Article

Not peer-reviewed version

Rare-Variant Genome-Wide Association and Polygenic Score Assessment of Vitamin D Status in the Qatari Population

Nagham Nafiz Hendi , Umm-Kulthum Ismail Umlai , Omar Albagha * , Georges Nemer *

Posted Date: 26 August 2025

doi: 10.20944/preprints202508.1838.v1

Keywords: vitamin D deficiency; endocrinology; GWAS; rare variants; polygenic score; middle eastern genetics

Article

Rare-Variant Genome-Wide Association and Polygenic Score Assessment of Vitamin D Status in the Qatari Population

Nagham Nafiz Hendi [1], Umm-Kulthum Umlai [2], Omar Albagha [2,3,*] and Georges Nemer [2,4,*]

[1] Faculty of Pharmacy, Applied Science University, P.O. Box 11937, Amman, Jordan

[2] College of Health and Life Sciences, Hamad Bin Khalifa University, Qatar Foundation, P.O. Box 34110, Doha, Qatar

[3] Diabetes Research Center, Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Qatar Foundation, P.O. Box 34110, Doha, Qatar

[4] Department of Biochemistry and Molecular Genetics, American University of Beirut, P.O. Box 110236, Beirut, Lebanon

* Correspondence: nemer@hbku.edu.qa (G.N.); oalbagha@hbku.edu.qa (O.A.)

Abstract

Vitamin D deficiency is highly common in the Middle East despite abundant sunlight; however, most genetic studies have concentrated on common variants in Europeans. We analyzed whole-genome sequences from 13,808 Qatar Biobank participants, testing rare variants (minor allele frequency 0.01–0.0001) for quantitative 25-hydroxyvitamin D (25(OH)D) levels and deficiency risk (<20 ng/mL) in independent discovery (n=5,885) and replication (n=7,767) cohorts, followed by meta-analyses. Discovery analysis identified 41 genome-wide significant rare variants, including *CD36* rs192198195 ($P = 2.48 \times 10^{-8}$), and replication found 46, including *SLC16A7* rs889439631 ($P = 2.19 \times 10^{-8}$), implicating lipid metabolism pathways. Binary analysis revealed *RAP1GAP* rs577185477 ($P = 4.5 \times 10^{-8}$), linked to endothelial and immune regulation, while the meta-analysis uncovered *SLC25A37* rs952825245 ($P = 5.15 \times 10^{-12}$), associated with vitamin D signaling. Rare-variant polygenic scores derived from discovery cohort significantly predicted both continuous ($R^2 = 0.146$, $P = 9.08 \times 10^{-12}$) and binary traits (AUC = 0.548, OR = 0.99, $P = 9.22 \times 10^{-6}$) in replication cohort. This is the first rare-variant GWAS of vitamin D in Middle Easterners identifies novel loci and pathways, underscores the contribution of ancestry-specific rare alleles, and supports integrating rare and common variants for precision prevention in high-burden populations.

Keywords: vitamin D deficiency; endocrinology; GWAS; rare variants; polygenic score; middle eastern genetics

1. Introduction

Vitamin D plays a crucial role in maintaining calcium homeostasis, promoting skeletal health, and supporting various physiological functions. Deficiency in serum 25-hydroxyvitamin D (25(OH)D)—defined as levels below 20 ng/mL (50 nmol/L)—is a widespread global health concern. It has been linked to increased risks of osteoporosis, cardiovascular disease, immune dysfunction, and cancer [1,2]. Paradoxically, despite year-round sunlight, vitamin D deficiency remains alarmingly prevalent in Qatar and the wider Middle East, underscoring a multifaceted interplay of environmental, behavioral, and genetic factors [3,4].

Large genome-wide association studies (GWAS) in predominantly European populations (e.g., UK Biobank) and in regional cohorts like the Qatar Biobank (QBB) have identified several loci associated with serum 25(OH)D concentrations, like *GC* (Group-specific component, encodes vitamin

D-binding protein), *CYP2R1* (encodes 25-hydroxylase), and *DHCR7* (encodes 7-dehydrocholesterol reductase) [3,5,6]. However, many of these identified variants are common (minor allele frequency, MAF >5%), and collectively explain only a modest portion of the trait's heritability. This leaves a significant proportion of the genetic contribution to vitamin D status unaccounted for—often referred to as the "missing heritability"[7]. Rare variants (MAF ≤1%) have largely been overlooked in prior GWAS; however, population genetics theory suggests they may play a pivotal role, particularly due to their potential deleterious effects maintained through purifying selection [8].

Recent studies have shown that rare, high-impact variants can substantially influence human traits and disease risk, including Mendelian disorders and monogenic forms of common diseases [9]. For example, exome sequencing has revealed rare non-synonymous and putative loss-of-function variants in *CYP2R1*—a key vitamin D 25-hydroxylase—linked to altered vitamin D metabolism and increased susceptibility to rickets [10]. Similarly, rare variants in genes such as *AGO4* (Argonaute Component 4) and ATP-related pathways have been implicated in vitamin D regulation in Korean and European cohorts [11,12]. However, such findings remain underexplored in Middle Eastern populations, which possess distinct demographic histories and genetic structures.

Whole-genome sequencing (WGS) offers a powerful tool to uncover population-specific rare variants that may be absent or underrepresented in global reference panels [3]. Leveraging the deeply characterized QBB cohort, this study aims to perform a large-scale rare variant association analysis targeting variants with MAF between 0.01 and 0.0001. By integrating quantitative 25(OH)D phenotypes and replicating key findings within the QBB dataset, we aim to identify novel rare variants that contribute to vitamin D levels and risk of deficiency in the Qatari population. Given that genetics explain up to 50% of 25(OH)D variability [3,5], this study will clarify the genetic basis of vitamin D regulation and support precision health approaches tailored to Middle Eastern populations and their disease burden, including cancer.

2. Results

2.1. Study Population Characteristics

An overview of the study design is presented in Figure 1. The overall mean age of participants was 40.1 years (±13.1), with similar age distributions across both groups. Females constituted the majority of the cohort (55.5%), while males made up 44.5%. The average body mass index (BMI) across the entire cohort was 29.6 kg/m² (±6.1), placing most individuals in the overweight or obese range (BMI ≥25). The mean serum 25(OH)D concentration was 19.5 ng/mL (±11.1), and more than 60% of participants were classified as vitamin D deficient. An additional 26.8% were considered insufficient (20–30 ng/mL), while only 12.8% had levels in the normal range. Participant demographics and phenotype distributions are described in Table 1.

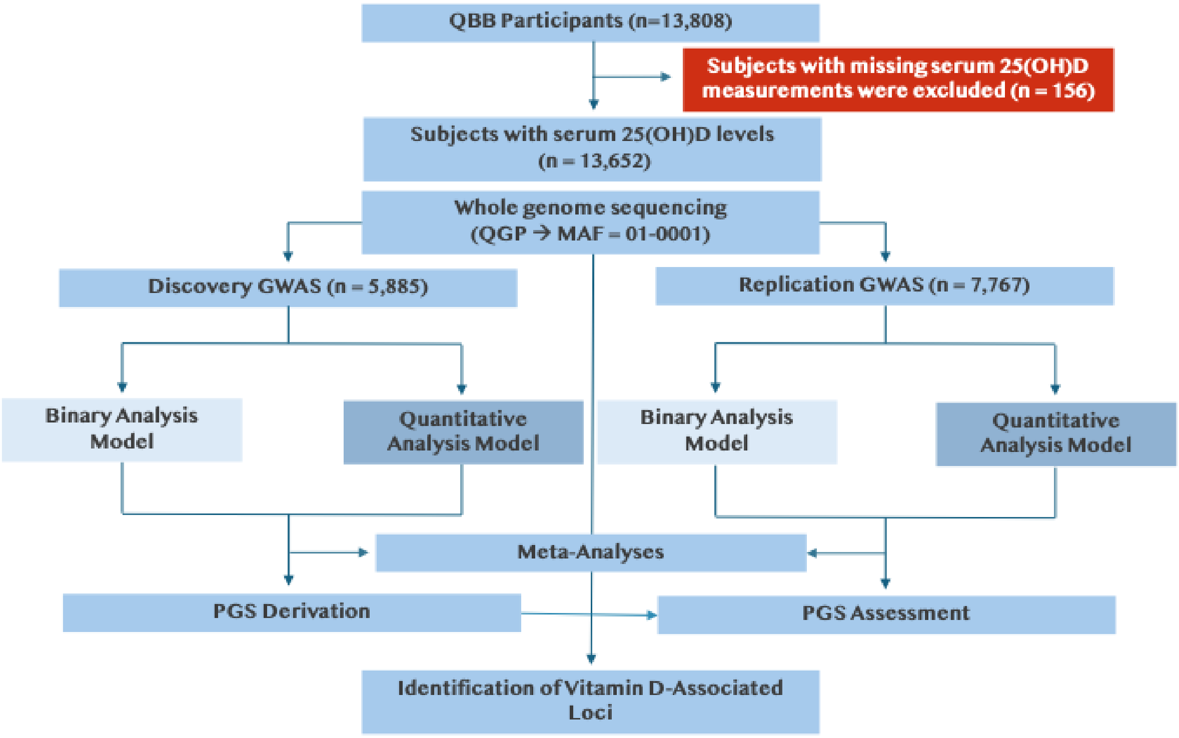
QBB Participants (n=13,808)
Subjects with missing serum 25(OH)D measurements were excluded (n = 156)
Subjects with serum 25(OH)D levels (n = 13,652)
Whole genome sequencing (QGP → MAF = 01-0001)
Discovery GWAS (n = 5,885)
Replication GWAS (n = 7,767)
Binary Analysis Model
Quantitative Analysis Model
Binary Analysis Model
Quantitative Analysis Model
Meta-Analyses
PGS Derivation
PGS Assessment
Identification of Vitamin D-Associated Loci

Figure 1. Study design for rare-variant genome-wide association analysis of vitamin D. Data were obtained from 13,808 Qatar Biobank (QBB) participants, of whom 13,652 had available serum vitamin D measurements. Vitamin D deficiency was defined as serum 25-hydroxyvitamin D (25(OH)D) levels < 20 ng/mL. Whole-genome sequencing (WGS) was performed through the Qatar Genome Program (QGP). Genome-wide association studies (GWAS) were conducted separately in a discovery cohort (n = 5,885 of 6,013) and a replication cohort (n = 7,767 of 7,795) using SAIGE (Scalable and Accurate Implementation of GEneralized mixed model), under both quantitative (inverse-normal–transformed 25(OH)D) and binary (deficient vs. sufficient) models. Results were combined in fixed-effects meta-analysis using PLINK. Polygenic scores (PGS) were derived from discovery GWAS results and evaluated in the replication cohort.

Table 1. Characteristics of study participants.

Cohort	Discovery	Replication	All Subjects
Sample size	5,885	7,767	13,652
Male n (%)	2,567 (43.6)	3,508 (45.2)	6,075 (44.5)
Female n (%)	3,318 (56.4)	4,259 (54.8)	7,577 (55.5)
Mean age ± SD	39.75 ± 12.83	40.38 ± 13.37	40.11 ± 13.14
BMI (kg/m≤)	29.38 ± 6.05	29.69 ± 6.14	29.55 ± 6.10
Vit D (ng/mL) ± SD	19.36 ± 11.12	19.52 ± 11.14	19.45 ± 11.13
Normal Vit D n (%)	675 (11.5)	1,073 (13.8)	1,748 (12.8)
Insufficient Vit D n (%)	1,612 (27.4)	2,053 (26.4)	3,665 (26.8)
Deficient Vit D n (%)	3,598 (61.1)	4,641 (59.8)	8,239 (60.4)

Descriptive statistics of the Qatar Biobank (QBB) cohort used in the discovery (batch 1) and replication (batch 2) analyses. Continuous variables are presented as mean ± standard deviation (SD), and categorical variables are presented as number (percentage). Vitamin D status was categorized based on serum 25-hydroxyvitamin D (25(OH)D) levels as follows: normal (>30 ng/mL), insufficient (20–30 ng/mL), and deficient (<20 ng/mL). BMI = body mass index (kg/m^2). The Discovery and Replication cohorts correspond to samples collected in separate batches for the genomic analysis.

2.2. Rare Variants Associated with 25(OH)D Levels Identified in Quantitative Trait GWAS

GWAS analysis was conducted to test the association of 8,517,321 rare single-nucleotide polymorphisms (SNPs, MAF 0.0001 to 0.01) with inverse-normal–transformed 25(OH)D concentrations in the discovery dataset (*n* = 5,885 participants). The results of the discovery GWAS are presented as Manhattan and Quantile-quantile (Q–Q) plots (Figure 2). The distribution of observed versus expected *P*-values demonstrated minimal deviation under the null, with a genomic inflation factor (λ_{GC}) of 1.04 (Figure 2a), indicating appropriate control for confounding. In total, 41 rare variants reached genome-wide significance ($P < 5 \times 10^{-8}$) in the discovery phase (Figure 2b). Several of the top discovery-stage associations were rare variants located in genes implicated in lipid metabolism and vitamin D transport (Table S1). These included rs374799245 near *NFIC* (Nuclear Factor I C, effect size (*β)* = –3.42 (standard error (SE) = 0.57), $P = 1.67 \times 10^{-9}$), rs758713488 in *TMPRSS9* (Transmembrane Serine Protease 9, *β* = –4.04 (0.69), $P = 6.21 \times 10^{-9}$), and rs192198195 in *CD36* (Cluster of Differentiation 36, *β* = –1.94 (0.35), $P = 2.48 \times 10^{-8}$).

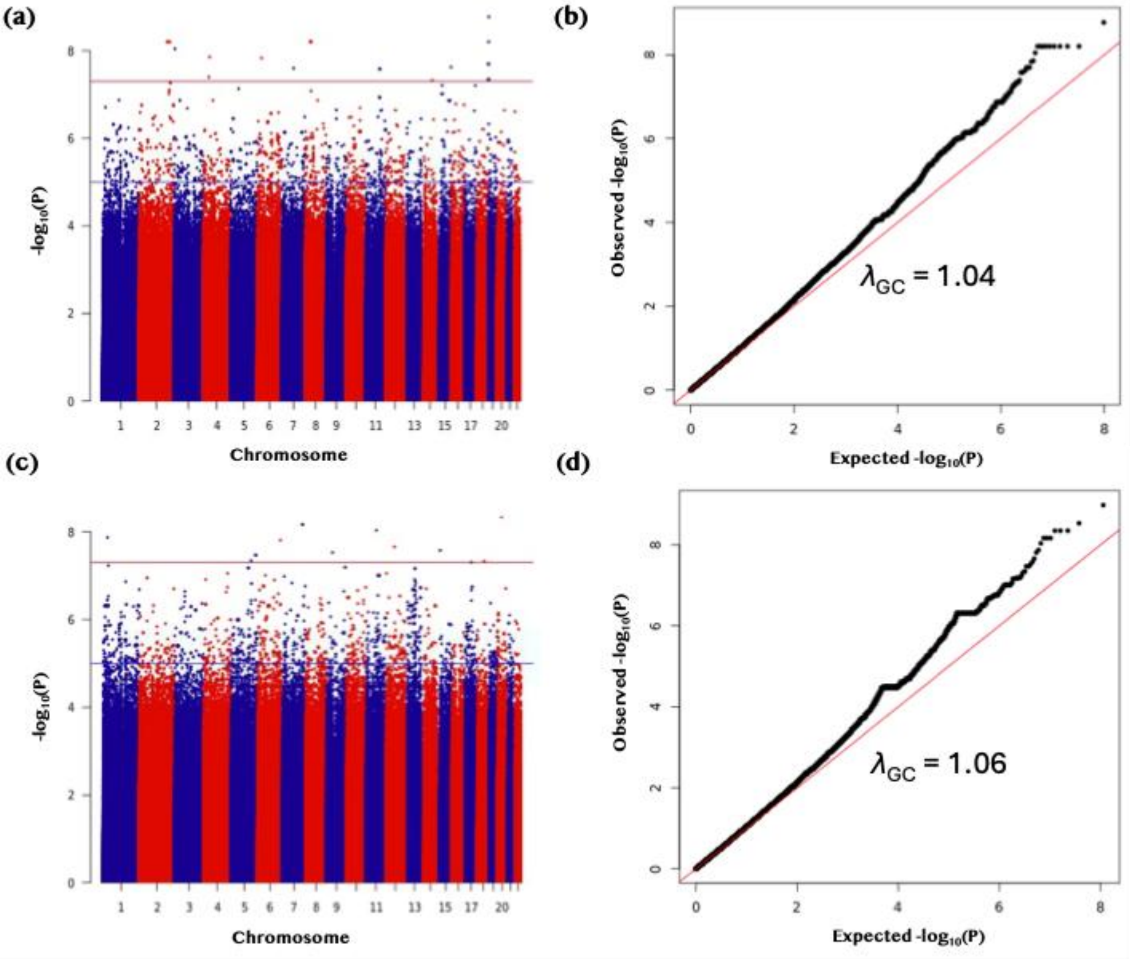

Figure 2. *Genome-wide association analysis of serum 25(OH)D concentrations in discovery and replication cohorts.* **(a)** Manhattan plot of GWAS results for the discovery cohort (QGP6013; autosomes only) using variants with minor allele frequency (MAF) between 0.0001 and 0.1. Serum 25(OH)D concentrations were inverse-normal transformed for analysis. The red horizontal line denotes the genome-wide significance threshold ($P < 5 \times 10^{-8}$). **(b)** Quantile–quantile (Q–Q) plot for the discovery cohort GWAS showing minimal genomic inflation (λ_{GC} = 1.04). **(c)** Manhattan plot of GWAS results for the replication cohort (QGP7795; autosomes only) applying the same filtering and transformation as in the discovery cohort. The red horizontal line denotes the genome-wide significance threshold ($P < 5 \times 10^{-8}$). **(d)** Q–Q plot for the replication cohort GWAS showing minimal genomic inflation (λ_{GC} = 1.06).

These findings were tested for validation in an independent replication cohort (*n* = 7,767 participants) using identical model specifications. Replication analysis yielded 46 genome-wide significant associations (Table S2), also with minimal genomic inflation (λ_{GC} = 1.06; Figure 2a, b). One of the top signals was rs889439631 in *SLC16A7* (solute carrier family 16A7; β = –2.69, (0.48), P = 2.19 × 10^{-8}) and chr11:87081213:G:T in *TMEM135* (Transmembrane Protein 135; β = –3.62, (0.68), P = 9.86 × 10^{-8}). Notably, cross-cohort comparison demonstrated strong concordance, with 9,192 out of 15,857 variants (58.0%) demonstrating a consistent direction of effect at nominal significance (P < 0.05) within a ±250 kb window between the discovery and replication cohorts (Figure 3a; Table S3). Effect sizes for the top variants ($P \leq 4.0 \times 10^{-05}$) showed strong agreement between cohorts, as illustrated by a high correlation in effect weights (coefficient of determination (R^2) = 0.869; regression slope = 0.869, 95% confidence intervals (CI): 0.863–0.876, $P < 2.2 \times 10^{-16}$) (Figure 3b). Similarly, allele frequencies were highly correlated between discovery and replication samples (R^2 = 0.988; regression slope = 1.008, 95% CI: 1.006–1.010, $P < 2.2 \times 10^{-16}$; Figure 3c).

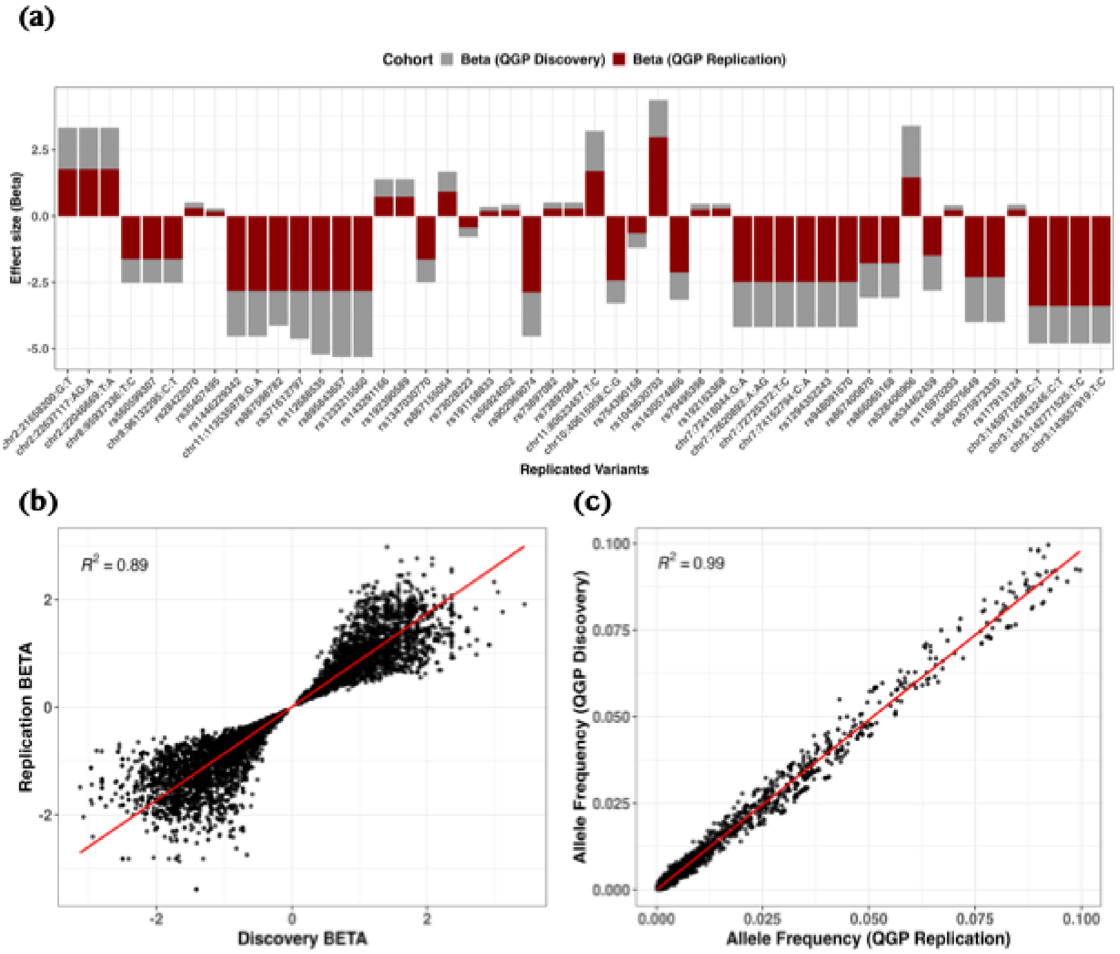

Figure 3. Replication of rare variant associations for serum 25(OH)D levels in the QGP cohorts. **(a)** Bar plots comparing effect estimates (β) between discovery (grey) and replication (dark red) cohorts, highlighting variants meeting a suggestive significance threshold ($P < 3.0 \times 10^{-5}$) with consistent effect direction. **(b)** Scatter plot of β estimates between cohorts, with the regression line shown in red and R^2 indicating the coefficient of determination. **(c)** Scatter plot of allele frequencies between cohorts with corresponding regression statistics. Displayed variants met the nominal significance threshold (P < 0.05) and had a consistent direction of effect.

A fixed-effects inverse-variance meta-analysis identified 54 rare variants that reached genome-wide significance, despite showing only suggestive evidence of association ($P \leq 5.0 \times 10^{-5}$) in the individual datasets, with minimal between-study heterogeneity (Cochran's Q-test P-heterogeneity > 0.05). Prominent genome-wide significant meta-analysis signals were observed in *CNTN3* (1-Contactin 3; rs115651661, β = 0.73, P = 1.48 × 10^{-08}) and *EBF1* (Early B-cell Factor 1; rs536115678, β = –1.74, P = 1.57 × 10^{-08}). A complete summary of meta-analysis statistics is provided in Table 2.

Table 2. Rare variants significantly associated with vitamin D levels discovered from a meta-analysis of the Replication and Discovery QGP Quantitative GWAS studies.

SNP	CHR	Position (BP)	Mapped Gene	HGVS ID	Consequence	A1	A2	GWAS in Replication QGP (n = 7,767)			GWAS in Discovery QGP (n = 5,885)			Meta-analysis		
								MAF (A1)	Beta (SE)	*P*-value	MAF (A1)	Beta (SE)	*P*-value	*P*-value	BETA	*P*-hat
rs115651661	3	74272255	*CNTN3*	NC_000003.12:g.74272255T>C	Intron	C	T	0.0040	0.48 (0.12)	0.00007	0.0027	0.73 (0.17)	2.46E-05	1.48E-08	0.56	0.22
rs536115678	5	158884335	*EBF1*	NC_000005.10:g.158884335C>A	Intron	A	C	0.0003	-1.84 (0.48)	0.00013	0.0003	-1.67 (0.40)	3.03E-05	1.57E-08	-1.74	0.79
chr21:43954055:C:T	21	43954055	*AGPAT3*	-	Intron	T	C	0.0004	1.26 (0.39)	0.00138	0.0007	1.57 (0.34)	6.10E-06	3.65E-08	1.43	0.55
chr21:43790823:A:G	21	43790823	*RRP1*	-	Upstream gene	G	A	0.0004	1.26 (0.39)	0.00138	0.0007	1.57 (0.34)	6.10E-06	3.65E-08	1.43	0.55
rs550626115	15	63022380	*TPM1-AS*	NC_000015.10:g.63022380C>T	Intron	T	C	0.0001	-2.41 (0.68)	0.00039	0.0002	-2.9 (0.69)	2.40E-05	4.07E-08	-2.67	0.59
rs1014490316	20	59592010	*PHACTR3*	NC_000020.11:g.59592010A>G	Intron	G	A	0.0001	2.33 (0.68)	0.00063	0.0002	2.98 (0.69)	1.73E-05	5.04E-08	2.65	0.50

Presented data are the same variants in Qatar Biobank (Vitamin D rare GWAS) with the P-het (P-value for Cochran's Q heterogeneity statistic) > 0.05 and P-value < 7E-08 of the meta-analysis. Abbreviations: A1, effect allele; A2, reference allele; AC_Allele2, allele count for reference allele (A2); AF_A2, allele frequency of reference allele (A2); AF_A1, allele frequency of effect allele (A1); BETA, effect size of A1 allele on 25(OH)D levels; SE, standard error of BETA; Mapped Genes from ANNOVAR; HGVS ID from Ensmebl; GWAS, genome-wide association study; QGP, Qatar Genome Project.

2.3. Rare Variant Associated with Vitamin D Deficiency Identified in Binary Trait GWAS

To assess the contribution of rare variants (MAF 0.0001–0.01) to the risk of vitamin D deficiency, we conducted a binary genome-wide association study using SAIGE logistic mixed models in the discovery cohort (2,287 controls; 3,598 cases), adjusting for age, sex, and population structure. No variants reached genome-wide significance, and the test statistics were well-calibrated (λ_{GC} = 1.00; Figure 4a). Nonetheless, 95 rare variants exhibited suggestive associations ($P < 5.0 \times 10^{-5}$; Table S4, Figure 4b).

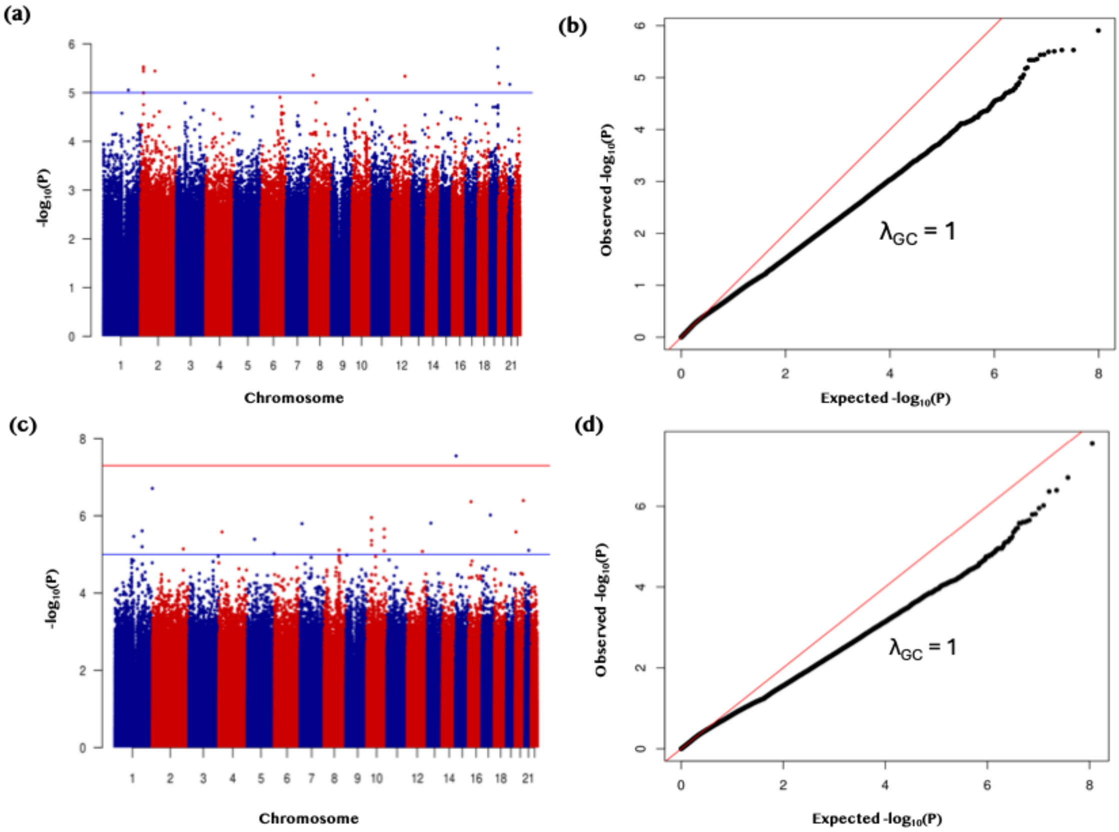

Figure 4. Binary genome-wide association study (GWAS) of vitamin D deficiency in discovery and replication cohorts. (a) Manhattan plot of GWAS results for the discovery cohort (QGP6013; autosomes only) filtered for variants with MAF 0.0001–0.1, assessing the binary vitamin D deficiency trait. The blue horizontal line denotes the genome-wide suggestive significance threshold ($P < 5 \times 10^{-5}$). (b) Quantile–quantile (Q–Q) plot for the discovery GWAS, indicating minimal genomic inflation ($\lambda_{GC} = 1.00$). (c) Manhattan plot of GWAS results for the replication cohort (QGP7795; autosomes only) filtered for variants with MAF 0.0001–0.1, assessing the binary vitamin D deficiency trait. The red horizontal line denotes the genome-wide significance threshold ($P < 5 \times 10^{-8}$). (d) Q–Q plot for the replication GWAS, indicating minimal genomic inflation ($\lambda_{GC} = 1.00$).

These variants were subsequently tested in the replication cohort using the identical model specifications (3,126 controls; 4,641 cases), which again showed strong genomic control (λ_{GC} = 1.00; Figure 4c). One variant, rs577185477 in *RAP1GAP* (Rap1 GTPase Activating Protein), exceeded genome-wide significance ($P = 4.5 \times 10^{-8}$), and an additional 115 variants reached suggestive significance (Figure 4d, Table S5). Of the 24,503 overlapping variants, 14,866 (60.6%) showed consistent effect directions at nominal significance (Figure 5a, Table S6). Effect estimates between cohorts were highly concordant ($R^2 = 0.805$, slope = 0.889; 95% CI: 0.882–0.896; $P < 2.2 \times 10^{-16}$; Figure 5b), and allele frequencies were closely aligned ($R^2 = 0.991$, slope = 0.9936; 95% CI: 0.992–0.995; Figure 5c).

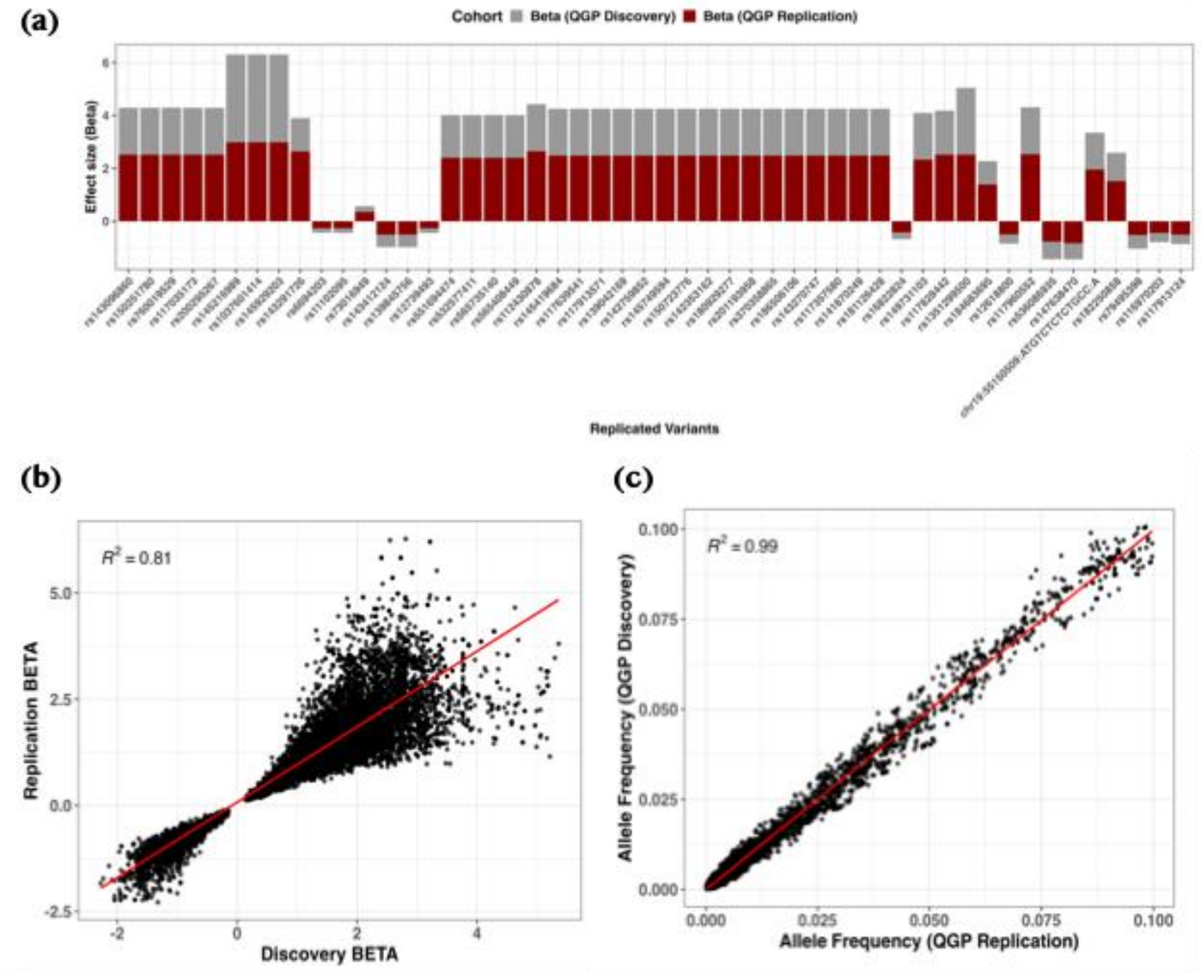

Figure 5. Replication analysis of rare variant associations for a binary trait in the QGP cohorts. **(a)** Bar plots comparing effect estimates (odds ratios, OR) between discovery (grey) and replication cohorts (dark red), highlighting variants meeting a suggestive significance threshold ($P < 3.0 \times 10^{-5}$) with consistent effect direction. **(b)** Scatter plot of OR between cohorts, with the red line indicating the best-fit linear regression. **(c)** Scatter plot of allele frequencies between cohorts, with the red line representing the regression fit. R^2 denotes the coefficient of determination from correlation analysis, and 95% confidence intervals (CI) are shown for regression slopes. All displayed variants met the nominal significance threshold ($P < 0.05$) and had a consistent direction of effect.

While neither cohort alone yielded multiple genome-wide significant associations, combining results through fixed-effects meta-analysis enhanced statistical power. This approach identified 93 rare variants that achieved genome-wide significance, with minimal heterogeneity between datasets. These variants included signals across both coding and regulatory regions, with notable mapping to biologically relevant genes, included rs140456089 in *PPP1R12C* (Protein Phosphatase 1 Regulatory Subunit 12C; odds ratios (OR) = 1.55 (0.30), $P = 3.12 \times 10^{-13}$), rs1268647997 near *RDH13* (Retinol Dehydrogenase 13; OR = 1.62 (0.33), $P = 2.13 \times 10^{-12}$), rs952825245 in *SLC25A37* (solute carrier family 25A37; OR = 2.30, (0.47), $P = 5.15 \times 10^{-12}$), and rs1454700296 in *NT5C2* (Cytosolic 5′-Nucleotidase II; OR = 2.15 (0.45), $P = 6.45 \times 10^{-11}$) (Table 3 and Table S7).

Table 3. Rare variants significantly associated with binary vitamin D deficiency discovered from a meta-analysis of the Replication and Discovery QGP GWAS studies.

SNP	Gene Mapped	CHR	Position (BP)	HGVS ID	Consequence	A1	A2	Binary GWAS for Vitamin D in Replication QGP (n=7795)			Binary GWAS for Vitamin D in Discovery QGP (n=6013)			Meta-analysis (n=2)		
								MAF (A1)	Beta (SE)	*P*-value	MAF (A1)	Beta (SE)	*P*-value	*P*-value	OR	*P*-het
rs140456089	*PPP1R12C*	19	55096117	NC_000019.10:g.55096117G>A	synonymous	A	G	0.0042	1.55 (0.30)	2.55E-07	0.0042	1.55 (0.30)	2.55E-07	3.12E-13	4.70	1
rs1268647997	*RDH13*	19	55065320	NC_000019.10:g.55065320G>A	upstream gene	A	G	0.0036	1.62 (0.33)	6.76E-07	0.0830	0.36 (0.07)	6.77E-07	2.13E-12	5.07	1
rs62122090	*HS1BP3*	2	20628549	NC_000002.12:g.20628549C>T	intron	T	C	0.0830	0.36 (0.07)	6.77E-07	0.0036	1.62 (0.33)	6.76E-07	2.13E-12	1.44	1
rs73916930	*HS1BP3*	2	20628921	NC_000002.12:g.20628921G>T	intron	T	G	0.0830	0.36 (0.07)	7.21E-07	0.0830	0.36 (0.07)	7.21E-07	2.41E-12	1.44	1
rs4426492	*HS1BP3*	2	20633127	NC_000002.12:g.20633127G>A	intron	A	G	0.0823	0.36 (0.07)	7.38E-07	0.0823	0.36 (0.07)	7.38E-07	2.53E-12	1.44	1
rs1185902565	*ANKRD36B*	2	97591789	NC_000002.12:g.97591790del	upstream gene	T	TC	0.0022	2.09 (0.43)	8.48E-07	0.0022	2.09 (0.43)	8.48E-07	3.31E-12	8.11	1
rs73916931	*HS1BP3*	2	20628926	NC_000002.12:g.20628926G>A	intron	A	G	0.0826	0.36 (0.07)	8.58E-07	0.0826	0.36 (0.07)	8.58E-07	3.39E-12	1.43	1
rs952825245	*SLC25A37*	8	23537529	NC_000008.11:g.23537529C>T	intron	T	C	0.0018	2.30 (0.47)	1.06E-06	0.0018	2.30 (0.47)	1.06E-06	5.15E-12	9.95	1
rs143947667	*ATP2B1-AS1*	12	89934704	NC_000012.12:g.89934704A>G	intron, NCT	G	A	0.0060	1.24 (0.25)	1.11E-06	0.0060	1.24 (0.25)	1.11E-06	5.68E-12	3.45	1
rs150425221	*ATP2B1-AS1*	12	89951086	NC_000012.12:g.89951086T>C	intron, NCT	C	T	0.0060	1.24 (0.25)	1.11E-06	0.0060	1.24 (0.25)	1.11E-06	5.68E-12	3.45	1
rs150021601	*ATP2B1-AS1*	12	90017542	NC_000012.12:g.90017542A>C	intron, NCT	C	A	0.0060	1.24 (0.25)	1.11E-06	0.0060	1.24 (0.25)	1.11E-06	5.68E-12	3.45	1

rs143313202	-	20	5666007	NC_000020.11:g.5666007A>G	intergenic	G	A	0.0144	0.81 (0.17)	1.62E-06	0.0144	0.81 (0.17)	1.62E-06	1.18E-11	2.25	1
rs73195003	*BTG3*	21	17610076	NC_000021.9:g.17610076A>G	intron	G	A	0.0339	-0.52 (0.11)	1.71E-06	0.0339	-0.52 (0.11)	1.71E-06	1.32E-11	0.59	1
rs140599862	-	1	167559027	NC_000001.11:g.167559027T>C	intron, NCT	C	T	0.0032	1.71 (0.36)	2.32E-06	0.0032	1.71 (0.36)	2.32E-06	2.39E-11	5.53	1
rs80111761	*HS1BP3*	2	20615440	NC_000002.12:g.20615440A>C	downstream gene	C	A	0.0374	0.50 (0.11)	2.69E-06	0.0374	0.50 (0.11)	2.69E-06	3.21E-11	1.65	1
rs62125675	*HS1BP3*	2	20616068	NC_000002.12:g.20616068C>T	downstream gene	T	C	0.0374	0.50 (0.11)	2.69E-06	0.0374	0.50 (0.11)	2.69E-06	3.21E-11	1.65	1
rs73776179	*LAMA2*	6	129485004	NC_000006.12:g.129485004A>G	intron	G	A	0.0049	1.39 (0.30)	3.42E-06	0.0049	1.39 (0.30)	3.42E-06	5.11E-11	4.01	1
rs1454700296	*NT5C2*	10	103270756	NC_000010.11:g.103270756A>G	intron	G	A	0.0017	2.15 (0.46)	3.85E-06	0.0017	2.15 (0.46)	3.85E-06	6.45E-11	8.56	1
rs1315965692	*NT5C2*	10	103270757	NC_000010.11:g.103270757G>A	intron	A	G	0.0017	2.15 (0.46)	3.85E-06	0.0017	2.15 (0.46)	3.85E-06	6.45E-11	8.56	1
rs867934853	-	8	41159441	NC_000008.11:g.41159441G>T	downstream gene	T	G	0.0011	2.42 (0.53)	0.0000045	0.0011	2.42 (0.53)	4.50E-06	8.79E-11	11.26	1
rs563431181	*PXK*	3	58386670	NC_000003.12:g.58386670A>G	intron	G	A	0.0021	1.89 (0.41)	4.61E-06	0.0021	1.89 (0.41)	4.61E-06	9.21E-11	6.64	1
rs6708069	*HS1BP3*	2	20616814	NC_000002.12:g.20616814C>A	downstream gene	A	C	0.0360	0.50 (0.11)	5.07E-06	0.0360	0.50 (0.11)	5.07E-06	1.11E-10	1.64	1

Please refer to Table 1 for parameter detalis. Abbreviations: NCT, non-coding transcript.

2.4. Discovery-derived Rare Variant Polygenic Score Predicts 25(OH)D Levels and Deficiency

Using genome-wide significant and suggestive rare variants from the discovery GWAS, we constructed four polygenic scores (PGS) applying a clumping-and-thresholding approach. Each panel retained 355,852 variants (0.32% of tested SNPs) in the replication dataset (Table 4).

Table 4. Performance of Candidate polygenic scores (PGS) for vitamin D levels in the replication cohort.

PGS Score	PGS Name	Available variants/Variants in score (%)	Adjusted R^2	Correlation			
				(95% CI)	BETA (SE)	*P*-value	Rho
P<5x10-5_r2<0.2	Q-PGS1	355,852/112,176,302 (0.32%)	0.146	(0.133 - 0.162)	0.0004 (0.0076)	9.08E-12	0.0721
P<5x10-6_r2<0.2	Q-PGS2	355,852/112,176,302 (0.32%)	0.146	(0.133 - 0.162)	0.0004 (0.0076)	9.08E-12	0.0721
P<5x10-7_r2<0.3	Q-PGS3	355,852/112,176,302 (0.32%)	0.146	(0.133 - 0.162)	0.0004 (0.0076)	9.08E-12	0.0721
P<5x10-8_r2<0.2	Q-PGS4	355,852/112,176,302 (0.32%)	0.146	(0.133 - 0.162)	0.0004 (0.0076)	9.08E-12	0.0721

PGS, Polygenic Score; CI, Confidence interval; SE, Standard Error; Rho, Spearman correlation coefficient.

In linear regression adjusted for age, sex, and principal components (PCs) 1–4, all PGS demonstrated identical predictive performance for quantitative 25(OH)D levels in the replication cohort ($R^2 = 0.146$, 95% CI: 0.133–0.162, $\beta = 0.0004$ (SE = 0.0076, $P = 9.08 \times 10^{-12}$) (Figure S1). Spearman correlation coefficients between the PGS and observed vitamin D levels were 0.0721 across all thresholds ($P = 2.11 \times 10^{-11}$, Figure 6a, Figure S2). For the binary trait (vitamin D deficiency, <20 ng/mL), logistic regression models yielded positive classification accuracy (AUC = 0.548, $P = 9.22 \times 10^{-6}$; OR = 0.9995, 95% CI: 0.9993–0.9997) across all thresholds (Figure 6b, Figure S3).

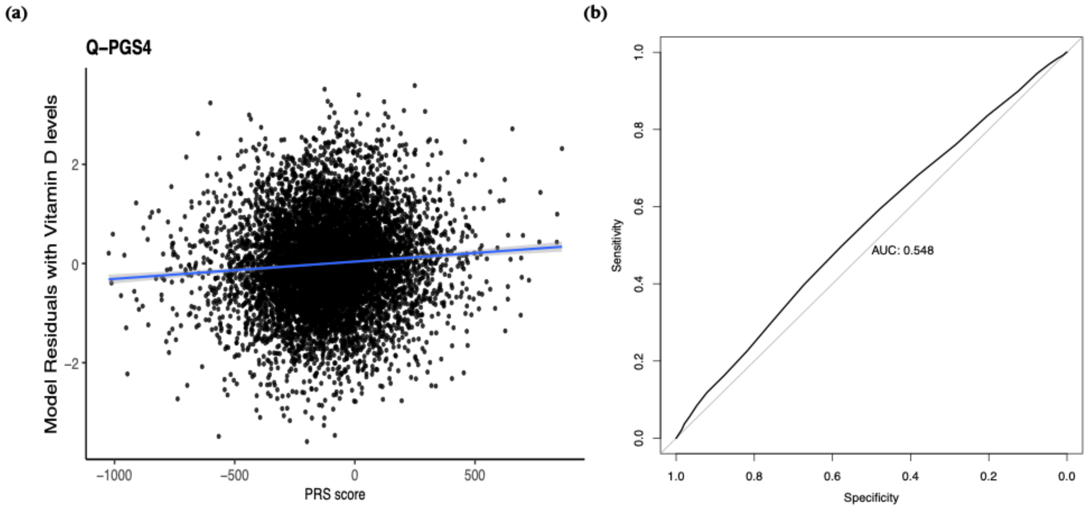

Figure 6. Predictive performance of discovery-derived polygenic risk scores (PRS) in the replication cohort. **(a)** Linear regression of inverse-normalized 25(OH)D levels on a representative PRS ($P < 5{\times}10^{-8}$), adjusted for age, sex, and the first four genetic principal components. **(b)** Receiver operating characteristic (ROC) curve for the same PRS predicting vitamin D deficiency (25(OH)D < 20 ng/mL; AUC = 0.548, $P = 9.22 \times 10^{-6}$; odds ratio = 0.9995; 95% CI: 0.9993–0.9997). Results for additional PRS thresholds ($P < 5{\times}10^{-6}$, $P < 5{\times}10^{-7}$, $P < 5{\times}10^{-8}$) are shown in the Supplementary Figures 1 and 2.

3. Discussion

This study presents the first GWAS in a Middle Eastern population to identify rare variants (MAF 0.01–0.0001) influencing continuous 25(OH)D concentrations and clinical vitamin D deficiency, leveraging high-coverage WGS data from over 13,000 QBB participants. The notably high prevalence of vitamin D deficiency in our cohort (>60%), consistent with earlier QBB and regional reports [3,4], underscores the public health relevance of identifying genetic determinants contributing to this trait. Despite abundant sunlight, deficiency remains widespread, a paradox attributed to a combination of cultural, lifestyle, metabolic, and genetic predisposition [1]. Most previous vitamin D GWAS have predominantly examined common variants in European cohorts [5,6], which may overlook ancestry-specific or low-frequency signals relevant to Middle Eastern populations. Our rare-variant analysis complements earlier QBB work on common-variant associations, identifying multiple genome-wide significant signals replicated across independent Qatari cohorts, thereby broadening the known genetic architecture of vitamin D and demonstrating the value of WGS in underrepresented groups.

Across quantitative and binary GWAS, we observed strong concordance of allele frequencies and effect sizes between discovery and replication cohorts, with over 60% of overlapping variants showing consistent effect directions. This reproducibility underscores the robustness of association signals across independent QBB subsets, even when individual cohort analyses lacked genome-wide significance. For binary traits, the magnitude of β estimates reflects the per-allele change in log-odds of deficiency; for rare variants, such effect sizes may correspond to substantial individual-level risk differences despite limited impact at the population level [7]. Mapping vitamin D deficiency as a binary trait poses inherent challenges, as dichotomizing a continuous biomarker reduces statistical power and environmental influences introduce misclassification [1,13]. The modest yield observed in single-cohort analyses in our study is therefore more likely attributable to these methodological and biological constraints than to technical bias.

Our quantitative GWAS identified 41 and 46 genome-wide significant rare variants in the discovery and replication cohorts, respectively, with several mapping to biologically plausible genes involved in lipid metabolism and nutrient transport. For example, *CD36*, a class B scavenger receptor implicated in lipid uptake, has also been linked to intestinal absorption and systemic transport of vitamin D [14], providing a potential mechanistic link to our observed associations. The rare variant rs192198195 in CD36 was associated with a 0.86-fold lower inverse-normalized 25(OH)D concentration per effect allele, potentially reducing systemic vitamin D bioavailability.

Similarly, *SLC16A7* (rs889439631) and *TMEM135* (chr11:87081213:G:T) variants were associated with an estimated ~0.95-fold reduction in vitamin D concentration per effect allele. *SLC16A7* and *TMEM135* encode membrane transporters involved in monocarboxylate and lipid handling, pathways relevant to the intracellular trafficking of lipophilic compounds like vitamin D [15,16]. These mechanisms are distinct from the canonical *GC*, *CYP2R1*, *DHCR7*, *MGAM*, and *PHF2* loci [3–6], suggesting that rare variants may capture regulatory and transport processes under stronger selective constraint. The combined meta-analysis further identified strong signals in *CNTN3* (rs115651661) and *EBF1* (rs536115678), associated with ~2.1-fold higher and ~0.18-fold lower vitamin D levels per allele, respectively. While the role in vitamin D metabolism is unclear, *EBF1* may influence vitamin D via metabolic–endocrine regulation [17], highlighting the potential impact of rare variants acting through noncanonical pathways.

In the binary deficiency analysis, no genome-wide significant variants were detected in the discovery cohort; however, replication revealed a significant association at *RAP1GAP* (rs577185477), with a 67% higher odds of deficiency per risk allele. *RAP1GAP* negatively regulates RAP1, a GTPase influencing endothelial integrity and cell adhesion—processes linked to vitamin D biology through vascular and immune pathways [18]. Given evidence connecting vitamin D deficiency to endothelial dysfunction, and supplementation can improve vascular health [19], this variant may act via endothelial stability pathways and warrants further investigation.

Meta-analysis across cohorts substantially improved power, uncovering rare-variant associations beyond single-cohort detection— paralleling European vitamin D GWAS where low-frequency variants with larger effects complement established *GC*, *CYP2R1*, and *DHCR7* loci [3,4].

Notably, *RDH13*, encoding mitochondrial retinol dehydrogenase, connects retinoid metabolism to the VDR–retinoid X receptor (RXR) transcriptional complex, a core regulator of vitamin D-responsive gene expression [20]. Additionally, rs952825245 in *SLC25A37* (Mitoferrin-1), a mitochondrial iron importer, was associated with ≥50% higher odds of deficiency per allele. Given iron's role in immune and endocrine function, and reported interplay with vitamin D metabolism [21], this association may reflect an indirect yet biologically relevant mechanism.

Although several PGS for vitamin D have been developed from common-variant GWAS in European cohorts [6], no published PGS currently incorporates rare variants. In our previous work, we evaluated the performance of a European-derived PRS [22], which demonstrated markedly reduced predictive performance in Qataris (R = 0.098) compared to the R ≈ 0.46 reported in Europeans, and achieved only modest discrimination for vitamin D deficiency [3]. Similar findings were observed in our Lebanese cohort, where European-derived PRS performance was also diminished [4]. In both cases, predictive power was far lower than in the original European populations, underscoring the limitations of cross-ancestry portability for common-variant PRS.

In the present study, we focused exclusively on evaluating the performance of rare-variant-based PGS derived from our Qatari discovery dataset. Using genome-wide significant and suggestive rare variants from the discovery GWAS, we constructed four population-specific PRS panels, retaining over 350,000 variants in the replication dataset. Despite the low frequency of contributing alleles, these PRS explained a ~14.6% of variance in continuous 25(OH)D levels—substantially exceeding the <2% variance explained by a European-derived PRS when applied to regional data [3,4]. This improvement is consistent with broader evidence that ancestry-specific models outperform those developed in other populations, largely due to differences in linkage disequilibrium (LD) structure, allele frequency, and variant architecture [3,23]. While predictive accuracy for binary deficiency was modest (AUC = 0.548)—consistent with European PRS studies where common variants explain most variance (AUC ≈ 0.59–0.61) [22]—our findings demonstrate that incorporating ancestry-specific rare alleles captures meaningful genetic risk for vitamin D deficiency. Integrating these rare-variant signals with common-variant predictors may further enhance risk stratification in Middle Eastern populations.

Our findings provide novel insights into the genetic architecture of vitamin D status in a large, underrepresented Middle Eastern population, yet several avenues remain for further investigation. Further efforts should explore ultra-rare variants (MAF <0.0001), expand analyses to diverse ancestries, and incorporate detailed environmental and lifestyle data to refine estimates of genetic effects. Functional validation of key candidates, such as *RAP1GAP*, *RDH13*, and *SLC25A37*, will be essential to clarify causal mechanisms in vitamin D regulation. Broader cross-ancestry replication, gene–environment interaction studies, and improved polygenic models integrating rare and common variants could ultimately enhance risk prediction and guide precision strategies for preventing vitamin D deficiency.

In summary, this rare-variant GWAS in a Middle Eastern population identifies novel genetic determinants of vitamin D status, including *RAP1GAP*, *RDH13*, and *SLC25A37*, implicating diverse metabolic and regulatory pathways. While clinical prediction of deficiency remains challenging, these findings establish a robust genetic foundation for functional validation and for developing precision medicine strategies tailored to populations with a high burden of deficiency.

4. Materials and Methods

4.1. Study Population and Ethical Approvals

This study was conducted using data from the QBB linked to the Qatar Genome Program (QGP) WGS. QBB enrolls adult participants aged 18 years and older who are Qatari nationals or long-term residents (≥15 years). All participants undergo standardized clinical assessments, provide detailed lifestyle and medical history information through questionnaires, and contribute biological samples, including blood, urine, and saliva. QBB recruitment, sample handling, and data access procedures are described previously [3,24].

The current analysis focused on 13,808 Qatari individuals from the QBB cohort with high-quality whole-genome sequencing from the Qatar Genome Program, of whom 13,652 with available serum 25(OH)D measurements were included in the final analysis. This dataset was randomly split into two non-overlapping subsets for independent analysis: a discovery cohort (QGP6013, *n* = 5,885) and a replication cohort (QGP7795, *n* = 7,767), to allow internal validation of genetic findings within the same national cohort. Participants were recruited following ethical approval from the QBB Institutional Review Board (IRB project number, QF-QGP-RES-ACC-00075), and all individuals gave written informed consent before participation.

4.2. Phenotype Measurements and Related Covariates

Serum 25(OH)D levels were quantified in the diagnostic laboratories of Hamad Medical Corporation using a standardized chemiluminescent immunoassay (CLIA) platform (LIAISON, DiaSorin, Germany). Briefly, blood samples were centrifuged for serum separation and stored at −80°C before biochemical analysis. Full methodological details, including assay protocol and instrument calibration, have been previously described.

Two phenotype definitions were used in this study. First, for quantitative genetic analysis, raw serum 25(OH)D concentrations (in ng/mL) were normalized using rank-based inverse-normal transformation implemented in R (v3.4.0). This transformation normalizes the distribution, minimizes skewness, and reduces the influence of outliers in association models. Second, a binary vitamin D deficiency phenotype was defined using serum 25(OH)D concentrations, with individuals classified as deficient if 25(OH)D < 20 ng/mL and as sufficient controls if 25(OH)D > 30 ng/mL. These clinically relevant thresholds are consistent with international guidelines and have been widely adopted in nutritional epidemiology and previous vitamin D GWAS.

Anthropometric measures, including weight and height, were obtained during physical examinations using standardized equipment (Seca 284 stadiometer and balance), and BMI was calculated as weight (kg) divided by the square of height (m^2).

4.3. Whole-Genome Sequencing and Quality Control

The procedures for genomic DNA extraction and WGS have been detailed in a prior publication [25]. Concisely, DNA quantity and integrity were assessed using the Quant-iT dsDNA Assay Kit (Invitrogen, USA) and FlexStation 3 reader (Molecular Devices, USA). High-coverage WGS (30× depth) was performed at Sidra Medicine Genomics Facility, Qatar, on the Illumina HiSeq X Ten (Illumina, USA). Raw reads were quality-checked with FastQC (v0.11.2), aligned to the human reference genome GRCh38 using BWA-MEM (v0.7.12), and variants were called using GATK HaplotypeCaller (v3.4). Joint genotyping was conducted on consolidated gVCFs via GenomicsDB, and variants were filtered via GATK's Variant Quality Score Recalibration (VQSR), retaining only "PASS" variants for analysis.

Stringent quality control (QC) procedures were applied at both the sample and variant level to ensure high confidence in genetic findings through PLINK (v2.0) [26]. Samples were excluded if they had genotype call rates <95%, ambiguous or mismatched gender, excess heterozygosity (>±4 standard deviations from the mean), or were identified as duplicates. Population structure was assessed using multidimensional scaling (MDS) and pairwise identity-by-state (IBS) analysis, based on a pruned set of independent autosomal SNPs selected using an LD threshold of $r^2 < 0.05$ within a sliding window of 200 SNPs. Individuals deviating more than ±4 standard deviations from the first two MDS components were flagged as population outliers and excluded. Variant-level QC excluded SNPs with MAF outside the rare range of interest (i.e., <0.0001 or >0.01), genotype call rate <90%, those on the X chromosome, or those deviating from Hardy-Weinberg equilibrium (P-value < 1×10^{-6}). The final analysis comprised 49,260,795 high-quality rare variants (MAF 0.0001–0.01) in the discovery dataset and 56,600,172 in the replication dataset, analyzed separately.

4.4. Genome-Wide Association Analyses

Genome-wide association analyses were conducted using a two-stage design to identify rare genetic variants (MAF 0.0001–0.01) associated with serum 25(OH)D levels. The discovery stage included 5,884 participants from the QGP6013 cohort, while the replication stage involved 7,767 independent participants from the QGP7795 cohort. Association testing was performed separately in each cohort using the SAIGE/R package (Scalable and Accurate Implementation of GEneralized mixed model), a computationally efficient mixed-model regression framework that accounts for sample relatedness and case–control imbalance [27].

For the quantitative trait (rank-based inverse-normalized 25(OH)D), linear mixed models were applied; for the binary trait of vitamin D deficiency, logistic mixed models were used. All models included covariates for age, sex, and the first four genetic PCs to correct for population stratification. As all samples were collected during a comparable sunny season in Qatar, seasonality was not considered a confounding factor in the analysis. Genome-wide significance was defined as $P < 5\times10^{-8}$, suggestive significance as $P < 1\times10^{-5}$, and nominal significance as $P < 0.05$.

To evaluate replication and assess the correlation of effect sizes and allele frequencies between cohorts, we first identified loci associated with 25(OH)D that were driven by the same lead variants in both the discovery and replication datasets. We then examined variants within a ±250 kb region flanking each genome-wide significant signal from the discovery cohort to identify additional replicated associations. Variants passing genome-wide significance ($P < 5 \times 10^{-8}$) in discovery were considered replicated if they showed nominal significance ($P < 0.05$) and had a consistent direction of effect (based on beta coefficients) in the replication cohort. To assess cross-cohort concordance, we performed linear regression analyses comparing allele frequencies and effect sizes between cohorts, reporting the regression slope, 95% CI, and R^2.

Subsequently, a fixed-effects inverse-variance-weighted meta-analysis was implemented in PLINK (v2.0) to combine results across both datasets [26]. Heterogeneity across cohorts was assessed using Cochran's Q test [28]. The impact of associated variants was evaluated based on the magnitude and direction of their beta coefficients. Q–Q plots, Manhattan plots, and λ_{GC} were generated using R (v4.4.1). Our discovery dataset demonstrated sufficient statistical power (≥95%) to detect variants with an effect size of $\beta = -3.03$ at the genome-wide significance threshold ($P < 5 \times 10^{-8}$).

4.5. Polygenic Score Construction and Evaluation

PGS were constructed to estimate individual genetic predisposition to serum 25(OH)D levels based on the cumulative effect of associated rare variants. We derived the PGS using summary statistics from the discovery GWAS dataset, applying the clumping-and-thresholding (C+T) approach as implemented in PLINK (v1.9) [29]. This method identifies sets of independent variants by grouping SNPs in LD around index SNPs based on a predefined LD threshold ($r^2 = 0.2$) and applying varying P-value cutoffs to include only statistically relevant variants.

PGS were generated across a range of P-value thresholds (from 5×10^{-8} to 5×10^{-1}), resulting in multiple rare-variant-based PGS panels, sequentially labeled from Q-PGS1 to QGP-4 according to decreasing statistical stringency. Each PGS was tested for predictive performance in the replication cohort (QGP7795, $n = 7{,}767$) using linear regression models for the continuous trait and logistic regression for vitamin D deficiency. Models were adjusted for age, sex, and the first four PCs to account for population structure. The optimal PGS panel was identified based on the highest adjusted R^2 for the quantitative phenotype and AUC for the binary phenotype.

4.6. Variant Annotation and Functional Characterization

Genome-wide significant variants identified in GWAS and meta-analyses were annotated using the Ensembl Variant Effect Predictor (VEP; GRCh38, release 114) [30]. Functional consequences (e.g., missense, intronic, intergenic), gene proximity, and regulatory annotations were also extracted.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1. Predictive performance (adjusted R²) of rare-variant polygenic scores for quantitative 25(OH)D across multiple P-value thresholds in the replication cohort; Figure S2: Linear regression of inverse-normalized baseline vitamin D levels on weighted PRS derived from a discovery dataset; Figure S3: ROC curves evaluating the predictive performance of the discovery-derived PRS in the replication cohort for vitamin D deficiency. Table S1: Rare variants identified in quantitative genome-wide analyses for circulating 25-hydroxyvitamin D levels in Discovery cohort; Table S2. Rare variants identified in quantitative genome-wide analyses for circulating 25-hydroxyvitamin D levels in replication cohort; Table S3. Replication of rare SNP associations for 25-hydroxyvitamin D levels between Discovery and Replication cohorts; Table S4. Rare variants identified in binary genome-wide analyses for vitamin D deficiency in discovery cohort; Table S5. Rare variants identified in binary genome-wide analyses for vitamin D deficiency in replication cohort; Table S6. Replication of rare SNP associations for vitamin D deficiency between Discovery and Replication cohorts; Table S7. Rare variants significantly associated with vitamin D levels discovered from a meta-analysis of the Replication and Discovery QGP Quantitative GWAS studies.

Author Contributions: Conceptualization, NNH, OA, and GN; Methodology, NNH and U-KU; Software, NNH and U-KU; Validation, NNH, OA, and GN; Formal Analysis, NNH and U-KU; Investigation, NNH, OA, and GN; Resources, OA and GN; Data Curation, NNH, U-KU; Writing – Original Draft Preparation, NNH; Writing – Review & Editing, NNH, U-KU, OA, and GN; Visualization, NNH and U-KU; Supervision, OA and GN; Project Administration, OA and GN; Funding Acquisition, OA and GN.

Funding: The open-access publication was supported by the Qatar National Library.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Qatar Biobank (IRB project number: QF-QGP-RES-ACC-00075, January 1, 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to licensing and access restrictions. Due to data privacy regulations, raw whole-genome sequencing data from Qatar Biobank cannot be deposited in public repositories. Access to QBB/QGP phenotype and sequencing data is available through an ISO-certified process, requiring submission of a project application at https://www.qatarbiobank.org.qa/research/how-apply and approval by the QBB Institutional Review Board. The GWAS summary statistics generated in this study are available in the NHGRI-EBI GWAS Catalog (https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90654001-GCST90655000/) under accession numbers GCST90654670, GCST90654671, GCST90654669, and GCST90654667. The corresponding polygenic score (PGS) data have been deposited in the PGS Catalog (https://www.pgscatalog.org) under accession number insert PGS ID (under process).

Acknowledgments: We acknowledge the Qatar Biobank (QBB) and Qatar Genome Program (QGP) for providing access to genomic and phenotypic data and thank all participants for their contributions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

25(OH)D	25-hydroxyvitamin D
AUC	Area under the curve
AGO4	Argonaute Component 4
BMI	Body mass index
CI	Confidence interval
CLIA	Chemiluminescent immunoassay
C+T	Clumping and thresholding
GC	Group-specific component (vitamin D-binding protein)
GWAS	Genome-wide association study

DHCR7	7-dehydrocholesterol reductase
IBS	Identity-by-state
IRB	Institutional Review Board
LD	Linkage disequilibrium
MAF	Minor allele frequency
MDS	Multidimensional scaling
OR	Odds ratio
PC	Principal component
PGS	Polygenic score
QC	Quality control
QBB	Qatar Biobank
QGP	Qatar Genome Program
QGS	Qatar Genome sequencing
Q–Q	Quantile–quantile
R^2	Coefficient of determination
RAP1GAP	Rap1 GTPase activating protein
RDH13	Retinol dehydrogenase 13
RXR	Retinoid X receptor
SAIGE	Scalable and Accurate Implementation of GEneralized mixed model
SE	Standard error
SNP	Single nucleotide polymorphism
SLC	Solute carrier
TMEM	Transmembrane protein
VDR	Vitamin D receptor
VEP	Variant Effect Predictor
WGS	Whole-genome sequencing

References

1. Hendi, N.N.; Nemer, G. Epigenetic regulation of vitamin D deficiency. *Epigenomics* **2023**, *15*, 653-655, doi:10.2217/epi-2023-0246.
2. Hendi, N.N.; Nemer, G. SDR42E1 modulates vitamin D absorption and cancer pathogenesis: insights from an in vitro model. *Front Endocrinol (Lausanne)* **2025**, *16*, 1585859, doi:10.3389/fendo.2025.1585859.
3. Hendi, N.N.; Al-Sarraj, Y.; Ismail Umlai, U.K.; Suhre, K.; Nemer, G.; Albagha, O. Genetic determinants of Vitamin D deficiency in the Middle Eastern Qatari population: a genome-wide association study. *Front Nutr* **2023**, *10*, 1242257, doi:10.3389/fnut.2023.1242257.
4. Hendi, N.N.; Chakhtoura, M.; Al-Sarraj, Y.; Basha, D.S.; Albagha, O.; Fuleihan, G.E.; Nemer, G. The Genetic Architecture of Vitamin D Deficiency among an Elderly Lebanese Middle Eastern Population: An Exome-Wide Association Study. *Nutrients* **2023**, *15*, doi:10.3390/nu15143216.
5. Revez, J.A.; Lin, T.; Qiao, Z.; Xue, A.; Holtz, Y.; Zhu, Z.; Zeng, J.; Wang, H.; Sidorenko, J.; Kemper, K.E.; et al. Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat Commun* **2020**, *11*, 1647, doi:10.1038/s41467-020-15421-7.
6. Manousaki, D.; Mitchell, R.; Dudding, T.; Haworth, S.; Harroud, A.; Forgetta, V.; Shah, R.L.; Luan, J.; Langenberg, C.; Timpson, N.J.; et al. Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci. *Am J Hum Genet* **2020**, *106*, 327-337, doi:10.1016/j.ajhg.2020.01.017.
7. Lee, S.; Abecasis, G.R.; Boehnke, M.; Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **2014**, *95*, 5-23, doi:10.1016/j.ajhg.2014.06.009.
8. Consortium, U.K.; Walter, K.; Min, J.L.; Huang, J.; Crooks, L.; Memari, Y.; McCarthy, S.; Perry, J.R.; Xu, C.; Futema, M.; et al. The UK10K project identifies rare variants in health and disease. *Nature* **2015**, *526*, 82-90, doi:10.1038/nature14962.
9. Kingdom, R.; Beaumont, R.N.; Wood, A.R.; Weedon, M.N.; Wright, C.F. Genetic modifiers of rare variants in monogenic developmental disorder loci. *Nat Genet* **2024**, *56*, 861-868, doi:10.1038/s41588-024-01710-0.

10. Xie, Y.; Acosta, J.N.; Ye, Y.; Demarais, Z.S.; Conlon, C.J.; Chen, M.; Zhao, H.; Falcone, G.J. Whole-Exome Sequencing Analyses Support a Role of Vitamin D Metabolism in Ischemic Stroke. *Stroke* **2023**, *54*, 800-809, doi:10.1161/STROKEAHA.122.040883.
11. Kim, Y.A.; Yoon, J.W.; Lee, Y.; Choi, H.J.; Yun, J.W.; Bae, E.; Kwon, S.H.; Ahn, S.E.; Do, A.R.; Jin, H.; et al. Unveiling Genetic Variants Underlying Vitamin D Deficiency in Multiple Korean Cohorts by a Genome-Wide Association Study. *Endocrinol Metab (Seoul)* **2021**, *36*, 1189-1200, doi:10.3803/EnM.2021.1241.
12. Martin, R.J.; McKnight, A.J.; Patterson, C.C.; Sadlier, D.M.; Maxwell, A.P.; Warren, U.K.G.S.G. A rare haplotype of the vitamin D receptor gene is protective against diabetic nephropathy. *Nephrol Dial Transplant* **2010**, *25*, 497-503, doi:10.1093/ndt/gfp515.
13. Fedorov, V.; Mannino, F.; Zhang, R. Consequences of dichotomization. *Pharm Stat* **2009**, *8*, 50-61, doi:10.1002/pst.331.
14. Oh, J.; Weng, S.; Felton, S.K.; Bhandare, S.; Riek, A.; Butler, B.; Proctor, B.M.; Petty, M.; Chen, Z.; Schechtman, K.B.; et al. 1,25(OH)2 vitamin d inhibits foam cell formation and suppresses macrophage cholesterol uptake in patients with type 2 diabetes mellitus. *Circulation* **2009**, *120*, 687-698, doi:10.1161/CIRCULATIONAHA.109.856070.
15. Wang, Y.; Liu, L.; Liu, X.; Tan, X.; Zhu, Y.; Luo, N.; Zhao, G.; Cui, H.; Wen, J. SLC16A7 Promotes Triglyceride Deposition by De Novo Lipogenesis in Chicken Muscle Tissue. *Biology (Basel)* **2022**, *11*, doi:10.3390/biology11111547.
16. Chornyi, S.; L, I.J.; van Roermund, C.W.T.; Wanders, R.J.A.; Waterham, H.R. Peroxisomal Metabolite and Cofactor Transport in Humans. *Front Cell Dev Biol* **2020**, *8*, 613892, doi:10.3389/fcell.2020.613892.
17. Griffin, M.J.; Zhou, Y.; Kang, S.; Zhang, X.; Mikkelsen, T.S.; Rosen, E.D. Early B-cell factor-1 (EBF1) is a key regulator of metabolic and inflammatory signaling pathways in mature adipocytes. *J Biol Chem* **2013**, *288*, 35925-35939, doi:10.1074/jbc.M113.491936.
18. Song, Y.S.; Jamali, N.; Sorenson, C.M.; Sheibani, N. Vitamin D Receptor Expression Limits the Angiogenic and Inflammatory Properties of Retinal Endothelial Cells. *Cells* **2023**, *12*, doi:10.3390/cells12020335.
19. Reynolds, J.A.; Haque, S.; Williamson, K.; Ray, D.W.; Alexander, M.Y.; Bruce, I.N. Vitamin D improves endothelial dysfunction and restores myeloid angiogenic cell function via reduced CXCL-10 expression in systemic lupus erythematosus. *Sci Rep* **2016**, *6*, 22341, doi:10.1038/srep22341.
20. Mianesaz, H.; Goczi, L.; Nagy, G.; Poliska, S.; Fadel, L.; Bojcsuk, D.; Penyige, A.; Szirak, K.; AlHaman, F.; Nagy, L.; et al. Genomic regions occupied by both RARalpha and VDR are involved in the convergence and cooperation of retinoid and vitamin D signaling pathways. *Nucleic Acids Res* **2025**, *53*, doi:10.1093/nar/gkaf230.
21. Azizi-Soleiman, F.; Vafa, M.; Abiri, B.; Safavi, M. Effects of Iron on Vitamin D Metabolism: A Systematic Review. *Int J Prev Med* **2016**, *7*, 126, doi:10.4103/2008-7802.195212.
22. Sinnott-Armstrong, N.; Tanigawa, Y.; Amar, D.; Mars, N.; Benner, C.; Aguirre, M.; Venkataraman, G.R.; Wainberg, M.; Ollila, H.M.; Kiiskinen, T.; et al. Author Correction: Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **2021**, *53*, 1622, doi:10.1038/s41588-021-00956-2.
23. Gunn, S.; Wang, X.; Posner, D.C.; Cho, K.; Huffman, J.E.; Gaziano, M.; Wilson, P.W.; Sun, Y.V.; Peloso, G.; Lunetta, K.L. Comparison of methods for building polygenic scores for diverse populations. *HGG Adv* **2025**, *6*, 100355, doi:10.1016/j.xhgg.2024.100355.
24. Al Thani, A.; Fthenou, E.; Paparrodopoulos, S.; Al Marri, A.; Shi, Z.; Qafoud, F.; Afifi, N. Qatar Biobank Cohort Study: Study Design and First Results. *Am J Epidemiol* **2019**, *188*, 1420-1433, doi:10.1093/aje/kwz084.
25. Thareja, G.; Al-Sarraj, Y.; Belkadi, A.; Almotawa, M.; Qatar Genome Program Research, C.; Suhre, K.; Albagha, O.M.E. Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nat Commun* **2021**, *12*, 1250, doi:10.1038/s41467-021-21381-3.
26. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, 7, doi:10.1186/s13742-015-0047-8.
27. Zhou, W.; Nielsen, J.B.; Fritsche, L.G.; Dey, R.; Gabrielsen, M.E.; Wolford, B.N.; LeFaive, J.; VandeHaar, P.; Gagliano, S.A.; Gifford, A.; et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **2018**, *50*, 1335-1341, doi:10.1038/s41588-018-0184-y.

28. Huedo-Medina, T.B.; Sanchez-Meca, J.; Marin-Martinez, F.; Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods* **2006**, *11*, 193-206, doi:10.1037/1082-989X.11.2.193.
29. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **2007**, *81*, 559-575, doi:10.1086/519795.
30. Dyer, S.C.; Austine-Orimoloye, O.; Azov, A.G.; Barba, M.; Barnes, I.; Barrera-Enriquez, V.P.; Becker, A.; Bennett, R.; Beracochea, M.; Berry, A.; et al. Ensembl 2025. *Nucleic Acids Res* **2025**, *53*, D948-D957, doi:10.1093/nar/gkae1071.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.