

The HUA BioResource: A National Infrastructure for Genetic and Phenotypic Insights into BMI and Obesity in Greece

Panagiotis G. Symianakis, Ioanna Panagiota Kalafati, Charalampia Amerikanou, Despoina Brekou, Maria Dimitriou, Aliko-Eleni Farmaki, Olga Giannakopoulou, Panagiota Giardoglou, Effimia Grigoriou, Maria Kafyra, Stavroula Kanoni, Vassiliki-Eirini Kariakli, Efthymia A. Katsareli, Eirini Marouli, Ioanna Ntalla, Natia A. Pogossova, Eirini V. Theodoraki, Maria Tsota, Andreas Alexandrou, Anastasia Amfilochiou, Erifili Hatzigelaki, Andriana C Kaliora, Alexander Kokkinos, Genovefa Kolovou, Loukianos S. Rallidis, Kimon Stamatelopoulos, Nikolaos Tentolouris, Anastasia Thanopoulou, Konstantinos Triantafyllou, George Trovas, Antigone S. Dimas, Vasiliki Mamakou, Konstantinos Rouskas, Panos Deloukas, Eleftheria Zeggini, George V. Dedoussis

P. G. Symianakis, I.P. Kalafati, C. Amerikanou, D. Brekou, A-E. Farmaki, P. Giardoglou, E. Grigoriou, M. Kafyra, AC Kaliora, V-E Kariakli, E. A. Katsareli, I. Ntalla, N. A. Pogossova, E. V. Theodoraki, M. Tsota, and G. V. Dedoussis are with the Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, 17671 Athens, Greece

M. Dimitriou is with the Department of Nutritional Science and Dietetics, School of Health Sciences, University of the Peloponnese

O. Giannakopoulou is with the William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, EC1M 6BQ

S. Kanoni is with the Clinical Pharmacology and Precision Medicine, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, UK

E. Marouli is with the William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK

A. Alexandrou is with the First Department of Surgery, National and Kapodistrian University of Athens, "Laikon" General Hospital, 115 27 Athens, Greece

A. Amfilochiou is with the Sleep Unit, Sismanoglio General Hospital of Attica, Region of Athens, Athens, Greece

E. Hatzigelaki is with the Medical School, National and Kapodistrian University of Athens, Greece and the School of Medicine, European University Cyprus

A. Kokkinos is with the First Department of Propaedeutic Internal Medicine and Diabetes Center, Laiko General Hospital, Medical School, National and Kapodistrian University of Athens, 11527 Athens, Greece

G. Kolovou is with the Metropolitan Hospital, Department of Preventive Cardiology, 9, Ethn. Makariou & 1, El. Venizelou, N. Faliro, 185 47, Athens, Greece

L.S. Rallidis is with the Second Department of Cardiology, School of Medicine, National and Kapodistrian University of Athens, Attikon University General Hospital, Athens, Greece

K. Stamatelopoulos is with the Department of Clinical Therapeutics, National and Kapodistrian University of Athens

N. Tentolouris is with the First Department of Propaedeutic Internal Medicine, National and Kapodistrian University of Athens, Medical School, Laiko General Hospital, Athens, Greece

A. Thanopoulou is with the Diabetes Center, Second Department of Internal Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece

K. Triantafyllou is with the Hepatogastroenterology Unit, 2nd Department of Internal Medicine, Propaedeutic, Medical School, National and Kapodistrian University of Athens, "Attikon" University General Hospital, Athens, Greece.

G. Trovas is with the Th. Garofalidis Laboratory for Research of the Musculoskeletal System, Medical School, National and Kapodistrian University of Athens, Athens, Greece

A.S. Dimas is with the Institute for Bioinnovation, Biomedical Sciences Research Center 'Alexander Fleming', Fleming 34, 16672, Vari, Greece

Abstract—The HUA BioResource (HUABR) is one of the first large-scale multicohort resources in Greece, integrating genetic and phenotypic data across 16 national studies. This paper presents the first comprehensive description of the HUA BioResource (HUABR), outlining its structure, data collection methodology, and research potential. The repository serves as a critical resource for investigating the interplay between genetics, lifestyle, and body mass index (BMI), enabling advanced analyses of obesity and obesity-related cardiometabolic disorders. Initial findings highlight significant associations between BMI and genetic predisposition as well as lifestyle, contextual and clinical factors, with polygenic risk scores (PRSs) enhancing obesity prediction models. By creating one of the first databases with unified, multidisciplinary data for Greek citizens, the HUA Bioresource fosters international collaborations and advances precision medicine, public health, and targeted prevention strategies.

Index Terms—HUABR, HUA BioResource, Genetics, PRS, Body mass index, BMI, Diet, Lifestyle, Personalized medicine, Greek

I. INTRODUCTION

IN recent years, biobanks have emerged as crucial resources for advancing research in genetics, biology and personalized counselling in both nutrition and medicine. They are repositories which collect, store, and manage biological

V. Mamakou is with the Department of Internal Medicine, Dromokaiteio Psychiatric Hospital, Athens, Greece

K. Rouskas is with the Institute of Applied Biosciences, Center for Research and Technology Hellas, 57001 Thessaloniki, Greece

P. Deloukas is with the William Harvey Research Institute & NIHR Barts Biomedical Research Centre, Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK, the Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University, Dammam 34212, Saudi Arabia and the Greek Institute of Human Genomics, Foundation for Research and Technology – Hellas, Athens, Greece

E. Zeggini is with the Institute of Translational Genomics, Computational Health Center, Helmholtz Zentrum München – German Research Center for Environmental Health, 85764 Neuherberg, Germany and the Technical University of Munich (TUM), TUM University Hospital, TUM School of Medicine and Health, 81675 Munich, Germany

samples, such as blood, saliva, tissues, cells, or DNA/RNA, along with associated health, environmental, and lifestyle data from volunteers. The present work introduces the construction of one of the first comprehensive repositories in Greece, a pioneering initiative that performs a systematic, retrospective harmonization of 16 disparate national cohorts, collectively encompassing over 13,900 participants. Harokopio University of Athens BioResource (HUABR) represents a significant milestone for biomedical research in the country, providing an unprecedented resource for investigating genetic factors, as well as phenotypic characteristics (i.e. socio-demographic data, anthropometric data, medical history, lifestyle data and deep phenotyping through medical imaging) and their role in health and disease. The database includes an extensive catalogue of genetic data, such as Single Nucleotide Polymorphisms (SNPs), coupled with rich phenotypic information, covering a wide range of health metrics, lifestyle behaviours, and other environmental exposures. This holistic approach positions the infrastructure as an invaluable tool for exploring the complex interplay between genetic variants and phenotypic traits, shedding light on the underlying biological and environmental mechanisms that drive complex diseases, thereby facilitating earlier prevention and more effective interventions. The herein presented bioresource is among the first of its kind in Greece, addressing a critical gap in the nation's research infrastructure, as it enables researchers to assess population-specific genetic and environmental influences on health. Moreover, it aligns Greece with global biobanking initiatives, opening opportunities for international collaborations and comparative studies. By leveraging this robust resource, the HUABR not only fosters advancements in personalized medicine and nutrition but also paves the way for targeted prevention strategies and therapeutic innovations tailored to the unique characteristics of the Greek population. This paper marks the first comprehensive description of the dataset, offering an in-depth overview of its scope, structure, and potential applications in future research.

Harokopio University of Athens (HUA), in order to establish its sources, has started collecting the data since 2006 and its records are constantly updated. Several researchers have participated in this effort through sixteen individual studies which are:

- | | |
|--------------------------|------------------------|
| 1) ATLAS [1] | [11] |
| 2) GENDAI [2] | 10) iMPROVE [12]–[14] |
| 3) GOMAP [3] | 11) NAFLD [15]–[18] |
| 4) Greek Metabo [4], [5] | 12) OSTEOS [19] |
| 5) Greek NASH-PI [6] | 13) SLEEPSTUDY |
| 6) GRMIC [7] | 14) TEENAGE [20], [21] |
| 7) HELIC-MANOLIS [8] | 15) THISEAS [22], [23] |
| 8) HELIC-POMAK [9] | 16) WATERPOLO |
| 9) HYPER MASTIC [10], | |

The primary novelty of the HUABR lies in the transformation of these 16 asynchronous and heterogeneous datasets into a single, searchable, and dynamic bioresource. By standardizing diverse phenotypic definitions and mapping genetic data onto a common analytical pipeline, we have enabled

high-powered, cross-cohort analyses that were previously impossible due to the temporal and methodological differences between the original studies. Beyond this technical integration, a core contribution of the HUABR is the establishment of a robust sample of individuals from diverse socioeconomic strata and geographic regions across Greece. While the repository does not strictly aim for a perfectly representative snapshot of the general population, it provides a high-quality framework that is constantly updated to facilitate longitudinal research. This allows for the continuous monitoring of health evolution and disease progression over time, transitioning the infrastructure from a static collection of data into a living laboratory for the study of gene-environment interactions in the Greek population. Among these, obesity stands out as a critical and growing public health concern in Greece and globally. According to the World Health Organization (WHO), in 2022, 2.5 billion adults worldwide were classified as overweight, of which 890 million were living with obesity [24]. This growing epidemic underscores the urgent need for comprehensive research resources to better understand its causes, consequences, and potential interventions. HUABR provides unique opportunities for advancing obesity research, since it is designed to support detailed investigations into the genetic, health, and lifestyle factors contributing to complex diseases, including obesity. The aim of this paper is to introduce the HUABR by describing its sample population and data collection methodology, while also presenting the resource's initial findings on obesity using metrics such as Body Mass Index (BMI) to support population-level analyses and ultimately inform targeted healthcare strategies. Additionally, the transferability of Polygenic Risk Scores (PRS) for BMI prediction evaluated in the population.

II. RESOURCE CONSTRUCTION

Figure 1 provides a schematic overview of the HUABR, from sample and data acquisition through to downstream analytics and clinical applications. The collection process draws on a diverse set of subgroups of Greek populations, to ensure a comprehensive and representative dataset. The main focus of the research teams was on gathering biological samples and health-related data from individuals across four population groups: the pooled cohorts, isolated populations, children and athletes. This diversity of sources provides an opportunity for broad-based research, including the study of common as well as rare genetic variants, health conditions, and it lays a solid foundation for future studies into gene-environment interactions.

Figure 2 represents the origin of the volunteers participating in HUABR separated in the 13 regions of Greece. The establishment of a Bioresource comprising genetic samples from volunteers across the entirety of Greece serves as a valuable and comprehensive resource for future genetic research. HUABR reflects the genetic heterogeneity of the Greek population, enabling researchers to investigate population-specific genetic traits, rare variants, and disease susceptibility with greater precision.

The Bioethics Committee of Harokopio University of Athens

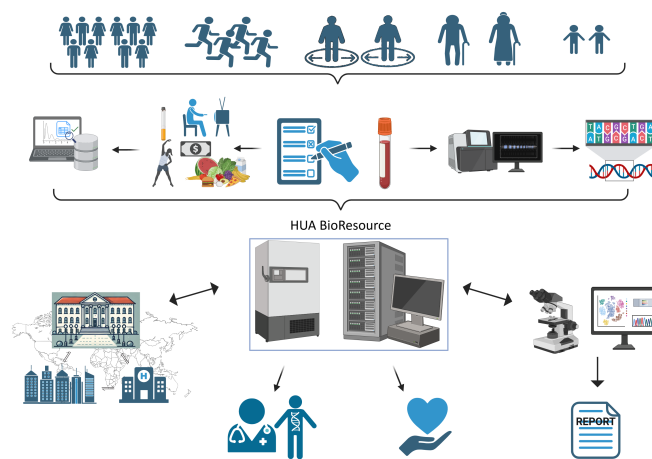


Fig. 1: Overview of the HUA BioResource. Created in Biorender.

The infrastructure acquires samples from a diverse demographic spectrum, including the pooled cohorts, athletes, isolated populations, elderly individuals, and children. The deep data collection encompasses biochemical profiles, anthropometric measurements, dietary and lifestyle habits, and comprehensive genetic sequencing. By integrating these physical samples with secure digital repositories, the HUABR is specifically designed to support advanced downstream analytics, foster global research collaborations, and drive clinical applications in precision health.

approved each protocol used, with the approval numbers available in the cited articles. All subjects provided voluntary informed consent, in accordance with the standards of the World Medical Association's Declaration of Helsinki for ethical principles [25] governing medical research involving human subjects. To facilitate ongoing research, HUABR operates under a dynamic ethical framework (Approval No. Γ-2643/12-06-2025) that explicitly allows for participant re-contact. This ensures that longitudinal follow-up studies can be conducted—collecting phenotypic, clinical, and lifestyle data consistent with the baseline assessments to enable robust temporal comparisons—provided strict data pseudonymization protocols are maintained. All participants were interviewed by trained personnel using validated questionnaires. The questionnaires included clinical evaluation and anthropometric measurements, as well as questions on socio-demographic and lifestyle characteristics, detailed medical history, physical activity, smoking and dietary habits. The medical history and the corresponding medications were self-reported and cross-checked by trained clinicians, while in the THISEAS study, the diagnosis of coronary artery disease was confirmed through coronary angiography.

A. Pooled Cohorts

For a broader representation of the population, we partnered with municipal and urban healthcare institutions, such as hospitals and health centers, where individuals of various socioeconomic and demographic backgrounds were approached, representing both individuals with chronic conditions and healthy volunteers. The GOMAP case-control study (Genetic



Fig. 2: Geographic Distribution of HUABR's Studies

The map highlights the spatial distribution of the primary recruitment regions, with colors corresponding to specific study cohorts. The concentration of samples in the Attica region (red, $n=9,478$) reflects the country's urban demographic distribution. Additionally, the biore-source captures broader regional, genetic, and lifestyle heterogeneity through the inclusion of geographically isolated populations—namely the HELIC-MANOLIS cohort in Crete (green, $n=1,564$) and the HELIC-POMAK cohort in Thrace (yellow, $n=1,964$)—alongside extensive multi-regional recruitment across the rest of Greece via the OSTEOS cohort (blue, $n=993$).

overlap between metabolic and psychiatric diseases) focuses on patients with hospitalized schizophrenia (SCZ), including those with type 2 diabetes (T2D). This is the first large-scale investigation of SCZ-T2D comorbidity in the Greek population [3]. Greek Metabo [4], [5] study involved patients with class III obesity undergoing bariatric surgery and normal weight individuals undergoing other surgical procedures (e.g. appendectomy). The Greek NASH-PI study developed cost-effective, stepwise algorithms using noninvasive methods and vibration-controlled transient elastography (VCTE) to detect and refer patients with nonalcoholic fatty liver disease (NAFLD) [6]. The Greek Recurrent Myocardial Infarction Cohort (GRMIC) is a prospective study which investigated patients with recurrent myocardial infarctions (re-MI), focusing not only in environmental but also in genetic risk factors and prevention strategies [7]. Moreover, the HYPER-MASTIC trial is a double-blind, randomized study evaluating the effects of Chios mastic supplementation on hypertension in Greek patients with controlled cardiovascular risk factors. The results suggested that Chios mastic may improve hypertension without a history of heart failure or recent cardiovascular events [10], [11]. The iMPROVE study constituted a dietary intervention in adults with overweight or obesity who participated in a randomized clinical trial for weight loss proposing adherence to either a high-carbohydrate or a high-protein hypocaloric diet for

three months [12]. The NAFLD case-control study investigated the relationship between dietary and lifestyle factors and the odds of NAFLD as well as related biomarkers in a Greek sample [16]. OSTEOS is an observational cross-sectional study that recruited volunteers from a variety of regions in Greece, including Central Greece, the Peloponnese, Thessaly, the Aegean islands, and Macedonia, expanding the origin of individuals outside Attica [19]. OSTEOS aimed to assess the prevalence of vitamin D insufficiency or deficiency in the Greek population, examine the impact of lifestyle factors on vitamin D levels and association with quantitative ultrasound parameters. SLEEPSTUDY is an ongoing survey which aims to evaluate the potential of miRNAs as a biomarker for sleep apnea severity and its clinical implications. Finally, The Hellenic study of Interactions between Single nucleotide polymorphisms and Eating in Atherosclerosis Susceptibility (THISEAS) [22] is a case-control study which investigated the relationship between dietary and lifestyle factors and the risk of coronary artery disease (CAD). In summary, the described cohorts provide valuable information on the Greek population, enabling detailed analysis of demographic, clinical, and lifestyle factors and their associations with health outcomes.

B. Isolated Populations

To capture unique genetic as well as lifestyle data from isolated populations, we collaborated closely with local health centers and hospitals located in geographically remote and socio-culturally distinct regions. One of them was the Minoan Isolates (MANOLIS) [8] target population with adults originating from the mountainous region of Mylopotamos, on Mount Idi in Crete, which includes Anogia Municipality, Zoniana Municipal Area and Livadia Local Community. The second study aimed at describing the geographically isolated Pomak [9] population and its particular dietary patterns in relationship to cardiovascular risk factors. The data collected in a population-based cohort in a cross-sectional study, with detailed anthropometric, biochemical, clinical, and lifestyle parameter information.

C. Children

Data collection in children was conducted through two cohorts, GENDAI and TEENAGE, both targeting students in the Attica region. The first cohort focused on fifth and sixth graders from 40 public primary schools, also randomly selected from urban regions [2]. TEENAGE involved adolescents aged 13–15 years from 24 randomly selected public secondary schools in urban areas of Attica [20], [21]. Under the appropriate consent and ethical safeguards in place, this group enables research into development-related health patterns, early markers of disease, and long-term studies on genetic factors influencing growth and development.

D. Athletes

To explore the effects of structured physical activity on health, the repository includes a group of athletes from the ATLAS and WATERPOLO studies. These cross-sectional

studies comprise data from healthy Greek competitive athletes collected between 2014 and 2023 [1]. Standardized data collection was conducted by a multidisciplinary team of dietitians, doctors, and biologists at appropriate athletic facilities (e.g., track and field stadiums, swimming pools, and university laboratories). The cohorts encompass participants from a diverse range of disciplines—including field sports (football, track and field, volleyball, martial arts) and water sports—representing several major national athletic clubs.

III. DATA HARMONIZATION

The data analysis framework of HUABR is designed not only to support diverse research applications, but also to fulfill its central aim: the creation of a comprehensive, high-quality database of phenotypic, genetic and lifestyle information that accurately represents the Greek population.

A. Phenotypic Data

A comprehensive assessment of human health and well-being requires an integration of diverse data types. To transform the 16 original, asynchronous studies into a single, high-quality resource, a systematic retrospective harmonization was performed. A centralized data dictionary was established by identifying commonalities across all participating cohorts, and variables originally recorded using disparate instruments were mapped to a unified ontology. This included the standardization of biochemical markers to consistent units (e.g., mg/dL) and the re-coding of categorical variables, such as smoking status and education level, into uniform classes based on specific clinical and sociological cut-offs.

Anthropometric measurements, including body weight and height, BMI [calculated as $\text{weight (kg)}/\text{height}^2 (m^2)$] [26], waist-to-hip ratio (WHR), as well as body composition measurements such as fat and muscle mass, provide critical insights into physical health and nutritional status. Herein, waist measurements were taken using a stretch-resistant measuring tape at the point midway between the lower costal margin and the iliac crest and were recorded to the nearest 0.5 cm. Hip measurements were taken at the widest portion of the buttocks. Waist and hip circumferences have been measured by well-trained dietitians, have been repeated twice and the average measurement has been recorded. Body composition measurements have been performed using bioelectrical impedance analysis. Complementing these, harmonized socioeconomic data, e.g. education level, profession, income, living arrangements, and marital status, offer context on the social determinants of health. Environmental and lifestyle factors, including smoking habits, physical activity habits, sedentary activities, sleeping patterns and urban versus rural living environments, significantly influence health outcomes.

Meanwhile, dietary data, derived from 24h dietary recalls, food diaries and/or food frequency questionnaires (FFQs), categorized into food groups such as fruits, vegetables, and dairy products, highlight dietary behaviours and patterns. Lastly, clinical and biomarker data, encompassing laboratory-measured parameters such as blood glucose, lipid profiles, inflammatory markers, and hormone levels, alongside clinical

records of personal and family disease history, medication use, and diagnostic results, provide objective measures of physiological and pathological states. Together, these data streams form a multifaceted framework for studying health determinants and outcomes.

Beyond standard assessments, HUABR incorporates deep phenotyping via non-invasive imaging. Specific sub-cohorts feature quantitative ultrasound for bone health [19], alongside vibration-controlled transient elastography (VCTE) and hepatic ultrasound for NAFLD characterization [16]. This multi-modal approach enables high-resolution investigations of cardiometabolic and musculoskeletal disorders.

B. Genetic Data

High-resolution genotyping arrays collected across all cohorts provide a comprehensive coverage of genetic variation. HUABR includes more than 9,000 genotyped samples. Cohorts currently designated as 'Not genotyped yet' in Table I represent the second phase of HUABR's genomic expansion. While their comprehensive phenotypic data are already available for multi-modal analyses, their genetic profiles will be integrated in future releases to further augment the resource's analytical resolution and statistical power. Genome-wide association studies (GWAS) performed on these datasets have identified loci associated with diverse traits and diseases, enabling insights into genetic susceptibility, gene-environment interactions, and polygenic risk modeling for precision medicine [3], [8], [9], [16], [22].

1) *DNA Isolation and Genotyping*: Genomic DNA (gDNA) was isolated from human blood samples and concentrated leukocyte preparations (buffy coat) using standardized protocols and commercially available extraction kits in conjunction with automated purification systems, designed to deliver consistent, high-quality yields across all studies. Additionally, the salting-out method [27] has been used in some cohorts, providing a reliable, cost-effective means of obtaining high molecular weight DNA without the use of organic solvents. All procedures were carefully optimized to maintain DNA integrity, reduce contamination risk, and ensure compatibility with downstream applications such as genotyping and molecular analyses. Cohorts were genotyped using a range of genotyping arrays, as detailed in Table I. Two cohorts listed in Table I are pending genotyping and are scheduled for analysis in upcoming phases. Genotyping completion rates and the proportion of genotyped samples passing sample-level QC for each cohort are reported in Supplementary Table S8.

2) *Pre-imputation*: The genetic data processing workflow begun with pre-imputation quality control (QC) performed within each study. This typically included filtering SNPs with a call rate (CR) < 98%, samples with a CR < 95%, and variants deviating from Hardy-Weinberg equilibrium (HWE) at $p < 0.0001$. Also, individuals with heterozygosity outside the median $\pm 3 \cdot$ of IQR range were excluded. Monomorphic variants were also removed at this stage. After QC, the raw PLINK files (BED, BIM, FAM) [28] were harmonized using R [29] with the human reference genome (GRCh38/hg38 or GRCh37/hg19) to ensure allele alignment across datasets.

TABLE I: Arrays used in HUA BioResource.

Cohort	Population	Genotyped (n)	Array	Variants
ATLAS	204	136	Illumina GSAMDv3	623,511
GENDAI	1,003		Not genotyped yet	
GOMAP	2,874	2,173	Illumina HumanCoreExome	524,271
Greek Metabo	106	106	Illumina Omni 2.5/ Exome 24v1.1	1,158,987
Greek NASH-PI	58		Not genotyped yet	
GRMIC	396	210	Illumina Infinium CoreExome-24 v1.1	542,933
HELIC-MANOLIS	1,564	1,282	Illumina Infinium OmniExpressExome-8	702,284
HELIC-POMAK	1,964	1,019	Illumina Infinium OmniExpressExome-8	683,466
HYPER-MASTIC	119	106	Thermo Fisher Scientific Axiom PMDA	804,961
IMPROVE	202	197	Thermo Fisher Scientific Axiom PMRA	810,428
NAFLD	459	445	Illumina Infinium CoreExome-24 v1.1	330,188
OSTEOS	993	854	Thermo Fisher Scientific Axiom PMRA	310,472
SLEEPSTUDY	250	93	Thermo Fisher Scientific Axiom PMDA	715,197
TEENAGE	852	707	Illumina Infinium CoreExome-24 v1.4	525,036
THISEAS	2,889	1,887	Illumina MetaboChip	196,725
WATERPOLO	66	56	Thermo Fisher Scientific Axiom PMDA	765,478
Total	13,999	9,271		

GSAMD: Global Screening Array Multi-Disease

PMDA: Precision Medicine Diversity Array

PMRA: Precision Medicine Research Array

The harmonized data converted into VCF format, which was sorted, compressed, and indexed. To facilitate parallel processing, the VCF file was split by chromosome. These chromosome-specific files prepared for imputation.

3) *Imputation*: Imputation is an essential computational process in the analysis of genetic datasets, allowing for the prediction of untyped genetic variants by leveraging known haplotype reference panels. This methodology significantly enhances the genomic resolution and coverage of GWAS, thereby increasing statistical power and improving the detection of associations between genetic variants and phenotypic traits. The cohorts in HUABR have been imputed mostly in two reference panels, Haplotype Reference Consortium (HRC r1.1 2016) [30] and Trans-Omics for Precision Medicine (TOPMed) [31], [32] both hosted at Michigan Imputation Server [33]. Each cohort was imputed separately and then all datasets were merged by chromosome.

4) *Post-imputation*: Post-imputation, several critical steps were performed to clean and standardize the data. The post-imputation VCF files do not have IDs (usually dbSNP ids) for all variants. That is because the reference panel contains also novel variants, not yet verified and submitted to dbSNP [34]. Since we needed an ID for each variant, we annotated each chromosome's imputed VCF file to assign a unique ID to each variant using the format

chromosome:position:reference:alternative.

Variants with low imputation quality ($R^2 \leq 0.3$) were removed, and duplicate indels were resolved to avoid redundant entries. The cleaned files for all 23 chromosomes (22 autosomal chromosomes and X chromosome) were concatenated into a single merged file, which was compressed and indexed for efficient access. Genetic variant annotation was conducted using dbSNP [34] in Java [35]. Finally, the VCF was converted back into PLINK format (BIM, BED, FAM) to facilitate downstream analyses. Once converted, a post-imputation QC was applied. Variants were retained only if they had a call rate greater than 98%, and a HWE p-value exceeding 10^{-6} . Similarly, sample-level QC is performed to exclude individuals with a CR below 95% and heterozygosity outside the median $\pm 3 \cdot$ IQR range.

IV. DATA MANAGEMENT, GOVERNANCE, AND SUSTAINABILITY

To ensure the long-term preservation and integrity of both biological samples and digital records, rigorous storage conditions are implemented.

Biological materials — including plasma, serum, white blood cells, extracted DNA and RNA, as well as subcutaneous and visceral adipose tissue biopsies — are preserved in ultra-low temperature freezers maintained at -80°C . This standard biospecimen management practice halts enzymatic activity, minimizes degradation, and preserves molecular integrity over time [36].

On the digital side, all data generated from the samples — such as genetic sequences, biochemical markers, anthropometric measurements, lifestyle data and clinical records — are stored in a secure, high-performance server located in the Network and Computing Systems Center (NOC) of Harokopio University. The NOC ensures physical security (controlled access, fire safety, UPS with backup generator) and system reliability through the climate-controlled environment. The server runs Ubuntu Server 24.04 LTS with weekly security updates. Access governance is strictly role-based: users connect only via institutional accounts and VPN, with each researcher granted rights solely to their own project data. No public IP or HTTP(S) access is available, and only authorized team members who have signed confidentiality agreements and have undergone rigorous technical training can obtain dedicated credentials. To further ensure data integrity, these legally binding agreements explicitly prohibit the unauthorized movement, replication, or copying of data, ensuring sensitive information remains confined to approved, encrypted environments.

In addition, the servers are equipped with advanced encryption technologies to safeguard confidentiality and reduce the risk of breaches [37]. Regular backups are performed, both on-site and off-site, to prevent data loss in the event of hardware failures or other unforeseen incidents. Access is granted based on predefined roles and responsibilities, with multi-factor authentication required for entry into systems. Secure remote access is facilitated through Virtual Private Networks (VPNs), which provide encrypted communication channels between authorized users and the infrastructure. Stringent data protection protocols are maintained, including informed consent practices, data minimization strategies, and measures to protect individual privacy. Beyond safety, this digital infrastructure supports a wide range of biomedical research activities. Advanced tools for big data analytics, machine learning, and secure collaboration allow researchers to extract insights while adhering to strict data governance practices [38]. Finally, to promote long-term sustainability and global collaboration, data from several HUABR cohorts have been deposited in international archives such as the European Genome-Phenome Archive (EGA) [39], ensuring governance transparency and continued relevance to the broader scientific community.

V. STATISTICAL METHODS

The statistical framework was designed to accommodate the large-scale, multi-cohort nature of the HUABR. Normality was

assumed for the study population based on the Central Limit Theorem (CLT) [40], [41]. Due to their skewed distributions, fasting insulin (Ins) and triglyceride (TG) values were log-transformed prior to analysis to better approximate a normal distribution. Data points exceeding three standard deviations (SD) from the mean were identified as outliers and excluded from the final dataset. To assess BMI across various factors, Analysis of Variance (ANOVA) [42] was used for categorical variables, while Pearson Correlation Coefficients [43] measured the linear relationship between BMI and continuous variables. Pairwise differences were further explored using Tukey's Honest Significant Difference (HSD) test [44] or Dunn's test [45] with a Bonferroni correction to control for Type I error in multiple comparisons [46].

Beyond descriptive comparisons, a Generalized Linear Model (GLM) was implemented to examine the association between BMI, age, and sex, including an interaction term to assess gender-specific effects of aging. For the determination of adult BMI predictors, a backward stepwise regression was employed to identify the most significant predictors while controlling for potential confounding factors [47]–[49]. This model utilized the Akaike Information Criterion (AIC) [50] to optimize variable selection from a pool of demographic, lifestyle, and clinical measures, including blood pressure, glucose, and lipid profiles. To evaluate the robustness of the BMI regression models reported in Tables IV–VI to potential confounding by cohort-level heterogeneity across HUABR's 16 constituent studies, each model was additionally re-fitted as a linear mixed-effects model with study/cohort included as a random intercept (*lme4* R package [51]). Finally, Polygenic Risk Scores (PRS) were incorporated into linear regression models to quantify their association with BMI, with adjusted R^2 values and 95% confidence intervals (C.I.) reported via bootstrapping.

The integration and analysis of the HUABR utilized a specialized computational pipeline across R [29] and Python [52] environments. Genomic data management was handled using the *snpStats* [53] package to process large-scale genotype matrices in PLINK format [28]. To address population stratification, Principal Component Analysis (PCA) was conducted using merged genotype data from the HUABR and 1000 Genomes (1KG) [54]. PCA was executed using *FlashPCA* [55], with the top three components (PC1–PC3) used to visualize ancestry structure and adjust for population-level variation in downstream analyses; the percentage of variance explained by each PC was estimated from the FlashPCA eigenvalue output. The calculation of PRS was executed via *PRSice2* [56], which performs optimized matrix operations to compute weighted sums of risk alleles based on summary statistics and effect sizes from the PGS Catalog [57], [58]. To ensure genomic consistency, the pipeline included an automated allele-switch check to align effect alleles with the genotype coding of the imputed dataset. Data visualization was primarily conducted using the *ggplot2* [59] package in R, while Figure 4 and 6 were generated using Matplotlib library [60] in Python. Each regression model reported in this manuscript (summarised in full in Supplementary Table S9) was fitted as an independent, separate analysis; a nominal significance threshold of $p < 0.05$

was applied on a per-model basis throughout.

VI. RESULTS

Understanding the demographic and health characteristics of a study population is crucial for drawing meaningful conclusions. An analysis of the BMI distribution reveals that 27% of individuals are classified as people with normal weight, 40% as people with overweight, and 33% as people with obesity (a visual representation is provided in Supplementary Figure which illustrates the distribution of BMI categories, highlighting the proportion of individuals classified as people with normal weight, people with overweight, or people with obesity. Notably, nearly three-quarters (73%) of the population falls into the overweight or obese category, reflecting broader public health concerns.

HUABR includes a total of 13,999 individuals, with 5,492 males (39.2%), 6,217 females (44.4%), and 2,022 (14.4%) children while there is 1.9% of missing gender or age information. The age distribution of the participants spans from 9 to 99 years. The median age of the resource is 52 years, with a mean age of 48.29 years.

After removing outliers, the dataset was reduced to 11,715 individuals, 9,870 adults and 1,845 children. Adults consist of 4,438 males (44.9%) and 5,402 females (54.7%) while the age distribution of the remaining participants ranges from 18 to 97 years. The median age is 55 years, and the mean age is 54 years. The first quartile (Q_1) of age is 43 years, while the third quartile (Q_3) is 66 years. Some summary statistics for the variables included to HUABR are depicted in Table II. Figure 3 illustrates the distribution of key demographic and behavioral characteristics in the dataset, including education level, profession type, marital status, and smoking habits which helps to paint a picture of the population's traits. Moreover, the Figure 4 reveals that BMI tends to increase with

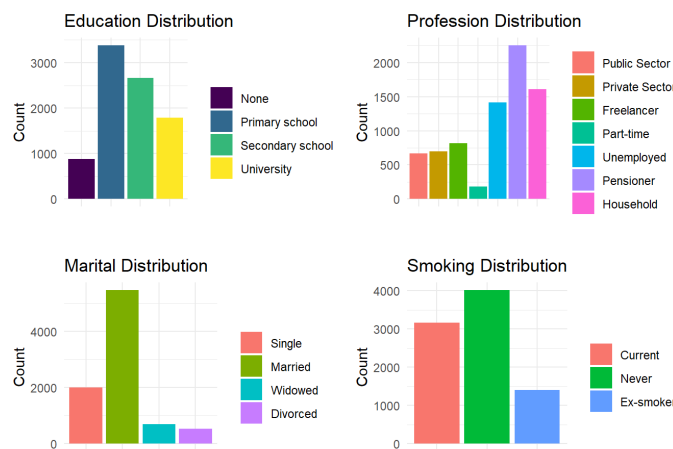


Fig. 3: Education level, profession type, marital status, and smoking habits

age for both males and females, but the rate of increase differs between genders. The regression model (Table:III) shows that while younger females generally have a lower BMI than their male counterparts, their BMI rises at a significantly faster rate over time. This leads to a narrowing gender gap in BMI as

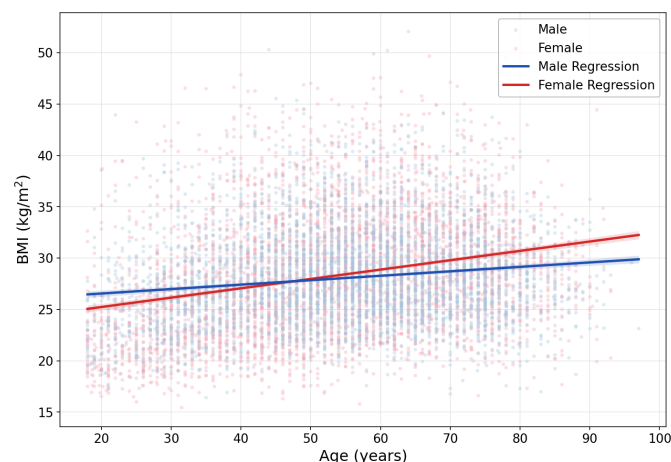


Fig. 4: Age and Gender Differences in BMI

individuals age. However, the model explains only 5.54% of the variation in BMI, so further analyses have been conducted incorporating additional factors.

The median BMI for females is slightly higher than that for males (From Boxplots in Figure 5(a)). The IQRs overlap, but females have a higher upper quartile and more extreme outliers. The simple regression model fitted (BMI vs Gender) indicates that gender has small, but statistically significant effect on BMI ($\beta = 0.263, p = 0.01$) which means that females have a BMI that is 0.263 kg/m² higher than males. Furthermore, the polyserial correlation [61] between BMI and education level is -0.144 ($p\text{-value} = 0 < 0.0001$), indicating a weak negative relationship between BMI and education level as it is shown in Figure 5(b). In Figure 5(c) it is outlined that urban residents tend to have slightly lower BMI compared to semi-urban and agricultural residents. The linear regression model confirmed significant differences: Semi-urban residents have 1.22 kg/m² higher BMI ($p < 0.001$) compared to urban residents and agricultural residents have 0.66 units higher BMI ($p < 0.0001$) compared to urban residents.

From Figure 5(d) it is observed that married individuals have a slightly higher BMI median compared to single individuals, while widowed and divorced individuals show more variability. The ANOVA results indicate a statistically significant difference in BMI across marital status groups (F-value: 83.9, $p = 2e - 16$). Tukey's HSD test [44] and Dunn's test [45], adjusted using the Bonferroni correction were used to further assess pair-wise differences. BMI differences among marital status groups were significant, except from the difference between divorced and married individuals. More detailed results are in Tables S10-S12 in supplementary files. For further analysis of factors contributing to obesity, stepwise regression conducted. The first model uses BMI as the response variable, and its results are presented in Table IV. To evaluate whether the associations reported in Tables IV - VI were confounded by cohort-level heterogeneity, these models were re-fitted as a linear mixed-effects model with cohort as a random intercept. For the Table IV Cohort membership accounted for approximately 5.5% of the residual variance in BMI. The direction and significance of most fixed effects were preserved, although

TABLE II: Summary statistics of key variables

Variable	n	Adults					Children				
		Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD
Age	13,256	18.00	97.00	54.00	55.00	16.10	9.00	17.10	12.4	12.20	1.52
Education years	6,569	0.00	24.00	10.20	10.00	4.75	11.00	12.00	11.70	12.00	0.577
Weight (kg)	12,883	36.00	130.00	78.30	77.20	15.60	24.70	90.30	50.50	49.50	12.20
Height (cm)	12,832	137.00	196.00	167.00	166.00	9.81	126.00	185.00	156.00	155.00	10.90
BMI	12,773	15.40	52.10	28.20	27.70	5.11	12.10	31.50	20.60	20.30	3.31
zBMI	1,845						-3.56	3.07	0.85	0.90	0.94
WHR	11,359	0.57	1.28	0.90	0.90	0.109	0.57	1.10	0.790	0.78	0.06
SBP	10,969	80.00	183.00	130.00	130.00	16.40	79.00	160.00	119.00	119.00	12.80
DBP	10,975	46.00	113.000	79.40	80.00	10.80	40.00	108.00	73.20	73.00	10.40
Pulses	6,925	43.00	134.00	78.80	78.000	12.00	48.00	125.00	83.10	83.00	13.30
Glu(mg/dL)	8,592	21.00	196.00	98.40	94.00	22.40	52.00	119.00	83.60	84.00	8.72
TC(mg/dL)	6,184	81.00	314.00	194.00	192.00	39.50	97.00	271.00	173.00	171.00	30.60
HDL-C(mg/dL)	8,500	12.00	89.50	48.60	47.00	13.10	28.00	90.00	54.30	53.00	11.30
LDL-C(mg/dL)	8,439	9.90	244.00	122.00	120.00	36.10	34.80	192.00	108.00	107.00	26.09
ln(TG)(mg/dL)	6,184	3.30	6.01	4.72	4.71	0.48	3.04	4.91	4.05	4.04	0.32
ln(Insulin)(μ U/mL)	6,340	1.90	48.00	12.10	10.00	8.21	0.00	29.00	8.99	8.00	4.94
Energy intake (kcal/day)	6,964	466.00	4926.00	2152.00	2030.00	811.00	380.00	3924.00	1817.00	1747.00	642.00

BMI: Body Mass Index, WHR: Waist-to-Hip Ratio, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, Glu: Glucose, TC: Total Cholesterol, HDL-C: High-Density Lipoprotein Cholesterol, LDL-C: Low-Density Lipoprotein Cholesterol, TG: Triglycerides

TABLE III: Regression results for BMI by Age and Gender

Variable	Estimate	Std. Error	p-value
BMI Males at Age 18	25.69	0.27	2e - 16
Effect of Age on BMI for Males	0.04	0.004	2e - 16
Female vs Male at Age 18	-2.22	0.35	1.48e - 10
Effect of Age on BMI for Females	0.048	0.006	2.54e - 14

TABLE IV: M1: Summary of Stepwise Regression for BMI, $R^2 = 0.36$, $F = 47.4$.

Variable	Estimate	Std. Error	p-value
Intercept	8.54	1.90	< 0.001 ***
Gender (Female vs Male)	1.96	0.30	< 0.001 ***
Age (years)	0.04	0.01	< 0.001 ***
Residency (Semi-urban vs Urban)	2.55	1.34	0.058 .
Residency (Agricultural vs Urban)	-1.96	0.50	< 0.001 ***
Education Years	-0.1	0.05	0.04*
SBP (mmHg)	0.09	0.009	2e - 16 ***
Glu (mg/dL)	0.02	0.01	0.04 *
TC(mg/dL)	0.09	0.003	0.015 *
HDL-C (mg/dL)	-0.08	0.01	< 0.001 ***
Energy Intake (kcal/day)	0.0004	0.0001	0.03 *
ln(Ins) (μ U/mL)	2.4	0.27	2e - 16 ***

SBP: Systolic Blood Pressure, Glu: Glucose, TC: Total Cholesterol, HDL-C: High-Density Lipoprotein Cholesterol, ln(Ins): Log-Insulin

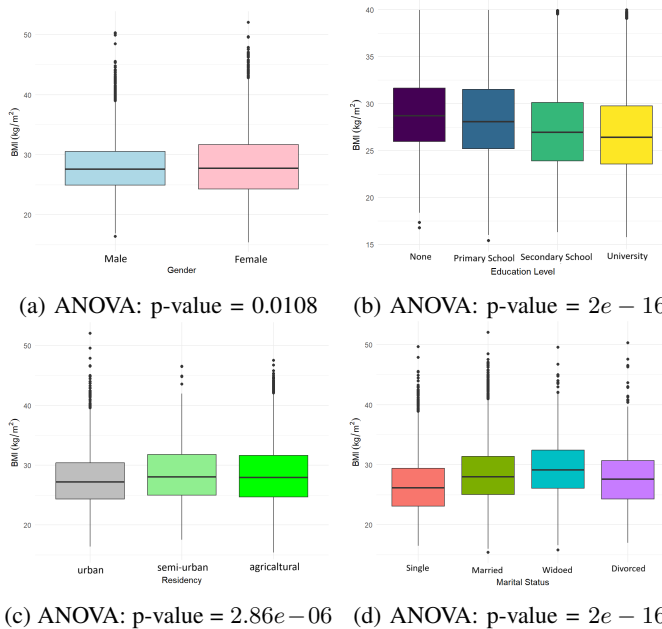


Fig. 5: Boxplots of BMI vs (a) Gender, (b) Education level, (c) Residency and (d) Marital Status

the effect of total cholesterol was markedly attenuated and residency was no longer a statistically significant predictor. Full results for all models are reported in Supplementary Tables S1-S4.

One more factor that is widely used to measure body fat

distribution is WHR. It is calculated as the ratio of the circumference of the waist to the circumference of the hips and it is greatly associated with BMI. A high WHR is linked to abdominal fat deposition. Adjusting BMI for WHR allows for a clearer understanding of how other variables (like gender, education, or lifestyle factors) affect BMI, after accounting for the influence of body fat distribution. The standardized residuals (ϵ) from the BMI vs WHR regression model were extracted and used to create a new variable called BMI adjusted for WHR. The results of the stepwise regression are represented in Table V.

Both models in Tables IV and V share several common independent variables. However, the first model includes age as a continuous variable but omits marital status. The first model has a slightly higher R^2 , indicating it explains more variability in BMI, but its residual standard error is higher, suggesting greater prediction error. Finally, the first model has a higher F-statistic, indicating better overall fit. Moreover, four separate linear regression models were applied to examine the factors associated with BMI across different demographic groups Table VI: males over 18 y.o., males under 18 y.o., females over 18 y.o., and females under 18y.o. The adjusted R^2 values

TABLE V: M2: Summary of Stepwise Regression for BMI Adjusted for WHR, $R^2 = 0.30$, $F = 38.11$

Variable	Estimate	Std. Error	p-value
(Intercept)	-14.70	2.07	$2.76e-12$ ***
Gender (Female vs Male)	3.70	0.29	$< 2e-16$ ***
Residency (Semi-urban vs Urban)	2.54	1.27	0.047 *
Residency (Agricultural vs Urban)	-1.17	0.31	0.00024 ***
SBP (mmHg)	0.077	0.009	$< 2e-16$ ***
Glu (mg/dL)	0.03	0.01	< 0.009 **
TC (mg/dL)	0.009	0.004	0.03 *
HDL-C (mg/dL)	-0.07	0.013	$6.84e-06$ ***
ln(TG)(mg/dL)	-0.69	0.39	0.08 .
ln(Ins) (μ U/mL)	1.74	0.27	$2.77e-10$ ***
Energy Intake (kcal/day)	0.0004	0.0001	0.011 *

SBP: Systolic Blood Pressure, Glu: Glucose, TC: Total Cholesterol, HDL-C: High-Density Lipoprotein Cholesterol, LDL-C: Low-Density Lipoprotein Cholesterol, ln(TG): Log-Triglycerides, ln(Ins): Log-Insulin

ranged from 0.18 to 0.38, indicating moderate explanatory power. Across all models, SBP levels, glucose levels, and insulin levels were consistently significant positive predictors of BMI. Among adult males, residing in an agricultural area was significantly associated with lower BMI, while ex-smokers and never-smokers had significantly higher BMI compared to current smokers. In contrast, for adult females, education years were negatively associated with BMI, and total cholesterol (TC) and LDL-C were significant predictors. For children under 18, higher SBP, glucose, and insulin levels remained key predictors, but lipid profiles played a more prominent role, with HDL-C and LDL-C showing significant associations. Interestingly, triglycerides (TG) were negatively associated with BMI in both female groups. These findings highlight the importance of metabolic and cardiovascular factors in BMI regulation across different age and gender groups, with sociodemographic factors playing a more pronounced role in adults.

In the 3D PCA plot- Figure 6, HUABR and HELIC samples

TABLE VI: Regression Results for BMI Prediction by Sex and Age Group.

Variable	Males		Females	
	≥ 18 y.o.	< 18 y.o.	≥ 18 y.o.	< 18 y.o.
(Intercept)	5.2*	11.31**	9.2***	12.24***
Residency (Semi-urban vs Urban)	1.23	-	2.87	-
Residency (Agricultural vs Urban)	-1.31**	-	-1.93*	-
Marital Status (Married vs Single)	0.47	-	-	-
Marital Status (Widowed vs Single)	-2.73	-	-	-
Marital Status (Divorced vs Single)	2.25*	-	-	-
Age(years)	-	0.34	0.09***	-
Education Years	-	-	-0.14*	-
Smoking (Never vs Current)	1.37***	-	-	-
Smoking (Ex-smoker vs Current)	1.7***	-	-	-
Energy Intake (kcal/day)	0.0003	0.0009**	0.0004	0.0009**
Pulses	-	-	0.028	0.29*
SBP (mmHg)	0.05**	0.036*	0.086***	0.047**
DBP (mmHg)	0.036	-	-	-
Glu (mg/dL)	0.03*	-	0.02	-
TC (mg/dL)	-0.06***	0.015	0.14*	-0.018**
HDL-C (mg/dL)	-	-0.083***	-0.08***	-
LDL-C (mg/dL)	0.07***	-	-	-
ln(TG)(mg/dL)	1.58**	-	-	1.57**
ln(Ins) (μ U/mL)	1.79***	2.13***	2.64***	1.54***
Adj. R ²	0.38	0.3	0.38	0.18
Residual Std. Error	3.08	2.72	4.27	2.71

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, Glu: Glucose, TC: Total Cholesterol, HDL-C: High-Density Lipoprotein Cholesterol, LDL-C: Low-Density Lipoprotein Cholesterol, ln(TG): Log-Triglycerides, ln(Ins): Log-Insulin

For participants aged < 18 years, the outcome variable is zBMI

cluster near each other, showing overall genetic similarity and overlap with the 1KG European reference population,

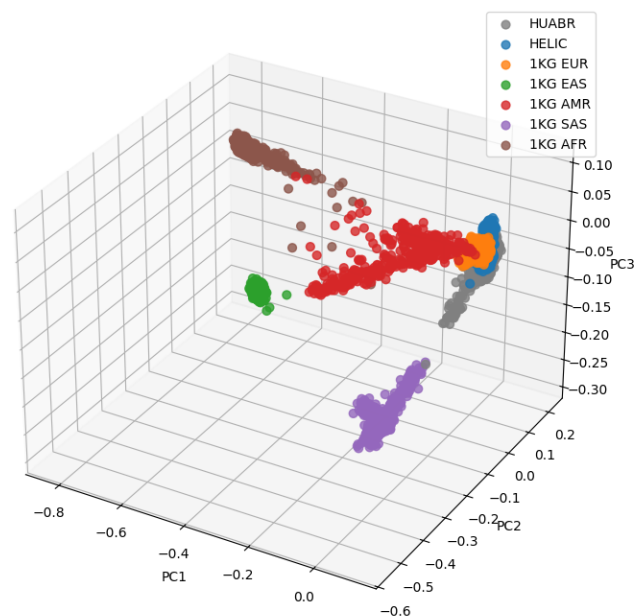


Fig. 6: Three-dimensional Principal Component Analysis (PCA) plot based on data from the HUABR and 1000 Genomes (1KG).

suggesting shared European ancestry. PC1, PC2, and PC3 explained 3.59%, 2.72%, and 2.16% of the total genetic variance, respectively (cumulative 8.47%). However, the HELIC samples form a more distinct and separate cluster, indicating genetic differentiation from HUABR. This separation reflects the greater genetic and social isolation of the HELIC population compared to the more admixed HUABR cohort.

PRSs were derived using summary statistics from the PGS Catalog [57], [58]. Specifically, PRS were computed using SNP weights from publicly available PRS models (e.g., PGS000027, PGS004733, etc.) that were curated based on previous GWAS related to BMI.

To account for potential differences in genetic influences on BMI across age groups, separate PRS were calculated for children and adults. This stratification ensured that PRS were appropriately tailored to each group, allowing for a more precise evaluation of their predictive performance. For each individual, PRS was calculated from HRC [30] imputed data, applied uniformly across all individuals regardless of their original genotyping array (Table I), as the weighted sum of risk alleles, where the effect sizes from the PGS Catalog were used as weights. SNPs that were unavailable in the dataset were excluded.

To assess the association between BMI and PRS, a generalized linear model (GLM) was fitted and the results are represented in Tables VII & VIII. Additionally, the Pearson correlation coefficient computed at the scaled scores [62] to quantify the strength of the association with the actual BMI.

The density plot (Figure 7) illustrates the distribution of PGS004733 across three BMI categories—Normal, Overweight, and Obese—in HUABR. This PRS was selected because it had more SNPs compared to PGS000910 (735,440 SNPs, 100% matched to the HUABR genotype data), while

TABLE VII: PRS performance summary for individuals aged 18 and over.

PGS ID	Pearson Cor.		Model A				Model B			
	r	P-value	R^2	RMSE	MAE	P-value	R^2	RMSE	MAE	P-value
PGS000027 [63]	0.19	$< 2.2e - 16$	0.038	5.293	4.051	3.11×10^{-59}	0.081	5.166	3.917	7.42×10^{-63}
PGS000910 [64]	0.24	$< 2.2e - 16$	0.059	5.235	4.007	4.24×10^{-92}	0.101	5.104	3.880	2.74×10^{-96}
PGS002679 [65]	0.22	$< 2.2e - 16$	0.050	5.260	4.024	4.11×10^{-78}	0.093	5.129	3.890	5.72×10^{-82}
PGS004066 [66]	0.21	$< 2.2e - 16$	0.042	5.281	4.044	4.23×10^{-66}	0.085	5.150	3.910	9.04×10^{-70}
PGS004733 [67]	0.24	$< 2.2e - 16$	0.059	5.235	4.011	3.30×10^{-92}	0.102	5.102	3.876	5.57×10^{-97}

Model A: BMI $\sim$ PRS only; Model B: BMI $\sim$ PRS + age + sex + 5PCs.
RMSE: Root Mean Squared Error, MAE: Mean Absolute Error

TABLE VIII: PRS performance summary for individuals aged under 18.

PGS ID	Pearson Cor.		Model A				Model B			
	r	P-value	R^2	RMSE	MAE	P-value	R^2	RMSE	MAE	P-value
PGS000027 [63]	0.28	$1.4e - 13$	0.076	3.394	2.647	1.38×10^{-13}	0.093	3.365	2.626	6.64×10^{-13}
PGS000716 [68]	0.21	$1.5e - 08$	0.045	3.451	2.664	1.46×10^{-8}	0.066	3.415	2.641	1.68×10^{-8}
PGS004610 [69]	0.18	$1.3e - 06$	0.032	3.473	2.703	1.34×10^{-6}	0.051	3.440	2.681	3.34×10^{-6}

Model A: zBMI $\sim$ PRS only; Model B: zBMI $\sim$ PRS + age + sex + 5PCs.
RMSE: Root Mean Squared Error, MAE: Mean Absolute Error

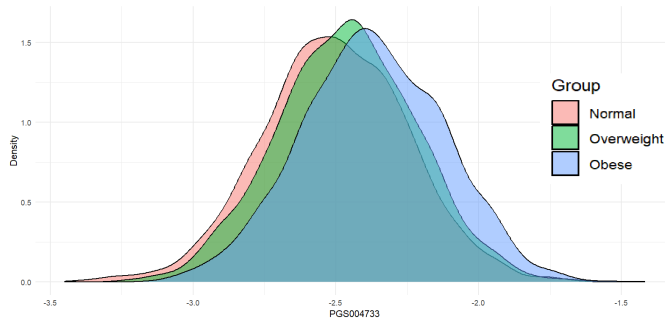


Fig. 7: Density Plot of PGS004733 for BMI Groups

maintaining comparable predictive performance in terms of R^2 and Pearson correlation. PGS004733 also achieved the lowest p-value among all evaluated adult PRS models (Table VII). SNP matching rates for all evaluated PRS models, relative to their original PGS Catalog variant counts, are reported in Supplementary Table S7. The plot demonstrates a clear rightward shift in PRS distribution from the Normal to the Overweight and Obese groups, indicating that individuals with higher genetic risk scores are more likely to belong to higher BMI categories. These results support the effectiveness of PGS004733 in stratifying BMI risk based on genetic predisposition within HUABR. This observation is further supported by the statistical comparisons presented in Supplementary Table S13, where all pairwise differences between BMI groups were significant, with the largest mean difference observed between the Obese and Normal groups (0.1309 , $p = 4.09 \times 10^{-12}$). Among the PRS models evaluated, PGS000910 and PGS004733 demonstrated the strongest performance in predicting BMI for adults, as indicated by the highest adjusted R^2 , lowest p-value and the strongest Pearson correlation coefficient. Similarly, for children, PGS000027 exhibited the best predictive power based on these metrics. These findings

suggest that PGS000910 and PGS004733 are the most informative PRS for BMI in adults, while PGS000027 is the most relevant predictor in children. The results highlight the role of genetic predisposition in BMI across different age groups, supporting the utility of age-specific PRS models. However, as BMI is influenced by both genetic and environmental factors, further validation in diverse populations is warranted to assess the generalizability of these findings.

The corresponding adjusted R^2 comparison with 95% confidence intervals is provided in Supplementary Figure S1. The full models consistently explained a greater proportion of variance in BMI, with minimal overlap in the 95% C.I., suggesting that PGS adds statistically meaningful predictive value. Notably, the adult sample showed slightly higher adjusted R^2 values across PGS, indicating stronger genetic contributions to BMI in this group.

VII. CONCLUSION

The establishment of the HUABR represents a significant milestone in biomedical research within Greece, providing a comprehensive and diverse dataset that facilitates a deeper understanding of genetic, environmental, and lifestyle factors influencing health outcomes. Our present and future findings will contribute valuable insights into the complex interplay between these factors, particularly concerning obesity and other complex traits. One of the primary strengths of the HUABR is its breadth of harmonized genetic, phenotypic, and lifestyle data across a large, geographically diverse sample, enhancing the applicability of findings across multiple population subgroups and enabling detailed depiction of population-specific genetic traits within those subgroups. In HUABR's adult population, overweight/obesity prevalence was 16.3%/16.0% among females and 17.8%/10.5% among males—lower than the Hellenic National Nutrition

and Health Survey (HNNHS) [70]–[72] (25.0%/15.3% females, 41.6%/16.6% males), the National Survey of Morbidity and Risk Factors (EMENO) [73] (29.8%/34.1% women, 44.9%/30.6% men), and, most markedly, World Health Organization (WHO) estimates [74] (30.8%/25.4% females, 44.0%/24.2% males) for Greece.

One potential explanation for the lower prevalence in our findings is the composition of the HUABR. A significant proportion of participants were recruited through case-control studies targeting chronic conditions like osteoporosis, cardiovascular disease and others. Individuals affected by these conditions are often advised to improve their lifestyle habits—such as adopting healthier diets and increasing physical activity—which can lead to better weight management [75] [76]. In addition, the repository includes two socially and genetically isolated populations, HELIC - MANOLIS and POMAK, which differ significantly in lifestyle practices, dietary habits, and cultural norms. Notably, one of these populations adheres to a distinct religion that influences daily routines and includes specific dietary restrictions and fasting traditions. Also, the age distribution of the HUABR at the period of the studies may partly explain the lower obesity prevalence. For instance, the HUABR has an older average age (56 years) compared to the EMENO study (49.2 years), while the data collection of HUABR started in 2006 but EMENO's duration was from 2013 until 2016.

HUABR reveals a concerning picture for children's weight status [74], with 21.9% overweight and 19.1% obese—markedly higher than HNNHS [77] (12.2%/2.0%) and the WHO European COSI initiative [78] (29%/12% across 33 countries; 42%/18% for Greece specifically among 7-9 year-olds). Several contextual factors likely contribute to this gap: HUABR draws primarily from urban Athens school populations, where dietary habits, sedentary behavior, and fast-food exposure may differ from the rural/island regions HNNHS also sampled; HUABR includes children older than 9 years (unlike HNNHS/COSI's 7-9-year focus), and since overweight/obesity prevalence rises with age, this likely inflates HUABR's rates; and although 10% of HUABR children are athletes, the elevated overall prevalence may reflect growing inequality in health behaviors within this urban population, compounded by increased screen time and reduced physical education. These findings underscore the need for targeted, school-based interventions in urban youth settings. Another important consideration is the predictive power of PRSs in HUABR compared to other large-scale biobanks. In our analysis, the best-performing PRS for adults (e.g., PGS000910 and PGS004733) achieved an adjusted R^2 of approximately 0.06, while for children, PGS000027 reached an R^2 of 0.08. These values are comparable to those reported in other European ancestry cohorts. For instance, in the UK Biobank [79], a PRS for BMI constructed using the GIANT consortium [80] GWAS as base data explained approximately 5% of BMI variance in over 388,000 individuals, as detailed in the tutorial "A guide to performing Polygenic Risk Score analyses" [81]. PGS000027 was tested in two different populations according to PGS Publications. In the UK Biobank [79], PGP000017 [63] (based on GIANT consortium [80] data and evaluated in 288,016

individuals of European ancestry) reports an R^2 of 0.085 using a model adjusted only for genetic principal components. In contrast, PGP000349 [82], tested in 832 individuals from the Bogalusa Heart Study (BHS) [83], reports an R^2 of 0.073 for a model that included age, sex, principal components, and PRS. These findings underscore that while PRS alone explain a moderate portion of BMI variance, their predictive power can be modestly improved through the integration of basic demographic variables.

Despite the robust dataset and methodological rigor, certain limitations must be acknowledged. The primary limitation of the current analysis is its cross-sectional nature, which restricts causal inference and limits the ability to track changes in health parameters over time. However, longitudinal follow-up studies are already ongoing to address this issue, providing more dynamic insights into disease progression and health outcomes. These follow-ups will allow for better understanding of temporal trends and causality in disease development, enhancing the overall impact of the data. Additionally, some of the studies included in the HUABR have utilized a case-control design, therefore the total sample is not fully representative of the general population. As a result, the findings may not be generalizable to the broader Greek population. Future analyses should consider adjusting for this limitation by incorporating a representative sample and exploring longitudinal designs to strengthen causal inferences and generalizability.

Future work will focus on enriching predictive models with additional variables and types of data, such as dietary patterns, physical activity and medical history. The HUABR holds immense potential for expanding research in precision medicine and public health. The integration of machine learning techniques incorporating real-time health monitoring through wearable devices and leveraging artificial intelligence for data integration will further improve model accuracy and applicability. These enhancements will support more comprehensive risk assessments and tailored intervention strategies for disease prevention and management.

In conclusion, the HUABR serves as a pivotal resource for advancing our understanding of genetic and environmental determinants of health in Greece. The findings of this study provide a strong foundation for future research aimed at developing targeted prevention and intervention strategies to combat obesity and related metabolic disorders. Continued investment in BioResource infrastructure, coupled with interdisciplinary collaborations, will be critical in harnessing the full potential of this valuable research asset.

VIII. DATA AVAILABILITY

The data used in this study were obtained from a series of projects and were consolidated into a structured bioresource. Portions of this dataset are publicly available through the European Genome-phenome Archive (EGA) (<https://ega-archive.org/>) under controlled access: EGAD00010000516, EGAD00010000522, EGAD00010000630, EGAD00010001005, EGAD00010001484, EGAD0001001636, EGAD0001001897. Due to ethical and consent-related restrictions, other parts of the dataset are not publicly available but can be shared upon request.

IX. CONFLICT OF INTEREST

I.N. was a PhD student when TEENAGE and GENDAI studies were conducted, and is now an employee of Novartis Pharmaceuticals UK, Ltd. She reports stocks from Gilead Sciences, Inc and Novartis Pharmaceuticals. O.G. is an employee and stock owner of UCB Pharma; this work was conducted before employment by UCB.

X. ACKNOWLEDGMENT

We gratefully acknowledge the generous contribution of all volunteers, who provided their data and consent, making the creation of HUABR possible. We also extend our sincere thanks to the students and researchers whose effort was essential to this work.

In particular, we thank : V. Andreou, L. Angelidakis, A-A. Kyriazopoulou Korovesi, G. Arnaoutis, A. Athanasiadis, D. Bampatsias, D. Borsa, T. Diamantis, A. C. Dimopoulos, I. Dontas, E. Evagelidaki, P. Feritoglou, C. Fotakis, I. E. Gavra, G. Georgiopoulos, M. Giannakopoulou, K. Giannitsopoulou, V. Gkesou, S. Glentis, E. Glezou, I. Griniatsos, V. Harokopos, I. Hatzopoulou, A. Kalantzi, G. Kapravelou, A. Karabela, M. Karaleftheri, S. Katopodis, C. Kiagiadaki, P. Kokkoris, C. Konstantaki, V. Kontaxakis, I. G. Kontele, E. Koukouloti, G. Kourlaba, S. D. Ladas, S. Lambrinou, K. Lamnissou, P. Makras, C. Makridi, K. Makris, E. Maneta, E. Manios, M. Marougka, A. Matchan, E. Mentsiou-Nikolaou, G. Mikou, E. More, P. Moulos, K. Ntaoutidou, V. K. Pachi, N. Panousis, M. Papaioannou, N. Papaioannou, C. Papoutsakis, I. Petropoulos, A. Planchon, E. Prountzou, J. Ragoussis, M. Reczko, K. Revenas, I. Serafetinidis, K. P. Skenderi, A. Skoulakis, I. Smyrnioudis, P. Spiliopoulou, G. Stefanou, E.T. Dermitzakis, S. Tournis, E. Tsafantakis, A. Tsekoura, E. Tsompanaki, M. Tzirkalli, I. Varlamis, N. Vidra, J. Vlachogiannakos, P. Vlachou M. Yannakoulia and E. Zengini. Their collective expertise and collaboration was critical to the success of this effort. We also acknowledge the support of technical staff, data managers, and institutional partners involved throughout the process. P.G.Symianakis was supported by the European Union's Horizon Europe research and innovation programme "Preventing obesity through biologically and behaviorally tailored interventions for you" (BETTER4U) under Grant Agreement No 101080117.

REFERENCES

- [1] N. A. Pogosova, D. Brekou, I. E. Gavra, E. A. Katsareli, E. More, P. G. Symianakis, M. Kafyra, I. P. Kalafati, G. Arnaoutis, and G. V. Dedoussis, "Atlas study: Design, athletic performance, and sex-specific regression models for muscle strength in the greek population," *Sports Medicine and Health Science*, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666337624001203>
- [2] C. Papoutsakis, N. V. Vidra, I. Hatzopoulou, M. Tzirkalli, A.-E. Farmaki, E. Evagelidaki, G. Kapravelou, I. G. Kontele, K. P. Skenderi, M. Yannakoulia, and G. V. Dedoussis, "The gene-diet attica investigation on childhood obesity (gendai): overview of the study design," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 45, no. 3, pp. 309–315, 2007. [Online]. Available: <https://doi.org/10.1515/CCLM.2007.080>

- [3] V. Mamakou, S. Hackinger, E. Zengini, E. Tsompanaki, E. Marouli, I. Serafetinidis, B. Prins, A. Karabela, E. Glezou, L. Southam, N. W. Rayner, K. Kuchenbaecker, K. Lamnissou, V. Kontaxakis, G. Dedoussis, F. Gonidakis, A. Thanopoulou, N. Tentolouris, and E. Zeggini, "Combination therapy as a potential risk factor for the development of type 2 diabetes in patients with schizophrenia: the gomap study," *BMC Psychiatry*, vol. 18, no. 1, p. 249, Aug 2018. [Online]. Available: <https://doi.org/10.1186/s12888-018-1826-4>
- [4] E. Katsareli, C. Amerikanou, K. Rouskas, A. Dimopoulos, T. Diamantis, A. Alexandrou, J. Griniatsos, S. Bourgeois, E. Dermitzakis, J. Ragoussis *et al.*, "A genetic risk score for the estimation of weight loss after bariatric surgery," *Obesity surgery*, vol. 30, pp. 1482–1490, 2020. [Online]. Available: <https://doi.org/10.1007/s11695-019-04320-6>
- [5] K. Rouskas, E. A. Katsareli, C. Amerikanou, A. C. Dimopoulos, S. Glentis, A. Kalantzi, A. Skoulakis, N. Panousis, H. Ongen, D. Bielser *et al.*, "Identifying novel regulatory effects for clinically relevant genes through the study of the greek population," *BMC genomics*, vol. 24, no. 1, p. 442, 2023. [Online]. Available: <https://doi.org/10.1186/s12864-023-09532-w>
- [6] M. Romero-Gómez, F. Pollarsky, S. G. Rey, R. Gallego-Durán, V. García-Fernández, I. Fernández-Lizaranzu, F. J. A. Martín, P. Remon, C. Lara-Romero, M. C. Roque-Cuellar, I. Dominguez, Y. M. Sanchez-Torrijos, I.-P. Kalafati, S. Kahl, M. Roden, P. P. García-Luna, J. Schattenberg, G. Dedoussis, D. Fortin, A. Mourad, P. Carrieri, P. Marino, and J. V. Lazarus, "Wed-195 nash-pi: a randomized trial to compare cost-effectiveness across strategies for screening, referring and management of masld patients in clinical practice," *Journal of hepatology*, vol. 80, pp. S508–S509, 2024. [Online]. Available: <https://inserm.hal.science/inserm-05002721>
- [7] P. Giardoglou, "Functional analysis of genes involved in cardiovascular disease," Ph.D. dissertation, Harokopio University of Athens. School of Health Sciences and Education. Department of Nutrition and Dietetics, 2022. [Online]. Available: <https://estia.hua.gr/entities/publication/2af8ef0f-f473-46db-81f7-fc627ebde2c5>
- [8] A.-E. Farmaki, N. W. Rayner, A. Matchan, P. Spiliopoulou, A. Gilly, V. Kariakli, C. Kiagiadaki, E. Tsafantakis, E. Zeggini, and G. Dedoussis, "The mountainous cretan dietary patterns and their relationship with cardiovascular risk factors: the hellenic isolated cohorts manolis study," *Public Health Nutrition*, vol. 20, no. 6, p. 1063–1074, 2017. [Online]. Available: <https://doi.org/10.1017/S1368980016003207>
- [9] A.-E. Farmaki, N. W. Rayner, M. Kafyra, A. Matchan, K. Ntaoutidou, P. Feritoglou, A. Athanasiadis, A. Gilly, V. Mamakou, E. Zengini, M. Karaleftheri, E. Zeggini, and G. Dedoussis, "A dietary pattern with high sugar content is associated with cardiometabolic risk factors in the pomak population," *Nutrients*, vol. 11, no. 12, 2019. [Online]. Available: <https://www.mdpi.com/2072-6643/11/12/3043>
- [10] I. P. Kalafati, A. A. K. Korovesi, M. Papaioannou, C. Konstantaki, L. Angelidakis, V. K. Pachi, E. Maneta, D. Bampatsias, I. Petropoulos, E. Manios *et al.*, "Chios mastic improves night-time diastolic pressure in well-controlled patients with hypertension; preliminary results of the hyper-mastic clinical trial," *Journal of Hypertension*, vol. 41, no. Suppl 3, p. e274, 2023. [Online]. Available: https://journals.lww.com/jhypertension/fulltext/2023/06003/chios_mastic_improves_night_time_diastolic.742.aspx
- [11] M. Tsota, P. Giardoglou, E. Mentsiou-Nikolaou, P. Symianakis, I. P. Kalafati, A.-A. Kyriazopoulou-Korovesi, L. Angelidakis, M. Papaioannou, C. Konstantaki, H.-M. Consortium *et al.*, "Investigation of antihypertensive properties of chios mastic via monitoring microrna-21 expression levels in the plasma of well-controlled hypertensive patients," *Non-Coding RNA*, vol. 10, no. 3, p. 33, 2024. [Online]. Available: <https://doi.org/10.3390/ncrna10030033>
- [12] M. Kafyra, I. P. Kalafati, E. A. Katsareli, S. Lambrinou, I. Varlamis, A. C. Kaliora, and G. V. Dedoussis, "The improve study; design, dietary patterns, and development of a lifestyle index in overweight and obese greek adults," *Nutrients*, vol. 13, no. 10, p. 3495, 2021. [Online]. Available: <https://doi.org/10.3390/nu13103495>
- [13] M. Kafyra, I. P. Kalafati, G. Stefanou, G. Kourlaba, P. Moulos, I. Varlamis, A. C. Kaliora, and G. V. Dedoussis, "A dietary intervention in adults with overweight or obesity leads to weight loss irrespective of macronutrient composition," *Nutrients*, vol. 16, no. 17, 2024. [Online]. Available: <https://www.mdpi.com/2072-6643/16/17/2842>
- [14] M. Kafyra, I. P. Kalafati, I. Varlamis, A. C. Kaliora, P. Moulos, and G. V. Dedoussis, "A randomized clinical trial of a dietary intervention and mental health associations in adults with increased genetic risk for obesity," *Scientific reports*, vol. 15, no. 1, p. 14188, 2025. [Online]. Available: <https://doi.org/10.1038/s41598-025-99347-4>

- [15] I.-P. Kalafati, D. Borsas, M. Dimitriou, K. Revenas, A. Kokkinos, and G. V. Dedoussis, "Dietary patterns and non-alcoholic fatty liver disease in a greek case-control study," *Nutrition*, vol. 61, pp. 105–110, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S089990071830460X>
- [16] I. Kalafati, M. Dimitriou, D. Borsas, J. Vlachogiannakos, K. Revenas, A. Kokkinos, S. Ladas, and G. Dedoussis, "Fish intake interacts with tm6sf2 gene variant to affect nafld risk: results of a case-control study," *European journal of nutrition*, vol. 58, pp. 1463–1473, 2019. [Online]. Available: <https://doi.org/10.1007/s00394-018-1675-4>
- [17] C. Fotakis, A. I. Amanatidou, M. Kafyra, V. Andreou, I. P. Kalafati, M. Zervou, and G. V. Dedoussis, "Circulatory metabolite ratios as indicators of lifestyle risk factors based on a greek nafld case-control study," *Nutrients*, vol. 16, no. 8, 2024. [Online]. Available: <https://www.mdpi.com/2072-6643/16/8/1235>
- [18] C. Fotakis, I.-P. Kalafati, A. I. Amanatidou, V. Andreou, M. Matzapetakis, M. Kafyra, I. Varlamis, M. Zervou, and G. V. Dedoussis, "Serum metabolomic profiling unveils distinct sex-related metabolic patterns in nafld," *Frontiers in Endocrinology*, vol. Volume 14 - 2023, 2023. [Online]. Available: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1230457>
- [19] E. V. Grigoriou, G. Trovas, N. Papaioannou, P. Makras, P. Kokkoris, I. Dontas, K. Makris, S. Tournis, and G. V. Dedoussis, "Serum 25-hydroxyvitamin d status, quantitative ultrasound parameters, and their determinants in greek population," *Archives of osteoporosis*, vol. 13, pp. 1–9, 2018. [Online]. Available: <https://doi.org/10.1007/s11657-018-0526-5>
- [20] I. Ntalla, M. Giannakopoulou, P. Vlachou, K. Giannitsopoulou, V. Gkesou, C. Makridi, M. Marougka, G. Mikou, K. Ntaoutidou, E. Prountzou et al., "Body composition and eating behaviours in relation to dieting involvement in a sample of urban greek adolescents from the teenage (teens of attica: Genes & environment) study," *Public health nutrition*, vol. 17, no. 3, pp. 561–568, 2014. [Online]. Available: <https://doi.org/10.1017/s1368890013000074>
- [21] M. Kafyra, I. P. Kalafati, I. Gavra, S. Siest, and G. V. Dedoussis, "Associations of vegf-a-related variants with adolescent cardiometabolic and dietary parameters," *Nutrients*, vol. 15, no. 8, 2023. [Online]. Available: <https://www.mdpi.com/2072-6643/15/8/1884>
- [22] M. Dimitriou, L. S. Rallidis, E. V. Theodoraki, I. P. Kalafati, G. Kolovou, and G. V. Dedoussis, "Exclusive olive oil consumption has a protective effect on coronary artery disease; overview of the theiseas study," *Public Health Nutrition*, vol. 19, no. 6, pp. 1081–1087, 2016. [Online]. Available: <https://doi.org/10.1017/s1368890015002244>
- [23] M. Dimitriou, P. Moulos, I. P. Kalafati, G. Saranti, L. S. Rallidis, and G. V. Dedoussis, "Evaluation of polygenic risk scores for prediction of coronary artery disease in a greek case-control study," *Journal of Personalized Medicine*, vol. 14, no. 6, 2024. [Online]. Available: <https://www.mdpi.com/2075-4426/14/6/565>
- [24] World Health Organization, "Obesity and overweight," March 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>
- [25] W. M. Association., "World medical association declaration of helsinki. ethical principles for medical research involving human subjects," *Bulletin of the World Health Organization*, vol. 79, no. 4, pp. 373 – 374, 2001. [Online]. Available: <https://doi.org/10.1001/jama.2013.281053>
- [26] G. Eknoyan, "Adolphe quetelet (1796–1874)—the average man and indices of obesity," *Nephrology Dialysis Transplantation*, vol. 23, no. 1, pp. 47–51, 09 2007. [Online]. Available: <https://doi.org/10.1093/ndt/gfm517>
- [27] S. Miller, D. Dykes, and H. Polesky, "A simple salting out procedure for extracting dna from human nucleated cells," *Nucleic Acids Research*, vol. 16, no. 3, pp. 1215–1215, 02 1988. [Online]. Available: <https://doi.org/10.1093/nar/16.3.1215>
- [28] S. Purcell and C. Chang, "Plink 1.9." [Online]. Available: www.cog-genomics.org/plink/1.9/
- [29] R Core Team, R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: <https://www.R-project.org/>
- [30] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau et al., "A reference panel of 64,976 haplotypes for genotype imputation," *Nature genetics*, vol. 48, no. 10, pp. 1279–1283, 2016. [Online]. Available: <https://doi.org/10.1038/ng.3643>
- [31] D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang et al., "Sequencing of 53,831 diverse genomes from the nhlbi topmed program," *Nature*, vol. 590, no. 7845, pp. 290–299, 2021. [Online]. Available: <https://www.biorxiv.org/content/early/2019/03/06/563866>
- [32] C. Fuchsberger, G. R. Abecasis, and D. A. Hinds, "minimac2: faster genotype imputation," *Bioinformatics*, vol. 31, no. 5, pp. 782–784, 10 2014. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btu704>
- [33] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue et al., "Next-generation genotype imputation service and methods," *Nature genetics*, vol. 48, no. 10, pp. 1284–1287, 2016.
- [34] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.
- [35] K. Arnold, J. Gosling, D. Holmes, and D. Holmes, *The Java programming language*. Addison-wesley Reading, 2000, vol. 2.
- [36] S. B. Lee, K. C. Clabaugh, B. Silva, K. O. Odigie, M. D. Coble, O. Loreille, M. Scheible, R. M. Fournay, J. Stevens, G. R. Carmody et al., "Assessing a novel room temperature dna storage medium for forensic biological samples," *Forensic Science International: Genetics*, vol. 6, no. 1, pp. 31–40, 2012.
- [37] S. W. Smith and D. Safford, "Practical server privacy with secure coprocessors," *IBM Systems Journal*, vol. 40, no. 3, pp. 683–695, 2001.
- [38] J. Eder, H. Gottweis, and K. Zatloukal, "It solutions for privacy protection in biobanking," *Public Health Genomics*, vol. 15, no. 5, pp. 254–262, 06 2012. [Online]. Available: <https://doi.org/10.1159/000336663>
- [39] I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J. D. Spalding, S. Ur-Rehman, G. Saunders, J. Kandasamy, M. Caccamo, R. Leinonen et al., "The european genome-phenome archive of human data consented for biomedical research," *Nature genetics*, vol. 47, no. 7, pp. 692–695, 2015.
- [40] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the Practice of Statistics*. WH Freeman New York, 2009, vol. 4.
- [41] G. Casella and R. Berger, *Statistical inference*. CRC press, 2024.
- [42] R. Fisher, "Statistical methods and scientific induction," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 17, no. 1, pp. 69–78, 1955.
- [43] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.
- [44] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.
- [45] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.
- [46] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze*, vol. 8, pp. 3–62, 1936.
- [47] J. Chambers and T. Hastie, *Statistical Models in S*, ser. Wadsworth & Brooks/Cole computer science series. Wadsworth & Brooks/Cole Advanced Books & Software, 1992. [Online]. Available: <https://books.google.gr/books?id=uYfVAAAAAAAJ>
- [48] B. Thompson, "Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial," *Educational and Psychological Measurement*, vol. 55, no. 4, pp. 525–534, 1995. [Online]. Available: <https://doi.org/10.1177/0013164495055004001>
- [49] P. E. Green and J. D. Carroll, *Analyzing multivariate data*. Dryden Press Hinsdale, IL, 1978, vol. 519.
- [50] H. Akaike, "Factor analysis and aic," *Psychometrika*, vol. 52, pp. 317–332, 1987.
- [51] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
- [52] G. van Rossum and J. de Boer, "Interactively testing remote servers using the python programming language," *CWI Quarterly*, vol. 4, no. 4, pp. 283–304, Dec. 1991.
- [53] D. Clayton, *snpStats: SnpMatrix and XSNpMatrix classes and methods*, 2024, r package version 1.54.0. [Online]. Available: <https://bioconductor.org/packages/snpStats>
- [54] S. Fairley, E. Lowy-Gallego, E. Perry, and P. Flicek, "The international genome sample resource (igsr) collection of open human genomic variation resources," *Nucleic Acids Research*, vol. 48, no. D1, pp. D941–D947, 10 2019. [Online]. Available: <https://doi.org/10.1093/nar/gkz836>
- [55] G. Abraham, Y. Qiu, and M. Inouye, "Flashpca2: principal component analysis of biobank-scale genotype datasets," *Bioinformatics*, vol. 33, no. 17, pp. 2776–2778, 05 2017. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btx299>
- [56] S. W. Choi, P. F. O'Reilly, Monsanto pinheiro, wenjiany, and V. de Souza, "choishingwan/prscice: New compilation and unit testing," Mar. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3703335>

- [57] S. A. Lambert, B. Wingfield, J. T. Gibson, L. Gil, S. Ramachandran, F. Yvon, S. Saverimuttu, E. Tinsley, E. Lewis, S. C. Ritchie, J. Wu, R. C´acute;acute;novas, A. McMahon, L. W. Harris, H. Parkinson, and M. Inouye, "Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization." 2024. [Online]. Available: <https://www.repository.cam.ac.uk/handle/1810/373114>
- [58] S. A. Lambert, L. Gil, S. Jupp, S. Ritchie, Y. Xu, A. Buniello, G. Abraham, M. Chapman, H. Parkinson, J. Danesh, J. A. L. MacArthur, and M. Inouye, "The polygenic score catalog: an open database for reproducibility and systematic evaluation," *medRxiv*, 2020. [Online]. Available: <https://www.medrxiv.org/content/early/2020/05/23/2020.05.20.20108217>
- [59] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: <https://ggplot2.tidyverse.org>
- [60] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [61] F. Drasgow, "Polychoric and polyserial correlations," *Encyclopedia of statistical sciences*, 2004.
- [62] R. Becker, J. Chambers, and A. Wilks, *The New S Language: A Programming Environment for Data Analysis and Graphics*, ser. Computer science series. Wadsworth & Brooks/Cole Advanced Books & Software, 1988. [Online]. Available: <https://books.google.gr/books?id=pYcZAQAIAAJ>
- [63] A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, K. G. Aragam, E. S. Lander, G. D. Smith, H. Mason-Suares, M. Fornage, M. Lebo, N. J. Timpson, L. M. Kaplan, and S. Kathiresan, "Polygenic prediction of weight and obesity trajectories from birth to adulthood," *Cell*, vol. 177, no. 3, pp. 587–596.e9, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0092867419302909>
- [64] A. I. Campos, A. Mulcahy, J. G. Thorp, N. R. Wray, E. M. Byrne, P. A. Lind, S. E. Medland, N. G. Martin, I. B. Hickie, and M. E. Renteria, "Understanding genetic risk factors for common side effects of antidepressant medications," *Communications medicine*, vol. 1, no. 1, p. 45, 2021.
- [65] O. Weissbrod, M. Kanai, H. Shi, S. Gazal, W. J. Peyrot, A. V. Khera, Y. Okada, A. R. Martin, H. K. Finucane *et al.*, "Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores," *Nature Genetics*, vol. 54, no. 4, pp. 450–458, 2022.
- [66] R. Monti, L. Eick, G. Hudjashov, K. Läll, S. Kanoni, B. N. Wolford, B. Wingfield, O. Pain, S. Wharrie, B. Jermy, A. McMahon, T. Hartonen, H. Heyne, N. Mars, S. Lambert, K. Hveem, M. Inouye, D. A. van Heel, R. Mägi, P. Marttinen, S. Ripatti, A. Ganna, and C. Lippert, "Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning," *The American Journal of Human Genetics*, vol. 111, no. 7, pp. 1431–1447, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S000292972400209X>
- [67] B. Truong, L. E. Hull, Y. Ruan, Q. Q. Huang, W. Hornsby, H. Martin, D. A. van Heel, Y. Wang, A. R. Martin, S. H. Lee, and P. Natarajan, "Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases," *Cell Genomics*, vol. 4, no. 4, p. 100523, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666979X2400065X>
- [68] T. G. Richardson, E. Sanderson, B. Elsworth, K. Tilling, and G. Davey Smith, "Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study," *BMJ*, vol. 369, 2020. [Online]. Available: <https://www.bmj.com/content/369/bmj.m1203>
- [69] G. Hawkes, R. N. Beaumont, J. Tyrrell, G. M. Power, A. Wood, M. Laakso, L. Fernandes Silva, M. Boehnke, X. Yin, T. G. Richardson *et al.*, "Genetic evidence that high bmi in childhood has a protective effect on intermediate diabetes traits, including measures of insulin sensitivity and secretion, after accounting for bmi in adulthood," *Diabetologia*, vol. 66, no. 8, pp. 1472–1480, 2023.
- [70] E. Magriplis, D. Panagiotakos, E. Papakonstantinou, A.-V. Mitsopoulou, D. Karageorgou, I. Dimakopoulos, I. Bakogianni, M. Chourdakis, R. Micha, G. Michas, T. Ntouriou, S.-M. Tsaniklidou, K. Argyri, G. Dimitriadis, A. Zampelas, E. Fappa, E.-M. Theodoraki, E. Trichia, T.-E. Sialvera, A. Varytimiadi, E. Spyrelli, A. Koutelidakis, G. Karlis, S. Zacharia, A. Papageorgiou, G. Chrousos, G. Dedoussis, Y. Manios, and E. Roma, "Prevalence of type 2 diabetes mellitus in a representative sample of greek adults and its association with modifiable risk factors: results from the hellenic national nutrition and health survey," *Public Health*, vol. 197, pp. 75–82, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0033350620304443>
- [71] E. Magriplis, D. Panagiotakos, I. Kyrou, C. Tsioufis, A.-V. Mitsopoulou, D. Karageorgou, I. Dimakopoulos, I. Bakogianni, M. Chourdakis, R. Micha, G. Michas, T. Ntouriou, S.-M. Tsaniklidou, K. Argyri, and A. Zampelas, "Presence of hypertension is reduced by mediterranean diet adherence in all individuals with a more pronounced effect in the obese: The hellenic national nutrition and health survey (hnnhs)," *Nutrients*, vol. 12, no. 3, 2020. [Online]. Available: <https://www.mdpi.com/2072-6643/12/3/853>
- [72] E. Magriplis, I. Dimakopoulos, D. Karageorgou, A.-V. Mitsopoulou, I. Bakogianni, R. Micha, G. Michas, T. Ntouriou, S.-M. Tsaniklidou, K. Argyri *et al.*, "Aims, design and preliminary findings of the hellenic national nutrition and health survey (hnnhs)," *BMC medical research methodology*, vol. 19, pp. 1–13, 2019.
- [73] G. S. Stergiou, A. Ntineri, A. Menti, N. Kalpourtzi, C. Vlachopoulos, E. N. Liberopoulos, L. Rallidis, D. Richter, M. Gavana, A. Vantarakis *et al.*, "Twenty-first century epidemiology of dyslipidemia in greece: Emeno national epidemiological study," *Hellenic Journal of Cardiology*, vol. 69, pp. 1–8, 2023.
- [74] World Health Organization, "Greece: Country profile," European Health Information Gateway, 2024, accessed: 10 March 2025. [Online]. Available: <https://gateway.euro.who.int/en/country-profiles/greece/>
- [75] J. Strelitz, A. L. Ahern, G. H. Long, M. J. Hare, G. Irving, C. E. Boothby, N. J. Wareham, and S. J. Griffin, "Moderate exercise and weight-control following diabetes diagnosis and 10 year incidence of cardiovascular disease and mortality," *Diabetologia*, vol. 62, pp. 1391–1402, 2019.
- [76] C.-Y. Hsu, C.-Y. Huang, C.-H. Hsieh, P.-C. Chien, C.-C. Chen, S.-Y. Hou, and S.-C. Wu, "Regular exercise and weight-control behavior are protective factors against osteoporosis for general population: A propensity score-matched analysis from taiwan biobank participants," *Nutrients*, vol. 14, no. 3, 2022. [Online]. Available: <https://www.mdpi.com/2072-6643/14/3/641>
- [77] E. Magriplis, G. Michas, E. Petridi, G. P. Chrousos, E. Roma, V. Benetou, N. Choleopoulos, R. Micha, D. Panagiotakos, and A. Zampelas, "Dietary sugar intake and its association with obesity in children and adolescents," *Children*, vol. 8, no. 8, 2021. [Online]. Available: <https://www.mdpi.com/2227-9067/8/8/676>
- [78] W. H. Organization *et al.*, "Who european childhood obesity surveillance initiative (cosi)," *Report on the fourth round of data collection*, vol. 2017, 2015.
- [79] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.
- [80] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, and the GIANT Consortium, "Meta-analysis of genome-wide association studies for height and body mass index in ≈ 700000 individuals of european ancestry," *Human Molecular Genetics*, vol. 27, no. 20, pp. 3641–3649, 10 2018. [Online]. Available: <https://doi.org/10.1093/hmg/ddy271>
- [81] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly, "Tutorial: a guide to performing polygenic risk score analyses," *Nature protocols*, vol. 15, no. 9, pp. 2759–2772, 2020.
- [82] M. Shi, W. Chen, X. Sun, L. A. Bazzano, J. He, A. C. Razavi, C. Li, L. Qi, A. V. Khera, and T. N. Kelly, "Association of genome-wide polygenic risk score for body mass index with cardiometabolic health from childhood through midlife," *Circulation. Genomic and precision medicine*, vol. 15, no. 4, p. e003375, August 2022. [Online]. Available: <https://www.ahajournals.org/doi/pdf/10.1161/CIRCGEN.121.003375>
- [83] T. A. Foster and G. S. Berenson, "Measurement error and reliability in four pediatric cross-sectional surveys of cardiovascular disease risk factor variables—the bogalusa heart study," *Journal of Chronic Diseases*, vol. 40, no. 1, pp. 13–21, 1987. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0021968187900920>