Genetic Determinants of Vitamin D Deficiency in the Middle Eastern Qatari Population by a Genome-Wide Association Study

Georges Nemer (✉ gnemer@hbku.edu.qa)  
Hamad Bin Khalifa University  
https://orcid.org/0000-0003-2157-5279

Nagham Hindi  
Hamad Bin Khalifa University

Yasser Al-Sarraj  
Qatar Genome Program  
https://orcid.org/0000-0002-1505-6360

Umm-Kulthum Umlai  
Hamad Bin Khalifa University

Karsten Suhre  
Weill Cornell Medicine-Qatar

Omar Albagha  
Hamad Bin Khalifa University

Article

Keywords: 25-hydroxyVitamin D, Vitamin D deficiency, genome-wide association study, genetic predispositions, Meta-analysis, polygenic risk score, Middle Eastern

Posted Date: February 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2585804/v1

License: ☒ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Epidemiological studies have revealed that Middle Eastern countries have the highest incidence of Vitamin D deficiency with severe complications. However, the impact of Vitamin D polymorphisms and the performance of polygenic models have been studied primarily in European with little knowledge of the Middle Eastern. Here, we conducted the first genome-wide association study to identify genetic determinants of Vitamin D levels in Middle Eastern populations using a whole genome sequencing approach in 6,047 discovery subjects. We discovered a novel variant, rs2298850 ($P$-value = $1.71 \times 10^{-08}$, effect size (Beta) = -0.1285), in a region of a known locus for the group-specific component gene ($GC$). We also confirmed the association of Vitamin D to several variants, including rs11723621 ($P$-value = $1.93 \times 10^{-08}$, Beta = -0.12574) and rs4588 ($P$-value = $8.06 \times 10^{-08}$, Beta = -0.1188) in the $GC$. A GWAS meta-analysis combining results from our Qatari cohort and previous European data identified novel variants in known loci, including rs67609747 and rs1945603 on chromosome 11. We found a moderately low heritability of Vitamin D (estimated at 18%) compared to Europeans. Finally, a low predictive performance of European ancestry-derived polygenic scores was observed when applied to the Qatari individuals. These results emphasize the diversity in the genetic architecture and its impact on preventive and precision medicine across different populations. Our findings offer novel perspectives on the physiological mechanisms and genetic factors contributing to the variation of Vitamin D levels in the Qatari population.

Introduction

Vitamin D is a steroid hormone and nutrient that modulates calcium and phosphorus homeostasis. The serum 25-hydroxy-Vitamin D (25(OH)D), also known as calcidiol) is the predominant circulating metabolite of Vitamin D and a reversible biomarker of Vitamin D status. Concentrations of 25(OH)D less than 20 ng/mL (50 nmol/L) are considered the most common nutritional deficiency worldwide (1). Of vital importance, this deficiency can lead to severe clinical manifestations, including osteoporosis, with rickets being the main expression in children. Worldwide Vitamin D deficiency is attributed mainly to environmental and clinical factors, namely deprived exposure to solar ultraviolet B radiation due to latitude or cultural reasons, obesity (body mass index greater than 30 kg/m$^2$), skin pigmentation, female gender, and advanced age. In addition to the conventional risk factors, genetic predisposition contributes strongly to 25(OH)D levels (2). Over the last few years, twin and familial studies have observed a wide range of heritability estimates indicating that 23–90% of the variance in 25(OH)D levels might be from genetic variations (3–5). Moreover, the evolutionary perspectives have provided potent evidence about the significance of genetics for Vitamin D status (6). Although much is known about the epidemiology of Vitamin D, how genetic markers affect Vitamin D levels among global populations with downstream pathways still needs to be clarified (7).

Genome-wide association studies (GWAS) have contributed to significant apprehension of the genetic architecture underlying 25(OH)D levels. Until the present, multiple single-nucleotide polymorphisms (SNPs) have been identified at or near loci of genes encoding proteins responsible for Vitamin D
transportation (GC, APOA1), biosynthesis (DHCR7, NADSYN1), metabolism (CYP2R1, RRAS2, PDE3B, CYP24A1, CYP27A1, CYP27B1, CYP11A1, SSTR4/FOXA2, AMDHD1, and SEC23A), and activity (VDR and RXRa, RXRβ, RXRγ) (8–12). Around 70 Mendelian randomization analyses have suggested a causal influence estimation of Vitamin D through genetic variations (13–19). Detection of common genetic signatures (minor allele frequency (MAF) greater than 5%) that regulate Vitamin D status aids in providing detailed insight into physiological roles and mechanisms and in predicting individuals at risk of Vitamin D deficiency. Better risk estimation could be achieved by applying an individualized risk assessment that classifies individuals according to their Vitamin D deficiency risk. Polygenic risk scores (PRS) estimated by aggregating the additive effects of selected common SNPs, could predict the influence of susceptible loci on phenotype, which is often weighted by effect size (20). Hence, genetic profiling and risk categorization might be implemented clinically to guide personalized therapeutic strategies for high-risk individuals. By way of illustration, clinical practice guidelines recommend a daily Vitamin D-enriched diet and supplementation to minimize multiple sclerosis development in high-risk individuals based on recent genetic epidemiological evidence (21).

Despite the ample and prolonged sunshine in the Middle East regions, the highest prevalence of Vitamin D deficiency with major complications has been reported in these countries (22), with an incidence of up to 90% in Qatar (23), as in other Gulf states. Previous GWAS have focused primarily on identifying Vitamin D polymorphisms and evaluating the performance of aggregated models in the European population (8, 10–12). No previous studies have been conducted on Middle Eastern individuals to date. Given the high incidence reported in sunny regions, characterizing the genetic architecture underlying Vitamin D pathways in these populations is crucial. We conducted the first GWAS of Vitamin D in Middle Eastern individuals to accomplish this objective using a whole-genome sequencing approach. We performed a post-GWAS analysis to gain further insight into our findings by determining human tissue expression and exploring its frequency in global populations. Compared with the European observations, we examined the frequency and effect size of high-impact alleles in genes linked to Vitamin D in the Qatari population. We combined the current GWAS results with a previous large GWAS data (10) consisting of 345,923 individuals in a meta-analysis. Based on this, we validated the replication of previously identified genetic determinants of 25(OH)D in the Middle Eastern population. Finally, we assessed the association between genetic markers related to Vitamin D and phenotype severity by using polygenic scores from the UK Biobank (10) and assessed their performance in the Qatari cohorts.

**Methods**

**Study participants**

Data used in the present study were obtained from the Qatar Biobank (QBB) dataset. The QBB cohort is the first population-based prospective cohort study that included participants of Qatari nationals or long-term residents (living in Qatar for ≥15 years), aged 18 years and older, and appearing phenotypically healthy. Physical measurements for all participants were collected, and each participant completed a standardized questionnaire reporting lifestyle and medical history information. In addition, detailed...
baseline sociodemographic data, phenotypic data, clinical biomarkers, and biochemical tests were covered for the study participants. More details of the QBB project are explained previously (24).

Hamad Medical Corporation Ethics Committee approved the QBB study design (MOPH-AQBB-000222). All QBB participants signed informed consent waivers before participation. We submitted a request to access the QGP and QBB data (https://www.qatarbiobank.org.qa/research/how-to-apply), which was approved by the QBB IRB (IRB project number, QF-QGP-RES-ACC-00075). The first QBB dataset release (N = 6,218 individuals) was used to carry out the current GWAS analysis. A large GWAS published recently using European, African, and South Asian participants were used in performing the meta-analysis and replication analyses (N = 363,228 individuals) (10).

Circulating 25(OH)D and dependent covariates

Blood samples were collected, centrifuged for serum separation, and immediately stored at -80°C for all participants. Quantitative evaluations of serum 25(OH)D concentrations were analyzed using a fully automated chemiluminescent immunoassay (CLIA), DiaSorin LIAISON, Germany, in the diagnostic laboratories at Hamad Medical City. Briefly, serum Vitamin D was dissociated from Vitamin D binding protein, and a labeled tracer was added. A washing step was performed to remove any unbound protein before initiating the CLIA reaction. 25(OH)D levels were determined using a photomultiplier. Serum concentration of 25(OH)D were available for 5885 subjects. Before the statistical analyses, the phenotype was normalized using rank-based inverse standard transformation by R (version 3.4.0). The anthropometric measurements, such as body weight and height, were performed by the Seca 284 stadiometer and balance. Body mass index (BMI) was calculated by dividing the weight (kg) by the square of height (m²).

Whole genome sequencing and bioinformatics analysis

Genomic DNA was isolated from whole peripheral blood using an automated QIASymphony SP instrument following the Qiagen MIDI kit protocol's instructions (Qiagen, Germany). Quantification was performed on the FlexStaion 3 (Molecular Devices, USA) using Quant-iT dsDNA Assay (Invitrogen, USA). Whole genome sequencing was conducted on the Illumina HiSeq X Ten (Illumina, USA) platform with an average coverage of 30x at Sidra Clinical Genomics Laboratory Sequencing Facility as previously described(25). Briefly, FastQC data (version 0.11.2) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were aligned to the human reference genome GRCh37 (hs37d53) using Burrow-Wheeler Aligner (version 7.12) (BWA, https://github.com/lh3/bwa/tree/master/bwakit). Variant calling was obtained using HaplotypeCaller provided by Genome Analysis Toolkit (version 3.4) (GATK, https://software.broadinstitute.org/gatk/documentation/article?id=3238). Joint calling was conducted on all individual intermediate genomic variant call files (gVCF) to create a joint multi-samples VCF file for all the samples. The process consisted of two steps. We first applied GenomicsDB to combine regions for
all samples. We then utilized GenotypeGVCFs using SNP/Indel recalibration to merge all regions. Variants with the PASS filter were only included for downstream analysis following the GATK VQSR filtering steps.

A comprehensive quality control (QC) assessment was performed to minimize population structure and genetic diversity in our data using PLINK (version 2.0) (26). SNPs with genotyping call rate < 90%, the minor allele frequency (MAF) < 1%, or the Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$ were excluded. Additionally, we excluded samples with call rate < 95% ($N = 1$), duplicates ($N = 10$), excess heterozygosity ($N = 8$), and gender ambiguity ($N = 65$). We also performed multidimensional scaling (mds) analysis to identify population ancestry outliers with PLINK (26). A set of pruned independent autosomal SNPs ($N = 62,475$) was used to determine the pairwise identity-by-state (IBS) matrix through a window size of 200 SNPs and LD threshold of $r^2 = 0.05$. Population outliers were considered and excluded if they deviated from the mean of the first two mds components by four standard deviation units or more ($\pm 4$ SD). The genome-wide association analysis was conducted on 7,880,618 genetic variants obtained from 6,047 participants.

**Genome-wide association analysis**

GWAS analysis under a variance component-based linear model was performed using GRAMMAR-Gamma (27) within the GenABEL/R package (version 1.8-0) (28) to study the association of each variant and 25(OH)D levels. This method corrects for relatedness and genetic substructure by using the genomic kinship matrix. Considering the mixture of the Qatari population, we performed principal components analysis through PLINK software (26). The first four PCs were used as covariates in the association model to minimize bias from population stratification. Further, we adjusted the regression model for age (years) and gender. Genome-wide significance cutoff was defined as $P < 5 \times 10^{-8}$ and the nominal significance level as $P < 0.05$ (29). The quantile-quantile (Q-Q) plot, Manhattan plot, and genomic inflation factor were generated using R (version 3.4.0). Heritability ($h^2$) was identified as part of the polygenic model in GenABEL to estimate the degree of variation in the 25(OH)D levels due to inter-individual genetic variation in a population (30).

**Meta-analysis**

We combined the results of the QBB GWAS and a recent large GWAS (GCST90019526) from the UK Biobank by the N Sinnott-Armstrong et al (10) to derive a combined meta-analysis for the suggestively associated loci. The GCST90019526 GWAS models were characterized with the same phenotype and methods following the correction of relevant covariates, including age, sex, and genotype PCs (10). Summary statistics for study GCST90019526 (10) were obtained from the NHGRI-EBI GWAS Catalog (31) taken on 02 December 2022. We confirmed matching A1 and A2 alleles with the alternate and reference alleles in the GCST90019526 study. We also canonicalized our association statistics as our reported effect size based on the alternate allele in the reference genome. Notably, 25(OH)D measurements were inversely normalized using rank-based transformation in both GWASs. PLINK (version 1.9) was utilized to perform an inverse-variance weighted meta-analysis and estimate heterogeneity of effects analysis (26).
Validation of previous association with 25(OH)D

We compared our findings to those of the GCST90019526 GWAS study on Vitamin D from the UK Biobank project (10) to evaluate the extent of replication and correlation of effect size and allele frequency for common signals. To test the possibility that SNPs observed in our meta-analysis were previously reported in the UK biobank at \( P < 5.0 \times 10^{-8} \), each SNP was checked in the GWAS catalog (EFO_0004631) of Vitamin D measurement trait from the November 2022 released data that was accessed on 02 December 2022 (31). We first examined the locus associated with Vitamin D levels in QBB and UK biobank driven by the same variants. We then determined markers within a 250-kb upstream and downstream region of the GWAS catalog signals to identify significant variants associated with Vitamin D.

Polygenic scores analysis

We tested the performance of European-derived polygenic risk score (PRS) on the QBB cohort to estimate genetic liability to Vitamin D using PLINK (version 1.9) (26). Polygenic scores consist of combined SNPs by the sum of risk alleles, which are weighted by corresponding effect sizes predicted by GWAS results. We used polygenic scores derived from one of the largest Vitamin D genome-wide association studies carried out in European populations (PGS000702: \( n = 255,256 \) individuals and 8,012 variants (10)). The scoring files of the study were obtained via the Polygenic Score Catalog (https://www.PGSCatalog.org) (32) accessed on 09 December 2022. Pearson's correlation (R) between the inverse normalized Vitamin D levels, and European-derived PRS was computed with adjustment of age, gender, and first four PCs using the R software to evaluate the performance of the models on the Qatari population. We also used the area under the receiver operating characteristic (ROC) curve, also known as the "area under curve" (AUC). The AUC ranges from 0.5 (no distinction) to 1 (complete distinction), indicating the effectiveness of the derived PRSs in identifying those with Vitamin D deficiency, defined as serum 25(OH)D levels below 20 ng/mL and Vitamin D insufficiency when 25(OH)D levels between 21 to 30 ng/mL.

SNP annotation and functional analysis

The identified Vitamin D associates from the GWAS and meta-analysis were annotated using the Ensembl Variant Effect Predictor release 108 (VEP, https://grch37.ensembl.org/index.html) (33). We used the Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org) and Allele Frequency Aggregator (ALFA, www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/) to compare the frequencies of the identified Vitamin D variants with those in the global populations.

Results

Study description

The present study is established from the whole genome sequence data of Qatari participants. The average (± SD) age of QBB participants at the time of study enrollment was 40 (± 12.8) years.
(interquartile range: 18 to 88 years), and 56.3% of the filtered cohort was female (n = 3,318). Remarkably, we reported that approximately 50% of the participants had Vitamin D baseline levels below 20 ng/mL. Statistically significant associations were observed between 25(OH)D and both age (Pearson's coefficient of correlation (R) = 0.26, P-value = 2.2 × 10^{-16}) and sex (R = 0.072, P-value = 1.3 × 10^{-8}). The mean BMI (in kg/m^2) was approximately similar between both genders, 29.38 (± 6.05), with no significant link to serum 25(OH)D levels. Detailed characteristics of the study participants and phenotype assessment are provided in Table 1 and Table 2.

Genome-wide association study on 25(OH)D

We conducted a genome-wide association study and statistical fine-mapping analysis to identify genetic architecture and putative causal genes for Vitamin D in the Middle Eastern Qatari population. The association with circulating 25(OH)D was examined in 6,045 participants who passed quality control (QC). We restricted our examination to common and low-frequency risk alleles (MAF > 1%; N = 7,880,618) using linear mixed models correcting for age, sex, population principal components, and relatedness (full details in Methods). Quantile–quantile (Q-Q) and Manhattan plots of genetic associations for circulating 25(OH)D concentrations are shown in Fig. 1. The genomic inflation factor of our GWAS did not reveal evidence for widespread inflation (λ_{GC} = 1.01, standard error (SE) = 2.36 × 10^{-07}), suggesting no substantial effects of population stratification or cryptic relatedness as revealed in the Q-Q plot (Fig. 1A). The Manhattan plot of the GWAS showed a single genome-wide significant signal on chromosome 4q12 as a potential risk locus for 25(OH)D levels in the Qatari cohort (Fig. 1B).

We identified three genome-wide significant SNPs at the 4q12 locus (chromosome 4: 72,607,410 – 72,671,237) with a P-value of less than 5.0 × 10^{-8} (Table 3). The top-hit markers were all located within the region harboring the group-specific component gene (GC), which encodes a Vitamin D binding protein (VDBP). Of them, the only novel variant was rs2298850, located in intron 11 of GC, which showed the most significant association with 25(OH)D at a P-value of 1.71 × 10^{-08}. The other two SNPs in the GC were rs11723621 (P-value = 1.93 × 10^{-08}), followed by rs4588 (P-value = 8.06 × 10^{-08}) (Table 3). Both variants have been previously and consistently associated with the absolute concentrations of 25(OH)D among different populations (8, 10, 34–36). Polymorphisms suggestively linked to 25(OH)D concentrations with a P-value of less than 5 × 10^{-05} are presented in Supplementary Table 1. We further characterized the GWAS SNPs attribution to 25(OH)D variation (SNP-heritability, h^2) in the Qatari population. The heritability of 25(OH)D using all filtered SNPs was estimated as 18%.

Evaluating Replication of Known Loci

We evaluated the extent of replication by comparing our findings to prior published work on Vitamin D from the UK Biobank (GCST90019526) (10). We chose this study as it is the largest and most comprehensive GWAS on Vitamin D. Furthermore, the GCST90019526 study normalized the Vitamin D levels similar to our analysis using reverse-based inverse normalization (10), facilitating the comparison of effect sizes for identified loci. Most of the loci that reached the GWAS threshold in GCST90019526
(10) were replicated significantly ($P$-value $< 5.0 \times 10^{-8}$) or nominally ($P$-value $\leq 0.05$) in our study (Supplementary Table 2). For example, the GC rs2282679 showed a marginally significant association in Qatari cohort ($P$-value $= 2.61 \times 10^{-07}$) but had a significant association in the GCST90019526 study ($P$-value $= 1.0 \times 10^{-1268}$).

We then tested the correlation of effect directions and effect size for the replicated SNPs at a significant genome association threshold ($n = 58$). Most of them showed a consistent directionality (Fig. 2B), with slightly smaller effect sizes than the previous report of the GCST90019526 ($n = 43$, $R = 0.8$, regression slope $= 0.79$, 95% CI $= 0.63$ to $0.96$, $P$-value $< 0.0001$, Fig. 2C). The remaining SNPs showed a reverse association direction compared to QBB, possibly due to limited power. In the GCST90019526 study, the allele frequencies were available for only 40 of the replicated variants, which displayed a Pearson's coefficient (R) of 0.6 with a $P$-value of 0.002 (as seen in Fig. 2A).

The frequency of the most significant Vitamin D variants in the QBB cohort were compared with control populations from the gnomAD and ALFA browsers. The frequency of rs2298850 was similar, but rs11723621 and rs4588 were lower in the Qatari population compared to the European population in gnomAD and ALFA (Table 4). We identified 43 matching variants with consistent effect sizes for the Vitamin D trait in the European data.

GWAS Meta-analysis for Vitamin D

To detect potential novel variants that have a genome-wide significant association in the QBB GWAS, we combined our GWAS data with a comprehensive European GWAS by N Sinnott-Armstrong et al. of similar phenotype ($n = 363,228$ individuals) (10). Details of the replication GCST90019526 study are described previously (10). Our meta-analysis confirmed the replication of several variants reported similarly in the GWAS Catalog ($n = 18$ variants, Supplementary Table 3) or located in the same loci of known variants ($n = 42$ variants). We discovered 28 new variants with genome-wide significance in known loci related to Vitamin D. The top-hit variants were rs13361160 in CYP2R1 (cytochrome P450 2R1) (11:14910234 A $<$ G; Beta $= 0.0889$, $P$-value $= 6.38 \times 10^{-282}$), rs12504112 (4:72718873 T $>$ C; Beta $= -0.09$, $P$-value $= 8.78 \times 10^{-189}$), followed by rs10832256 (11:14442875 G $>$ A; Beta $= -0.07$, $P$-value $= 4.12 \times 10^{-146}$). Among the 28 novel variants, two were below the genome significance threshold in the GCST90019526 study and reached the GWAS threshold after combining with the QBB cohort, rs67609747 (11:15125750 T $>$ C; Beta $= -0.02$, $P$-value $= 1.18 \times 10^{-08}$) and rs1945603 (11:15275158 G $>$ A; Beta $= 0.02$, $P$-value $= 2.56 \times 10^{-08}$). The summary of meta-analysis results for discovery and replication cohorts is presented in Supplementary Table 4.

Polygenic score estimation

We tested the performance of European-derived polygenic scores represented by panel PGS000702 (10) in our QBB cohort against 5,885 individuals with available Vitamin D measurement data, consisting of 3,318 females and 2,567 males. Out of the 8,012 variants in panel PGS000702, 6,326 were deemed valid
predictors. The scoring process omitted 1,624 variants, with 1620 being disregarded due to a discrepancy in the variant identifier and 4 being disregarded due to an allele code mismatch. The performance of the European-derived polygenic scores on the Qatari population for Vitamin D is illustrated in Fig. 3.

We found that European-derived polygenic scores have a lower predictive performance on the QBB cohort (R = 0.098, 95% CI = 0.073 to 0.124, P-value = 4.60 × 10^{-14}) compared to the previously reported R values of 0.46 in the PGS000702 study (10). Vitamin D Polygenic risk were significantly associated with the risk of Vitamin D deficiency with an AUC of 0.680 (P-value of 4.71 x10^{-9}, Odds Ratio = 0.0935, 95% CI = 0.0420 − 0.2071). The risk of Vitamin D insufficiency and deficiency was predicted with an AUC of 0.6385 (P-value = 2.28 x10^{-5}, Odds Ratio = 0.0832, 95% CI = 0.0259 − 0.2646).

**Discussion**

This study presents the first genome-wide association analysis of Vitamin D deficiency in a large cohort of Middle Eastern individuals (n > 6200). Previous studies have estimated the heritability of 25(OH)D in Europeans to range from 23–40% (37). However, it is crucial to consider that these estimates can vary based on the population, methods used, and environmental factors affecting 25(OH)D levels. In our QBB cohort, the SNP-based heritability of 25(OH)D was found to be lower than that estimated in the UK Biobank participants, at approximately 18%. This difference is mainly due to the geographical location and population-specific genetic architecture, as heritability estimates increase with greater exposure to sunlight (11). Therefore, our study is crucial in uncovering novel and replicating common genomic markers of Vitamin D in Middle Eastern populations.

Our GWAS analysis revealed three SNPs linked to Vitamin D in a general Qatari population, with a minor allele reduction associated with reduced serum 25(OH)D. The top-associated markers were in the GC locus, which encodes VDBP, an essential member of the albumin family that synthesizes in the liver and transports Vitamin D and its metabolites (37). We identified a new genomic marker, rs2298850, on chromosome 4, associated with Vitamin D levels at a genome-wide significance level. This common-frequency variant (MAF = 0.177 (C)) is located in intron 11 of the GC gene. The lack of new loci discoveries reinforces the established understanding of known genes and their interactions in the Vitamin D pathway.

Several SNPs reported in previous studies have also been confirmed to be associated with Vitamin D in our QBB cohort, including chromosome 4:rs11723621 and chromosome 4:rs4588 in GC gene. Both GC variants showed consistent and robust associations with 25(OH)D concentrations in individuals from East Asia (34, 38) and Europe (8, 39). Furthermore, candidate gene studies have also shown strong associations between SNPs in the GC gene and 25(OH)D concentrations in Middle Easterners (40, 41). However, the association of GC rs4588 was only suggestive in the GWAS study by N Sinnott-Armstrong et al. (P-value of 2.463 × 10^{-05}) with no detection of GC rs11723621 (10). While all studies converge on the role of the GC gene in Vitamin D deficiency, further investigation of Vitamin D’s genomic background and
biological pathways is necessary to improve its clinical management and precision medicine applications.

To increase the statistical power of our findings, we combined data from a large European GWAS (10) with the QBB observations. It is noteworthy that individual alleles may have different genetic backgrounds across populations due to the significant variation in allele frequency and effect size among lineages. Therefore, we analysed the replication and correlation of identified variants with the UK Biobank data to examine the consistency of our observations. Our meta-analysis uncovered 28 new SNPs in known loci associated with Vitamin D levels that were not previously reported as genome-wide significant in the GWAS catalog. These SNPs include rs13361160 in \textit{CYP2R1}, which codes a key enzyme in the Vitamin D metabolism pathway (42), and rs10832256 near the \textit{SPON1} (Spondin 1) gene, previously implicated in regulating Vitamin D metabolism (8, 11).

Two new genetic variations, rs67609747 and rs1945603, have been identified as reaching the genome-wide association threshold through the combination of the GCST90019526 study (10) and the QBB cohort. Both variants are located in an intergenic region of the same locus on chromosome 11 (11:15125650–15275258), downstream of the \textit{CALCB} (calcitonin-related polypeptide beta) gene. Previous GWAS have established a strong association between \textit{CALCB} and Vitamin D, which regulates the calcium regulatory hormone calcitonin (8, 10, 11). The detected signals in our analysis are most likely due to better coverage of whole genome sequencing and slightly higher allele frequencies in the QBB data compared to the published GWAS, which was based on SNP arrays followed by imputation.

The UK Biobank cohort is extensive and comprehensive, offering the potential for driving PRS on the QBB population from the European populations. Nevertheless, the predictive performance of European-derived PRS was lower in the Qatari people, which may be attributed to several factors, including the number of variants included, GWAS sample size, the allele frequency of causal variants, and different variant weights among populations. The polygenic score model for Vitamin D was able to predict deficiency with slightly improved accuracy but not statistically significant when compared to predicting both deficiency and insufficiency. This finding highlights the necessity of a more comprehensive GWAS specific to Middle Eastern people to enhance the accuracy of polygenic score predictions.

In conclusion, our GWAS has identified for the first time the primary genetic determinant of Vitamin D predisposition in Middle Eastern individuals as a polymorphism in the \textit{GC} gene. Our analysis confirmed previous findings of shared genetic factors among diverse ethnic groups and revealed consistent patterns in the effect size and allele frequency of common variants. Our combined analysis of Middle Eastern and UK Biobank data has led to the discovery of 2 novel markers linked to Vitamin D and the replication of many previously known loci. The poor performance of the European-derived polygenic score when applied to Middle Eastern individuals underscores the importance of a more comprehensive investigation of Vitamin D genomic variations in non-European populations. The findings provide valuable understanding into the underlying mechanisms of Vitamin D and the relationship between an individual’s 25(OH)D status and related health issues across Middle East populations.
Declarations

Conflict of Interest

All the authors declared that there are no competing interests.

Code availability

All the analyses were done using publicly available software tools, which are mentioned in both the main text and the Methods section.

Data availability

The data analyzed herein are subject to the following licenses/restrictions: the raw whole-genome sequence data from Qatar Biobank are protected and are not available for deposition into public databases due to data privacy laws. Access to QBB/QGP phenotype and whole-genome sequence data can be obtained through an ISO-certified protocol, which involves submitting a project request at https://www.qatarbiobank.org.qa/research/how-apply, subject to approval by the Institutional Review Board of the QBB.

Acknowledgements

N.N.H is supported by a Ph.D. scholarship from Hamad Bin Khalifa University (HBKU) funded by the Qatar Foundation. The Qatar biobank (QBB) and Qatar Genome Program (QGP) are Research, Development, and Innovation’s entities within Qatar Foundation for Education, Science and Community Development.

Funding

Open access funding provided by the Qatar National Library.

Author information

Authors and Affiliations

Division of Biological and Biomedical Sciences, College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar

Naghm Nafiz Hindi

Division of Genomics and Translational Biomedicine, College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar

Georges Michel Nemer, Umm-Kulthum Ismail Umlai, Yasser Al-Sarraj, & Omar Albagha

Qatar Genome Program (QGP), Qatar Foundation Research, Development and Innovation, Qatar Foundation (QF), Doha, Qatar
Yasser Al-Sarraj  

Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar  

Karsten Suhre  

Author contribution  

NNH, OA, and GN conceived the idea and designed the model. NNH, YA, UIU, and KS developed the codes for analysis of the meta-analyses and polygenic risk score. NNH did the data analysis and interpretation of results as well as the write-up of the first draft. All authors contributed to the final write-up and agreed on its content.  

Ethics declarations  

Competing interests  

The authors declare no conflict of interests.  

References  


37. Gozdzik A, Zhu J, Wong BY, Fu L, Cole DE, Parra EJ. Association of vitamin D binding protein (VDBP) polymorphisms and serum 25(OH)D concentrations in a sample of young Canadian adults of


Tables

Tables 1 to 4 are available in the Supplementary Files section

Figures

![Manhattan Plot Dihydroxyvitamin_D_Total_InvNorm](imageA)

![Q-Q Plot Dihydroxyvitamin_D_Total_InvNorm](imageB)

Figure 1
Genome-wide association results of serum 25-hydroxyVitamin D levels presented as Manhattan and quantile-quantile plots. (A) Manhattan plot for the GWAS performed using linear mixed models correcting for age, gender, the first four principal components, and relatedness. The chromosomal positions of genetic variants (N = 7,880,618) plotted against –log10 P-value. The genome-wide significance threshold (P < 5 x 10^-8) is presented as a horizontal red line. (B) Q-Q plot shows the observed –log10 P-values and the expected –log10 P values.

Figure 2

Comparison of allele frequency and effect size for common variants between QBB and UK Biobank. (A) Risk allele frequency correlation for overlapped variants in QBB and European study (r = 0.60). (B) Effect weight correlation for overlapped variants in QBB and European study (r = 0.80). The best fit line from linear regression analysis is presented as a red line. (C) The effect weight (Beta) of SNPs shows replication after correction in the Qatari population (QBB; red bars) compared to the European population (UK Biobank, grey bars).
Figure 3

Performance of the European-derived polygenic score on the Qatari population. (A) Correlation regression of inverse-normalized baseline Vitamin D levels and weighted polygenic risk scores (PRS) derived from a large European dataset (PGS000702: R = 0.098, P = 4.60 × 10⁻¹⁴). The blue line represents the best fit of linear regression analysis. Receiver operating characteristic (ROC) curve of the PRS for the prediction of Vitamin D deficiency (25(OH)D < 20 ng/mL) (B) and Vitamin D insufficiency and deficiency (25(OH)D < 30 ng/mL) (C).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplemetaryMaterial.pdf
- Table1.pdf
- Table2.pdf
- Table3.pdf
- Table4.pdf