scientific reports

OPEN

A comprehensive analysis of genetic risk for metabolic syndrome in the Egyptian population via allele frequency investigation and Missense3D predictions

Mahmoud Bassyouni[1,2], Mohamed Mysara[1,3], Inken Wohlers[4,5], Hauke Busch[4,6], Maha Saber-Ayad[7,8]✉ & Mohamed El-Hadidi[1,9]✉

Diabetes mellitus (DM) represents a major health problem in Egypt and worldwide, with increasing numbers of patients with prediabetes every year. Numerous factors, such as obesity, hyperlipidemia, and hypertension, which have recently become serious concerns, affect the complex pathophysiology of diabetes. These metabolic syndrome diseases are highly linked to genetic variability that drives certain populations, such as Egypt, to be more susceptible to developing DM. Here we conduct a comprehensive analysis to pinpoint the similarities and uniqueness among the Egyptian genome reference and the 1000-genome subpopulations (Europeans, Ad-Mixed Americans, South Asians, East Asians, and Africans), aiming at defining the potential genetic risk of metabolic syndromes. Selected approaches incorporated the analysis of the allele frequency of the different populations' variations, supported by genotypes' principal component analysis. Results show that the Egyptian's reference metabolic genes were clustered together with the Europeans', Ad-Mixed Americans', and South-Asians'. Additionally, 8563 variants were uniquely identified in the Egyptian cohort, from those, two were predicted to cause structural damage, namely, CDKAL1: 6_21065070 (A > T) and PPARG: 3_12351660 (C > T) utilizing the Missense3D database. The former is a protein coding gene associated with Type 2 DM while the latter is a key regulator of adipocyte differentiation and glucose homeostasis. Both variants were detected heterozygous in two different Egyptian individuals from overall 110 sample. This analysis sheds light on the unique genetic traits of the Egyptian population that play a role in the DM high prevalence in Egypt. The proposed analysis pipeline -available through GitHub- could be used to conduct similar analysis for other diseases across populations.

Egypt is one of the top ten countries with the highest prevalence of diabetes mellitus, according to the International Diabetes Federation (IDF). The number of diabetic patients in the Middle East and North Africa (MENA) is projected to increase from 34.6 million in 2013 to 67.9 million in 2035, with a 96% increase. In

[1]Bioinformatics Group, Center for Informatics Sciences (CIS), School of Information Technology and Computer Science (ITCS), Nile University, Giza, Egypt. [2]Bioscience Research Laboratories Department, MARC for Medical Services and Scientific Research, 6th of October, Jiza, Egypt. [3]Microbiology unit, Belgian Nuclear Research Centre (SCK CEN), Mol, Belgium. [4]Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology, and Institute for Cardiogenetics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany. [5]Biomolecular Data Science in Pneumology, Research Center Borstel, 23845 Borstel, Germany. [6]University Cancer Center Schleswig-Holstein, University Hospital of Schleswig-Holstein, Campus Lübeck, 23538 Lübeck, Germany. [7]Department of Clinical Sciences, College of Medicine, University of Sharjah, 27272, Sharjah, UAE. [8]Pharmacology Department, College of Medicine, Cairo University, Cairo 12613, Egypt. [9]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham Dubai Campus, Dubai, United Arab Emirates. ✉email: msaber@sharjah.ac.ae; melhadidi@nu.edu.eg

Egypt, individuals between the age of 20 and 79 years have a diabetes prevalence of around 20.9%, and the disease accounted for 122,684 deaths in the year 2021as reported in the IDF Atlas 10th edition 2021[1]. In the same report, the IDF calculated that 10.93 million Egyptians reported having diabetes, while 6.8 million have undiagnosed form of the disease. In addition, estimates suggest that 62% of Egyptians with diabetes and the majority of those with prediabetes are probably undiagnosed. The dramatic rise in diabetes prevalence in Egypt from around 7.3 million in 2011 to 10.9 million in 2021 over a very short period is concerning, especially that by the year 2045, it is anticipated that this number would increase to 19.9 million. Apart from being a significant public health issue, diabetes is predicted to have cost the Middle East area $13.6 billion in 2013 (14% of its overall health care expenditures), which represents just 2.5% of the disease's global expenditures. The economic burden of Type 2 Diabetes (T2D) in Egypt was assessed by cost experts to be $1.29 billion in 2010, not considering expenses connected to prediabetes or lost productivity that is likely to be doubled by the year 2030[2].

Several risk factors have been attributed to the high prevalence of diabetes—particularly in Egypt- including physical inactivity and obesity, particularly visceral adiposity[2]. The Egypt National STEPwise Survey for Non-communicable Diseases Risk Factors Report (2017), a national household survey on citizens aged 15–69 years old, reported that approximately 35.7% of adults were obese (BMI > 30 kg/m^2), with a prevalence of 48.8% among women and 24.8% among men[3]. In the same report, it has been shown that obesity percentage increased from 31.3% at 2012 and reached 35.7% in 2017. However, according to a more recent survey -100 million health survey- conducted in Egypt in 2019 including 49.7 million adult Egyptian citizens (≥ 18 years old), 39.8% suffered from obesity (BMI ≥ 30 kg/m^2). Prevalence of obesity in females was revealed to be more than males from the same age group, with 49.5% for adult females compared to 29.5% of males[4].

A group of disorders known as metabolic syndrome increases the chance of developing heart disease, stroke, and type 2 diabetes. Hypertension, Obesity, and Hyperlipidemia are among these problems. According to Abd Elaziz et al., a high metabolic syndrome prevalence of 55% was found in Egyptians, 85.6% among diabetics, and 76.6% among hypertensive patients[5]. The prevalence of hypertension in Egypt is notably high, while the rates of awareness, treatment, and control are comparatively low. The coexistence of other cardiovascular risk factors exacerbates hypertension in 60% of patients, leading to heightened cardiovascular morbidity and mortality[6]. According to the World Health Organisation (WHO) 2020 hypertension Egypt report, it is estimated that 17.8 million of the Egyptian population in 2017 had hypertension, with 15.4 million not having it under control[7]. Furthermore, the occurrence of dyslipidemia exhibited variability amongst the overall population of Egypt, with a range of 19.2–36.8%. However, the incidence of dyslipidemia was comparatively greater in individuals diagnosed with acute coronary syndrome (ACS) at 50.9 and 52.5%, and those with coronary artery disease at 58.7%. According to a national report, dyslipidemia screening was conducted on 8.6% of the overall population. However, there is a lack of data regarding the rates of diagnosis and treatment[8].

The high prevalence of DM and the other metabolic traits are characterized by a wide range of DNA sequence variants, that play a role in phenotypical/pathophysiology of the diseases[9]. High-impact variants, with allele frequencies below 0.5%, cause monogenic and syndromic metabolic diseases like MODY and familial hypercholesterolaemia[10,11]. These diseases typically appear early in life and have a clear familial pattern. Late-onset metabolic diseases, on the other hand, are primarily influenced by common variants, with allele frequencies greater than 5%. Genome-wide association studies have identified hundreds to thousands of these common variants, each with a minor effect on disease risk[12–15]. Recent research with larger sample sizes and sequence data has shown that the genetic predisposition to prevalent metabolic traits in later life is primarily due to a broad distribution of common allele effects that gradually decrease in impact[16,17].

In this work, we aim to assess the potential genetic risk of metabolic syndrome and identify the commonalities and distinctive genetic traits between the Egyptian and the 1000-genome subpopulations (Europeans, Ad-Mixed Americans, South Asians, East Asians, and Africans). For this purpose, the metabolic syndrome using the allele frequency of different populations' variations, supported by genotypes' principal component analysis, was assessed. Additionally, the analysis steps are presented as a single pipeline -through GitHub- to analyse similar diseases among various populations.

Results

Intersection analysis of diabetes type II with obesity, hypertension and hyperlipidemia: annotation and characterization of associated genes

Here we explore the genetic overlap between T2D, Hyperlipidemia, Obesity, and Hypertension. Using the search criteria described in the methods, we were able to identify 4 datasets from Ensembl. In total 3164, 633, 380, and 1064 single nucleotide variants were found associated with T2D, Hypertension, Obesity, and Hyperlipidemia, respectively. A total of 18 Single Nucleotide Variants (SNVs) were shared between the T2D and the other disorders. Among these, two SNVs were shared with Hyperlipidemia, seven SNVs were shared with Obesity, and nine SNVs were shared with Hypertension (Fig. 1A). Those 18 variants were identified to be linked to various genes using the Variant Effect Predictor (VEP) tool from Ensembl. When analyzing the impact of these variants on upstream and downstream effects, we found that they affect a total of 19 genes, 14 of which were protein coding (Supplementary Table S1), while the other five were pseudogenes and RNA genes (Supplementary Table S2). Results of the Hypergeometric tests of the STRING database network analysis for the protein coding genes (Fig. 1B) showed significant interactions with Protein–Protein Interaction (PPI) enrichment p-value of 3.92×10^{-12} with most of the disease-gene associations be T2D, Familial Hyperlipidemia, and Acquired Metabolic Disease.

Since all the shared variations were single nucleotide variants (SNVs), the first intersection (Fig. 1A), yielded the variants ***rs268*** and ***rs328***. These variants are associated with the LPL gene and exhibit missense and stop-gained effects, respectively (Table 1). The second intersection included the variants ***rs1421085***, ***rs1800437***,

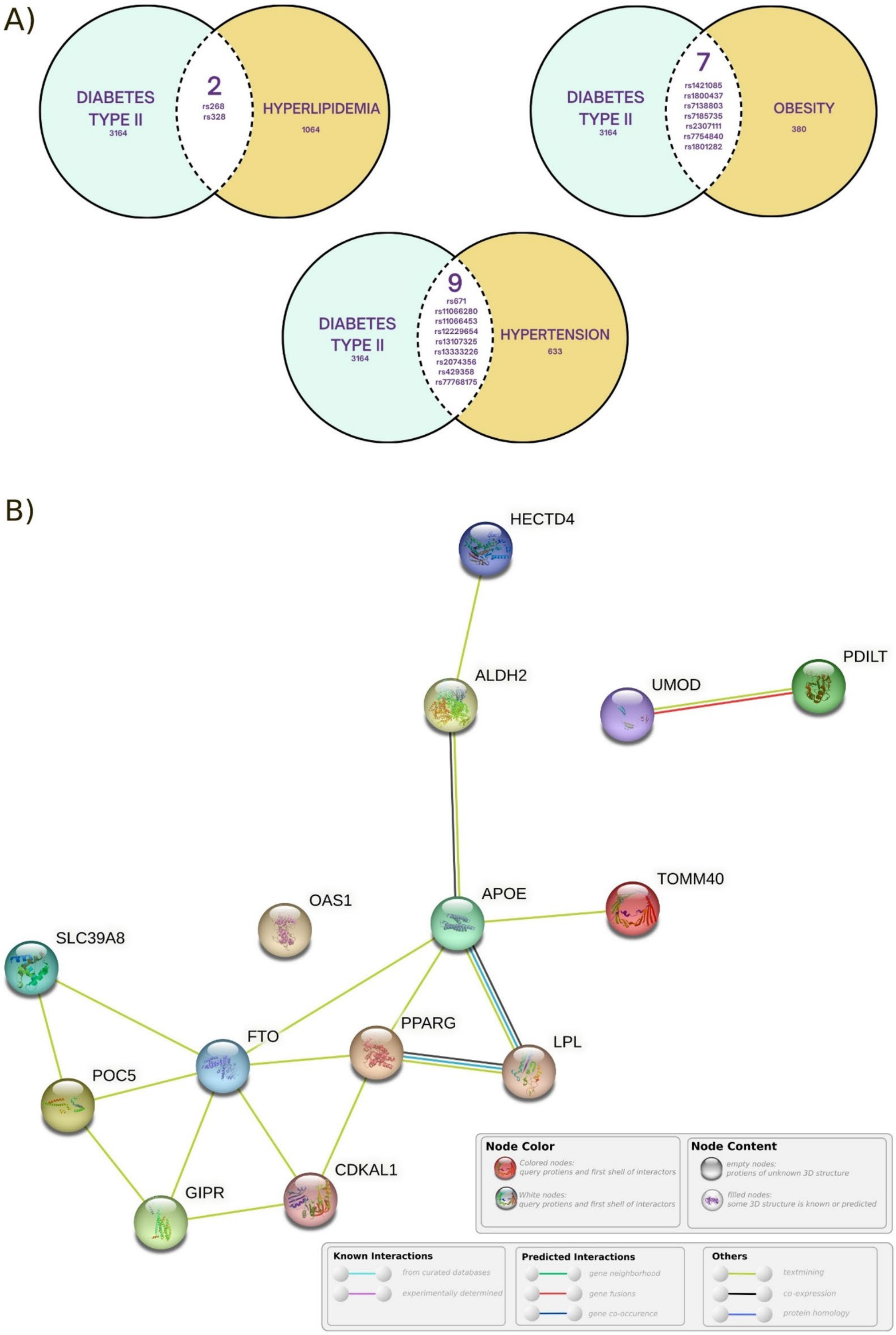


Figure 1. Protein-Protein Interaction network analysis from STRING database and the intersection results of the metabolic diseases variants with T2D. (**A**) Venn Diagrams represent the genetic overlap between T2D and each of Hyperlipidemia, Obesity, and Hypertension as -collectively- the figure illustrates the shared variants between these conditions. Two variations are shared between T2D and Hyperlipidemia. In a similar fashion, Obesity and Hypertension share 7 and 9 variants, respectively. (**B**) Genes network analysis created by STRING database, showing the possible interactions between each and every protein coding gene. The light green colour edge that nearly links every two adjacent nodes refers to the text mining legend which means that they are co-mentioned in PubMed Abstracts. More edges consequently mean stronger relationships between the two genes.

Variant ID	Gene	Consensus	Position	Allele	Global MAF
rs11066280	HECTD4	Intron variant	chr12:112379979	T > A/T > G	0.04 (A)
rs11066453	OAS1	Intron variant	chr12:112927816	A > G	0.025 (G)
rs12229654	Intergenic variant		chr12:110,976,657	T > G	0.032 (G)
rs13107325	SLC39A8	Missense variant	chr4:102267552	C > A/C > T	0.024 (T)
rs13333226	UMOD	Intron variant	chr16:20354332	A > G	0.024 (G)
rs1421085	FTO	Intron variant	chr16:53767042	T > C	0.23 (C)
rs1801282	PPARG	Missense variant	chr3:12351626	C > G/C > T	0.07 (G)
rs2074356	HECTD4	Intron variant	chr12:112207597	G > A	0.026 (A)
rs2307111	POC5	Missense variant	chr5:75707853	T > A/T > C	0.38 (T)
rs268	LPL	Missense variant	chr8:19956018	A > G	0.005 (G)
rs328	LPL	Stop gained	chr8:19962213	C > A/C > G	0.09 (G)
rs429358	APOE	Missense variant	chr19:44908684	T > C	0.15 (C)
rs671	ALDH2	Missense Variant	chr12:111803962	G > A	0.035 (A)
rs7185735	FTO	Intron variant	chr16:53788739	A > G/A > T	0.34 (G)
rs7754840	CDKAL1	Intron variant	chr6:20661019	G > A/G > C/G > T	0.4 (C)
rs77768175	HECTD4	Intron variant	chr12:112298314	A > G	0.032 (G)
rs7138803	Intergenic variant		chr12:49,853,685	G > A/G > T	0.26 (A)
rs1800437	GIPR	Missense variant	chr19:45678134	G > C	0.16 (C)

Table 1. Provides the list of intersected variants' annotation along with relevant details about each one. The variants are designated by their variant ID, and the genes to which they correspond are provided. In the consensus column, the type of variation, such as Intron Variant, Missense Variant, or Stop Gained, is indicated. In the Position column, the chromosomal location of the variant is specified. The Allele column displays the observed allele variations, including the reference and alternative alleles. The Global MAF column displays the results from the 1000 Genomes Project Phase 3.

rs7138803, ***rs7185735***, ***rs2307111***, ***rs7754840***, and ***rs1801282***. Among these variants, ***rs1421085*** and ***rs1800437*** correspond to the FTO and GIPR genes, respectively, and are characterized as intronic and missense variants. The remaining variants in this intersection are found in intergenic regions or exhibit intronic or missense effects. The third intersection, which involved Hypertension, encompassed the variants ***rs671***, ***rs11066280***, ***rs11066453***, ***rs12229654***, ***rs13107325***, ***rs13333226***, ***rs2074356***, ***rs429358***, and ***rs77768175***. Notably, ***rs671, rs13107325***, and ***rs429358*** are associated with the ALDH2, SLC39A8, and APOE genes -respectively- and represent missense variants. The other variants in this intersection are either intronic or of intergenic nature. These identified variants provide further insights into the genetic junctions between T2D and the aforementioned metabolic diseases.

Genotypes PCA and heatmap analysis of allele frequencies matrix reveal population clustering

The examination of the metabolic genetic variants across the current range of populations encompassing Egyptians, East Asians, Europeans, South Asians, Ad-Mixed Americans, and Africans, elucidated interesting patterns and interrelationships. The upset (Fig. 2B) and Venn intersections (Supplementary Fig. S1) revealed that Africans had the largest number of unique variants with 27,794 followed by East and South Asians with 17,308 and 16,718, respectively. However, the European population showed 10,192 unique variants, Egyptians and Ad-Mixed Americans were the closest and the least in uniqueness with 8563 and 8034 with the latter being lesser. It is worth mentioning that all the sets shared 11,958 unique variants. Notably, the Egyptians demonstrated significant genetic overlap with Africans, with 1,678 variants in common. In contrast, a smaller number of common variants was observed between East Asians, Europeans, and South Asians, indicating a degree of genetic distinctiveness for Egyptians in relation to these populations.

The intricate genetic structure of the Egyptian population, suggesting both shared ancestry and potential unique evolutionary trajectories, is further emphasized by the analysis of intersection patterns. Notably, the intersection between Egyptians, Ad-mixed Americans, and Africans encompassed 3,788 variants, highlighting shared genetic components among these groups. Conversely, Egyptians exhibited limited intersections with East Asians, Europeans, and South Asians, indicating a relative genetic differentiation from these populations.

Unlike the variants intersection, the results obtained from the genotypes and allele frequencies principal component analysis (PCA) and heatmap demonstrated a noticeable clustering pattern, whereby individuals of European, Ad-Mixed American, and South Asian descent were observed to cluster together, with Egyptians exhibiting a subsequent clustering tendency. Distinct clusters were observed among individuals of African and East Asian ancestry. A heatmap and a PCA generated for the allelic frequency (Fig. 2A–E) data exhibited analogous clustering patterns to those observed in the genotypes PCA (Fig. 2C–D). The observed clustering of the Egyptians AF and genotypes with those of Europeans, Ad-Mixed Americans, and South Asians is noteworthy. Conversely, the genotypes of Africans and East Asians showed distinct genetic characteristic reflecting the complex interplay of genetic diversity, historical migrations, and population dynamics.

In summary, the analysis of variant set intersections, allele frequencies, and genotypes' principal component analysis reveals a contrasting picture in the genetic relationships involving Egyptians. While the variant set intersections indicate a substantial sharing of variants with Africans, the AF and PCA analyses demonstrate a closer clustering of Egyptians with Europeans, Ad-Mixed Americans, and South Asians. These contrasting results suggest a complex genetic landscape for Egyptians, characterized by both shared genetic affinity with Africans and genetic similarities with Europeans, Ad-Mixed Americans, and South Asians.

Analysis of two protein-coding SNPs with structural damage potential in CDKAL1 and PPARG genes

To assess the potential functional and structural implications of protein coding variants that are unique to the Egyptian cohort, an analysis was conducted using the Missense3D database, in combination with protein structure data obtained from the AlphaFold database. The information regarding the VEP output and the filtered SNVs can be found in (Supplementary Table S3) and (Supplementary Table S4), respectively. Results of the analysis of a total of 60 protein-coding SNPs (Supplementary Table S5) predicted two SNVs in CDKAL1 and PPARG to be causing structural damage. It is noteworthy that both variants were identified as heterozygous in two Egyptian individuals out of a total of 110 Egyptians with a Minor Allele Frequency (MAF) of 0.0045 represented by a half genotype. However, to accurately determine the prevalence of the identified variants and rule out the potential for their rarity or exclusivity within the Egyptian population, it is imperative to obtain larger cohorts of individuals with Egyptian ancestry.

The detailed characteristics and outcomes of this analysis are presented in (Table 2). The table encompasses various information, including the chromosome and position of each SNV, the alternative allele observed, the gene symbol associated with the variant, existing variations as reported by the VEP, the protein position affected by the SNV, the resultant amino acid change, and the findings derived from the Missense3D analysis. Notably, the Missense3D analysis revealed specific consequences for each SNV, such as buried hydrogen bond breakage for the CDKAL1 variant (represented as rs756851756) at protein position 360 and altered cavity configuration for the PPARG variant at position 23, resulting in an amino acid change from alanine (A) to valine (V).

The structural impact of these variants was further visualized through the comparison of wild-type and mutant structures of the respective proteins, as illustrated in (Fig. 3A–B) and Figure (Fig. 3C–D). The observed structural alterations, including the disruption of hydrogen bonds and contraction of protein cavities, provide crucial insights into the potential functional consequences of these protein-coding variants specific to the Egyptian population.

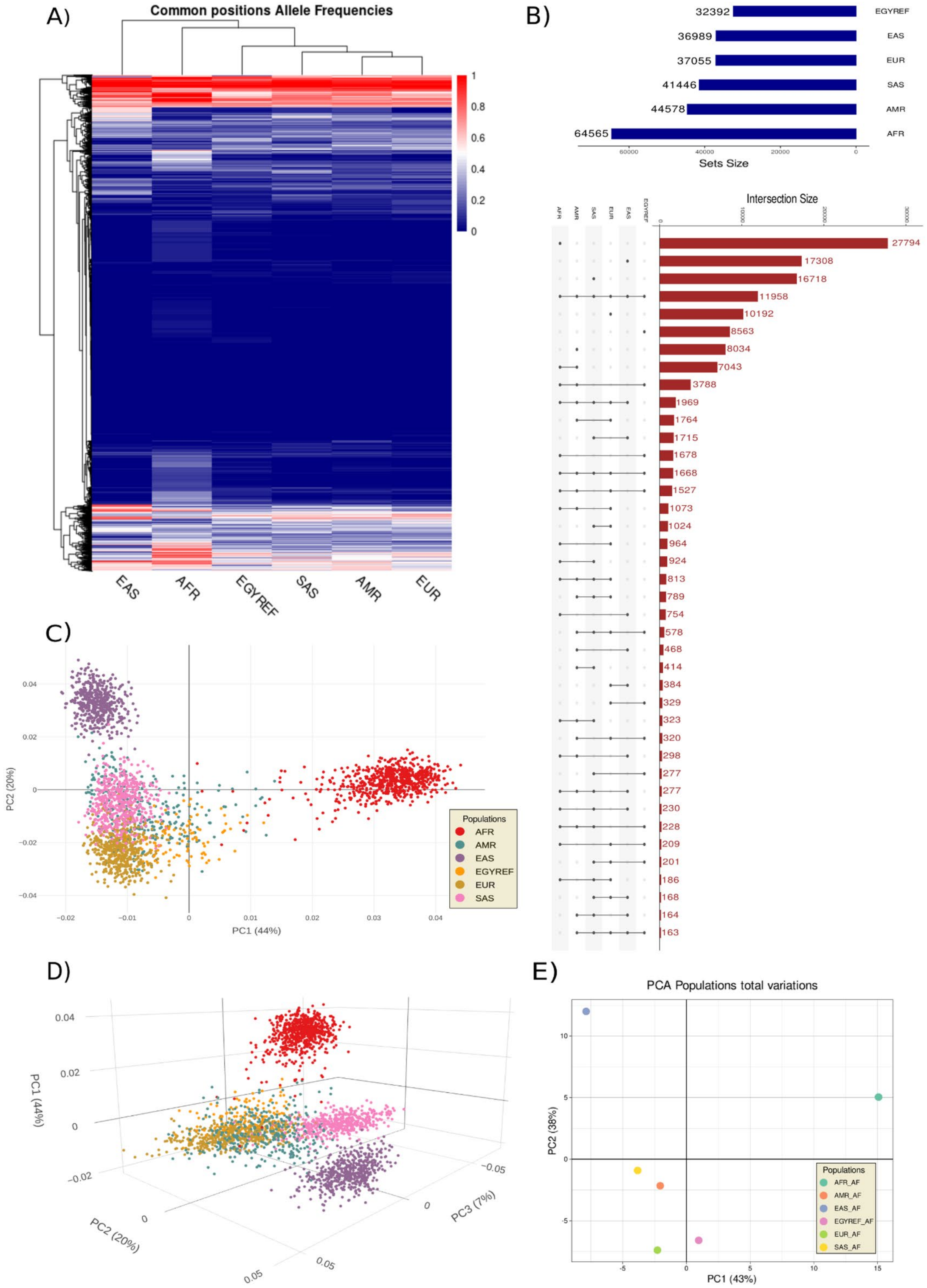
Discussion and conclusion

Metabolic syndrome diseases including Hypertension, Hyperlipidemia, and Obesity are highly linked to T2D. Those are believed to be linked to genetic factors that are different across populations. In this work we are trying to identify the unique genetic traits in the Egyptian population that contribute to the high prevalence of metabolic syndrome and its connection with the T2D. Our analysis revealed that genes linking T2D with different other components of metabolic syndrome (namely: obesity, hypertension, and dyslipidemia) exhibit high number of SNPs in the Egyptian cohort. We will discuss the genes of interest at the pathophysiological level.

Our current study revealed that FTO variants are shared among diabetes and obesity. Certain genetic variants in the fat mass and obesity-associated (FTO), were reported to be associated with obesity[18,19] in addition to their effect on food preference patterns such as increased total energy intake, in particular carbohydrate consumption[20,21]. The FTO gene was strongly linked with the development of obesity[22]. Among tens of loci identified to be associated with obesity, FTO gene has been identified as one of the influential genes with significant impact[23]. The FTO gene variants has been reported to associate with increased energy, fat, and protein intake[23,24]. The link of FTO and diabetes (and its complications) has been previously investigated. The study of Bego et al. revealed a significant association of FTO genetic variant rs8050136 A > C with the major markers of insulin resistance, obesity, and inflammation, in the population of West Balkan region[25]. Our previous study reported an association between FTO rs9939609 "A" allele and the impaired fasting glucose and insulin resistance in Emirati population[26]. Obesity and insulin resistance are clearly underlying the pathophysiology that leads to T2D.

Also, GIPR was shared between obesity and diabetes. The Gastric inhibitory polypeptide receptor (also called Glucose-dependent insulinotropic peptide) has been identified as one of two incretin hormones, linking nutrient intake to systemic metabolism[26,27]. The other incretin is Glucagon-like peptide-1 (GLP-1). Intriguingly, a paradox has emerged upon contradictory findings on the GIPR agonists versus antagonists as potential anti-obesity therapies. Targeting the pathways of endogenous nutrient-stimulated hormones allowed for better efficacy with acceptable safety, according to recent research using long-acting glucagon-like peptide-1 (GLP-1) receptor agonists[26,28]. Another attractive target is GIP, as it controls energy balance in the brain and adipose tissue via signalling through cell-surface receptors (GIP)[26,29]. An agonist that combines GIP and GLP receptor stimulation, Tirzepatide, has been recently approved by the FDA as a new treatment of diabetes mellitus type 2[30]. In addition, it led to sustained remarkable reductions in body weight, according to the SURMOUNT-1 trial, NCT04184622[31]. Our findings emphasize the link between obesity and diabetes through GIPR SNVs. We postulate that such SNVs may lead to GIP dysfunction, with impaired insulin secretion in response to the incretin effect in such patients.

Furthermore, peroxisome proliferator-activated receptor gamma is a nuclear receptor and drug target for insulin sensitizers and hypolipidemic agents. The encoded protein of *PPAR-γ* gene is a regulator of adipocyte differentiation. Moreover, *PPAR-γ* has been involved in the pathogenesis of obesity, diabetes, atherosclerosis, and cancer. A class of insulin sensitizers (thiazolidinediones) are basically agonists of *PPAR-γ*. PPAR-ligands support the storage of fatty acids in fat depots and control the expression of hormones released by adipocytes that


A)
Common positions Allele Frequencies
1
0.8
0.6
0.4
0.2
0
EAS
AFR
EGYREF
SAS
AMR
EUR
B)
32392
36989
37055
41446
44578
64565
EGYREF
EAS
EUR
SAS
AMR
AFR
60000
40000
20000
0
Sets Size
Intersection Size
AFR
AMR
SAS
EUR
EAS
EGYREF
0
10000
20000
30000
27794
17308
16718
11958
10192
8563
8034
7043
3788
1969
1764
1715
1678
1668
1527
1073
1024
964
924
813
789
754
578
468
414
384
329
323
320
298
277
277
230
228
209
201
186
168
164
163
C)
0.04
0.02
0
−0.02
−0.04
PC2 (20%)
−0.02
−0.01
0
0.01
0.02
0.03
0.04
PC1 (44%)
Populations
AFR
AMR
EAS
EGYREF
EUR
SAS
D)
0.04
0.02
0
−0.02
PC1 (44%)
0
0.05
PC2 (20%)
−0.05
0
0.05
PC3 (7%)
E)
PCA Populations total variations
10
5
0
-5
PC2 (38%)
-5
0
5
10
15
PC1 (43%)
Populations
AFR_AF
AMR_AF
EAS_AF
EGYREF_AF
EUR_AF
SAS_AF

◂**Figure 2.** Genetic Variation of the metabolic syndrome genes Among Six Populations: Insights from Allele Frequencies and Genotypes. (**A**) and (**E**) Allele Frequencies (AF) Heatmap and Principal Component Analysis (PCA): In this panel, the analysis was based on the allele frequencies (AF) of shared variants. The PCA and heatmap analyses were conducted to investigate the relationships and clustering patterns among individuals. The AF heatmap (**A**) and principal component analysis (**E**) revealed distinct clustering patterns. Individuals of European, Ad-Mixed American, and South Asian descent clustered, while Egyptians demonstrated a subsequent tendency to cluster. Africans and East Asians formed distinct clusters, implying their genetic profiles are distinct. (**B**) The plot represents the genetic variation present in six distinct populations, namely Africans, Ad-Mixed Americans, East Asians, Egyptians, Europeans, and South Asians. The present study employed the UpSetR package to investigate the intersection of variants, thereby elucidating the count of distinct variants for every population. The dots within the matrix symbolize the active involvement of a specific population in an inclusive intersection. Meanwhile, the edges connecting these dots indicate the participation of other populations in this intersection. The height of the dots represents the number of elements within that set, with taller columns indicating larger sets. (**C**) 2D and (**D**) 3D Genotypes Principal Component Analysis (PCA) Visualization. Genotypes were pruned for MAF of 0.05 with indep-pairwise of 50 5 0.5. The consistent clustering patterns observed in the genotypes' PCA and AF heatmap indicates a strong correlation. Particularly noteworthy is the grouping of Egyptian genotypes with those of Europeans, Ad-Mixed Americans, and South Asians, which suggests significant genetic similarities among these populations. In contrast, African genotypes exhibited separate clustering from both East Asians and the remaining genotypes, emphasizing the distinct genetic profiles of both Africans and East Asians. 2D and 3D HTML interactive plots can be found in the (Supplementary Files S1 and S2), respectively. A Bar plot for the proportional variances of the PCA is depicted in (Supplementary Fig. S2).

affect glucose homeostasis. Improved insulin sensitivity is the result of the PPAR-ligands' pleiotropic activities[32]. In general, PPARs control the expression of genes involved in the inflammation through controlling different pathways of inflammatory response[33]. The link of obesity, T2D, inflammation and cancer has been thoroughly investigated and reported to involve, at least in part, PPAR and their associated pathways[34].

In our analysis, LPL was the single gene shared between T2D and hyperlipidemia, with an obvious functional link. LPL gene encodes lipoprotein lipase; an enzyme expressed in the heart, muscle, and adipose tissue. LPL functions as a homodimer and has the dual functions of triglyceride hydrolase and ligand/bridging factor for receptor-mediated lipoprotein uptake. Interestingly, LPL mutations lead to a spectrum of LPL deficiency with the severest form known as type I hyperlipoproteinemia. An epidemiological study linked Lipoprotein lipase to insulin resistance, vitamin D and T2D in the Chinese[35]. Interestingly, Puri et al. found out that a significant metabolic reprogramming occurs when the diabetic heart cannot use lipoprotein lipase to control its own FA supply. This occurs with increasing diabetes severity and is linked to how the heart may handle excess fatty acids coming from adipose tissue. This change causes a cardiac metabolic profile that includes oxidative stress, triglyceride accumulation, mitochondrial FA excess, and cell death[36].

Several genes were shared between T2D and hypertension. In a study by Andreassen et al., HECTD4 was identified as an independent Locus associated with systolic blood pressure and high LDL through conditional False Discovery Rate (FDR; <0.01)[36]. HECTD4 SNPs were reported in association with alcohol consumption, with significant increased risk of type 2 diabetes[37]. Moreover, uromodulin is the most abundant protein in mammalian urine under physiological conditions. This protein may act as a constitutive inhibitor of calcium crystallization in renal fluids. Its excretion in urine may provide defence against urinary tract infections caused by uropathogenic bacteria. Uromodulin was reported to have a causal and adverse effect on kidney function, being involved in salt reabsorption via the NKCC2 ($Na^{+-}K^{+-}2Cl^{-}$ cotransporter) of the loop of Henle. Salt sensitivity is an important factor in the pathophysiology of hypertension[38], and uromodulin may therefore represent a shared point of the common complications of diabetes and hypertension. In addition to this, mutations in the APOE gene result in type III hyperlipoproteinemia (familial dysbetalipoproteinemia), with impaired chylomicron and VLDL remnant clearance and consequent increased plasma cholesterol and triglycerides. APOE, hypertension and Diabetes significantly increase the risk of dementia, including Alzheimer's disease[39,40]. A link has been established through the action of APOE4 that leads to endosomal entrapment of insulin receptors[41]. In addition to this, mutations in the APOE gene result in type III hyperlipoproteinemia (familial dysbetalipoproteinemia), with impaired chylomicron and VLDL remnant clearance and consequent increased plasma cholesterol and triglycerides. APOE, hypertension and Diabetes significantly increase the risk of dementia, including Alzheimer's disease[39,40]. A link has been established through the action of APOE4 that leads to endosomal entrapment of insulin receptors[41].

Clustering for the AFs of the Egyptian Reference metabolic genes goes along with the same populations' proximity in Wohlers et al.[42]. Intriguingly, South-Asians close stratification with the Europeans has been discussed before in C. Chambers et al.[43] with the low genetic distance value "F_{ST}". Meanwhile, the PCA results confirm the same information with the Egyptian genotypes clustered nearly at the center with the European and closely relating to the Ad-Mixed American, and South-Asian ones. Interestingly, the African population's genotypes clustered distinctly despite the high numbers of variants they shared with the Egyptian ones. The observed phenomenon may be attributed to the possibility that the common genetic variations shared by Egyptians and Africans could be localized in particular genomic regions that underwent minimal differentiation, thereby being eliminated during the linkage disequilibrium filtering process. Moreover, results from the allele frequency PCA showed distinctive characteristics of the African and East Asian population suggesting genetic divergence for both of them. These findings shed light on the multifaceted nature of human genetic diversity for the genes responsible for the metabolic syndrome, unveiling shared genetic traits and distinctive characteristics across diverse populations. Moreover, these findings emphasize the need for comprehensive investigations to unravel

Chromosome	Position	Alt Allele	Gene symbol	VEP existing variation	Protein position	Amino acid change	Missense3D analysis result
6	2106570	T	CDKAL1	rs756851756	360	T/S	Buried H-bond breakage
3	12351660	T	PPARG		23	A/V	Cavity altered

Table 2. Missensense3D structural damage prediction analysis result for the protein coding variants. Two SNPs found in two different samples out of the 110 Egyptian cohort with a minor allele frequency of 0.0045 were identified as potentially causing structural damage. These SNPs were found in the CDKAL1 and PPARG genes. The table includes the chromosome number, position, alternative allele, gene symbol, existing variation according to Variant Effect Predictor (VEP), protein position, amino acid change, and the results of Missense3D analysis.

the underlying mechanisms driving these contrasting patterns and gain a deeper understanding of the genetic structure and history of the Egyptian population.

Finally, the effect of the predicted structural damage of the protein coding variants is noteworthy. Besides, the thermostability of proteins against denaturation is greatly influenced with the number of the H-bonds formed[44]. Wild type of the protein Threonylcarbamoyladenosine TRNA Methylthiotransferase shows Hydrogen bonding between the side chain 360 Threonine (THR) and the main chain 320 Isoleucine (ILE) with 2.75Å. Another bond happens to be between 360 THR and ASP 361 with 3.67 Å. While the at the main chain, 322 Valine (VAL) forms a hydrogen bond (H-bond) with the 360 THR with 2.89Å. Conversely, the prediction result didn't find any H-bond at the mutant structure with the 360 serine (SER). Furthermore, it is worth noting that the SNP "rs756851756" showed uniqueness to the Egyptian Reference in our dataset, however in the ALFA Allele frequency it expresses an AF of 0.00007 with 1 European subpopulation allele out of 14,050[45]. Additionally, cavity contraction of the peroxisome proliferator-activated receptor gamma was predicted as for the SNP for the allele "T" at the position 12,351,660 on Chromosome 3, that happened to be causing amino acid change for the 23rd Alanine (ALA) to Valine (VAL). This change caused contraction for the cavity volume by 191.592 Å^3. A biological validation is required to confirm the pathogenicity of such volume contraction.

In conclusion, our study identified several genes that contribute to the high prevalence of metabolic syndrome, particularly in the context of T2D, in the Egyptian population. In this study, we introduce a comprehensive pipeline to investigate the allele frequency variations among different populations and supported our findings through genotypic PCA analysis. Our results demonstrated close clustering of the reference metabolic genes of Egyptians with those of Europeans, Ad-Mixed Americans, and South Asians. Additionally, we identified 8563 variants unique to the Egyptian cohort. Further analysis of these distinctive variants unveiled two missense variants that were found to be heterozygous in two different samples out of the 110 Egyptian cohort, CDKAL1: 6_21065070 (A > T) and PPARG: 3_12351660 (C > T) and were predicted to cause structural damage according to Missense3D. These findings suggest that the reference metabolic genes of Egyptians may exhibit population-specific alterations, potentially impacting disease susceptibility and tailored therapeutic approaches. However, Larger cohorts of Egyptian ancestry are necessary to assess the prevalence of the variants and exclude the possibility that the detected variants are ultra-rare or private also in the Egyptian population. Moreover, the emerging field of precision medicine (PM) bears promise for combating Egypt's growing diabetes epidemic. Population-scale genetic, clinical, and lifestyle data can be analysed to develop targeted strategies for specific subgroups, thereby augmenting outcomes, extending lifespan, and decreasing healthcare expenditures. Through this analysis, we hope to pave the way for advancements in healthcare practises, enhanced disease management, and an overall improvement in Egypt's public health outcomes. Future work is warranted to explore the underlying factors contributing to the observed genetic similarities and differences among populations, providing valuable insights into the genetic history and relationships of the Egyptian population within the broader context of human metabolic genetic diversity.

Methods

Data collection and pre-processing

The datasets selection was based on searching the 1000-genome project Ensembl database, using the name of the diseases representing the components of the metabolic syndrome, namely, Type 2 Diabetes, Hypertension, Obesity, and hyperlipidemia. In the data availability section, the links to these four datasets were documented. The Ensembl database utilizes ontologies such as HPO (Human Phenotype Ontology) and EFO (Experimental Factor Ontology) to map and display disease phenotypes, facilitating data integration across domains[46,47].

The common variants between Type II Diabetes and the other three diseases were identified and the Ensembl's Variant Effect Predictor (VEP) -release 109[48] was employed to analyze the resulting list of common variants and identify the associated genes. The reported variants by VEP were filtered by the removal of duplicates and the dbSNP database -build 156[49] was utilized to obtain the annotation of the common variants.

To annotate the protein-coding genes linked to the common variants and explore their potential interactions, the STRING database v11.5[50] was employed. This allowed for the examination of possible networks among these genes. Additionally, the GeneCards database v5.15[51] was used to annotate the genes while particularly focusing on identifying RNA genes and pseudogenes.

To collect information on the VEP shortlisted genes, we used the available variant data from an Egyptian Genome study[42] which integrated previously generated raw data[52]. Several procedures were taken to prepare

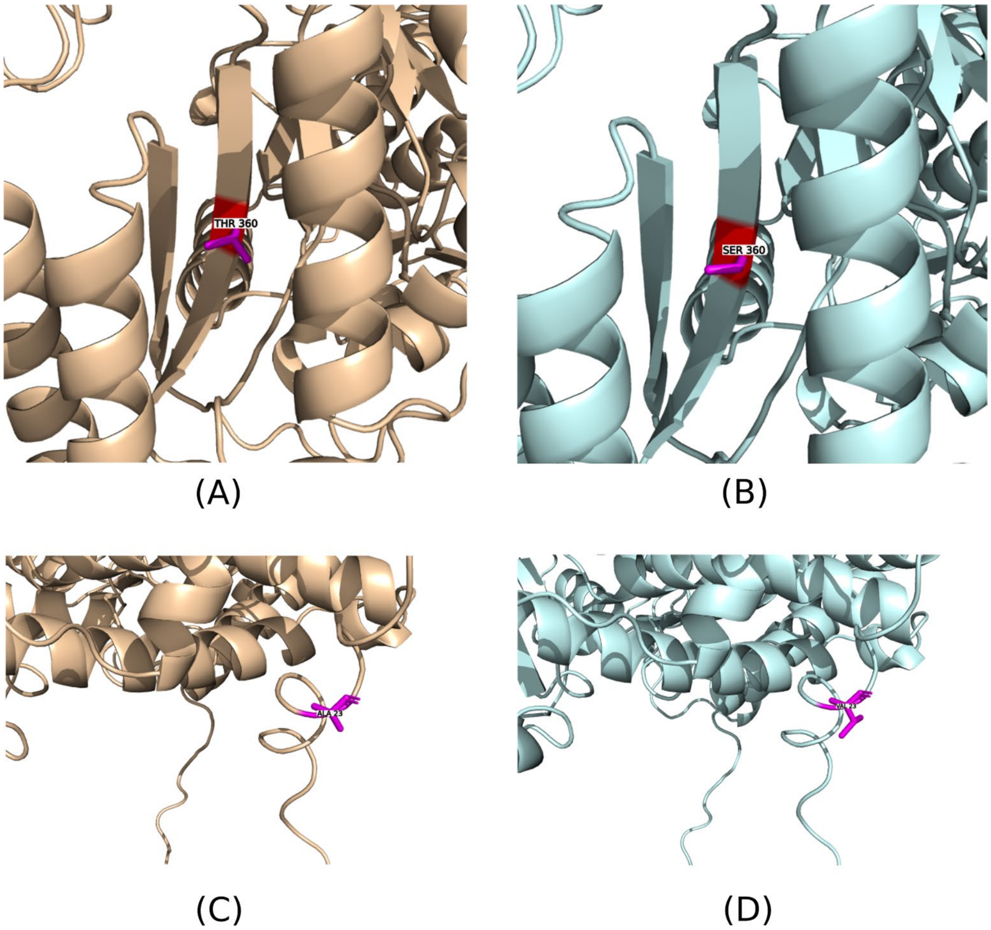


Figure 3. Missense3D Predicted damaged Structures for the mutant types compared to the wild ones. (**A**) Wild Type, (**B**) Mutant Type. Structural Damage predicted for the protein Threonylcarbamoyladenosine TRNA Methylthiotransferase (CDKAL1) as for the SNP rs756851756. This substitution disrupts all side-chain/side-chain H-bond. (**C**) and (**D**) Structural Damage predicted for the protein peroxisome_proliferator-activated_receptor_gamma (PPARG) as for the SNP 3_12351660_T. (**C**) Wild Type, (**D**) Mutant Type. The substitution leads to the contraction of cavity volume by 191.592 Å^3.

the genomic data Variant Calling Files (VCF) of the Egyptian and the 1000 Genome samples for analysis. These procedures included concatenating VCFs with BCFtools v1.16[53], removing duplicate entries, normalizing multi-allelic sites to biallelic sites, and extracting Allele Frequencies (AFs). Separate lists containing chromosome, position, and alternative allele information were generated for each subpopulation in the 1000 Genome sample. Similarly, pertinent data were extracted from the Egyptian genome VCF and stored in a separate list. Using these techniques, standardized formats of genetic variation within the genes of interest were generated, thereby facilitating insightful analysis. This approach allowed us to obtain insights into the common genetic variants associated with Type II Diabetes, Hyperlipidemia, Hypertension, and Obesity, thereby enhancing our understanding of the genetic basis of these complex diseases.

Genetic analysis and visualization of populations

In this investigation, a genetic analysis was performed to examine the genetic differences between the six populations: Africans, Ad-Mixed Americans, East Asians, Egyptians, Europeans, and South Asians. Multiple stages were required to preprocess the data and extract meaningful insights.

First, the data were sorted, annotated, and then merged using the "plyr" package v1.88[54] in R to assure a standardized ID column format. Non-genotype columns have been eliminated, and the output has been converted to VCF format. BCFtools were used to further process the sorted file to ensure data accuracy.

The populations' genetic information was then extracted using principal component analysis (PCA). Using Plink v2.0[55,56], an input bed file was generated, and variants were pruned based on a MAF of 0.05 as well as an indep-pairwise parameter of 50 5 0.5 (Window size in SNPs of 50, with a step of 5 SNPs to shift the window, and the r^2 threshold to be 0.5) to assure independence between variants. This step is intended to eliminate closely related variants to maintain the accuracy of subsequent analyses.

The extracted PCA data offered eigen vectors and eigen values for 2614 samples, based on 1961 pruned variants. These elements captured the genetic diversity present in the populations and laid the foundation for further analysis.

The analysis was then transferred to the R for data visualization. The subpopulation listings of the populations were intersected with the Egyptian list to comprehend the overlap and shared variants. The "UpSetR" package v.1.4[57], which generated interactive and flexible plots of intersecting sets, and an online tool from the Van de Peer Lab at the University of Ghent in Belgium (accessed via: https://bioinformatics.psb.ugent.be/webtools/Venn/) were used to generate a non-symmetrical Venn diagram. Using the "join" function from the R "plyr" package, the genomic data from the 1000 Genome and Egyptian genome files were combined to visualize the allele frequencies among the populations. This procedure of merging datasets enabled a comprehensive analysis of the genetic variation present in both datasets. Using the "pheatmap" package v1.0.12[58], a heatmap was generated to visually depict the similarities and differences in allele frequency among the populations.

In addition, principal component analysis (PCA) was performed using the "prcomp" function in R to analyze the allele frequency discrepancies among populations. Using the "plot_ly" function from the "plotly" v4.10.1[59] package, the first two components that accounted for a substantial portion of the cumulative variation were plotted. This additional visualization revealed the genetic distinctions among the populations based on their scores on the principal component.

Moreover, a genotypes PCA visualization was performed using R. The previously resulted eigenvalues and eigenvectors were used to calculate proportional variances, which were then displayed as a bar chart. To comprehend the role of geography and other demographic/clinical factors, the metadata of the 1000 Genome samples[60] and Egyptian samples[42] were utilized. The first two and three principal components were plotted against one another, providing a clearer picture of the genetic similarities and differences between the populations, and allowing the identification of prospective patterns in the data.

This thorough analysis and representation of genetic variation provided valuable insights into the diversity of the studied populations.

Unique Egyptians' single nucleotide variants (SNVs) and prediction of protein structural and functional damage

Additional research was conducted to learn more about the unique single nucleotide variants (SNVs) discovered in the Egyptian population. The Ensembl's VEP was used to analyze the functional impact of these SNVs and predict the amino acid changes corresponding to them. The output of VEP was then filtered based on the genes of interest, with a particular emphasis on SNVs that were protein-coding and predicted to have amino acid alterations.

In addition, the AlphaFold database v2.0[61,62] was used to acquire insight into the potential impact of identified SNVs on protein structure. This database contains predicted protein structures for protein-coding genes that have been filtered. The predicted structure in the PDB file format, as well as the anticipated amino acid change and its position, were input into the Missense3D database v1.5.4[63,64]. Missense3D utilizes computational algorithms to predict potential missense structural modifications resulting from amino acid substitutions.

This study aimed to shed light on the potential functional implications of genetic variations found in the Egyptian population by combining a unique SNV analysis with the prediction of protein structural damage.

Data availability

1k genome variants data was obtained from the December 2021 Ensemble database. Diabetes data was accessed from: (http://dec2021.archive.ensembl.org/Homo_sapiens/Phenotype/Locations?oa=EFO:0001360). Obesity data was accessed from: (https://dec2021.archive.ensembl.org/Homo_sapiens/Phenotype/Locations?ph=3816). Hypertension data was accessed from: (http://dec2021.archive.ensembl.org/Homo_sapiens/Phenotype/Locations?oa=EFO:0000537). Hyperlipidemia data was accessed from: (http://dec2021.archive.ensembl.org/Homo_sapiens/Phenotype/Locations?oa=HP:0003077). The Egyptian Genome VCF files were obtained from the European Genome-Phenome Archive under dataset ID EGAD00001006039: (https://ega-archive.org/datasets/EGAD00001006039).

Code availability

The code to reproduce the pipeline from this work can be found at GitHub: (https://github.com/Mahmoudbassuoni/EGYREF-Metabolic).

Received: 24 July 2023; Accepted: 6 November 2023
Published online: 22 November 2023

References

1. International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: 2021. Available at: https://www.diabetesatlas.org. (n.d).
2. Hegazi, R., El-Gamal, M., Abdel-Hady, N. & Hamdy, O. Epidemiology of and risk factors for type 2 diabetes in Egypt. *Ann. Glob. Health* **81**, 814–820. https://doi.org/10.1016/J.AOGH.2015.12.011 (2015).
3. World Health Organization. Egypt STEPS Survey 2017 Facts & Figures (2017).
4. Aboulghate, M. *et al.* The burden of obesity in Egypt. *Front. Public Health* **9**, 1247. https://doi.org/10.3389/FPUBH.2021.718978/BIBTEX (2021).
5. Elaziz, K. M. A., Gabal, M. S., Aldafrawy, O. A., Seif, H. A. A. A. & Allam, M. F. Prevalence of metabolic syndrome and cardiovascular risk factors among voluntary screened middle-aged and elderly Egyptians. *J. Public Health.* **37**, 612–617. https://doi.org/10.1093/PUBMED/FDU097 (2015).
6. Ibrahim, M. M. Problem of hypertension in Egypt. *Egypt. Heart J.* **65**, 233–234. https://doi.org/10.1016/J.EHJ.2013.03.005 (2013).

7. World Health Organization (WHO). Hypertension Egypt 2020 country profile (2020).
8. Reda, A., Ragy, H., Saeed, K. & Alhussaini, M. A. A semi-systematic review on hypertension and dyslipidemia care in Egypt—highlighting evidence gaps and recommendations for better patient outcomes. *J. Egypt. Public Health Assoc.* **96**, 200. https://doi.org/10.1186/S42506-021-00096-9 (2021).
9. Barroso, I. & McCarthy, M. I. The genetic basis of metabolic disease. *Cell* **177**, 146. https://doi.org/10.1016/J.CELL.2019.02.024 (2019).
10. Defesche, J. C. *et al.* Familial hypercholesterolaemia. *Nat. Rev. Dis. Prim.* **3**, 200. https://doi.org/10.1038/NRDP.2017.93 (2017).
11. Hattersley, A. T. & Patel, K. A. Precision diabetes: Learning from monogenic diabetes. *Diabetologia* **60**, 769–777. https://doi.org/10.1007/S00125-017-4226-2 (2017).
12. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005. https://doi.org/10.1038/ng.2385 (2012).
13. 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).
14. Klarin, D. *et al.* Genetics of blood lipids among ~ 300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523. https://doi.org/10.1038/s41588-018-0222-9 (2018).
15. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513. https://doi.org/10.1038/s41588-018-0241-6 (2018).
16. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**, 1114. https://doi.org/10.1038/ng.3390 (2015).
17. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41. https://doi.org/10.1038/nature18642 (2016).
18. Merritt, D. C., Jamnik, J. & El-Sohemy, A. FTO genotype, dietary protein intake, and body weight in a multiethnic population of young adults: A cross-sectional study. *Genes Nutr.* **13**, 1–10. https://doi.org/10.1186/S12263-018-0593-7/FIGURES/1 (2018).
19. Bayer, S., Winkler, V., Hauner, H. & Holzapfel, C. Associations between genotype-diet interactions and weight loss—a systematic review. *Nutrients* **12**, 2891. https://doi.org/10.3390/NU12092891 (2020).
20. Qi, Q. *et al.* FTO genetic variants, dietary intake and body mass index: Insights from 177 330 individuals. *Hum. Mol. Genet.* **23**, 6961–6972. https://doi.org/10.1093/HMG/DDU411 (2014).
21. Naja, F. *et al.* Dietary patterns and their associations with the FTO and FGF21 gene variants among Emirati adults. *Front. Nutr.* **8**, 211. https://doi.org/10.3389/FNUT.2021.668901/BIBTEX (2021).
22. Speakman, J. R. The "fat mass and obesity related" (FTO) gene: Mechanisms of impact on obesity and energy balance. *Curr. Obes. Rep.* **4**, 73–91. https://doi.org/10.1007/S13679-015-0143-1/TABLES/3 (2015).
23. Merkestein, M. *et al.* FTO influences adipogenesis by regulating mitotic clonal expansion. *Nat. Commun.* **6**, 1–9. https://doi.org/10.1038/ncomms7792 (2015).
24. Saber-Ayad, M. *et al.* The FTO genetic variants are associated with dietary intake and body mass index amongst Emirati population. *PLoS One* **14**, e0223808. https://doi.org/10.1371/JOURNAL.PONE.0223808 (2019).
25. Bego, T. *et al.* Association of FTO gene variant (rs8050136) with type 2 diabetes and markers of obesity, glycaemic control and inflammation. *J. Med. Biochem.* **38**, 153. https://doi.org/10.2478/JOMB-2018-0023 (2019).
26. Saber-Ayad, M. *et al.* The FTO rs9939609 "A" allele is associated with impaired fasting glucose and insulin resistance in Emirati population. *Gene* **681**, 93–98. https://doi.org/10.1016/J.GENE.2018.09.053 (2019).
27. Campbell, J. E. Targeting the GIPR for obesity: To agonize or antagonize? Potential mechanisms. *Mol. Metab.* **46**, 101139. https://doi.org/10.1016/J.MOLMET.2020.101139 (2021).
28. Wilding, J. P. H. *et al.* Once-weekly semaglutide in adults with overweight or obesity. *N. Engl. J. Med.* **384**, 989–1002. https://doi.org/10.1056/NEJMOA2032183/SUPPL_FILE/NEJMOA2032183_DATA-SHARING.PDF (2021).
29. Samms, R. J., Coghlan, M. P. & Sloop, K. W. How may GIP enhance the therapeutic efficacy of GLP-1?. *Trends Endocrinol. Metab.* **31**, 410–421. https://doi.org/10.1016/J.TEM.2020.02.006 (2020).
30. Karagiannis, T. *et al.* Management of type 2 diabetes with the dual GIP/GLP-1 receptor agonist tirzepatide: A systematic review and meta-analysis. *Diabetologia* **65**, 1251–1261. https://doi.org/10.1007/S00125-022-05715-4 (2022).
31. Jastreboff, A. M. *et al.* Tirzepatide once weekly for the treatment of obesity. *N. Engl. J. Med.* **387**, 205–216. https://doi.org/10.1056/NEJMOA2206038/SUPPL_FILE/NEJMOA2206038_DATA-SHARING.PDF (2022).
32. Rangwala, S. M. & Lazar, M. A. Peroxisome proliferator-activated receptor γ in diabetes and metabolism. *Trends Pharmacol. Sci.* **25**, 331–336. https://doi.org/10.1016/J.TIPS.2004.03.012 (2004).
33. Blanquart, C., Barbier, O., Fruchart, J. C., Staels, B. & Glineur, C. Peroxisome proliferator-activated receptors: Regulation of transcriptional activities and roles in inflammation. *J. Steroid Biochem. Mol. Biol.* **85**, 267–273. https://doi.org/10.1016/S0960-0760(03)00214-0 (2003).
34. Mirza, A. Z., Althagafi, I. I. & Shamshad, H. Role of PPAR receptor in different diseases and their ligands: Physiological importance and clinical implications. *Eur. J. Med. Chem.* **166**, 502–513. https://doi.org/10.1016/J.EJMECH.2019.01.067 (2019).
35. Huang, Y. *et al.* Lipoprotein lipase links vitamin D, insulin resistance, and type 2 diabetes: A cross-sectional epidemiological study. *Cardiovasc. Diabetol.* **12**, 1–8. https://doi.org/10.1186/1475-2840-12-17/FIGURES/1 (2013).
36. Puri, K. *et al.* Diabetes mellitus severity and a switch from using lipoprotein lipase to adipose-derived fatty acid results in a cardiac metabolic signature that embraces cell death. *J. Am. Heart Assoc.* https://doi.org/10.1161/JAHA.119.014022 (2019).
37. Lee, Y. J. *et al.* The potential effects of HECTD4 variants on fasting glucose and triglyceride levels in relation to prevalence of type 2 diabetes based on alcohol intake. *Arch. Toxicol.* **96**, 2487. https://doi.org/10.1007/S00204-022-03325-Y (2022).
38. Mary, S. *et al.* Role of uromodulin in salt-sensitive hypertension. *Hypertension* **79**, 2419–2429. https://doi.org/10.1161/HYPERTENSIONAHA.122.19888 (2022).
39. Sutherland, G. T., Lim, J., Srikanth, V. & Bruce, D. G. Epidemiological approaches to understanding the link between type 2 diabetes and dementia. *J. Alzheimers Dis.* **59**, 393–403. https://doi.org/10.3233/JAD-161194 (2017).
40. Sierra, C. Hypertension and the risk of dementia. *Front. Cardiovasc. Med.* https://doi.org/10.3389/FCVM.2020.00005 (2020).
41. Zhao, N. *et al.* Apolipoprotein E4 impairs neuronal insulin signaling by trapping insulin receptor in the endosomes. *Neuron* **96**, 115–129. https://doi.org/10.1016/J.NEURON.2017.09.003 (2017).
42. Wohlers, I. *et al.* An integrated personal and population-based Egyptian genome reference. *Nat. Commun.* **11**, 1–10. https://doi.org/10.1038/s41467-020-17964-1 (2020).
43. Chambers, J. C. *et al.* The south Asian genome. *PLoS One* **9**, 102645. https://doi.org/10.1371/JOURNAL.PONE.0102645 (2014).
44. Ahmed, Z., Zulfiqar, H., Tang, L. & Lin, H. A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *Int. J. Mol. Sci.* https://doi.org/10.3390/IJMS231710116 (2022).
45. Phan, L., Jin, Y., Zhang, H., Qiang, W., Shekhtman, E., Shao, D. *et al.* ALFA: Allele frequency aggregator. National Center for Biotechnology Information, US National Library of Medicine (2020).
46. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217. https://doi.org/10.1093/NAR/GKAA1043 (2021).
47. Malone, J. *et al.* Modeling sample variables with an experimental factor ontology. *Bioinformatics* **26**, 1112–1118. https://doi.org/10.1093/BIOINFORMATICS/BTQ099 (2010).
48. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995. https://doi.org/10.1093/NAR/GKAB1049 (2022).

49. Sherry, S. T. *et al.* dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308. https://doi.org/10.1093/NAR/29.1.308 (2001).
50. Szklarczyk, D. *et al.* The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638. https://doi.org/10.1093/NAR/GKAC1000 (2023).
51. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**, 1–30. https://doi.org/10.1002/CPBI.5 (2016).
52. Pagani, L. *et al.* Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96**, 986–991. https://doi.org/10.1016/J.AJHG.2015.04.019 (2015).
53. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* https://doi.org/10.1093/GIGASCIENCE/GIAB008 (2021).
54. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29. https://doi.org/10.18637/JSS.V040.I01 (2011).
55. Shaun Purcell, Christopher Chang. PLINK 2.0 (n.d).
56. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7. https://doi.org/10.1186/S13742-015-0047-8/2707533 (2015).
57. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940. https://doi.org/10.1093/BIOINFORMATICS/BTX364 (2017).
58. Kolde R. pheatmap: Pretty Heatmaps. R package version 1.0. 12 (2019).
59. Inc. PT. Collaborative data science (2015).
60. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74. https://doi.org/10.1038/nature15393 (2015).
61. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444. https://doi.org/10.1093/NAR/GKAB1061 (2022).
62. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873), 583–596. https://doi.org/10.1038/s41586-021-03819-2 (2021).
63. Ittisoponpisan, S. *et al.* Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?. *J. Mol. Biol.* **431**, 2197–2212. https://doi.org/10.1016/J.JMB.2019.04.009 (2019).
64. Khanna, T., Hanna, G., Sternberg, M. J. E. & David, A. Missense3D-DB web catalogue: An atom-based analysis and repository of 4M human protein-coding genetic variants. *Hum. Genet.* **140**, 805–812. https://doi.org/10.1007/S00439-020-02246-Z/FIGURES/3 (2021).

Author contributions

Conceptualization and design, M.B., M.S.A., H.B. and M.E.; data acquisition and curation, I.W.; data analysis, M.B., M.M., M.E. and I.W.; results visualization, M.B. and M.M.; results interpretation, M.B., I.W., M.M., M.E., H.B. and M.S.A; writing–original draft, M.B. and M.S.A.; writing–review and editing, I.W. M.M. and H.B.; supervision, M.S.A, H.B. and M.E.; All authors read and approved the final manuscript.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46844-z.

Correspondence and requests for materials should be addressed to M.S.-A. or M.E.-H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.