes. The dataset presented valid fluorescence measures for 33,297 probes (M columns) mapping 19,654 unique genes across the genome. The number of individuals presenting valid gene expression data was 24 on the VLCD group and 22 on the LCD group. Since all available time records were merged into a single primary database, each individual presented three consecutive entries in the database (long format), corresponding to its gene expression profile at each temporal point (baseline, after WL and after the WS period). The final number of rows in the database was N=138.

Data were normalized using RMA and submitted to feature-selection. The original number of probes was thus reduced to those DE genes by time interval and experimental condition. Time intervals corresponded to the WL period (comprising the end of WL vs baseline), the WS period (comprising the end of WS vs the end of WL) and the dietary intervention (DI) period (comprising the end of WS vs baseline). As a result, 431 probes matching 398 unique genes were selected for further analyses.

Remaining GED for the 431 probes were then submitted to data discretization. In each experimental group, a discretized matrix A of 431 probes was obtained, in which each probe by time interval may have one of three discrete states: 2, 1 and 0, meaning 'increase', 'decrease' and 'nochange' respectively. Once discretized, two sequence databases were constructed (one database per diet group) where each sequence corresponded to a subject and each event represented the change in the gene expression of a certain probe during a particular time interval (WL period or WS period). An example of the general structure of each constructed sequence database in the discovery case of study is presented in the Supplementary Fig 3. Details for each sequence database are presented in the caption of the figure.

Knowledge-extraction was conducted by CMRules in each experimental group separately and also by contrast (extracting only those association patterns exclusive for each experimental group). With the aim of avoiding losses of information, only results derived from the mining of each

group separately were employed for functional validation. CMRules results in form of sequential rules were thus submitted to functional validation and the five biological quality measures were computed by rule and added to the already present five frequentist quality metrics. Finally, derived output were visually represented by means of our hierarchical edge bundling visualization tool.

Field-experts received sequential rules results in the form of both tables and figures with traditional quality metrics and biological quality measurements included. The evaluation and interpretation of sequential rules and graphs by field-experts was committed following a few foundations: 1) Rules were ordered according to SRM quality metrics and biological quality measures; 2) Rules with very low values for SRM quality metrics were removed according to the reference values presented in the first *method subsection*. For this prune, the concept of very strong association rules was taken into account; 3) Correlation analyses were conducted between quality metrics and biological quality measures for remaining rules in order to evaluate the ability of CMRules to extract biologically relevant patterns; 4) Identification, on the help of visual representations, of interesting sequential rules and generation of particular gene-gene interaction hypotheses; and 5) Exploration of most interesting hypotheses (either by accessing to the list of GO annotation terms matches or by performing intensive literature search). Field-experts were selected from the research group *"CB12/03/30038"*, belonging to the Spanish research network CIBEROBN (Physiopathology of Obesity and Nutrition), Institute of Health Carlos III (ISCIII), Madrid, Spain.

In order to validate and contrast the insights derived from this discovery dataset, we further accessed temporal GED from WL interventions in three independent cohorts (GSE70529[49], GSE35411 [50] and GSE103766 [51]). Dataset details and main characteristics of the technical validation process are presented in **table 1**. Each validation dataset was processed and analyzed following exactly the same pipeline than the discovery population. Results and gene patterns discovered during the validation process are reported in a specific *result subsection*. Not restricted to the obesity field, our entire pipeline could be applied to any human long-term intervention with up to two experimental conditions (e.g. placebo and treatment).

Results

Discovery approach in the case of study

As we previously explained, with the aim of illustrating the performance of our method on human long-term intervention data, we accessed and downloaded a discovery dataset composed of 57 subjects with obesity participating in a long-term dietary program [47,48]. The dataset consisted on temporal GED collected in three different time records during the course of two dietary interventions (VLCD and LCD). In this dataset, we sought to mine sequential rules with the form of [gene A↑, gene B↓] → (time delay) [gene C↑, gene D↑, gene E↑], that could illustrate the WL-induced gene regulatory responses of AT in obesity. Before the application of our SRM

Table 1. Datasets details and problem description.

GEO Identifier	Design	Intervention Details	Time records available	N° subjects (female/male)	BMI at the beginning of the study	Age (years)	Sample tissue	Array Platform	N° mined Strong ARs	Network representation
GSE77962 (LCD group)* [47,48]	Dietary Intervention	1250 kcal/d during 12 weeks and a weight stable period of 4 weeks	3 (baseline, after weight reduction and after weight maintenance)	22 (12/10)	28–35 kg/m^2	51.8 ± 1.9	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Gene 1.1 ST*	40	Fig 3
GSE77962 (VLCD group)* [47,48]	Dietary Intervention	500 kcal/d during 5 weeks and a weight stable period of 4 weeks	3 (baseline, after weight reduction and after weight maintenance)	24 (13/11)	28–35 kg/m^2	50.7 ± 1.5	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Gene 1.1 ST*	301	Fig 4
GSE70529 [49]	Dietary Intervention	Low-calorie diet of self-prepared foods for consecutive 5,10 and 15% weightloss	4 (baseline, after 5, 10 and 15% weight reduction)	9 (8/1)	37.9 ± 4.3 kg/m^2	44 ± 12	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Gene 1.0 ST*	551	S5 Fig
GSE35411 [50]	Dietary Intervention	1200 kcal/d during 3 months and a weight stable period of 4 weeks	3 (baseline, after weight reduction and after weight maintenance)	9 (6/3)	42.7 ± 1.4 kg/m^2	40 ± 3.73	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Genome U133 Plus 2.0 Array*	83	S6 Fig
GSE103766 (WeightLosers group) [51]	Dietary Intervention + exercise counselling	800–1000 kcal/d during 6 weeks and a less restrictive diet plan + exercise counselling for 12 months	3 (baseline, after 5 months and after 12 months)	6 (3/3)	34.64 (0.7) kg/m^2	21–48	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Genome U133 Plus 2.0 Array*	870	S7 Fig
GSE103766 (WeightRegainers group) [51]	Dietary Intervention + exercise counselling	800–1000 kcal/d during 6 weeks and a less restrictive diet plan + exercise counselling for 12 months	3 (baseline, after 5 months and after 12 months)	13 (9/4)	34.65 (0.85) kg/m^2	20–45	Abdominal Subcutaneous Adipose Tissue	*Affymetrix Human Genome U133 Plus 2.0 Array*	70	S8 Fig

BMI and age data are presented as mean ± SEM, mean (SE) or rather as a range.

* Datasets employed as discovery population.

https://doi.org/10.1371/journal.pcbi.1007792.t001

methodology to the dataset, all described pre-processing stages were conducted. Once the data were properly formatted as described in *methods,* the data mining process was initiated. Specific minimum sequential support and sequential confidence thresholds were set by experimental condition during the knowledge-extraction stage (minSeqSup=0.45 and minSeqConf=0.4 for the VLCD group, and minSeqSup=0.4 minSeqConf=0.4 for the LCD group). Standard quality measures employed in SRM (lift, CF and conviction) were further computed in order to estimate the interestingness of each mined pattern. Beside conventional quality measures, the method also computed five new biological quality measures by rule based on external biological information (including functional and pathway annotations, and TF-target gene regulatory data). Through this strategy, interaction results were biologically pruned and placed within the context of already well-explored molecular systems [26,40,41]. Finally, the method applied a data visualization technique for the joint representation of output gene networks and all accessed biological information.

From the application of this pipeline to the discovery dataset, 50 output sequential rules were identified from the LCD group and 325 from the VLCD group (S1 and S2upplementary tables 1 and 2 respectively). With up to seven times more number of rules mined from the VCLD group than from the LCD group, our results are in accordance with previous findings from Vink et al. (2016) [47], which showed a higher impact in the gene expression of AT elicited by a rapid and aggressive WL in comparison to the effects derived from a light but more prolonged WL. From all extracted rules, only very strong sequential rules were considered for further evaluation (SeqSup > minSeqSup, Not(SeqSup) > (1- minSeqSup) and CF > 0), which constitute a suitable framework to discard misleading rules. During the evaluation process, output sequential rules were biologically assessed by means of the five new biological quality measures and graphically represented in two circular plots (**Figures 3** and **4**). Main descriptive statistics for the extracted rules by experimental condition are presented in **table 2**. In general, extracted rules presented robust values for all computed quality measures, which indicates a good performance of the algorithm during the gene association mining process. With robust quality metrics values, we refer to a minimum support higher than established threshold, a confidence higher to 0.8, a conviction value higher than 1, a lift higher than 1.1, and CF distinct of zero and as near as possible to 1 (see preliminary *method subsection*). According to mean values by group, slightly better metrics values were obtained for the VLCD than for the LCD, which probably was motivated on the higher impact elicited by this intervention in AT. In both groups, top rules (presenting higher values in the traditional quality measures) involved genes participating in molecular processes previously reported as part of the WL-induced AT response (e.g. mitochondrial function, angiogenesis, inflammation and lipid and glucose metabolism) [47,50] (Supplementary Fig 4). Of note, top sequential rules also presented good rates in the new proposed biological quality measures (**Figure 5**). Especially for the case of biological quality measures TF and BP, we showed significant correlations with the traditional quality metrics CF, conviction and confidence. This fact reflects a good performance of the knowledge-extraction process, where the best sequential patterns identified (from the ML perspective) are also the more biologically soundness. On the other hand, the fact of absence of correlation between some other biological quality measures and traditional metrics (**Figure 5**) reinforces the need for the functional validation of results. That is to say, although traditional metrics may indicate that some rules are good from the technical point of view, the biological information is not always what it could be expected.

In order to assess the biological utility of our gene networking strategy, obesity-field experts evaluated all extracted rules making use of the computed biological quality measures and the graphical representations as previously described (**Figures 3** and **4**). Since the most plausible mechanisms underlying gene regulation is the action of TFs on their target genes, the TF metric was the first measure employed by experts for filtering and evaluating output sequential rules. Through the application of a specific TF threshold (>=1), four rules were selected from the LCD group and sixty-two from the VLCD group. Among them, a subset of biologically meaningful

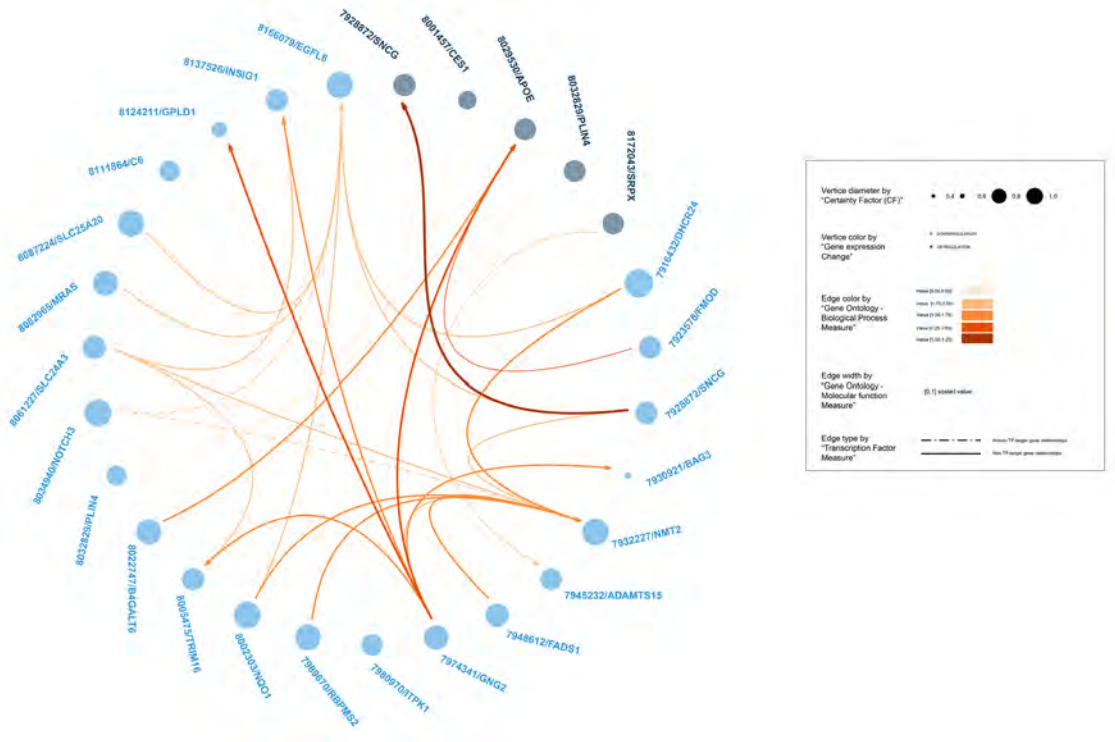


Figure 3. Visual representation of the sequential rules discovered by our method in the GSE77962 dataset (LCD group). Node names refer to (probe/gene).

rules are described in **table 3**. From both intervention groups, several similar rules were identified sharing the same TF gene *(Notch3)* as an LHS item. In these rules, the gene *Notch3* emerged as a TF factor whose downregulation (provoked by the WL intervention) elicited later secondary changes in the expression of other genes during the WS period. Since each group was mined independently during the knowledge-extraction process, the fact of finding similar rules from each diet speaks well of the performance and the validation ability of our method. *Notch3* is mammalian transmembrane protein that bind membrane-bound ligands expressed by adjacent cells in human tissues. By triggering intracellular proteolytic cleavages and through the release of active intracellular domains of Notch (NICD), *Notch3* controls the expression of a wide range of target genes participating in different obesity-related processes such as differentiation, proliferation, angiogenesis and apoptosis. Interestingly, several *Notch3* target sequences have been identified within and near the genomic sequences of a few of its RHS genes (such is the case of *Nmt2* and *Clmn)* [52]. In these cases, sequential rules illustrate how a downregulation of the *Notch3* is followed by a downregulation and upregulation (respectively) of mentioned RHS genes. Despite these interesting results, it is important to clarify that the identification of sequential rules including a TF as LHS does not necessarily imply a causal relationship between the TF and its reported RHS gene. In these cases, functional *in vitro* studies should be performed for validating proposed interactions.

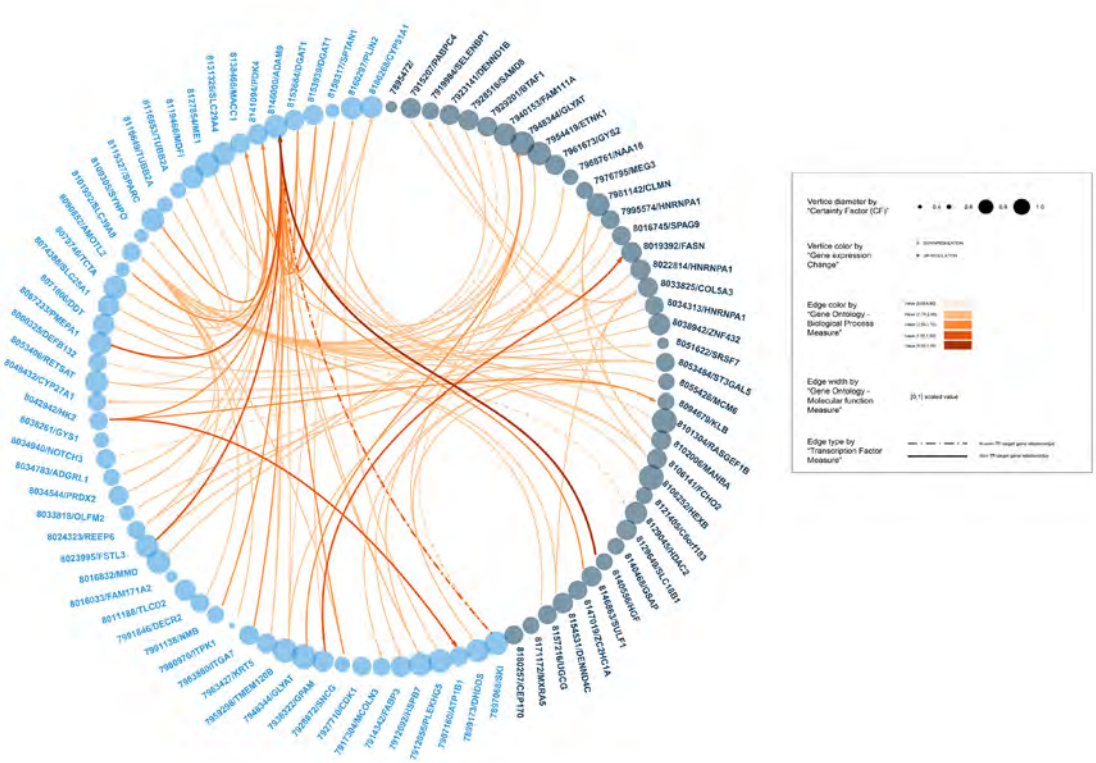


Figure 4. Visual representation of the sequential rules discovered by our method in the GSE77962 dataset (VLCD group). Node names refer to (probe/gene).

Table 2. Descriptive statistics on quality metrics for strong association rules discovered in the whole GSE77962 dataset (LCD and VLCD groups).

					VLCD					
	Support	Confidence	Lift	CF	Conviction	BP	CC	MF	SP	TF
n	301	301	301	301	301	301	301	301	301	301
Minimum	11.00	0.71	1.13	0.22	1.27	1.001	1.001	1.001	1.2	0
Mean	11.08	0.88	1.55	0.71	*Inf*	2.19	1.94	2.17	3.8	0.21
Standard Dev.	0.27	0.09	0.23	0.22	-	1.1	0.48	1.06	2.4	0.41
Median	11.00	0.85	1.56	0.69	3.25	1.91	1.9	1.91	6	0
Maximum	12.00	1.00	2.00	1	*Inf*	6	6	6	6	1
	[ort				LCD					
	[ort	Confidence	Lift	CF	Conviction	BP	CC	MF	SP	TF
n	40	40	40	40	40	40	40	40	40	40
Minimum	9.00	0.69	1.08	0.27	1.36	1.002	1.002	1.002	1.2	0
Mean	9.72	0.81	1.38	0.54	*Inf*	2.39	1.77	2.71	4.56	0.1
Standard Dev.	1.26	0.09	0.17	0.18	-	1.55	0.16	1.8	2.23	0.3
Median	9.00	0.82	1.38	0.55	2.21	1.84	1.84	1.84	6	0
Maximum	15	1	1.69	1	*Inf*	6	1.96	6	6	1

https://doi.org/10.1371/journal.pcbi.1007792.t002

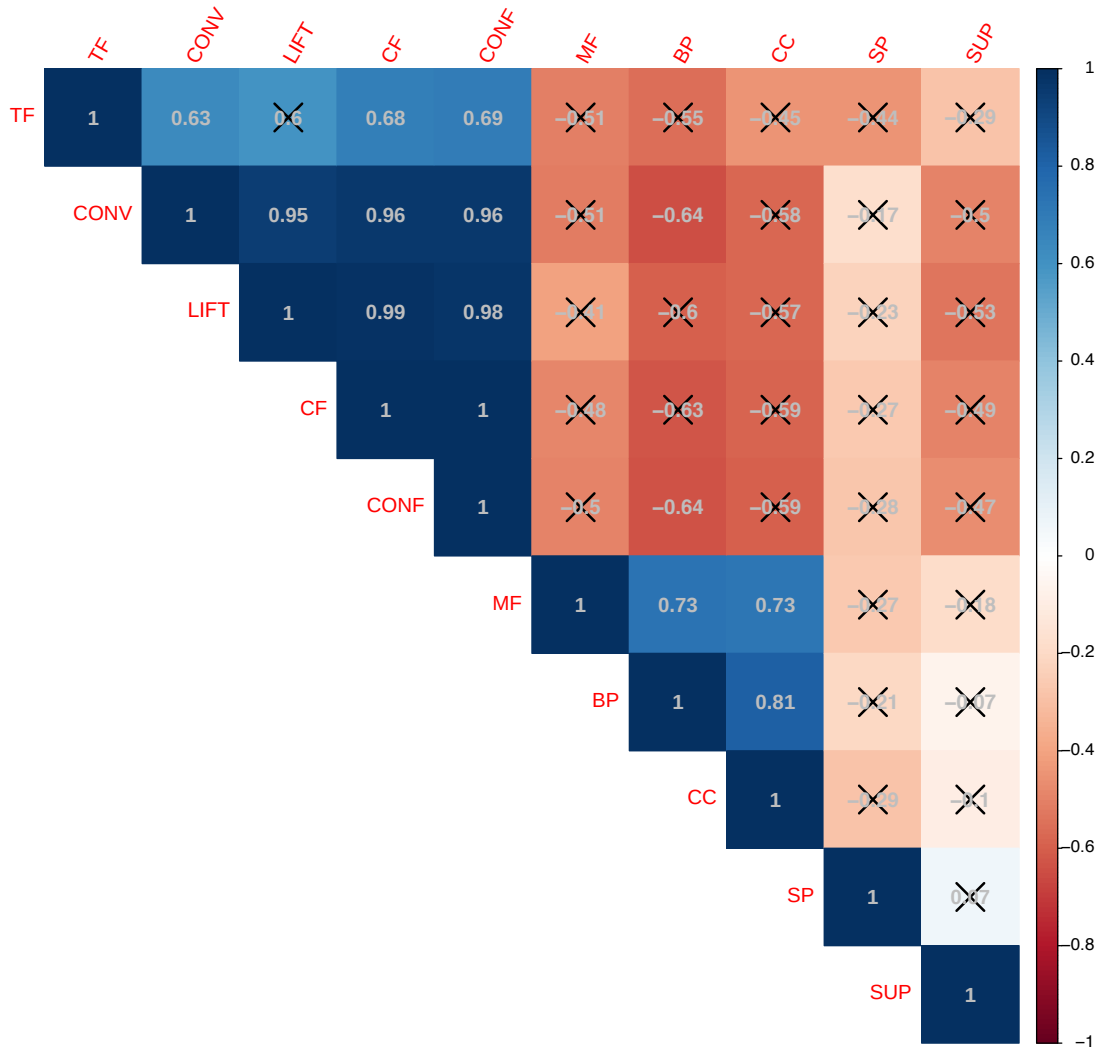


Figure 5. Correlation between traditional quality metrics and biological quality measures by rule in the sequential rules discovered from the whole GSE77962 dataset (LCD and VLCD groups). R2 values quantify the level of correlation for each pair of measures while the level of statistical significance (adjusted by Bonferroni multiple test correction) is evidenced with an X for P-values > 0.05 and nothing for P-values < 0.05.

From the LCD group it is also remarkable a rule with the *Notch3* as LHS and the *Egfl6* gene as RHS (**Figure 3**). Although no target sites for *Notch3* have been identified within the genetic sequence of *Egfl6*, a special functional connection has been evidenced between both genes in the context of obesity and angiogenesis [53,54]. In general, despite only a few previous evidences support an implication of the *Notch3* gene in obesity molecular pathways [55], the findings presented in this paper seem to point this TF as an important element for the proper regulation of AT cellular responses to WL.

Table 3. Subset of biologically meaningful extracted sequential rules in the whole GSE77962 dataset (LCD and VLCD groups).

Intervention Group	LHS	RHS	SUP	CONF	LIFT	CF	CONV	BP	MF	CC	SP	TF
VLCD	{8140556/HGF = 2}	{8146000/ ADAM9 = 1}	11.00	0.77	1.35	0.49	1.94	1.002	1.002	1.002	6.00	0.00
VLCD	{7897068/SKI = 1}	{8146000/ ADAM9 = 1}	12.00	0.92	1.58	0.81	5.42	1.39	1.39	1.39	1.20	1.00
VLCD	{8034940/NOTCH3 = 1}	{7981142/CLMN = 2}	11.00	0.85	1.45	0.63	2.71	1.79	1.79	1.79	6.00	1.00
LCD	{8034940/NOTCH3 = 1}	{8166079/EGFL6 = 1}	12.00	0.86	1.11	0.37	1.59	1.79	1.79	1.79	6.00	1.00
LCD	{7928872/SNCG = 1 & 8034940/NOTCH3 = 1}	{7932227/NMT2 = 1}	9.00	0.90	1.52	0.76	4.09	1.83	1.83	1.83	6.00	1.00
VLCD	{8131326/SLC29A4 = 1}	{8101992/ SLC39A8 = 1}	11.00	0.79	1.26	0.43	1.75	1.88	1.88	1.88	1.20	0.00
VLCD	{8129045/HDAC2 = 2}	{8101992/ SLC39A8 = 1}	12.00	0.92	1.48	0.79	4.87	1.93	1.93	1.93	6.00	1.00
VLCD	{7929201/BTAF1 = 2}	{8101992/ SLC39A8 = 1}	11.00	0.79	1.26	0.43	1.75	1.94	1.94	1.94	1.20	1.00
VLCD	{7929201/BTAF1 = 2 & 7940153/ FAM111A = 2}	{8101992/ SLC39A8 = 1}	11.00	0.92	1.47	0.78	4.50	1.94	1.94	1.94	1.20	1.00
VLCD	{7929201/BTAF1 = 2 & 8106141/FCHO2 = 2}	{8101992/ SLC39A8 = 1}	11.00	0.85	1.35	0.59	2.44	1.95	1.95	1.95	1.20	1.00
LCD	{8087224/SLC25A20 = 1& 8034940/ NOTCH3 = 1}	{8166079/EGFL6 = 1}	9.00	1.00	1.29	1.00	Inf	1.96	1.96	1.96	6.00	1.00
LCD	{7980970/ITPK1 = 1 & 8034940/ NOTCH3 = 1}	{8032829/PLIN4 = 2}	9.00	0.75	1.50	0.50	2.00	6.00	6.00	1.90	6.00	1.00

https://doi.org/10.1371/journal.pcbi.1007792.t003

For investigating the sequential rules discovered in the VLCD group (**Figure 4**), the use of other biological quality measures instead of TF (such as BP and MF) allowed experts to identify several interesting patterns. On the one hand, it highlights a sequential rule involving the loci *Fasn* (Fatty Acid Synthase) and the *Gpam* (Glycerol-3-Phosphate Acyltransferase 1, Mitochondrial); both of them genes coding for enzymes with a central role in the process of lipogenesis. The sequential rule between these genes was easily identified from its color intensity in the circular plot and may suggest a special relevance of the lipogenesis process as part of the responses of obese AT after a strong caloric restriction. Another interesting insight extracted from the graph is the fact that most of the gene expression changes elicited by the dietary intervention in the VLCD group ended in a later and secondary downregulation of the gene expression levels of *Adam9* during the WS period. Among all sequential rules illustrating this behavior, there are a few ones with special biological relevance (**Figure 4**); one highlighted by the BP metric and involving the gene *Hgf*, and another one including a TF-target gene regulatory relationship with the protooncogene protein Ski. *Adam9* is a cell-surface metalloprotease present in almost all cells and tissues of the body that participates in key processes such as cell migration, proliferation and cell-cell interactions. Mostly expressed by white cells, *Adam9* has been reported to get upregulated during many pathological processes including cancers. Regarding obesity, previous transcriptomics analyses have demonstrated how *Adam9* is significantly up-regulated in obese AT and how it plays a major mediating role in a chain of interactions that connect local inflammatory phenomena to the alteration of AT metabolic functions [56,57]. On this sense, the downregulation of Adam9 evidenced in our study might constitute a biologically meaningful finding with relevance for the understanding of the AT metabolic health amelioration achieved with dietary intervention in this case of study. All quality

metrics for the sequential patterns highlighted in this section have been resumed in **table 3**, while the full list of sequential rules identified can be explored in Supplementary tables 1 and 2. Taking all these into consideration, the model and its output results were considered by field-experts as an easily interpretable approach that could be successfully extended to other human long-term intervention datasets for the identification of biologically relevant molecular signatures.

Validation approach in independent cohorts:

In order to validate and contrast the insights derived from the discovery dataset, we accessed additional temporal GED from three WL interventions performed in independent cohorts (GSE70529[49], GSE35411 [50] and GSE103766 [51]). Dataset details and main characteristics of each population are presented in **table 1**. Although there were slight differences in the study design of each cohort, all studies constituted dietary interventions (caloric restriction programs) performed during a long-term intervention period in adult subjects with obesity. Each experimental group (in the case of datasets presenting more than one study condition) was again considered as an individual dataset. During the knowledge-extraction stage, minSeqSup=0.5 and minSeqConf=0.6 thresholds were set and only "very strong rules" were selected for subsequent evaluation. In **table 1**, we report the number of very strong association rules mined from each dataset. Visual representations of output gene patterns by dataset are presented in Supplementary Figs 5, 6, 7 and 8. Graphs illustrated again coherent gene-gene interactions within the context of obesity research (e.g. the gene association patterns governed by the locus Abca1 reported in the dataset GSE70529 (Supplementary Fig 5)) [58]. These figures display information following the same code of colors and format than previously presented **Figures 3** and **4**.

Very strong rules extracted from all datasets were pulled together for the identification of replicated patterns. During the process of contrasting rules between datasets, probe information was removed from each rule and only the locus tag of each item was considered. That is to say, we considered two sequential rules as replicates when they contain the same genes within LHS and RHS (but not necessarily the same probes). As a result, we found gene expression changes in 11 loci acting as trigger mechanisms (LHS items) concurrently in sequential rules extracted from different datasets (these were *C6*=Up-regulation, *Hnrnpa1*=Up-regulation, *Srsf7*=Up-regulation, *Gsap*=Up-regulation, *Sncg*=Downregulation, *Notch3*=Downregulation, *Srpx*=Up-regulation, *Itpka1*=Downregulation, *slc-transporters*=Downregulation, *Tmem-proteins*=Downregulation and *Znf-proteins*=Up-regulation). Interestingly, these validated trigger loci included TFs, splicing factors, mRNA processing molecules and cell surface transporters with a great implication in the control of the global gene expression cell profiles [59–62]. In the same manner, we found the gene expression change of 1 loci represented as a consequence (RHS item) in several ARs extracted from different datasets. This gene expression change corresponded to a downregulation of the locus *C6*, which encodes a component of the complement cascade with implication in the innate immune

system and inflammation pathways. Among all extracted rules, those containing at least one of the described common LHS and RHS loci were selected for further evaluation (Supplementary table 3). The graphical representation of all these rules allowed the identification of very interesting gene patterns and replicated interactions which have been shown in Supplementary Fig 9. On the one hand, we found rules from different datasets illustrating a sequential association between the downregulation of *Slc-transporter* genes and a subsequent downregulation of proteins from the *Adam*-family (Supplementary Fig 9A). In the same manner, we replicated a sequential relationship between the gene expression change of *Tmem* genes and the later modification in the expression of loci from the *Srsf*-family (Supplementary Fig 9A). Particularly, while a relationship of the type (down-regulation -> down-regulation) was found between these genes in the weight losers of the dataset GSE35411, a relationship of the type (up-regulation –> up-regulation) was found between these genes in the GSE103766 dataset (a cohort composed of weightregainers) (Supplementary Fig 9A). The target loci of these rules corresponded to serine/arginine-rich (SR) proteins, a conserved RNA-binding protein family, which consists of 12 members, serine/arginine-rich splicing factor (SRSF)1-12 in humans [60]. SR proteins have demonstrated multiple key roles in the control of gene expression, including constitutive and alternative pre-mRNA splicing, transcription, mRNA transport, mRNA stability and translation [60]. Therefore, these genes could perfectly be key regulatory points through which the WL intervention elicit long-term changes in adipocytes. Beside these replicated patterns, during the investigation of the set of rules containing common LHS or RHS (Ssupplementary table 3), we also noticed a rebound effect in the gene expression of certain loci during the dietary intervention program (Supplementary Fig 9B). Particularly, we observed how although certain genes experimented a downregulation of their gene expression in response to WL, these genes returned to their original gene expression status as soon as a normal-calorie diet was restored (exhibiting some kind of negative and positive feedback loop regulations of their own expression, which could be the explanation of fast transient dynamic changes or the maintaining in time of their expression levels). Altogether, these validated patterns might represent the sequence of genetic changes that occur in AT during a long-term weight loss intervention. Indeed, some of the identified loci have already been drawn as key genes or targets for the management of many complex diseases [62].

Discussion

Temporal gene networking has emerged as an effective approach for filling the missing heritability gap of complex human traits. Until date, several ML approaches have been proposed for the dynamic modelling of time course omics data, highlighting co-expression clustering methods [1]. Although they have yielded impressive results in terms of model accuracy and predictive ability, most of these applications are based on "Black-box" algorithms and more interpretable models have been claimed by the research community [10]. Especially during the reconstruction of gene

networks, one of the main concerns of biologists has been how to translate inferred networks into particular hypotheses that can be tested with real-life experiments. Fortunately, the recent XAI revolution offers a solution for this issue [63–65], were rule-based approaches are highly suitable for explanatory purposes [16,17]. Within this context, SRM approaches have emerged as an interesting XAI method for the modelling of temporal gene-gene interactions *in vitro* [15]. Some of the best characteristics of SRM methods for this task include the existence of statistical quality measures by interaction, the possibility of biological validation by relationship, the inclusion of time (causality) order information in networks or their ability to discover complex regulatory phenomena. Taking all these into account, and given the fact that temporal co-expression clustering methods present some drawbacks as described earlier, we propose that SRM could serve as an alternative of great interest and interpretability for mining particular temporal relations between genes in humans. The further integration of the data mining process along with functional annotation and pathway resources is an additional way towards more explanatory and biologically soundness models [4].

In the present study, we propose a full pipeline for extracting sequential rules from temporal GED through the application of SRM in longitudinal microarray human studies. As far as we concern, this is the first application of a naturally interpretable method for the modelling of temporal gene-gene relationships in humans. The whole pipeline of our method is illustrated in the Supplementary Fig 1. Gathered under open-source software, our proposal could be extended to any temporal GED human study, with special applicability in long-term interventions or clinical trials. The presented pipeline is organized on three main blocks:

1. Data Pre-processing stage, involving feature-selection and data discretization.
2. Knowledge-extraction stage, consisting of the adaptation of the algorithm CMRules to the problem of temporal GED.
3. Functional validation of results, in which we propose five new biological quality measures by rule and a tool for visualizing the results.

The two strategies adopted during the data pre-processing stage were intended to deal with some of the well-known human omics data complexities. As evidenced in the case of study, both strategies resulted useful for increasing model interpretability and for reducing the search space into a high quality data subset. During the second phase of the approach (knowledge-extraction), an SRM algorithm was adapted to the temporal GED problem given the previously proven ability of rule mining methods for extracting biologically meaningful gene association patterns both in static [14] and dynamics datasets [15]. Particularly, a method known as CMRules was chosen as a good technique for this task. CMRules implementation was accomplished following published recommendations in gene association analysis [14] and the biological knowledge played an important role during the mining process.

With this work, we have tried to move away from the "black-box" concept that is adopted in many of the current AI omics highthroutput applications, in which complex genetic networks are extracted from datasets without obtaining useful knowledge for the experts. An example of this kind of "poorly explanatory" models is the work recently published by Tareen *et al.* (2018) [66], with one of the datasets employed here. Although some interesting gene networks are reported in the work, the output format of co-expression networks and their visual representation are poorly explainable by itself, especially for the generation of particular hypotheses of gene-gene interactions. Moreover, the approach lacks of a method for the functional validation of established gene-gene relationships, thus hindering the biological interpretation of results.

In contrast, our approach presents a high *eXplainability*, which is mainly achieved by two consecutive ways: 1) given the type of employed knowledge-extraction algorithm, and 2) Thanks to the third proposal of the pipeline, which includes the functional validation and visual representation of results.

Regarding the knowledge-extraction algorithm, the chosen SRM method CMRules constitutes a methodological advance in comparison to previous SPM approaches. Moreover, it greatest virtue emanates from the format in which its results are presented. This is the form of rules: X -> (time delay) Y, where each interaction between two or more genes could be suggesting a causal time-lagged relationship between them. For example, sometimes, these interactions could be indicating how the increase in the amount of a TF causes a subsequent increase or decrease in the expression of a target gene, while other times they could suggest how two distinct genes (participating in a same metabolic route) increase their expression consecutively after an intervention. In the latter case, for example, the interaction would be illustrating how the activation of certain biological pathway is maintained over time in response to an intervention, after a first trigger event. Additionally, in other cases, when it is the same gene the one that occupies both LHS and RHS positions, rules could be suggesting negative or positive feedback phenomena (which could serve as explanation for fast transient dynamic changes or the maintaining in time of expression levels of certain genes in long-term interventions). Interestingly, all introduced types of relationships have been reported in our tested datasets (see results *section*) and assert with the two core ideas of XAI:

- Explainable models, while maintaining a high level of learning performance (e.g. support, confidence, conviction, CF, lift).
- Enabling human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent methods.

On the other hand, *eXplainability* is also achieved in our pipeline with the creation of five new biological quality measures by rule and by the visual representation of results. Although previous association rule mining studies have already employed the GO and KEGG resources for the functional validation of extracted rules [14], this is the first time specific biological quality measures

by rule are computed taking into consideration external molecular knowledge. The creation of biological quality measures by rule constitute an ideal initiative to sort and explore identified gene patterns according to any desired biological criterion. Moreover, it is also the first time that TF-target gene regulatory information has been taken into account for the functional interpretation of gene rules. Since TF gene regulation is the most plausible molecular mechanism underlying gene-gene and gene-environment interactions, this initiative could be extended to any other gene association analysis. Finally, although the results obtained directly from the CMRules algorithm in the form of sequential rules are perfectly interpretable themselves, their exploration from a global perspective may be somewhat complicated. For this reason, we though that the evaluation of these rules including their biological quality information would greatly benefit from a visual group approach, where all the rules could be studied together. Thus, we decided to incorporate our visualization tool as a stage for the functional validation of the results. We chose a type of plots whose greatest virtue lays in their ability to concentrate different types of information in single shot, which would allow practitioners an easy identification of the top and more coherent rules from the both the technical and biological points of view. On this matter, our visualization tool it is not intended as a network analysis tool (for generating networks from raw data) but only an additional value proposal in favor of model explicability and interpretability.

From the analyzed datasets, computed biological quality measures and the visualization tool demonstrated utility for the biological interpretation of results and the transference of large gene patterns to the expert eye. Thanks to them, field-experts were able to identify several rules corresponding to known biological relationships among genes. Moreover, although our CMRules algorithm does not strictly output a network-format like result such as co-expression approaches do, when one visualizes all sequential rules at a single shot, It is evidenced how SRM interestingly keep the scale-free network topology for inferred interactions. This network topology, which is also evidenced by co-expression approaches results, it is in tune with the concept of "good enough solutions" that seems to rule most biological systems and it consists on the existence of a few nodes with many connections ("hubs") and many nodes with few connections [18]. This, again, demonstrates the suitability of our SRM approach alongside the visualization tool for the modeling genetic interactions in humans.

Given the absence of a gold-standard which to compare with our approach, and taking into account that the important thing when inferring genetic networks in the biological field is to validate results in independent cohorts, in this work we decided to validate our proposal through its direct application to different cases of study. Within the context of the chosen research problem ("WL interventions in obesity"), this is the first study implementing a *XAI* analytic approach in temporal gene networking. Moreover, by incorporating data from up to 4 independent cohorts and 6 experimental groups (N=83 subjects), this analysis also constituted one of the biggest omics applications in the field of omics interventions. After applying our whole pipeline on these

datasets, we not only identified interesting gene networks within each of the mined datasets but also validated some of the patterns primarily extracted from the discovery sample. Altogether, these results further reinforce the goodness of our strategy for the mining of biologically relevant gene-gene temporal relations under different conditions and clinical designs. An exhaustive study of all the results from the case of study is needed otherwise to understand the concrete molecular patterns underlying WL-induced responses in obesity. In order to make this pipeline analysis extensible to any other temporal GED, we have implemented all described methods in open-source scripts and the codes have been shared online. Our method is not necessarily restricted to microarray data and could also be extended to RNAseq and other NGS technologies. For example, for applying our methodology to RNAseq datasets, it would be enough by simply counting on an *N x M* gene expression matrix of normalized reads. In future works, it would be of great interest testing our approach with RNAseq expression datasets given the stronger reliability of these data in term of the technical robustness of sequencing platforms. Approaches like this would greatly expand our knowledge for complex biological processes, with a special interest for long-term intervention experiments (such as clinical trials), in which gene regulatory mechanisms could reveal new drug targets.

The high dimensionality of microarray data is a permanent problem for this kind of approaches. For future analyses, it would also be advisable to test the effect of employing different "feature selection" and "discretization" strategies on the performance of the algorithm. In addition, it would be convenient that the biological quality measures could be computed at the same time that the rule extraction process, in such a way that they can guide the method within the search space. As a result, methods will be able to find fewer rules but with higher biological quality, which may otherwise remain hidden.

Finally, future works could also be focused on improving the computing of biological quality measurements based on GO ontology terms. For that purpose, we will combine our heuristic approach alongside the available tools that have been developed to evaluate the biological similarity of two genes based not only on the identical GO terms that they share, but also on the rest of GO terms that are annotated (not identical) [42]. In the future, a combined approach like this could be of great interest to improve the functional validation of our method and will be taken into consideration for the continuation of the work. Besides this modification, other future approaches could also consist of performing the visual representation of the rules with ontology terms representing nodes instead of genes. This would allow us to visualize networks in terms of functionality and to understand how cellular functions follow each other in human tissues after long-term interventions. In this case, the difficulty would be to identify which GO terms are the most characteristic for each gene in order to represent them within the network. Once achieved, the way in which the nodes of the network are connected could be different to our current representations and thereby reveal novel information extracted by the method that is not observed with our current approach.

Code availability

All data manipulation and processing steps as well as all secondary statistical GED analyses were conducted in R environment using the next list of libraries ("Matrix", "lattice", "fdrtool", "rpart", "affy", "oligo", "affydata", "ArrayExpress", "limma", "Biobase", "Biostrings", "genefilter", "affyQCReport", "affyPLM", "simpleaffy", "ggplot2", "dplyr", "pd.hugene.1.1.st.v1", "FGNet", "RGtk2", "RDAVIDWebService", "topGO", "KEGGprofile", "GO.db", "KEGG.db", "reactome.db", "org.Hs.eg.db", "arules", "arulesViz"). All employed codes have been gathered under a unique pre-processing R script, which is available online. The implementation of CMRules was carried out in Java using the open-source data mining library "SPMF" (http://www.philippe-fournier-viger.com/spmf/). The computing process for the five new biological quality measures was implemented in Python version 3.7 (http://www.python.org). The data visualization process instead was implemented in R environment. The codes for running all described processes (pre-processing, CMRules mining, computing of biological quality measures and the data visualization tool) have been shared online and can be easily extended to any other application. This software is distributed as open source software under the terms of the GNU Public License GPLv3 and it is hosted in the public hosting GitHub (https://github.com/AugustoAnguita/GeneSeqRules).

Supplementary data

Supplementary Data are available online at https://doi.org/10.1371/journal.pcbi.1007792.

Acknowledgement

The authors would like to thank the owners of datasets (GSE77962, GSE70529, GSE3541 and GSE103766) and the GEO repository for making science accessible for everyone and for encouraging the open data culture. The authors also acknowledge the University of Granada "Plan Propio de Investigacion 2016-Excellence actions: Unit of Excellence on Exercise and Health (UCEES)". This paper will be part of Augusto Anguita-Ruiz's doctorate, which is being completed at the University of Granada, Spain.

References

1. Liang Y, Kelemen A. Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. Brief Bioinform. 2018;19: 1051–1068. doi:10.1093/bib/bbx036
2. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446–450. doi:10.1038/nrg2809
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41: D991–D995. doi:10.1093/nar/gks1193
4. Liang Y, Kelemen A. Computational dynamic approaches for temporal omics data with applications to systems medicine. BioData Min. 2017;10: 1–20. doi:10.1186/s13040-017-0140-x
5. Lee W-P, Tzou W-S. Computational methods for discovering gene networks from expression data. Brief Bioinform. 2009;10: 408–23. doi:10.1093/bib/bbp028
6. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Res. 2018;46: D348–D353. doi:10.1093/nar/gkx842
7. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9: 796–804. doi:10.1038/nmeth.2016
8. Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv. 2017;abs/1708.0.
9. Castelvecchi D. Can we open the black box of AI? Nature. 2016;538: 20–23. doi:10.1038/538020a
10. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1: 206–215. doi:10.1038/s42256-019-0048-x
11. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv. 2017;abs/1702.0.
12. Agrawal R, Imielinski T, Swami A. Mining Association in Large Databases. ACM SIGMOD Rec. 1993, 22, 207–216.. doi:10.1145/170036.170072
13. Fournier-Viger P, Chun J, Lin W, Kiran RU, Koh YS, Thomas R. A Survey of Sequential Pattern Mining. 2017.
14. Alves R, Rodriguez-Baena DS, Aguilar-Ruiz JS. Gene association analysis: a survey of frequent pattern mining from gene expression data. Brief Bioinform. 2010;11: 210–224. doi:10.1093/bib/bbp042
15. Nam H, Lee KY, Lee D. Identification of temporal association rules from time-series microarray data sets. BMC Bioinformatics. 2009;10: 1–9. doi:10.1186/1471-2105-10-S3-S6
16. Truong-Chi T, Fournier-Viger P. A Survey of High Utility Sequential Pattern Mining. 2019. pp. 97–129. doi:10.1007/978-3-030-04921-8_4
17. Liu Y-C, Cheng C-P, Tseng VS. Mining differential top-k co-expression patterns from time course comparative gene expression datasets. BMC Bioinformatics. 2013;14: 230. doi:10.1186/1471-2105-14-230
18. Weiss JN, Karma A, MacLellan WR, Deng M, Rau CD, Rees CM, et al. "Good Enough Solutions" and the Genetics of Complex Diseases. Circ Res. 2012;111: 493–504. doi:10.1161/CIRCRESAHA.112.269084
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559
20. Li J, Lai Y, Zhang C, Zhang Q. TGCnA: temporal gene coexpression network analysis using a low-rank plus sparse framework. J Appl Stat. 2019. doi:10.1080/02664763.2019.1667311
21. Albrecht M, Stichel D, Müller B, Merkle R, Sticht C, Gretz N, et al. TTCA: An R package for the identification of differentially expressed genes in time course microarray data. BMC Bioinformatics. 2017;18: 1–11. doi:10.1186/s12859-016-1440-8
22. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Brief Bioinform. 2018;19: 878–892. doi:10.1093/bib/bbx017
23. Jiang Z, Zhou Y. Using gene networks to drug target identification. J Integr Bioinform. 2005;2: 48–57. doi:10.1515/jib-2005-14
24. Gallo CA, Cecchini RL, Carballido JA, Micheletto S, Ponzoni I. Discretization of gene expression data revised. Brief Bioinform. 2016;17: 758–770. doi:10.1093/bib/bbv074
25. Fournier-Viger P, Faghihi U, Nkambou R, Nguifo EM. CMRules: Mining sequential rules common to several sequences. Knowledge-Based Syst. 2012;25: 63–76. doi:10.1016/J.KNOSYS.2011.07.005
26. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46: D380–D386. doi:10.1093/nar/gkx1013

27. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016; bbw068. doi:10.1093/bib/bbw068

28. Geng L, Hamilton HJ. Interestingness measures for data mining. ACM Comput Surv. 2006;38: 9-es. doi:10.1145/1132960.1132963

29. Brin S, Motwani R, Ullman JD, Tsur S, Brin S, Motwani R, et al. Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Rec. 1997;26: 255–264. doi:10.1145/253262.253325

30. Berzal F, Blanco I, Sanchez D, Vila MA. Measuring the accuracy and interest of association rules: A new framework. Intelligent Data Analysis. 2002. pp. 221–235. doi:10.3233/ida-2002-6303

31. Gupta A, Shmueli O, Widom J. Proceedings of the Twenty-fourth International Conference on Very Large Databases, New York, NY, USA, 24-27 August, 1998. Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann; 1998.

32. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Math Biosci. 1975;23: 351–379. doi:10.1016/0025-5564(75)90047-4

33. Chattopadhyay S, Rakesh S, Land LPW, Acharya UR. Studying infant mortality rate: a data mining approach. Health Technol (Berl). 2011;1: 25–34. doi:10.1007/s12553-011-0005-0

34. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31: 15e – 15. doi:10.1093/nar/gng015

35. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008;9: 770–780. doi:10.1038/nrm2503

36. Garcia S, Luengo J, Sáez JA, López V, Herrera F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. IEEE Trans Knowl Data Eng. 2013;25: 734–750. doi:10.1109/TKDE.2012.35

37. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform. 2013;14: 469–490. doi:10.1093/bib/bbs037

38. Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol. 2003;4: R6. doi:10.1186/gb-2003-4-1-r6

39. Madeira SC, Arlindo P, Oliveira L. An Evaluation of Discretization Methods for Non-Supervised Analysis of Time-Series Gene Expression Data. INESC-ID Technical Report. 2005, 42/2005..

40. Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, et al. Gene ontology: tool for the unification of biology. {T}he {G}ene {O}ntology {C}onsortium. Nat Genet. 2000, 25, 25-9.

41. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012; 40, D109-14. doi:10.1093/nar/gkr988

42. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5. doi:10.1371/journal.pcbi.1000443

43. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019;47: W234–W241. doi:10.1093/nar/gkz240

44. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nat Methods. 2010;7: S56–S68. doi:10.1038/nmeth.1436

45. Pavlopoulos GA, Wegener A-L, Schneider R. A survey of visualization tools for biological network analysis. BioData Min. 2008;1: 12. doi:10.1186/1756-0381-1-12

46. Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on Visualization and Computer Graphics. 2006. pp. 741–748. doi:10.1109/TVCG.2006.147

47. Vink RG, Roumans NJ, Fazelzadeh P, Tareen SHK, Boekschoten M V, van Baak MA, et al. Adipose tissue gene expression is differentially regulated with different rates of weight loss in overweight and obese humans. Int J Obes. 2017;41: 309–316. doi:10.1038/ijo.2016.201

48. Vink RG, Roumans NJT, Arkenbosch LAJ, Mariman ECM, van Baak MA. The effect of rate of weight loss on long-term weight regain in adults with overweight and obesity. Obesity. 2016;24: 321–327. doi:10.1002/oby.21346

49. Magkos F, Fraterrigo G, Yoshino J, Luecking C, Kirbach K, Kelly SC, et al. Effects of Moderate and Subsequent Progressive Weight Loss on Metabolic Function and Adipose Tissue Biology in Humans with Obesity. Cell Metab. 2016;23: 591–601. doi:10.1016/j.cmet.2016.02.005

50. Johansson LE, Danielsson AP, Parikh H, Klintenberg M, Norström F, Groop L, et al. Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance. Am J Clin Nutr. 2012;96: 196–207. doi:10.3945/ajcn.111.020578

51. Bollepalli S, Kaye S, Heinonen S, Kaprio J, Rissanen A, Virtanen KA, et al. Subcutaneous adipose tissue gene expression and DNA methylation respond to both short-

and long-term weight loss. Int J Obes. 2018;42: 412–423. doi:10.1038/ijo.2017.245

52. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res. 2019;47: D100–D105. doi:10.1093/nar/gky1128
53. González-Muniesa P, Marrades M, Martínez J, Moreno-Aliaga M. Differential Proinflammatory and Oxidative Stress Response and Vulnerability to Metabolic Syndrome in Habitual High-Fat Young Male Consumers Putatively Predisposed by Their Genetic Background. Int J Mol Sci. 2013;14: 17238–17255. doi:10.3390/ijms140917238
54. Battle M, Gillespie C, Quarshie A, Lanier V, Harmon T, Wilson K, et al. Obesity induced a leptin-Notch signaling axis in breast cancer. Int J Cancer. 2014;134: 1605–1616. doi:10.1002/ijc.28496
55. Sandel DA, Liu M, Ogbonnaya N, Newman JJ. Notch3 is involved in adipogenesis of human adipose-derived stromal/stem cells. Biochimie. 2018;150: 31–36. doi:10.1016/j.biochi.2018.04.020
56. Henegar C, Tordjman J, Achard V, Lacasa D, Cremer I, Guerre-Millo M, et al. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. Genome Biol. 2008;9: R14. doi:10.1186/gb-2008-9-1-r14
57. Kawaguchi N, Sundberg C, Kveiborg M, Moghadaszadeh B, Asmar M, Dietrich N, et al. ADAM12 induces actin cytoskeleton and extracellular matrix reorganization during early adipocyte differentiation by regulating 1 integrin function. J Cell Sci. 2003;116: 3893–3904. doi:10.1242/jcs.00699
58. de Haan W, Bhattacharjee A, Ruddle P, Kang MH, Hayden MR. ABCA1 in adipocytes regulates adipose tissue lipid content, glucose tolerance, and insulin sensitivity. J Lipid Res. 2014;55: 516–523. doi:10.1194/jlr.M045294
59. Yu C-Y, Theusch E, Lo K, Mangravite LM, Naidoo D, Kutilova M, et al. HNRNPA1 regulates HMGCR alternative splicing and modulates cellular cholesterol metabolism. Hum Mol Genet. 2014;23: 319–32. doi:10.1093/hmg/ddt422
60. Zhou Z, Fu X-D. Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma. 2013;122: 191–207. doi:10.1007/s00412-013-0407-z
61. Lin S, Negulescu A, Bulusu S, Gibert B, Delcros J-G, Ducarouge B, et al. Non-canonical NOTCH3 signalling limits tumour angiogenesis. Nat Commun. 2017;8: 16074. doi:10.1038/ncomms16074
62. Lin L, Yee SW, Kim RB, Giacomini KM. SLC transporters as therapeutic targets: emerging opportunities. Nat Rev Drug Discov. 2015;14: 543–60. doi:10.1038/nrd4626
63. Runge J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos. 2018;28. doi:10.1063/1.5025050
64. Zhou D, Xiao Y, Zhang Y, Xu Z, Cai D. Granger causality network reconstruction of conductance-based integrate-and-fire neuronal systems. PLoS One. 2014;9. doi:10.1371/journal.pone.0087636
65. Abdul Razak F, Jensen HJ. Quantifying "causality" in complex systems: Understanding transfer entropy. PLoS One. 2014;9. doi:10.1371/journal.pone.0099462
66. Tareen SHK, Adriaens ME, Arts ICW, de Kok TM, Vink RG, Roumans NJT, et al. Profiling cellular processes in adipose tissue during weight loss using time series gene expression. Genes (Basel). 2018;9. doi:10.3390/genes9110525

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

general discussion and future perspectives

FROM THE FIRST SEQUENCING of the human genome in 2001, genetic technologies have rapidly emerged discovering batteries of SNPs as risk variants for chronic diseases. The use of such SNPs as risk markers initially appeared as an interesting clinical tool, since DNA sequence is a parameter that can be studied from the moment of birth, when there is still room for clinical actions. Likewise, SNPs were also proposed as interesting markers for assessing differential responses to pharmacological treatments in order to increase drug efficacy. The potential clinical utility of SNPs together with the rapid improvements and costs reductions of genotyping and sequencing technologies have made genetics to become one of the preferred tools in the individualisation or personalisation of clinical care.

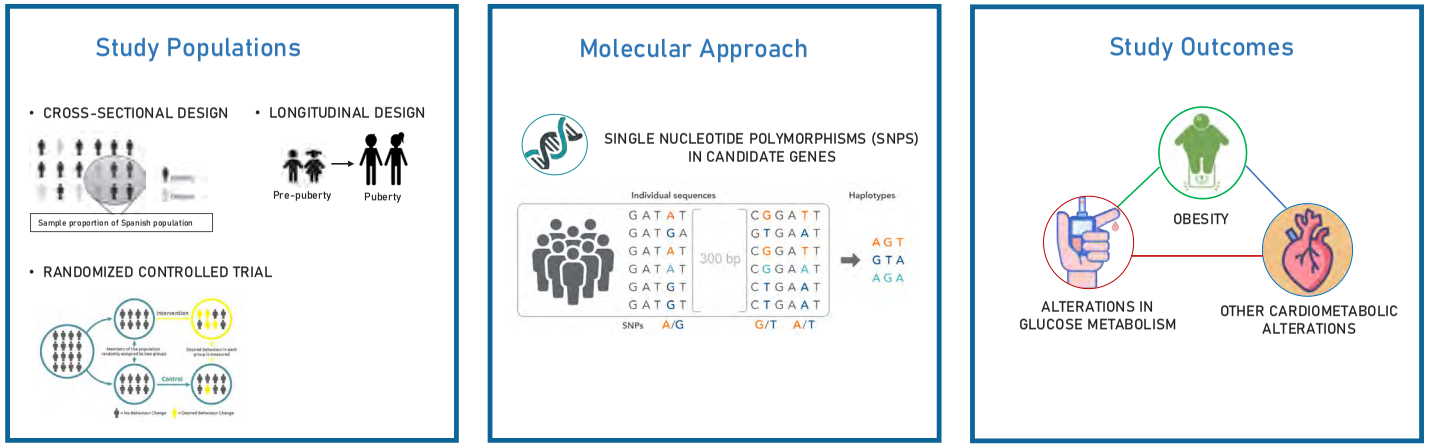
In the first section of the present Doctoral Thesis (**Figure 8**), SNPs in candidate genes are investigated as risk biomarkers for obesity and its metabolic alterations, as well as for their potential utility as pharmacogenetics signals of metformin response. Particularly, studies 1 and 2 show that SNPs in the X chromosome *TNMD* gene are associated with alterations in the glucose metabolism of children with obesity, and that the *TNMD* gen potentially plays interesting metabolic roles in adipocytes. Study 3, otherwise, provides evidence for 28 SNPs as promising pharmacogenetic regulators of metformin treatment response in terms of a wide range of anthropometric and biochemical outcomes, including glucose, lipid, and inflammatory traits. Finally, study 4 demonstrates that gathering obesity SNPs into a single obesity-predisposing GRS increases the robustness of their association with childhood BMI Z-Score, and that it could be used as a predictor of obesity longitudinal trajectories during puberty.

Despite the great findings reported by genetic research in obesity during the last decades, the SNPs identified so far explain only about 2–4% of the adult BMI variability, suggesting that there are still unknown genetic elements involved. This genetic gap or lacking knowledge has been termed as the 'missing heritability' problem, and is a common issue in complex disorders showing a polygenic architecture like the case of common obesity. According to several studies, part of this 'missing heritability' could be explained by the effect of SNPs that have not been yet discovered (for example, rare variants or SNPs mapping genomics regions neglected in previous GWASs) but could also be attributed to more complex and non-genetics phenomena. Indeed, obesity depends not just on genetics, but also on environment and the interaction between the two. On this regard, we might also be missing the existence of SNPs that increase the risk of obesity but only under certain conditions or environmental exposures.

As a common objective, the studies composing the first section of the present Doctoral Thesis address the 'missing heritability' issue in two ways. Primarily, they focus in the study of genomics regions previously neglected in obesity GWASs and that could harbour SNPs ever identified as associated with metabolic dysfunction. Particularly, they investigate SNPs mapping the *TNMD* gene in the X chromosome, which has been previously omitted from GWASs due to the complexity of its analysis. Secondly, they combine the effects of obesity SNPs as part of a more sophisticated tool, such is the case of a GRS, which allow for taking into account the cumulative polygenic nature of obesity. Interestingly, GRSs have been evidenced as a perfect tool for the investigation of gene-environment interactions, and thus are another efficient way of addressing the 'missing heritability' problem. As main results, these studies corroborate that potentially interesting SNPs and genes for obesity and metabolic dysfunction might remain unknown in the sexual chromosomes, and that the rationale combination of known obesity SNPs might increase the variability explained by the genetic component. Particularly, the obesity heritability attributable to assessed genetic markers in our study populations was estimated in 5.6%, which is higher than the 2–4% reported in adults. This is an interesting finding which could also be explained by the fact that hormonal and environmental modulatory effects on genetics during childhood may be softer than in adults. Therefore, it highlights the relevance of focusing on childhood for the study of the genetics basis of a disease. Finally, beyond prediction, the studies included in the first section of this Thesis, are also interesting since they contribute to the identification of new potential therapeutic molecular targets and biological pathways underlying the pathophysiology of disease. On this regard, it is not all about developing new predictive tools but simply increasing our knowledge about disease mechanisms, which at the end could open new lines of research in the development of novel and more effective anti-obesity drugs.

Despite the positive results and conclusions evidenced from this section, some limitations should be noticed. On the one hand, GRSs based on standard GWAS-discovered SNPs may not be the most suitable tool for investigating gene-environment phenomena, since many of the SNPs that truly interact with environmental exposures are not usually discovered as associated with obesity under the standard case-control obesity designs. In this sense, particular GRSs should be constructed based on environment-related SNPs in order to better asses the presence of gene-interactions in obesity. On the other hand, although there are reasons to hope that identified SNPs of obesity and diabetes will eventually lead to new preventive and therapeutic agents, this will take time because such developments require detailed mechanistic understanding of how an SNP influences phenotype. This involves the identification of the gene or genes whose expression is affected by alleles at the variant, and the mechanism (e.g., enhancer, repressor, epigenetic alteration) whereby the variant's alleles differentially affect expression. On this sense, genetic variants can introduce or delete methylation sites in CpG context, thereby inducing changes in DNA methylation at the SNP site. Moreover, SNPs located in cis or trans of a CpG site can alter the action of methylation enzymes. In order to test the existence of potential interactions between

Section 1: Study of genetic variants associated with childhood obesity and alterations in the glucose metabolism



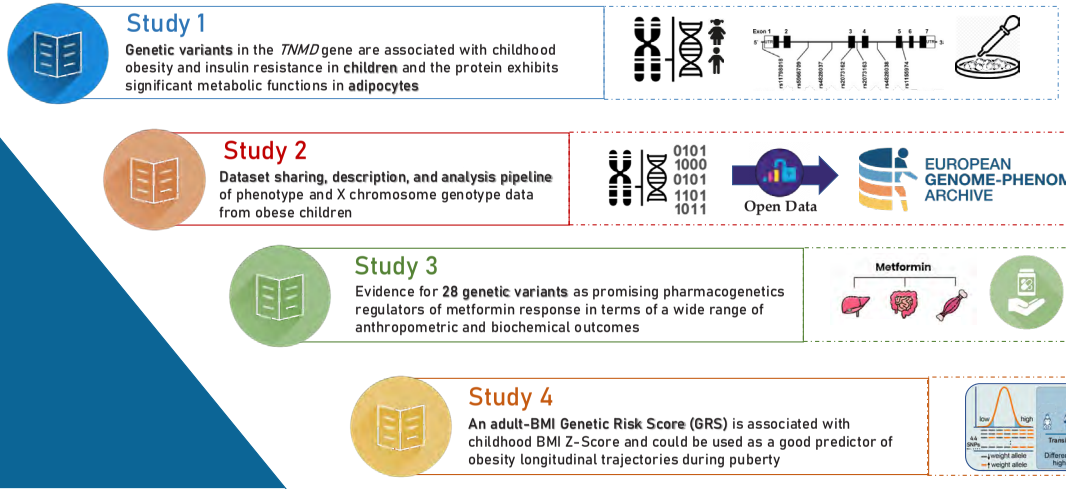


Figure 8. Graphical summary of the first section of the present doctoral thesis.

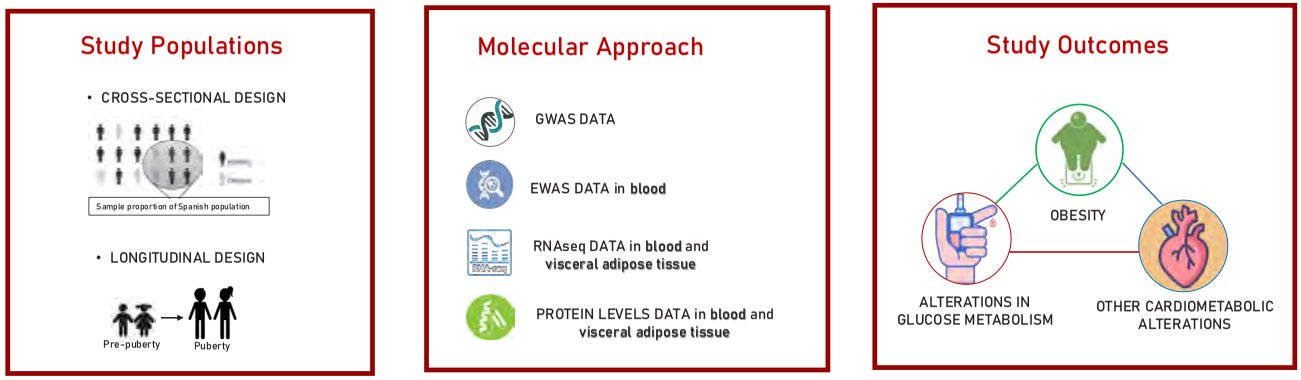
epigenetic profiles and genetic factors several studies have performed correlation analyses of DNA methylation and SNP genotypes (mQTLs). Other interesting approaches have involved the study of the correlation between DNA methylation and gene expression patterns (eQTMs).

The need for a better characterization of the molecular mechanisms underlying obesity and metabolic dysfunction was the motivation for the second section of this Thesis (**Figure 9**), in which two original researches are included (studies 5 and 6). In this second section, new multi-omics biomarkers for IR and cardiometabolic alterations are proposed in children with obesity during the metabolically critical period of puberty. Notably, this section focuses in the IR phenotype more than in obesity, since as mentioned in the general introduction, IR is one of the metabolic comorbidities of obesity that shows an earliest appearance in patients. Likewise, IR is the main pathophysiological mechanism leading to type 2 diabetes and CVD. For the investigation of such molecular patterns in IR and obesity, the second section of this Doctoral Thesis employs cross-sectional and longitudinal data derived from our Spanish cohorts within the context of puberty, a life stage of considerable metabolic risk for children with obesity.

Within this second section, the study 5, is an original research in which a potential candidate molecular marker for IR, the S100A4 protein, is investigated from a multi-omics perspective (involving the parallel analysis of EWAS, Transcriptomics Array and protein data). Otherwise, the study 6 is an ambitious research project analysing EWAS, GWAS, RNAseq and protein data in the PUBMEP longitudinal cohort. The study 6, which is the main research work conducted within the context of the present Doctoral Thesis, is a large-scale multi-omics integrative and longitudinal analysis aiming to unveil the molecular architecture and biological processes underlying IR in puberty, and the additional impact of obesity on these processes. As far as we concern, the study 6 is the first longitudinal multi-omics approach conducted to date for characterizing the molecular blood alterations for IR and obesity during the metabolically critical period of puberty.

In general, the results derived from the second section of this Doctoral Thesis shed new lights onto the molecular mechanisms and the epigenetic alterations of obesity and IR, and propose novel and promising multi-omics biomarkers for disease prevention. Particularly, two of the proteins highlighted in this section are the *S100A4* and the *VASN*, that had never been investigated in children with obesity. Interestingly, our results indicate that epigenetic signatures in these genes could be potentially useful as predictive tools for the appearance and development of IR in children with obesity when they enter into puberty. Moreover, given the strong changes in DNA methylation evidenced for these genes during puberty, they could also be part of the molecular mechanisms by which the obesogenic environment contributes to the IR prognosis in obesity. Given the importance and robustness of our findings, in the near future, additional in vitro and in vivo functional studies should be encouraged in order to clarify the exact role of these proteins in the pathophysiology of disease. An example of required studies can be seen in the first section of this Thesis, in which the TNMD genotyping analysis conducted in children was complemented

Section 2: Identification of new multi-omics biomarkers of insulin resistance and cardiometabolic alterations in childhood obesity during the metabolically critical period of puberty



Study 5

First report of an association between the **protein S100A4** and **insulin resistance (IR)** and **visceral adipose tissue** dysfunction in **prepubertal populations**. Beside, the change in plasma **S100A4** levels accompanies longitudinal trajectories of IR in children during **pubertal development**. **Epigenetic changes** in two methylation sites and an altered ***S100A4*** **adipose tissue expression** are proposed as plausible molecular mechanisms underlying this disturbance in obesity.

Study 6

First longitudinal **multi-omics** approach characterizing **molecular blood alterations** for insulin resistance (IR) and obesity during the metabolically critical period of **puberty**. Our results shed light on the molecular mechanisms underlying **epigenetic alterations** in obesity and propose novel and promising biomarkers for IR and metabolic alterations in children.

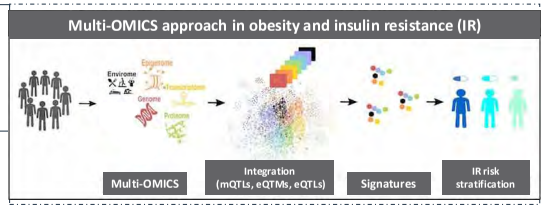


Figure 9. Graphical summary of the second section of the present doctoral thesis.

with a molecular biology approach involving the study of the metabolic functionality of the TNMD gene in adipocyte cultures. This kind of collaborative research is being more and more adopted by current biomedical laboratories and involve the work of multidisciplinary teams (molecular biologists, bioinformaticians, geneticists, physicians, etc). This type of research should be encouraged in the next years since it acts as a bridge between basic and clinical research and allow materializing the concepts of "bench to bedside" and "bedside to bench".

As we have previously mentioned, the aetiology of obesity and type 2 diabetes is multifactorial. Thus, identification of one specific factor associated with disease will most probably have limited prognostic or therapeutic value. Additionally, association does not imply causation and associations actually outnumber causations, where many of the reported associations are not reproduced in future studies. In this context, the multi-omics approach implemented in studies 5 and 6 allows for the identification of associated factors from different biological dimensions, i.e., DNA genetics variants, DNA methylation, gene expression, protein synthesis, etc., maximizing the available information, and thus, increasing the possibility of identifying the root causes of a disease. A second advantage of multi-omics analysis is the depth of the information it provides. For example, a single change in gene expression may be weakly associated with the pathophysiology of a multifactorial disease such as obesity. However, when this finding is further supported with alterations in DNA methylation and in protein concentration, the possibility that this gene or protein is an important factor in the pathogenesis of the disease increases. On this regard, the findings derived from the present Doctoral Thesis propose novel and reliable molecular mechanisms underlying the development of IR and reveal important pathways never associated with the aetiology of disease that merits additional attention.

Some limitations to highlight in this second section are, on the one hand, the low number of participants included in study populations and the fact that findings are mainly based on data from blood, which was the most accessible tissue, and may not be representative of other metabolically relevant organs such as live and adipose and muscle tissues. In this regard, there is a trend pointing to a correlation between the global state of methylation in blood and adipose tissue. This correlation might be explained by the abundant presence of white cells in both tissues and suggests that buffy coat might be a valid indicator of what happens at the methylation level in adipose tissue, especially for the case of inflammatory and immune system-related aspects. Another possible source of bias would be the difference in time elapsed between the two measurements (prepubertal and pubertal times) between the different participants of the PUBMEP cohort.

From presented results, it can be concluded that the concept of precision medicine is more than the use of genetic variants, and that the use of multi-omics approaches in the clinic would be extremely valuable. In spite of it, since the availability of multi-omics research approaches is still scarce, the clinical application of multi-omics signatures as predictive and prognostic biomarkers of disease will have to wait. One of the main reasons why multi-omics approaches are not yet

Section 3: Implementation of unsupervised machine learning (ML) models for the analysis of longitudinal omics data in obesity

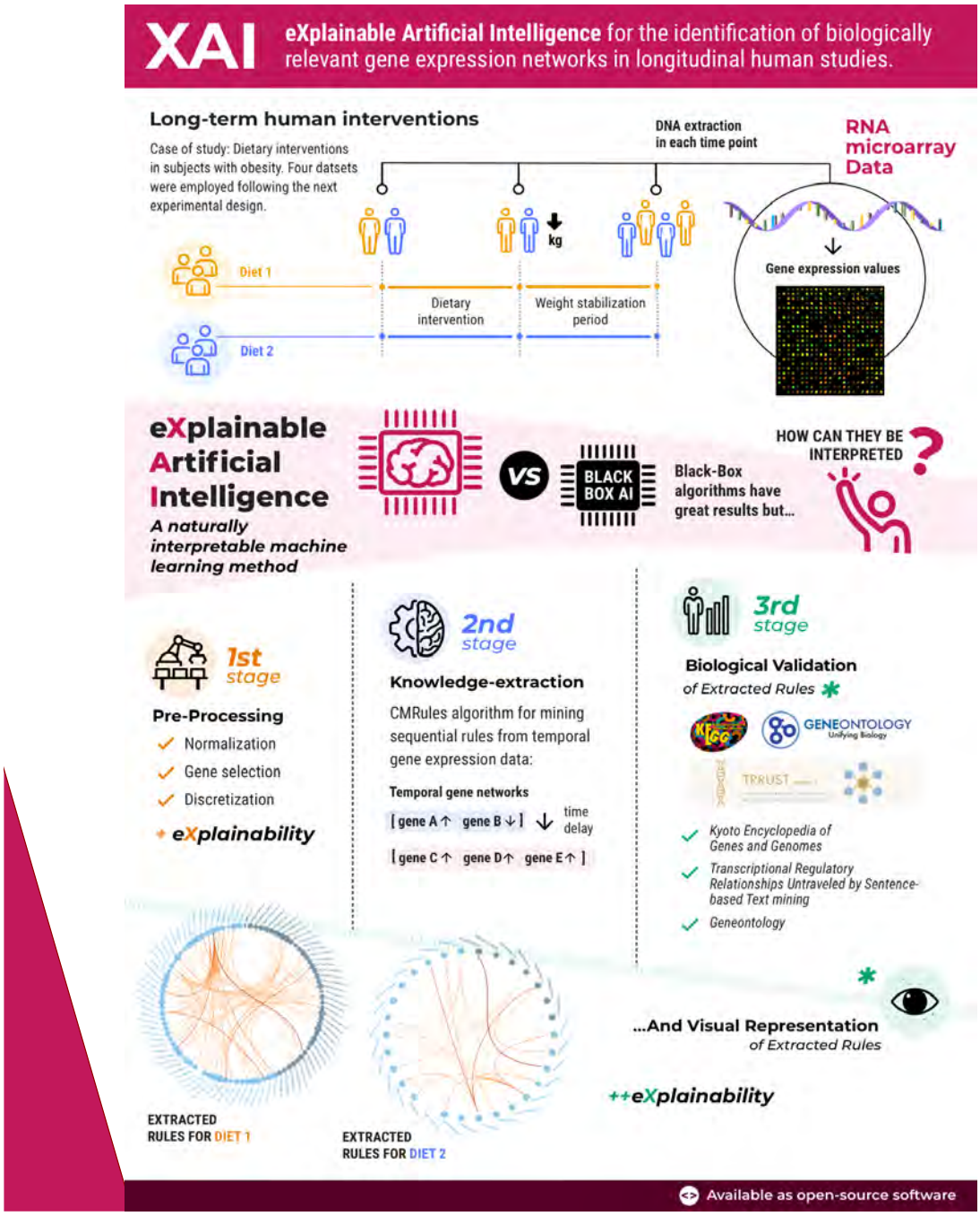


Figure 10. Graphical summary of the third section of the present doctoral thesis.

widespread in obesity research are the complexities in data analysis they involve (e.g., integration of multiple layers of complex data, high dimensionality datasets, existence of noise and spurious associations, and the need for an easy interpretation of findings). In the middle of this need, ML techniques, one of the areas of AI, have experienced a remarkable boost due to their ability to automatically obtain descriptive or predictive models from massive amounts of data (Big Data). These models not only allow us to improve our understanding of obesity, but also to improve our ability to predict with unprecedented accuracy.

The high potential of ML techniques and the need for new analytics approaches in multi-omics research was the motivation for the third and last section of the present Doctoral Thesis **(Figure 10**), in which an unsupervised ML model was proposed for the extraction of gene expression temporal patterns (study 7). Particularly, here, it was opted by the application of a method for the discovery of sequential associations in accordance with the current trend of making ML models more interpretable and explainable, giving rise to what is known as eXplainable Artificial Intelligence (XAI). Explainability in AI is a heavily debated topic with far-reaching implications that extend beyond the technical dimension, as in most cases, scientists do not understand how algorithms learn automatically from data and how they make decisions (the so-called "black box problem"). For fields such as health care, where mistakes can have devastating effects, the lack of interpretability in AI makes it even more difficult for physicians to trust it. XAI is also especially important in omics research, where one of the main concerns of biologists is how to translate inferred networks into particular hypotheses that can be tested with real-life experiments.

In the present Doctoral Thesis, this approach was implemented and materialized in the form of a methodological paper (including pre-processing, knowledge extraction and functional validation) based on sequential rule mining (study 7). The proposed method was validated in six datasets from obesity research (consisting of low-calorie diets interventions), where it was able to extract meaningful gene-gene temporal interactions with relevance in the aetiology of the disease. The application of such pipeline to other type of human temporal gene profiles would greatly expand our knowledge for complex biological processes, with a special interest for drug clinical trials, in which identified gene-gene regulatory interactions could reveal new therapeutic targets.

Ultimately, the clinical purpose of molecular sciences is to provide diagnoses and forecasts of future disease risk. Relatively simple statistical approaches such as GRSs have allowed for certain stratification ability for some common complex diseases, as it has been demonstrated in the present Doctoral Thesis. Nevertheless, they are still far from offering a true clinical utility. Complementarily, a few studies have attempted genomic prediction of complex human traits using AI algorithms, but most of those reported in the literature to date are probably overfit as they purportedly explain substantially more trait variance than should be possible on the basis of heritability estimates. In the future, an intelligent use of AI would probably come from the integration of a variety of

omics, health data types and risk factors as comprehensive predictors of disease risk. Future lines of research in the study populations collected in the present Doctoral Thesis thus might involve the creation of supervised AI predictive models based on prepubertal multi-omics and environmental data for the prediction of the future IR status in children with obesity. These tools would be of much interest for personalizing care in obesity and, at the end, would drastically reduce the associated deaths and economic costs of disease.

In summary, the results presented in the present Doctoral Thesis indicate that; 1) obesity is a complex disorder resulting from the interaction between genetic and environmental factors, 2) part of the missing heritability in obesity could be explained by the existence of neglected SNPs and rare variants in genomic regions such as the sexual chromosomes, 3) the creation of predictive tools based on the combination of small-risk effects SNPs is an interesting but simple strategy for predicting future obesity, 4) a multi-omics study of obesity is necessary to understand its complex underlying molecular mechanisms, and 5) the application of XAI ML models can help us to unravel the complex relationships between omics molecular elements. Further studies like those presented in the present Doctoral Thesis and as well as larger cohorts recruitments should be encouraged in order to validate presented findings. This will require a close collaboration between clinicians and basic researchers, and the creation of multidisciplinary teams, in which the presence of mixed bioinformatics profiles will be of great importance.

concluding

CONCLUDING REMARKS

remarks

General Conclusion

The molecular basis of childhood obesity represents a complex network of interactions between omic elements (SNPs, DNA methylation, RNA molecules, and proteins) and environmental factors. Only through the application of multi-omics research approaches and by employing complex analytical tools (such as bioinformatics and AI) we will be able to understand the complete molecular architecture of obesity, which at the end will allow us to design effective preventive tools or develop new personalised treatments.

Specific Conclusions

Section I. Study of genetic variants associated with childhood obesity and alterations in the glucose metabolism.

Study 1. Genetic variants within the *TNMD* gene in the X chromosome are associated with obesity and alterations in the glucose metabolism in children. Moreover, the TNMD protein might present significant metabolic functions in adipocytes and thus constitute a potential therapeutic target to improve the altered glucose metabolic status.

Study 2. X chromosome is an under-investigated genomic region with great potential for disease prevention. X-chromosome datasets and pipelines are crucial to get familiar with sex chromosome particularities and raise awareness of the importance of this genomic region. SNP and genotype data repositories should improve in some ways in order to provide data access without barriers in genetics.

Study 3. Genetic variants within previously-reported and well-known obesity loci, such as the *ADYC3* and the *BDNF-AS*, could explain part of the inter-individual variability in metformin response, and therefore clinically predict metformin efficacy based on genetics.

Study 4. Gathering obesity SNPs into a GRS increases the robustness of their association with childhood BMI Z-Score, and could be used as a predictor of obesity longitudinal trajectories during puberty. Otherwise, the GRS is not associated with cardio-metabolic comorbidities in children and certain environmental factors interact with the genetic predisposition to the disease.

Section II. Identification of new multi-omics biomarkers of IR and cardiometabolic alterations in childhood obesity during the metabolically critical period of puberty.

Study 5. The protein S100A4 is a novel and promising biomarker of IR in prepubertal and pubertal children with obesity, exhibiting altered multi-omics signatures in blood and metabolically relevant tissues. Particularly, we propose epigenetic changes in two methylation sites and an altered *S100A4* expression as plausible molecular mechanisms underlying this disturbance in obesity.

Study 6. Blood DNA methylation patterns significantly associate with IR longitudinal trajectories in children with obesity during pubertal maturation. Among identified genes, some new targets never reported in obesity research, such is the case of *VASN*, shed light onto new molecular multi-omics mechanisms underlying metabolic alterations in obesity and could serve as promising predictive biomarkers for IR.

Section III. Implementation of unsupervised machine learning (ML) models for the analysis of longitudinal omics data in obesity.

Study 7. Sequential rule mining is a type of unsupervised ML technique highly interpretable and self-eXplainable with a great potential for finding biologically relevant sequential patterns from longitudinal human gene expression data. The application of rule-based methods to other type of human temporal gene profiles would greatly expand our knowledge for complex biological processes, with a special interest for drug clinical trials, in which identified gene-gene regulatory interactions could reveal new therapeutic targets.

annexes

ANNEXES

Papers derived from the doctoral thesis

PUBLISHED/ACCEPTED PAPERS

Ruiz-Ojeda FJ*, Anguita-Ruiz A*, Rupérez AI, et al. Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism. **Sci Rep**. 2019;9(1):3979. doi:10.1038/s41598-019-40482-0. *Equal contributors. IF: 3.998, Q1 at MULTIDISCIPLINARY SCIENCES.

Anguita-Ruiz A, Plaza-Diaz J, Ruiz-Ojeda FJ, et al. X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline. **Sci Data**. 2019;6(1):130. doi:10.1038/s41597-019-0109-3. IF: 5.541, Q1 at MULTIDISCIPLINARY SCIENCES.

Anguita-Ruiz A, Pastor-Villaescusa B, Leis R, et al. Common Variants in 22 Genes Regulate Response to Metformin Intervention in Children with Obesity: A Pharmacogenetic Study of a Randomized Controlled Trial. **J Clin Med**. 2019;8(9):1471. doi:10.3390/jcm8091471. IF: 3.303, Q1 at MEDICINE, GENERAL & INTERNAL.

Anguita-Ruiz A, González-Gil EM, Rupérez AI, et al. Evaluation of the Predictive Ability, Environmental Regulation and Pharmacogenetics Utility of a BMI-Predisposing Genetic Risk Score during Childhood and Puberty. **J Clin Med**. 2020;9(6):1705. doi:10.3390/jcm9061705. IF: 4.241, Q1 at MEDICINE, GENERAL & INTERNAL.

Anguita-Ruiz A, Mendez-Gutierrez A, Ruperez AI, et al. The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity. **Metabolism**. 2020;105:154187. doi:10.1016/j.metabol.2020.154187. IF: 8.694, D1 at ENDOCRINOLOGY & METABOLISM.

Anguita-Ruiz A, Segura-Delgado A, Alcalá R, Aguilera CM, Alcalá-Fdez J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. **PLoS Comput Biol**. 2020;16(4):e1007792. doi:10.1371/journal.pcbi.1007792. IF: 4.475, Q1 at MATHEMATICAL & COMPUTATIONAL BIOLOGY.

PAPERS IN PREPARATION/SUBMITTED

Anguita-Ruiz A, Ruiz-Ojeda FJ, Alcalá-Fdez J, et al. Integrative analysis of blood cells DNA methylation, transcriptomics and genomics identifies novel epigenetic regulatory mechanisms of insulin resistance during puberty in children with obesity: a longitudinal study. Diabetologia. Submitted.

Curriculum Vitae

Augusto Anguita-Ruiz

Date of birth: 08/07/1992 | **Nationality:** Spanish | **Gender** Male | (+34) 628116908 |

augustoanguitaruiz@gmail.com | augustoanguita@ugr.es |

https://www.linkedin.com/in/augusto-anguita-ruiz-33458aa0 |

https://orcid.org/0000-0001-6888-1041 | https://github.com/AugustoAnguita |

Skype: https://join.skype.com/invite/imCX4RorKv8r |

"Cruz" Street, number 6, 18630, Granada, Spain

About me: Biological data scientist working on Obesity and Type II Diabetes (T2D). Biochemist by training, I am specialized in the analysis of complex biological datasets such as those composed of clinical, omics, biochemical, and environmental data. For the identification of early-life predictive and prognostic biomarkers in Obesity and T2D, I develop new analysis pipelines and implement existing algorithms able to handle complex omics data (genetics, epigenetics, transcriptomic...). My main technical skills include a strong statistical, programming and data visualization background, with special emphasis in the use of machine learning models. Summary at: https://youtu.be/dCYZ6xY69Z4

WORK EXPERIENCE

01/10/2020 – 01/02/2021 – Paris, France
BIOINFORMATICS SCIENTIST - VISITING PREDOCTORAL RESEARCHER (PARIS, FRANCE) – CLINICAL BIOINFORMATICS LAB - IMAGINE INSTITUTE - NECKER HOSPITAL "LES ENFANTS MALADES"

Public Spanish Government Grant nº MV19/00062 (M-AES modality)

01/01/2018 – CURRENT – Granada, Spain
BIOINFORMATICS SCIENTIST - PREDOCTORAL RESEARCHER (GRANADA, SPAIN) – INSTITUTE OF HEALTH CARLOS III & UNIVERSITY OF GRANADA

Public Spanish Government Grant nº IFI17/00048 (i-PFIS modality)

Thesis title: "**MULTI-OMICS INTEGRATION AND MACHINE LEARNING FOR THE IDENTIFICATION OF MOLECULAR MARKERS OF INSULIN RESISTANCE IN PREPUBERTAL AND PUBERTAL CHILDREN WITH OBESITY**"

GENERAL OBJECTIVES:

1. Get a deeper understanding of the underlying genetics and epigenetics architecture of obesity and its metabolic comorbidities. Identification of novel predictive biomarkers of insulin resistance in children.
2. Ascertain the influence of environmental and lifestyle factors on the omics-conferred risk for obesity and its complications in children.
3. Construction of Artificial Intelligence models integrating environmental and omics data for the identification of children presenting a high risk to develop obesity and insulin resistance during adulthood.

METHODS:

Massive genetics and epigenetics data processing and deep analysis; Statistical analysis; Machine Learning; eXplainable Artificial Intelligence.

Summary at: https://youtu.be/dCYZ6xY69Z4

01/01/2014 – 01/06/2017
CHIEF EXECUTIVE OFFICER AND CO-FOUNDER – NOVGEN SL

University of Granada (UGR) Spin-off; Start-up specialized in ***personalized medicine, health 2.0 and genetic counselling***.

Novgen is a start-up from the University of Granada specialized in personalized medicine and genetic counselling. Novgen has created a bioinformatics tool called "Novgen-Line" which works as a platform and e-health app for clinicians and health professionals. Novgen-Line provides doctors and patients with guidance in requesting and processing genetics analysis results, in the manner a geneticist would manage the process. Novgen-Line, therefore, promotes and enhances the translation of personalized medicine into everyday healthcare, providing both clinicians and patients with rigorous results and high-quality protocols that can improve not only doctors' services but also patients' health.

During my time as Chief executive officer in this company, Novgen has also developed a predictive genetic algorithm which, based on Kaspar technology, offers analysis services in the field of nutrigenomics. We have identified associations and interactions between a large number of polymorphisms affecting uptake, transport, metabolism and individual daily requirements for some essential nutrients. As a result, Novgen has developed a new algorithm for the study of how human genetic variation modifies our response to diet components, including micro-, macro-nutrients and toxins.

Business Achievements as CEO and co-founder of this company:

- **Winner of "Somos Empresa - We are enterprise" prize in the category "Somos Futuro - We are future" Emisor: Popular and PRISA Group (October 2015):** This prize recognizes Novgen as the young Company with the most potential in Spain. https://www.youtube.com/watch?v=f_yoxOP7SUk

- **Best project in the first edition of the Health-U Program by Sanofi Aventis. Emisor: Sanofi Aventis (October 2015):** Novgen was chosen to take part in the "Health-U" program organized by the pharmaceutical company Sanofi. This program consisted of a Startup Week which took place at Sanofi offices (Barcelona, October 2015). We shared experiences and activities with other five startups and we received formation and advising in the field of healthcare and Health 2.0. After exposing a pitch in front of a jury, Sanofi workers and managers, Novgen received the award of being the best project. https://youtu.be/EvHNfFiMMgo

- **Winner of the 5th Entrepreneurship Competition. Emisor: University of Granada (November 2015):** https://www.youtube.com/watch?v=TVdQXtMPO18

EDUCATION AND TRAINING

01/2015 – 2016 – Granada, Spain
MASTER OF SCIENCE "TRASLATIONAL RESEARCH AND PERSONALIZED MEDICINE" – University of Granada

EQF level 7

2010 – 2014 – Granada, Spain
BACHELOR'S DEGREE "BIOCHEMISTRY AND MOLECULAR BIOLOGY" – University of Granada

EQF level 6

14/02/2020 – 20/03/2020 – Granada, Spain
COURSE "PYTHON APPLIED TO THE SCIENTIFIC AND TECHNOLOGIC RESEARCH" – University of Granada

2019 – 2019 – Cambridge, United Kingdom
WINTER SCHOOL "5TH INTERNATIONAL WINTER SCHOOL ON BIG DATA" – University of Cambridge

2018 – 2018 – Granada, Spain
COURSE "ADVANCED R PROGRAMMING AND DATA ANALYSIS" – Doctoral School of Sciences, Technologies, and Engineering of the University of Granada

2018 – 2018
COURSE "STATISTICAL TECHNIQUES APPLIED IN THE FIELD OF NUTRITION AND HEALTH" – Doctoral School of Health Sciences - University of Granada

2017 – 2017 – Granada, Spain
COURSE "DATA SCIENCE IN THE BIG DATA ERA" – University of Granada

2013 – 2014 – Baltimore, United States
COURSE "INTRODUCTORY COURSE TO BIOINFORMATICS" – The Johns Hopkins University

2013 – 2013 – Washingtong, dc, United States
COURSE "INTRODUCTORY COURSE TO GENETICS AND GENOMICS" – Georgetown University

LANGUAGE SKILLS

Mother tongue(s): SPANISH

Other language(s):

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken production	Spoken interaction	
ENGLISH	C1	C2	C1	C1	C2
FRENCH	A2	B1	A1	A1	A1

Levels: A1 and A2: Basic user; B1 and B2: Independent user; C1 and C2: Proficient user

PUBLICATIONS

Selected Peer-Reviewed Publications

Anguita-Ruiz A, et al. Omics Approaches in Adipose Tissue and Skeletal Muscle Addressing the Role of Extracellular Matrix in Obesity and Metabolic Dysfunction. ***Int J Mol Sci***. 2021 Mar;22(5).

Llorente-Cantarero FJ,..., **Anguita-Ruiz A**, et al. Relationship between Physical Activity, Oxidative Stress, and Total Plasma Antioxidant Capacity in Spanish Children from the GENOBOX Study. ***Antioxidants*** (Basel, Switzerland). 2021 Feb; 10(2).

Gomez-Llorente MA,..., **Anguita-Ruiz A**, et al. A Multi-Omics Approach Reveals New Signatures in Obese Allergic Asthmatic Children. ***Biomedicines***. 2020 Sep;8(9).

Anguita-Ruiz A, Aguilera CM, Gil Á. Genetics of Lactose Intolerance: An Updated Review and Online Interactive World Maps of Phenotype and Genotype Frequencies. ***Nutrients***. 2020 Sep;12(9).

Rupérez AI, Mesa MD, **Anguita-Ruiz A**, et al. Antioxidants and Oxidative Stress in Children: Influence of Puberty and Metabolically Unhealthy Status. ***Antioxidants*** (Basel, Switzerland). 2020 Jul;9(7).

Anguita-Ruiz A, et al. Evaluation of the Predictive Ability, Environmental Regulation and Pharmacogenetics Utility of a BMI-Predisposing Genetic Risk Score during Childhood and Puberty. ***J Clin Med***. 2020 Jun;9(6).

Anguita-Ruiz A, Segura-Delgado A, Alcalá R, Aguilera CM, Alcalá-Fdez J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. ***PLoS Comput Biol***. 2020 Apr;16(4):e1007792.

Anguita-Ruiz A, et al. The protein S100A4 as a novel marker of insulin resistance in prepubertal and pubertal children with obesity. ***Metabolism***. 2020 Apr 1;105.

Leis R,..., **Anguita-Ruiz A**, et al. Cluster analysis of physical activity patterns, and relationship with sedentary behavior and healthy lifestyles in prepubertal children: Genobox cohort. ***Nutrients***. 2020;12(5).

Anguita-Ruiz A, et al. Common Variants in 22 Genes Regulate Response to Metformin Intervention in Children with Obesity: A Pharmacogenetic Study of a Randomized Controlled Trial. ***J Clin Med***. 2019 Sep;8(9).

Anguita-Ruiz A, et al. X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline. ***Sci data***. 2019 Jul;6(1):130.

Soto-Méndez MJ,..., **Anguita-Ruiz A**, et al. Role of Functional Fortified Dairy Products in Cardiometabolic Health: A Systematic Review and Meta-analyses of Randomized Clinical Trials. ***Adv Nutr***. 2019 May;10(suppl_2):S251–71.

Ruiz-Ojeda FJ, **Anguita-Ruiz A**, et al. Effects of X-chromosome Tenomodulin Genetic Variants on Obesity in a Children's Cohort and Implications of the Gene in Adipocyte Metabolism. ***Sci Rep***. 2019 Mar;9(1):3979.

Selected Book Chapter Publications

Anguita-Ruiz A; Segura-Delgado A; Alcala R; Aguilera CM; Alcalá-Fernández J. Describing sequential association patterns from longitudinal microarray data sets in humans. ***Bioinformatics and Biomedical Engineering***. pp. 17 - 40. Springer, 13/04/2019. ISBN 978-3-030-17935-9 DOI: https://doi.org/10.1007/978-3-030-17935-9_29

Rupérez-Cano A; **Anguita-Ruiz A**. Genetics of Oxidative Stress and Obesity-Related Diseases.Obesity. ***Oxidative Stress and Dietary Antioxidants***. pp. 17 - 40. Elsevier, 01/10/2018. ISBN 9780128125045 DOI: https://doi.org/10.1016/B978-0-12-812504-5.00002-7

PROJECTS

Participation in R&D and Innovation projects

PI20/00711., Enfoque Machine Learning y Big Data Multi-Ómico: Medicina Personalizada. Instituto de Salud Carlos III. Proyectos de Investigación en Salud de la convocatoria 2020 de la Acción Estratégica en Salud 2013-2016. Jesús Alcalá Fernández. (Universidad de Granada). 01/01/2021-31/12/2022. 27.830 €. Collaborator.

PI20/00563., Omicas e Inteligencia Artificial. Herramientas para entender los mecanismos moleculares de la resistencia a la insulina en niños obesos durante la pubertad. PI20/00563 Instituto de Salud Carlos III. Proyectos de Investigación en Salud de la convocatoria 2020 de la Acción Estratégica en Salud 2013-2016. Concepción María Aguilera. (Universidad de Granada). 01/01/2021-31/12/2022. 173.030 €. Collaborator.

P18-RT-2248., Explicabilidad de la Inteligencia Artificial para el Análisis Inteligente de Datos: Aplicaciones en Problemas de BioSalud y del Internet de las Cosas. P18-RT-2248. Junta de Andalucía. Ayudas a proyectos de I+D+I Programa Operativo FEDER 2014-2020 (PAIDI 2020) - Convocatoria 2018 (BOJA n.º 203, 18/10/2018). Jesús Alcalá Fernández. (Universidad de Granada). 01/01/2020-31/12/2022. 35.542 €. Team member.

PY18-4455., Transductores Moleculares del Ejercicio Físico y la Activación del Tejido Adiposo Pardo: en Busca de Nuevas Dianas Terapéuticas en la Comunicación Intercelular. PY18-4455. Junta de Andalucía. Ayudas a proyectos de I+D+I Programa Operativo FEDER 2014-2020 (PAIDI 2020) - Convocatoria 2018 (BOJA n.º 203, 18/10/2018). Concepción María Aguilera. (Universidad de Granada). 11/02/2020- 31/12/2021. 140.352 €. Collaborator.

PI18/00930, Mediterranean lifestyle in pediatric obesit prevention: MELI-POP. PI18/00930. Instituto de Salud Carlos III. Proyectos de Investigación en Salud de la convocatoria 2018 de la Acción Estratégica en Salud 2013-2016. Mercedes Gil Campos. (Instituto Maimónides de Investigación Biomédica de Córdoba). 01/01/2019-31/12/2021. 64.130 €. Team member.

PI16/00871, Puberty and metabolic risk in obese children. Epigenetic alterations and pathophysiological and diagnostic implications. PUBMEP Study. PI16/00871. Instituto de Salud Carlos III. Proyectos de Investigación en Salud de la convocatoria 2016 de la Acción Estratégica en Salud 2013-2016. Concepción María Aguilera García. (University of Granada). 01/01/2017-31/12/2020. 99.220 €. Team member.

Integration of omics and environmental data to develop artificial intelligence-based predictive tools for the early-life prevention of adulthood obesity-related chronic diseases Fundación Mapfre. AYUDAS A LA INVESTIGACIÓN IGNACIO H. DE LARRAMENDI en su edición 2017. Concepción Aguilera García. (University of Granada). 14/02/2018-18/04/2019. 35.000 €. Team member.

COMMUNICATION AND INTERPERSONAL SKILLS

Selected Conference Presentations

My academic and business careers have allowed me to demonstrate good oral communication skills in more than 8 International conferences (e.g. https://youtu.be/l0LQ9di2YsQ)

- Epigenetic changes in key metabolic genes accompany obesity and insulin-resistance trajectories during puberty. **European and international congress on obesity ECOICO**. 02/09/2020. (**Online**).
- A poor diet quality during puberty is able to induce epigenetic changes in key genes participating in pathogenic pathways of obesity and IR. ***FINUT 2020 International Virtual Conference.*** 11/10/2020. (**Online**).
- Towards a novel marker of insulin resistance in obesity: S100A4 in girls before and after pubertal onset — Evidences from the longitudinal study "PUBMEP" –. ***FENS 2019 ,13th European Nutrition Conference***. 15/10/2019. (**D ublin, Irlanda**).
- X chromosome genetic data in a Spanish children cohort: dataset analysis and pipeline. *"Better Science Through Better Data",* ***an international event organized by Springer Nature and the Welcome trust***. 15/10/2019. (**London, England**).
- Describing sequential association patterns from longitudinal microarray data sets in humans. ***7th International Work-Conference on Bioinformatics and Biomedical Engineering.*** 08/05/2019. (**Granada, Spain**).
- Association of an obesity-predisposing genetic risk score with a set of metabolic and inflammatory traits in a cohort of Spanish children. ***17th Conference of the Spanish Nutrition Society (SEÑ) and 10th Meeting of the Catalan Association of Food Science (ACCA)***. 27/06/2018. (**Barcelona, Spain**).
- Tenomodulin genetic variants on the X chromosome are associated with childhood obesity. ***IUNS-INC 21st International Congress of Nutrition*** *"From sciences to nutrition security"*. 15/10/2017. (**Buenos Aires, Argentina**).
- Genetic Polymorphism of PPAR Gamma modified the effects of metformin on BMI z-score in obese children. ***IUNS -INC 21st International Congress of Nutrition*** *"From sciences to nutrition security"*. 15/10/2017. (**Buenos Aires, Argentina**).

JOB-RELATED SKILLS

Computer Skills

Statistical language: R
Programming language: Python
Database Management System : R, Excel
Operating system: Unix/Linux shell
Data Science Skills (using R)

- Machine learning: Association Rule Mining, Sequential Rule Mining, linear regression, logistic regression, decision tree, random forest,Naive Bayes, k-means, PCA and k-nn.
- Data visualization: Shiny, plotly, ggplot, leaflet
- Data manipulation: dplyr, tidyr
- Text Mining: tm, rtweet, tidytext

ORGANISATIONAL SKILLS

Member of the organizing committee: V Meeting on Bioinformatics.

Affiliation entity: University of Granada City affiliation entity: GRANADA, Andalusia, Spain Start-End date: 24/02/2019 - 26/02/2019

Member of the organizing committee: IV Meeting on Bioinformatics. ISBN: 978-84-09-16565-0

Affiliation entity: University of Granada City affiliation entity: GRANADA, Andalusia, Spain Start-End date: 14/02/2020 - 15/02/2020

Member of the organizing committee: III Meeting on Bioinformatics. ISBN: 978-84-09-09196-6

Affiliation entity: University of Granada City affiliation entity: GRANADA, Andalusia, Spain Start-End date: 14/02/2019 - 05/02/2019

Member of the organizing committee: 1st Edition MOOC "Machine Learning y Big Data para la Bioinformática"

Organising entity: Universidad de Granada
Type of entity: University
Hours of teaching: 112
Teaching date: 22/02/2021

Acknowledgements

Resulta para mi imposible concentrar en apenas unas páginas el reconocimiento a todas esas personas que me han apoyado durante los años que me ha llevado terminar este trabajo.

Entre todas esas personas, si tengo que empezar por alguien, sin duda lo haría por la profesora Concepción Aguilera (**Chiqui**), mi directora, maestra, compañera, amiga, ¡Y hasta business angel!. GRACIAS, GRACIAS y GRACIAS, por tantas cosas. Nuestra relación comienza años antes de empezar el doctorado, cuando un compañero de la carrera y yo acudimos a ella, en busca de apoyo para iniciar una aventura empresarial con una todavía inmadura idea de negocio basada en la aplicación de la genética y la medicina personalizada. En ese mismo momento no lo dudó, y nos brindó todo su apoyo. Y es que, si algo la caracteriza, y así me lo ha enseñado durante todos estos años, es el dar a todo el mundo la oportunidad de demostrar lo que vale. Una oportunidad que no solo me permitió desarrollar un proyecto empresarial con el que aprendí infinidad de cosas, sino que también me abrió las puertas al mundo de la investigación biomédica. Un mundo al que nunca me había planteado dedicarme seriamente, pero que me cautivó desde un principio, despertando en mí una pasión desmedida por la genética humana y la bioinformática. Una pasión que Chiqui ha cultivado y animado cuidadosamente durante estos años; creyendo en mí, motivándome con su ejemplo, otorgándome responsabilidades dentro del grupo que me han hecho crecer como profesional, y procurándome un crecimiento académico y formativo excelente. ¡Por todo ello, gracias!. Además, más allá de lo profesional, también me gustaría agradecer su trato, su cariño, con el que ha conseguido convertirse en una persona importante para mí. En ella he encontrado una amiga a la que contarle mis preocupaciones e inquietudes, recibiendo siempre un consejo sincero y una simpatía inmejorables. Chiqui es una excelente científica, profesora y comunicadora, con una habilidad impresionante para la gestión de equipos, de la que he aprendido MUCHO, y con la que he trabajado inmensamente feliz durante el desarrollo de mi tesis. Por ello, y aunque se quede corto, con estas líneas quiero agradecértelo de corazón. Igualmente, espero que, de alguna forma u otra, me sigas acompañando y enseñando tanto en mi vida científica como personal durante muchos años más.

Si hay otro gran responsable de que me encuentre aquí hoy, ese es el profesor **Ángel Gil**, líder de mi grupo de investigación, y el primero en apoyar, junto con Chiqui, el proyecto empresarial que finalmente derivó en el desarrollo de mi tesis doctoral. Ángel es sin duda, una de las mentes más maravillosas que he conocido, con una capacidad de hacer BUENA ciencia incomparable. Aunque por el momento en el que he llegado al grupo de investigación, en el que Ángel se encuentra

culminando su carrera científica, no he tenido la suerte de poder compartir tantos momentos con él cómo los que me hubiera gustado, para mí ha sido un verdadero orgullo contar con su sabiduría y experiencia siempre que la he necesitado. Gracias por tus invalorables lecciones de bioestadística y bioquímica, que me acompañarán durante toda mi carrera. Gracias también por tus consejos y por hacer todo lo posible para que el grupo de investigación siga creciendo.

Mil millones de gracias también al profesor **Jesús Alcalá**, mi co-director de tesis, en quien he encontrado un inmejorable maestro, consejero y ejemplo a seguir. Gracias a él, me he formado en el apasionante mundo de la inteligencia artificial y he aprendido a aplicar adecuadamente técnicas de análisis de datos avanzadas para resolver complejos problemas biológicos. Ha sido un camino de aprendizaje que he disfrutado inmensamente gracias a su enorme capacidad de trabajo e instinto científico. Y es que, en Jesús he encontrado un excelente mentor, capaz de transmitir de forma sencilla complejos conceptos informáticos y matemáticos. Jesús, eres en gran parte responsable del bioinformático en el que me he convertido y por eso te doy las gracias. Te agradezco también por preocuparte por mi crecimiento profesional y mi futuro académico, dándome siempre los mejores consejos, permitiéndome participar en todos tus proyectos, y procurándome una formación inmejorable (cursos, estancias, etc.). Para mí ha sido un verdadero lujo contar con tu apoyo durante estos tres años intensos y he aprendido muchísimo. Gracias también por tu entusiasmo y motivación, que son realmente contagiosos para acometer cualquier proyecto de investigación, por complicado que pueda ser. En definitiva, gracias por regalarme algo tan valioso, tu tiempo. Igualmente, me gustaría agradecer a mi "co-director en la sombra", el profesor **Rafael Alcalá**, otro científico y mente maravillosa, quien con su siempre desinteresada disponibilidad y brillantes ideas ha sido una fuente de lecciones magistrales en las múltiples e interminables reuniones de trabajo en las que planteábamos posibles metodologías de análisis. ¡GRACIAS de corazón a los dos!

Me gustaría agradecer también a todos mis compañeros, quienes han hecho de mi día a día un verdadero placer. Empezando cronológicamente, gracias a **Belén Pastor** y a **Fran Ruiz** pues fuisteis mis primeros compañeros de trabajo y ejemplos a seguir. Gracias por vuestra disponibilidad para resolver dudas, vuestro apoyo en la gestión de muestras en los congeladores del laboratorio, y por contarme de primera mano todos los entresijos del mundo de la investigación en el que me metía. Gracias por esos congresos y viajes en los que me he divertido tanto como aprendido. Ha sido un lujo teneros como ejemplo de doctorandos en el laboratorio, y para mí ha sido una alegría ver como vuestras carreras investigadoras se han proyectado con tanto éxito. Gracias a **Julio**, quien, aunque siempre he visto como un investigador senior, se ha empeñado en ser uno más. Un compañero más del que aprender, tanto en el aspecto profesional como personal. Gracias por tu generosidad, por tu amabilidad, por tus bromas y tus sabios consejos. Sin duda tienes la capacidad para unir al grupo de investigación y de transmitir siempre tu sonrisa. Espero que en un futuro volvamos a compartir laboratorio y aventuras de congresos. Gracias a **Josune**,

por dedicarme tanto tiempo al principio de mi doctorado, y por transmitirme tu conocimiento, siempre en un tono dulce y amable. Gracias por supuesto a **M Cruz**, la alegría del laboratorio. Eres fuente inagotable de simpatía y ánimos. Gracias por obligarme a despegarme del ordenador y el trabajo, y por preocuparte de que desayunara cada día. Gracias por hacer de consejera y compartir tantas experiencias conmigo. Ha sido una suerte tenerte por allí tanto tiempo. Gracias a **Andrea**, ¡camarada doctoranda de fatigas! Hemos sido compañeros de despacho, congresos, docencia y preocupaciones, que se hacían más llevaderas al compartirlas. Gracias a todo el grupo de Marga, en quienes he encontrado compañeros tan buenos como los de mi propio grupo. A todo el equipo INYTA, **Carolina**, **MD**, **Marga** y **Jesús**, gracias por los "gallineros" (como dice Chiqui), que llenaban de risa y buen ambiente el laboratorio. Gracias por todos los desayunos y momentos de desconexión.

Gracias a todos los miembros del Departamento de Bioquímica y Biología Molecular II de la Facultad de Farmacia, en el que he tenido la envidiable oportunidad de impartir docencia. Gracias a **Luis**, quien además de haber sido un excelente jefe de grupo de investigación, ha sido un verdadero ejemplo docente en el que fijarme y apoyarme. Debo confesar que las prácticas de Patología son las que más me ha gustado enseñar de todas las asignaturas que he impartido durante el doctorado. Gracias a **M Carmen** y a **José Manuel**, que hacen que uno se sienta en el departamento como en casa. Y por supuesto, gracias al resto del equipo docente, **Paloma**, **Mercedes**, **Marina**, etc., de quienes he aprendido mucho y por darme la oportunidad de formar parte del mismo durante estos años. También a **Olga** y a **Fermín**, mis otros *business Angels*, a quienes estaré siempre agradecido por el apoyo inicial en la creación del proyecto Novgen.

Otros responsables directos de que esta tesis se haya podido realizar con éxito son el magnífico equipo CIBERobn-PUBMEP formado por los clínicos e investigadores del centro de investigación IMIBIC y los hospitales de Santiago de Compostela, Reina Sofía y Lozano Blesa (**Rosaura**, **Rocío**, **Gloria**, **Azahara**, **Esther**, **Luis Moreno**, **Mercedes**, **Katy**, **Fran**, **Juan Roa**, etc.). Gracias a todos por vuestro inestimable trabajo, sin el que no sería posible llevar a cabo este tipo de investigación. Gracias por vuestra profesionalidad y por vuestro entusiasmo, pues no es fácil compaginar las tareas asistenciales con todo el trabajo de investigación. Gracias igualmente por los buenos momentos en congresos y reuniones, ha sido un placer trabajar con vosotros.

Fuera del ámbito profesional, me gustaría agradecer también este logro a mi familia y amigos, quienes con su apoyo continuo me han dado las fuerzas necesarias para conseguirlo. Gracias a **Abel**, por los momentos de desconexión en bici y por tragarte charlas interminables sobre lo que iban mis investigaciones, a pesar de no entender ni la mitad, ¡y encima poniendo cara de que sí, y que interesaba!. Gracias a **Paco**, por tantas cosas, desde enseñarme una nueva manera de pensar y de hacer ciencia, hasta todas esas conversaciones filosóficas interminables. Gracias por acogerme en París como a un hermano, y por ser un excelente amigo a la vez que colega de profesión. Estoy seguro de que vas a ser (ya lo eres) un científico con una carrera brillante.

GRACIAS a **Inés**, mi compañera de camino, por aparecer en mi vida y llenarla de felicidad. ¡Eres fuente inagotable de energía y ánimo!. Gracias por ser una de las personas con más voluntad que conozco, y todo un ejemplo en el que fijarme. Durante toda esta etapa ha sido una SUERTE para mí el contar contigo; has sido la primera en escuchar mis presentaciones de congreso (siempre atendiendo con tu mejor cara de entusiasmo), has sido editora de revista científica revisando mis publicaciones y figuras en busca de erratas, a la vez que la mejor interlocutora científica cuando necesitaba exponer en voz alta mis ideas… Has sido eso, y MIL cosas más. Gracias por tu cariño, por tu comprensión, y por hacerme el mejor regalo que se le puede hacer a una persona, creer en mi incondicionalmente, que ha sido el mejor combustible para lograr metas que parecían imposibles. Gracias por ser TU y por hacer el camino tan sencillo y bonito.

GRACIAS a mis padres, mis héroes, **María José** y **Augusto**, por TODO. Gracias a mi padre por su empeño en hacer de mi loca cabeza una mente racional desde pequeño; por inculcarme la importancia de PENSAR y ANALIZAR, pues ha acabado siendo una de las mejores herramientas para desarrollar mi trabajo. Gracias también por ser un ejemplo de superación continuo y valentía. Gracias a mi madre por su continuo sacrificio, por hacerme entender desde pequeño la importancia del estudio y del esfuerzo, y por ser mi mayor confidente. Gracias por ser un ejemplo a seguir de trabajadora nata, valor y coraje. Gracias a los dos por darme las mejores condiciones para desarrollar mis ideas e inquietudes. Gracias por vuestro interés y pasión por mi trabajo, y por los momentos de felicidad que me regaláis cada día. GRACIAS por creer en mí. Sois los MEJORES padres del mundo, y sin vosotros nada de esto habría sido posible.

PROGRAMA DE DOCTORADO EN NUTRICION Y CIENCIAS DE LOS ALIMENTOS

Multi-Omics integration and machine learning for the identification of molecular markers of insulin resistance in prepubertal and pubertal children with obesity

AUGUSTO MIGUEL ANGUITA RUIZ