Title page

Title: Low Frequency Synonymous Coding Variation in *CYP2R1* has Large Effects on Vitamin D Level and Risk of Multiple Sclerosis.

Short title: Low-frequency Variant Confers Large Effect on Vitamin D levels

Authors: Despoina Manousaki[1,2,47], Tom Dudding[3,47], Simon Haworth[3,47], Yi-Hsiang Hsu [4,5,6 ,47], Ching-Ti Liu[7,47], Carolina Medina-Gómez [8,9,10,47] Trudy Voortman[9,10,47], Nathalie van der Velde[8,11,47], Håkan Melhus[12,47], Cassianne Robinson-Cohen[13,47], Diana L. Cousminer [14,15,47] Maria Nethander[16,17,47], Liesbeth Vandenput[16,47] Raymond Noordam[18,47], Vincenzo Forgetta[1,2], Celia MT Greenwood[1,2,19,20] Mary L. Biggs[21], Bruce M. Psaty[22,23], Jerome I. Rotter[24], Babette S. Zemel[25,26], Jonathan A. Mitchell[25,26], Bruce Taylor[27], Mattias Lorentzon[16,28,29], Magnus Karlsson[30] , Vincent V.W. Jaddoe [9,10] , Henning Tiemeier[9,10,31], Natalia Campos-Obando[8], Oscar H.Franco[10], Andre G. Utterlinden[8,9,10], Linda Broer[8], Natasja M. van Schoor[32], Annelies C. Ham[8], M. Arfan Ikram[10,33], David Karasik[4], Renée de Mutsert[34], Frits R. Rosendaal[34], Martin den Heijer[35] ,Thomas J. Wang[36], Lars Lind[12,48], Eric S.Orwoll [37,38,48], Dennis O. Mook-Kanamori[34,39,48], Karl Michaëlsson[40,48], Bryan Kestenbaum [13,48], Claes Ohlsson[16,48], Dan Mellström[16,28,48] , Lisette CPGM de Groot[41,48], Struan F.A. Grant[14, 25, 42,48] ,Douglas P. Kiel[4,5,6,43,48], M. Carola Zillikens[8,48], Fernando Rivadeneira[8,9,10,48], Stephen Sawcer[44,48] , Nicholas J Timpson[3,48] and J. Brent Richards[1,2,45,46,48]

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 1B1, Canada

[2]Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada

[3.]Medical Research Council Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, BS8 2BN, UK

[4]Institute for Aging Research, Hebrew SeniorLife, Boston, MA 02131, USA

[5]Harvard Medical School, Boston, MA 02115, USA

[6]Broad Institute of MIT and Harvard, Boston, MA 02142, USA

[7]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

[8]Department of Internal Medicine, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands

[9]The Generation R Study Group, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands

[10]Department of Epidemiology, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands

[11] Department of Internal Medicine, Section of Geriatrics, Academic Medical Center, Amsterdam, 1105 AZ, The Netherlands

[12]Department of medical sciences, Uppsala university, Uppsala, 751 85, Sweden

[13] Kidney Research Institute, Division of Nephrology, University of Washington, Seattle, WA 98195, USA

[14]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[15]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[16]Centre for Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, 40530, Sweden

[17]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, 41390, Sweden

[18]Department of Internal Medicine, section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, 2333 ZA, the Netherlands

[19]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC H3A 1A2, Canada

[20]Department of Oncology, McGill University, Montreal, QC H4A 3T2, Canada

[21] Cardiovascular Health Research Unit, Departments of Medicine and Biostatistics, University of Washington, Seattle, WA 98101, USA

[22]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA 98101, USA

[23]Kaiser Permanente Washington Health Research Unit, Seattle, WA 98101, USA.

[24]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics at Harbor-UCLA Medical Center, Torrance, CA 90502, USA

[25]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[26]Division of Gastroenterology, Hepatology and Nutrition, The Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA

[27] Menzies Institute for Medical Research University of Tasmania, Locked Bag 23, Hobart, Tasmania 7000, Australia

[28]Geriatric Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, 43180, Mölndal, Sweden

[29]Geriatric Medicine, Sahlgrenska University Hospital, 43180, Mölndal, Sweden.

[30]Clinical and Molecular Osteoporosis Research Unit, Department of Clinical Sciences, Lund University, and Department of Orthopaedics, Skåne University Hospital, 22241, Malmö, Sweden

[31]Department of Child and Adolescent Psychiatry/Psychology, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands

[32]Department of Epidemiology and Biostatistics and the EMGO Institute of Health and Care Research, VU University Medical Center, Amsterdam, 1081 HV, The Netherlands

[33]Department of Radiology and Nuclear Medicine, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands

[34]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[35]Department of Endocrinology, VU University Medical Center, Amsterdam, 1081 HV, The Netherlands

[36] Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[37]Bone and Mineral Unit, Oregon Health & Science University, Portland, OR 97239, USA

[38]Department of Medicine, Oregon Health & Science University, Portland, OR 97239, USA

[39]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[40]Department of surgical sciences, Uppsala university, 75105, Uppsala, Sweden

[41]Division of Human Nutrition, Wageningen University, Wageningen, 6708 WE, The Netherlands

[42]Division of Endocrinology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[43]Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

[44]University of Cambridge, Department of Clinical Neurosciences, Box 165, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0QQ, UK

[45]Department of Twin Research and Genetic Epidemiology, King's College London, London, WC2R 2LS, United Kingdom

[46]Department of Medicine, McGill University, Montreal, QC H3G 1Y6, Canada

[47]These authors contributed equally to this work

[48] These authors contributed equally to this work

Corresponding author:

Brent Richards, MD, MSc

Pavillon H-413, Jewish General Hospital

3755 Cote Ste Catherine

Montreal, QC, Canada, H3T 1E2

T: +1 514 340 8222 ext. 4362

F: +1 514 340 7529

brent.richards@mcgill.ca

Number of tables: 5, number of figures: 7. This article has Supplemental Data

Abstract

Vitamin D insufficiency is common, correctable and influenced by genetic factors, and it has been associated with risk of several diseases. We sought to identify low-frequency genetic variants that strongly increased the risk of vitamin D insufficiency and tested their effect on risk of multiple sclerosis, a disease influenced by low vitamin D concentrations. We used whole-genome sequencing data from 2,619 individuals through the UK10K program and deep imputation data from 39,655 genome-wide genotyped individuals. Meta-analysis of the summary statistics from 19 cohorts identified a low-frequency synonymous coding p.Asp120Asp variant (rs117913124[A], minor allele frequency=2.5%) in *CYP2R1* which conferred a large effect on 25-hydroxyvitamin D (25OHD) levels (-0.43 standard deviations of standardized natural log-transformed 25OHD, per A allele, P-value = 1.5×10^{-88}). The effect on 25OHD was four-times larger and independent of the effect of a previously described common variant near *CYP2R1*. By analyzing 8,711 individuals we showed that heterozygote carriers of this low-frequency variant have an increased risk of vitamin D insufficiency (OR=2.2, 95% CI 1.78-2.78, $P=1.26 \times 10^{-12}$). Individuals carrying one copy of this variant had also an increased odds of multiple sclerosis (OR=1.4, 95%CI 1.19-1.64, $P=2.63 \times 10^{-5}$) in a sample of 5,927 cases and 5,599 controls. In conclusion, we describe a low-frequency coding variant in *CYP2R1*, which exerts the largest effect upon 25OHD levels identified to date in the general European population and implicates vitamin D in the etiology of multiple sclerosis. **(235 words)**

Introduction

Vitamin D insufficiency affects approximately 40% of the general population in developed countries [1]. This may have important public health consequences, since vitamin D insufficiency has been associated with musculoskeletal consequences and several common diseases, such as multiple sclerosis (MIM:126200), types 1 and 2 diabetes (MIM:222100 and MIM:125853) and several cancers [2]. Further, repletion of vitamin D status can be achieved safely and inexpensively. Thus, understanding the determinants of vitamin D insufficiency, and their effects, can provide a better understanding of the role of vitamin D in disease susceptibility with potentially important public health benefits.

Approximately half of the variability in the concentration of the widely accepted biomarker for vitamin D status, 25-hydroxyvitamin D (25OHD), has been attributed to genetic factors in twin and family studies [3; 4]. Four common genetic variants (minor allele frequency [MAF] >5%) in loci near four genes known to be involved in cholesterol synthesis (*DHCR7* [MIM;602858]), hydroxylation (*CYP2R1* [MIM:608713]), vitamin D transport (*GC* [MIM:139200]) and catabolism (*CYP24A1* [MIM:126065]) are strongly associated with 25OHD levels, yet explain little of its heritability [5]. Low-frequency and rare genetic variants (defined as variants with a MAF of ≤5% and ≤1% respectively) have recently been found to have large effects on clinically relevant traits [6-8] providing an opportunity to better understand the biologic mechanisms influencing disease susceptibility in the general population.

Therefore, the principal objective of the present study was to detect low-frequency and rare variants with large effects on 25OHD levels, through a large-scale meta-analysis and describe their biological and clinical relevance. Similar to an earlier genome-wide association study (GWAS) studying common genetic variation (MAF ≥5%) by the SUNLIGHT Consortium [5], we sought to increase understanding of the genetic etiology of vitamin D variation within the general population, however, our current study focused on genetic variation with a MAF <5%. This has only recently been made possible through whole-genome sequencing and the use of improved genotype imputation for low frequency and rare variants, with the recent availability of large whole genome sequencing reference panels [9]. The second objective of this study was to better understand if low-frequency genetic variants with large effects on 25OHD could predict a higher risk of vitamin D insufficiency in their carriers, and whether vitamin D intake through diet may interact with such genetic factors to prevent, or magnify, vitamin D insufficiency. Finally, we sought to understand whether these

genetic determinants of 25OHD levels are implicated in multiple sclerosis, a disease influenced by low 25OHD levels[10].

To do so, we first undertook an association study of whole-genome sequence data and deeply imputed genome-wide genotypes to identify novel genetic determinants of vitamin D in 42,274 individuals. We next tested if these genetic variants conferred a higher risk of vitamin D insufficiency in 8,711 subjects and whether this insufficiency showed effect modification by dietary intake. Last we assessed their effect on multiple sclerosis in a separate sample of 5,927 cases and 5,599 controls.

Material and Methods

Cohorts

All human studies were approved by each respective institutional or national ethics review committees, and all participants provided written informed consent. To investigate the role of rare and low-frequency genetic variation on 25OHD levels in individuals of European descent, we used whole genome sequencing (WGS) data at mean read depth of 6.7x in 2,619 subjects from two cohorts in the UK10K project [11] with available 25OHD phenotypes **(Table 1).** We also used imputation reference panels to impute variants that were missing, or poorly captured, from previous GWAS in 39,655 subjects **(Table 1 and Figure 1).** The participating individuals were drawn from independent cohorts of individuals of European descent**.** Detailed description of each of the participating studies is provided in **Table S1.**

25OHD Measurements

The methods applied to measure 25OHD levels differed among the participating cohorts (**Tables S1 and S6)**. The four methods used were tandem mass spectrometry (in BMDCS, MrOS and BPROOF), combined high-performance liquid chromatography with mass spectrometry (in ALSPAC, BPROOF, CHS, ULSAM, NEO, Generation R), chemiluminescence immunoassay (DiaSorin, Inc, Stillwater, MN) (in TUK, PIVUS, FHS, MrOS Malmo, MrOS GBG and GOOD) and an electrochemiluminescence immunoassay (COBAS, Roche Diagnostics GmbH) (in RSI, RSII and RSIII). Detection limits for the different methods are provided in the **Table S6**.

Whole-Genome Sequencing, Genotyping and Imputation

ALSPAC WGS and TUK WGS cohorts had been sequenced at an average read depth of 6.7x through the UK10K consortium (www.UK10K.org) using the Illumina HiSeq platform, and aligned to the GRCh37 human reference using Burrows-Wheeler Aligner (BWA)31[12]. Single-nucleotide variant (SNV) calls were completed using samtools/bcftools[13], and VQSR[14] and GATK were used to recall these variants. The whole genome sequencing for the ALSPAC and TwinsUK cohorts has been described in detail in a previous publication from our group[7]. **Table S8** summarizes the data generation method for sequencing-based cohorts.

Participating studies separately genotyped samples and imputed them to WGS-based reference panels. The most recent imputation panels, such as the UK10K and 1000Genomes Project (v3) combined panel, which in total contained 7,562 haplotypes from the UK10K Project and 2,184 haplotypes from the 1000 Genomes Project [9], and the Haplotype Reference Consortium (HRC) panel, with 64,976 haplotypes[15], enabled more

accurate imputation of low frequency variants, when compared to the UK10K or the 1000Genomes reference panel alone[9]. Specifically, 11 out of the 17 participating cohorts were imputed to the UK10K and 1000 Genomes reference panel (total number of imputed individuals included in the meta-analysis N=25,589). Three of the participating cohorts were imputed using the HRC panel (total number of imputed individuals N=5,717). Finally, 2 cohorts were imputed to the 1000Genomes panel (N=7,536), and 1 cohort was imputed to the UK10K panel (N=863). **(Table S1**). Details on genotyping methods and imputation for the 17 participating cohorts are presented in **Table S6.** Info scores for the imputed SNVs per participating cohort are presented in **Table S7.** To assess the quality of imputation, we tested the non-reference discordance rate for the low frequency genome-wide significant SNVs and found this to be 0% (**Table S9**).

Association Testing for 25OHD levels and Meta-analysis

A GWAS was conducted separately by each cohort using an additive genetic model for 25OHD levels. Because 25OHD concentrations were measured using different methods, log-transformed 25OHD levels were standardized to z-scores, after being adjusted for age, sex, BMI, and season of measurement. Specifically, the phenotype for each GWAS study was prepared according to the following steps:1) 25OHD levels were log-transformed to ensure normality 2) Linear regression models were used to generate cohort-specific residuals of log transformed 25OHD levels adjusted for covariates (age, sex, BMI and season). Season was treated as a non-ordinal categorical variable (summer: July to September, fall: October to December, winter: January to March, and spring: April to June). 3) The mean of log transformed 25OHD levels was added to the residuals to create the adjusted 25OHD phenotype. 4) The above phenotype was then normalized within each cohort (mean of zero with SD of one) to make the phenotype consistent across cohorts, since 25OHD levels have been measured in different cohorts in our consortium using different methods. 5) Finally, outliers beyond 5 standard deviations were removed from step (4).

For comparison purposes, we computed the average 25OHD levels, adjusted for age, sex, BMI and season of measurement, in one cohort of our meta-analysis (TUK WGS) in carriers and non-carriers of the lead SNV(s).

The software used by each cohort to perform a GWAS is listed in **Table S1**. Single variant tests were undertaken for variants with MAF>0.1%, using an additive effect of the minor allele at each variant in each cohort. The type of software employed for single variant testing for each cohort is shown in **Table S1**.

Studies with related individuals used software that accounted for relatedness. Cohort-specific genomic inflation factors (lambdas) are also shown in **Table S1** (the mean lambda was 1.015).

We then meta-analyzed association results from all discovery cohorts (N total = 42,274). This stage included validation of results file format, filtering files by the above QC criteria, comparison of trait distributions among different studies, identification of potential biases (large betas and/or standard errors, inconsistent effect allele frequencies, extreme lambdas). Meta-analysis quality control of the GWAS data included the following SNV-level exclusion criteria: i) Info score <0.4, ii) HWE P-value $<10^{-6}$ iii) Missingness >0.05, and iv) MAF <0.5%. Alignment of the SNVs across studies was done using the chromosome and position information for each variant according to genome build hg19. SNVs in the X chromosome were not included in the meta-analysis. Fixed–effects meta-analysis was performed using the software package GWAMA[16] adjusting for genomic control. We tested bi-allelic SNVs with MAF ≥ 0.5% for association, declaring genome-wide statistical significance at $P \leq 1.2 \times 10^{-8}$ for variants present in more than one study. This stringent p-value threshold was set to adjust for all independent SNVs above the MAF threshold of 0.5%.[17]

Conditional analysis was undertaken for the four previously described lead vitamin D SNVs from the SUNLIGHT consortium using the GCTA package [18]. This method uses an approximate conditional analysis approach from summary-level statistics from the meta-analysis and linkage disequilibrium corrections between SNVs estimated from a reference sample. We used UK10K individuals as the reference sample to calculate the linkage disequilibrium information of SNVs. The associated regions flanking within 400kb of the top SNVs from SUNLIGHT were extracted and the conditional analyses were conducted within these regions. Conditional analyses of individual variants presented in **Table 2** and **Table S5** were conducted using GCTA v 0.93.9 using default parameters.

Haplotype block analyses were used for the candidate variants of interest by deriving phased haplotypes from 1013 individuals from the TUK WGS cohort using a custom R package.

Effects on Vitamin D insufficiency

To investigate the effect of genome-wide significant SNVs on vitamin D insufficiency (defined as 25OHD levels below 50 nmol/L), we used data from 4 cohorts: TUK Imputed, TUK WGS, BPROOF and MrOS (n_{total}=8,711). Logistic regression of this binary phenotype was performed against the SNVs, adjusting for the

following covariates: age, sex, BMI, and season of measurement. Meta-analysis of cohort-level summary statistics was performed in R [19] using the epitools [20] and metafor packages[21].

Interaction analysis with Vitamin D intake

We undertook an interaction analysis of our candidate SNV(s) with vitamin D dietary intake (continuous and tertiles) in 9,224 individuals from five of the cohorts participating in our discovery phase (Framingham, PIVUS, ULSAM, BPROOF and RSIII). A detailed description of the method to capture vitamin D intake in each one of the participating cohorts appears in **Table S6**. Linear regression was conducted in each of these studies under an additive genetic model. The following variables and co-variables were included in the model: log-transformed serum 25OHD as the dependent variable; SNV genotype (coded as 0, 1 or 2) as an independent variable; SNV (genotype)* dietary vitamin D intake (continuous or tertiles respectively) as an interaction term; age, sex, BMI, season of 25OHD measurement, dietary vitamin D intake (continuous or tertiles), supplemented vitamin D (yes/no), and total energy intake as covariates. The results from the 5 studies were meta-analyzed using a fixed-effects model using the metafor tool of the R statistical package.

Effects on Multiple Sclerosis

We tested the effect of the genome-wide significant SNVs on the risk of multiple sclerosis in 5,927 cases and 5,599 controls, assuming an additive genetic model. Controls were obtained from the UK Biobank [22] by random selection of participants without multiple sclerosis. The cases were obtained from UK Biobank[22], previously published MS GWAS[23; 24] and newly genotyped UK patients. Prior to genotype imputation of the genotyped cases, numerous quality control criteria were applied to ensure unbiased genotype calls between cohorts. These included retaining only SNVs with MAF > 1% and excluding SNVs or samples with high missingness [25]. Further, samples were assessed for population stratification using EIGENSTRAT [26; 27] and outliers were removed. Genotype data was then imputed using the Sanger Imputation Service[15] with the combined UK10K and 1000 Genomes Phase 3 reference panels[9; 28], the same reference panel used for the UK Biobank controls. Genotype data was phased using EAGLE2[29] and imputed using PBWT[30]. Association testing was undertaken using SNPTEST[31] on the combined case/control dataset, testing the additive effect of each allele on multiple sclerosis status, and including the top 10 principal components from EIGENSTRAT [26; 27] to adjust for population stratification and batch effects.

Results

GWAS

After strict quality control, the genomic inflation factor for the meta-analysis of 19 GWAS studies was 0.99, suggesting lack of bias due to population stratification (**Figure 2**). Through meta-analysis of 11,026,511 sequenced and imputed variants from our discovery cohorts **(Table 1)**, we identified a signal at the chromosome 11p.15.2 locus, harboring variants associated with 25OHD levels (lead low-frequency SNV p.Asp120Asp [rs117913124(A)], MAF = 2.5%, allelic effect size = -0.43 standard deviations of the standardized log-transformed 25OHD levels [SD], $P = 1.5\times10^{-88}$, **Figure 3 and Table 2)**. The direction of effect was consistent across all discovery cohorts **(Table 3 and Figure 3A)** and the mean imputation information score for the imputed studies was 0.97. This low-frequency synonymous coding variant is in exon 4 of the *CYP2R1* and is ~14 kb from the previously identified common *CYP2R1* variant, rs10741657 (r^2 between these two SNVs= 0.03) **(Figure 4).** To our knowledge**,** the rs117913124 SNV has not previously been associated with any vitamin D-related traits in humans.

A comparison of the average 25OHD levels, adjusted for age, sex, BMI and season of measurement, in non-carriers and heterozygote carriers of the A allele of rs117913124 in the TUK WGS appears in **Figure S1.** The average 25OHD levels, adjusted for age, sex, BMI and season of measurement were computed in 542 individuals from the Twins UK WGS cohort, among which 510 were no carriers and 32 were heterozygote carriers of the A allele of rs117913124 (no homozygote carriers present in this cohort). After removing outliers (adjusted 25OHD levels below and above 3 SD from the mean), we included in our analysis 449 non-carriers and 30 heterozygote carriers (for a total of 479 individuals). A linear regression model with the adjusted 25OHD levels as the dependent variable and the dose of the “A” allele of rs117913124 (numeric factor, 1 or 0) as the independent variable demonstrated a 8.3 nmol/L decrease in the adjusted 25OHD levels per “A” allele. The mean adjusted 25OHD levels were 64.3 nmol/L in non-carriers vs 56.0 nmol/L in heterozygote carriers.

Two-way conditional analysis between the *CYP2R1* common (rs10741657) and low-frequency (rs117913124) variants revealed that the two association signals are largely independent. Specifically, after conditioning on rs10741657, rs117913124 remained strongly associated with 25OHD level ($P_{cond} = 2.4x10^{-78}$); after conditioning on rs11791324, the effect of rs10741657 on 25OHD level remained significant ($P_{cond}= 4.0\ x10^{-33}$ versus $P_{pre\text{-}cond}= 8.8\ X\ 10^{-45}$) **(Table 2 and Table S5).** Further, no other low frequency variant in the region remained significant when conditioning on rs117913124 (**Table 2**). To further disentangle the role of rs117913124 from rs10741657 on 25OHD levels, we undertook a haplotype analysis based on WGS data from 3,781 individuals from the TUK WGS and ALSPAC WGS cohorts. We found that the 25OHD

decreasing allele A of rs117913124 was always transmitted in the same haplotype block with the 25OHD decreasing allele G of the common *CYP2R1* variant rs10741657. By using 25OHD data from the TUK WGS cohort, we compared the 25OHD levels among carriers of the various haplotype blocks. We observed evidence of decrease in the 25OHD levels in carriers of the A allele of the rs117913124 compared to non-carriers independent of the presence of the effect allele G of the common *CYP2R1* variant (**Table 4).**

No other low-frequency or rare variants were identified in the three previously described vitamin D-related loci at *DHCR7*, *GC* and *CYP24A1*. The mean effect size of the four previously reported common genome-wide significant SNVs (MAF ≥ 5%) from the SUNLIGHT consortium was -0.13 SD and the largest effect size was -0.25 SD (for the *GC* variant) in our meta-analysis **(Table S3 and Figure 3B).** The effect size of rs10741657(G), the known common *CYP2R1* variant, was -0.09 SD. Hence, the observed effect size of rs117913124 is 3-fold larger than the above mean, 4-fold larger that of the common *CYP2R1* variant and almost twice that of the largest previously reported effect of the *GC* variant. Last, the percentage of the variance of the 25OHD phenotype explained by the low-frequency *CYP2R1* variant was more than double than the percentage of the variance explained by the *CYP2R1* common variant (0.9% vs 0.4%).

We also identified 18 genome-wide significant low-frequency and rare SNVs on the same chromosome 11 region as rs117914124 located in the neighboring *PDE3B* (MIM:602047) **(Table 2, Table S4 and Figure 4B).** Signals from these SNVs in *PDE3B* were independent of the common variant at *CYP2R1* (**Table 2).** We then created haplotype blocks with rs117913124 and SNVs at *PDE3B* based on haplotype information from the 3,781 individuals from the TUK WGS and ALSPAC WGS cohorts (**Table S2**). We found that the 25OHD decreasing allele (A) of the rs117913124 was always inherited with the 25OHD decreasing allele (A) of its perfect proxy rs116970203 (r^2=1). Therefore, rs116970203 is not likely to have a distinct effect from rs117913124 on 25OHD levels. On the other hand, the 25OHD decreasing alleles of the remaining four low-frequency variants (all having a MAF of approximately 1.4%) were not always inherited in the same haplotype block as the rs117913124 and rs116970203 and were in moderate linkage disequilibrium with the rs117913124 (all r^2< 0.6, **Figure 4B and Figure 4C**). Each of the four alleles is in almost perfect linkage disequilibrium with the remaining three (all r^2 >0.96). This implied that these four SNVs might influence 25OHD levels independently of the rs117913124. Nevertheless, as mentioned above, when conditioning on the lead low-frequency *CYP2R1* SNV rs117913124, the P-values of the 4 *PDE3B* SNVs became non-significant and their betas decreased substantially (**Table 2**), demonstrating that they likely do not represent an independent signal at the chromosome 11 locus.

rs117913124 and risk of vitamin D insufficiency

To further investigate the clinical significance of the low-frequency *CYP2R1* variant rs117913124*,* we tested its effect on a binary outcome for vitamin D insufficiency (defined as 25OHD levels < 50 nmol/L) in 8,711 individuals from 4 studies (TUK WGS, TUK IMP, BPROOF and MROS). rs117913124 was strongly associated with an increased risk of vitamin D insufficiency (OR = 2.20*,* 95% CI 1.8-2.8, P =1.2 x 10^{-12})

(**Figure 5),** after control for relevant covariates as described in the Methods section**.**

Common 25OHD-associated SNVs

We report two additional loci associated with 25OHD levels (**Table 5**). Variants leading these associations were common and exerted a rather small effect on 25OHD: first, a variant in chromosome 12 (rs3819817[C], intronic to *HAL* [MIM:609457]), with a MAF of 45%, a beta of 0.04 and a P-value of 3.2×10^{-10}. Second, a variant in chromosome 14 (rs2277458[G], intronic to *GEMIN2* [MIM:602595]), with a MAF of 21%, a beta of -0.05 and a P-value of 6.0×10^{-9}. Both variants were present in all 19 studies, and the direction of the effect was the same among the 19 studies (**Figure 6**). Neither the *HAL* nor the *GEMIN2* loci are previously known to be associated with 25OHD levels. Of note, neither variant was present in the HapMap imputation reference used in the SUNLIGHT study.

Interaction analysis

CYP2R1 encodes the enzyme responsible for 25-hydroxylation of vitamin D in the liver [32], a necessary step in the conversion of dietary vitamin D and vitamin D oral supplements to the active metabolite, 1,25 dihydroxy-vitamin D. Therefore, we hypothesized that individuals heterozygous or homozygous for rs117913124 in *CYP2R1* would not show a response in their 25OHD levels to vitamin D intake compared to non-carriers. In other words, we expected carriers of the effect allele of rs117913124 to have steadily lower 25OHD levels, independently of their vitamin D intake. To investigate this hypothesis, we tested the presence of interaction of rs117913124 with vitamin D dietary intake (continuous values and tertiles) on 25OHD levels in 9,224 individuals from 5 studies **(Figure S2)**. We found no interaction between rs117913124 and dietary vitamin D intake (beta = -0.0002; P-value for interaction = 0.41 for continuous vitamin D intake and beta = 0.012; P-value = 0.60 for tertiles of vitamin D intake). Since the two common 25OHD-associated SNVs are located in genes (*HAL* and *GEMIN2*) with no known role in the processing of dietary vitamin D, we found no biological rationale for undertaking a gene-diet interaction analysis for these variants.

25OHD-assosiated variants and risk of multiple sclerosis

We tested whether the *CYP2R1* low-frequency variant rs117913124 and the common variants rsrs3819817 and rs2277458 in *HAL* and *GEMIN2*, respectively, influenced the risk of multiple sclerosis. In a sample of 5,927 multiple sclerosis cases and 5,599 controls, we found that the 25OHD decreasing allele at rs117913124[A], was associated with an increased odds of multiple sclerosis: OR = 1.40 (95%CI: 1.19-1.64);

P-value = 2.6 x 10^{-5}. By way of comparison, the OR of multiple sclerosis for the common *CYP2R1* variant was 1.03 (95%CI: 0.97-1.08); P-value 0.03 in the same multiple sclerosis study, and has previously been reported to be 1.05 (95%CI: 1.02-1.09); P-value 0.004 in a separate study [33]. Thus, the effect per allele of rs117913124 on multiple sclerosis was 12.4-fold larger than that attributed to the already known common variant at *CYP2R1*. With regards to the two common SNVs, the 25OHD decreasing allele [T] at the *HAL* variant rs3819817 was not clearly associated with risk of multiple sclerosis, however there was a trend in the expected direction: OR = 1.05 (95%CI: 1.00-1.11); P-value = 0.07. We found no association between the 25OHD decreasing allele [G] at the *GEMIN2* variant rs2277458 and risk of multiple sclerosis: OR = 1.03 (95%CI: 0.96-1.11); P-value = 0.34.

Discussion

Through the largest meta-analysis of genome-wide association studies for 25OHD levels in European populations to date, we have identified a low-frequency, synonymous coding genetic variant of large effect that strongly associates with 25OHD levels. This variant has an effect size four-fold larger than that described for the common variant in the same gene (*CYP2R1*) and is associated with two-fold increase in risk of vitamin D insufficiency and a 40% increase in the odds of developing multiple sclerosis. The biologic plausibility of these findings is supported by the fact that the low-frequency variant is located in *CYP2R1*, the major hepatic 25-hydroxylase for vitamin D [32]. These findings are of clinical relevance since 5% of the general European population carry this variant in either the homozygous or heterozygous state, and it is associated with a clinically relevant increase in the risk of multiple sclerosis.

Our study was enabled by large imputation reference panels (UK10K/1000 Genomes and HRC), which offer at least 10-fold more European samples than the 1000 Genomes reference panel alone. We did not identify genome-wide significant variants of large effect on 25OHD in novel genes in Europeans, although we found variants with smaller effects in two loci not previously known to be associated to 25OHD. Yet we did identify low-frequency variants in a known vitamin D related-gene with much larger effects than the previously described common variants.

CYP2R1 encodes the enzyme responsible for 25-hydroxylation of vitamin D, and is one of the two main enzymes responsible for vitamin D hepatic metabolism [32] (**Figure 7**). Rare mutations in *CYP2R1* have already been described to cause rickets (MIM: 27744) [32; 34]. Due to the important role of *CYP2R1* in the conversion of dietary vitamin D and vitamin D oral supplements to the active form of vitamin D, we hypothesized that carriers of the low-frequency *CYP2R1* variant might respond poorly to vitamin D replacement therapy. We tested this hypothesis by undertaking an interaction analysis between the *CYP2R1* low frequency variant and dietary vitamin D intake, which showed no clear interaction. However, we note that gene by environment interaction studies are generally underpowered, measurement error in dietary data is common, and this interaction was further limited by time differences between dietary intake assessment and measurement of 25OHD levels. Therefore, whether this genetic variant influences 25OHD response to vitamin D administration requires further study.

Although the aim of the present study was to describe variants of low MAF and large effect on 25OHD, we report two common genetic variants of small effect size on chromosome 12 (*HAL* gene) and chromosome 14 (*GEMIN2* gene) that reached genome-wide level significance in our meta-analysis. Although there is no existing evidence of implication of *GEMIN2* in vitamin D related physiologic pathways, *HAL* is expressed in the skin and is involved in formation of urocanic acid, a "natural sunscreen" [35; 36]. Thus, this could constitute a plausible pathophysiologic mechanism implicating *HAL* in vitamin D synthesis in the skin. Additional functional follow-up of the signals in chromosomes 12 and 14 is needed to characterize the genes and/or mechanisms underlying these associations.

Our findings may have clinical relevance for several reasons: First, individuals carrying at least one copy of the low-frequency *CYP2R1* variant have lowered levels of 25OHD by a clinically relevant degree. Specifically, the risk of vitamin D insufficiency is doubled in these individuals. Second, their risk of multiple sclerosis is also increased in accordance with previous evidence supporting a causal role for vitamin D in the risk of multiple sclerosis[10]. Third, these findings affect ~5% of individuals of European descent. And last, rs117913124 could be used as an additional genetic predictor of low 25OHD levels, along with the previously identified common vitamin D-related variants, in Mendelian randomization studies investigating the causal role of low vitamin D levels in human disease.

Our study also has its limitations. First, although the scope of our study was detection of low-frequency and rare variants, we opted to include in our meta-analysis two whole genome sequencing studies with a relatively low read depth of 6.7x, as well as three studies imputed to older imputation panels (1000Genomes and UK10K). These studies have a limited capacity to capture very rare variants, which might explain why we failed to identify such associations. The gene-diet interaction analysis, as mentioned above, may have lacked statistical power, in addition to the limitations arising from the time-difference between dietary vitamin D intake assessments and 25OHD measurements. Since our analysis is restricted to populations of European ancestry, we cannot make any assumptions concerning the effect of rs117913124 in non-European populations. Nonetheless, based on the 1000Genomes reference, this variant is rare in Africans (MAF = 0.3%) and has not been described in East Asians (MAF = 0%). Therefore, large sample sizes of these populations will be required to describe with any certainty the effect of this variant on 25OHD level in these populations. Finally, in the absence of functional experiments showing the exact function of the rs117913124 on *CYP2R1* and given that this synonymous polymorphism does not affect protein sequence, we cannot unequivocally confirm that this low-frequency variant is causal, however, given that this is a

coding variant in a well-documented 25OHD-associated gene, it seems most likely that it exerts its effect on *CYP2R1*.

In conclusion, our findings demonstrate the utility of whole-genome sequencing-based discovery and deep imputation to enable the characterization of genetic associations, offering an improved understanding of the pathophysiology of vitamin D, an enriched set of genetic predictors of 25OHD levels for future study, and enabling the identification of groups at increased risk for vitamin D insufficiency and multiple sclerosis.

Supplemental Data Description

Supplemental Data of this article include 2 figures, 9 Tables, Funding information, Author Information and Acknowledgements.

Acknowledgements

The authors have no conflicts of interest. Detailed acknowledgments are included in the Supplemental Data.

Web Resources

URL for Online Mendelian Inheritance in Man: http://www.omim.org

URL for the UK10K program: http://www.uk10k.org

URL for VQSLOD: http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

URL for GWAMA: http://www.geenivaramu.ee/en/tools/gwama

URL for GCTA: http://cnsgenomics.com/software/gcta/

REFERENCES

1. Forrest, K.Y., and Stuhldreher, W.L. (2011). Prevalence and correlates of vitamin D deficiency in US adults. Nutr Res 31, 48-54.
2. Rosen, C.J., Adams, J.S., Bikle, D.D., Black, D.M., Demay, M.B., Manson, J.E., Murad, M.H., and Kovacs, C.S. (2012). The nonskeletal effects of vitamin D: an Endocrine Society scientific statement. Endocr Rev 33, 456-492.
3. Shea, M.K., Benjamin, E.J., Dupuis, J., Massaro, J.M., Jacques, P.F., D'Agostino, R.B., Sr., Ordovas, J.M., O'Donnell, C.J., Dawson-Hughes, B., Vasan, R.S., et al. (2009). Genetic and non-genetic correlates of vitamins K and D. Eur J Clin Nutr 63, 458-464.

4. Livshits, G., Karasik, D., and Seibel, M.J. (1999). Statistical genetic analysis of plasma levels of vitamin D: familial study. Ann Hum Genet 63, 429-439.
5. Wang, T.J., Zhang, F., Richards, J.B., Kestenbaum, B., van Meurs, J.B., Berry, D., Kiel, D.P., Streeten, E.A., Ohlsson, C., Koller, D.L., et al. (2010). Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet 376, 180-188.
6. Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet 47, 1272-1281.
7. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112-117.
8. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305, 869-872.
9. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6, 8111.
10. Mokry, L.E., Ross, S., Ahmad, O.S., Forgetta, V., Smith, G.D., Goltzman, D., Leong, A., Greenwood, C.M., Thanassoulis, G., and Richards, J.B. (2015). Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. PLoS Med 12, e1001866.
11. 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. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.
12. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-2993.
13. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
14. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-498.
15. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48, 1279-1283.
16. Magi, R., and Morris, A.P. (2010). GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288.
17. Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., Greenwood, C.M., and Consortium, U.K. (2014). Estimating genome-wide significance for whole-genome sequencing studies. Genet Epidemiol 38, 281-290.
18. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88, 76-82.
19. Team, R.C. (2013). R: A language and environment for statistical computing. In. (R Foundation for Statistical Computing, Vienna, Austria.
20. Aragon T.J., Wollschlaeger D., Omidpanah A. (2017). epitools: Epidemiology Tools. https://cran.r-project.org/package=epitools
21. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. . Journal of Statistical Software 36, 1-48.
22. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying

the causes of a wide range of complex diseases of middle and old age. PLoS Med 12, e1001779.
23. International Multiple Sclerosis Genetics Consortium, Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., et al. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357, 851-862.
24. Australia, and New Zealand Multiple Sclerosis Genetics Consortium (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41, 824-828.
25. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat Protoc 5, 1564-1573.
26. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904-909.
27. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.
28. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
29. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48, 1443-1448.
30. Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics 30, 1266-1272.
31. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39, 906-913.
32. Cheng, J.B., Levine, M.A., Bell, N.H., Mangelsdorf, D.J., and Russell, D.W. (2004). Genetic evidence that the human CYP2R1 enzyme is a key vitamin D 25-hydroxylase. Proc Natl Acad Sci U S A 101, 7711-7715.
33. International Multiple Sclerosis Genetics Consortium, Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., et al. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 45, 1353-1360.
34. Casella, S.J., Reiner, B.J., Chen, T.C., Holick, M.F., and Harrison, H.E. (1994). A possible genetic defect in 25-hydroxylation as a cause of rickets. J Pediatr 124, 929-932.
35. Barresi, C., Stremnitzer, C., Mlitz, V., Kezic, S., Kammeyer, A., Ghannadan, M., Posa-Markaryan, K., Selden, C., Tschachler, E., and Eckhart, L. (2011). Increased sensitivity of histidinemic mice to UVB radiation suggests a crucial role of endogenous urocanic acid in photoprotection. J Invest Dermatol 131, 188-194.
36. Suchi, M., Sano, H., Mizuno, H., and Wada, Y. (1995). Molecular cloning and structural characterization of the human histidase gene (HAL). Genomics 29, 98-104.

Figure 1: Schematic of the discovery single variant meta-analysis

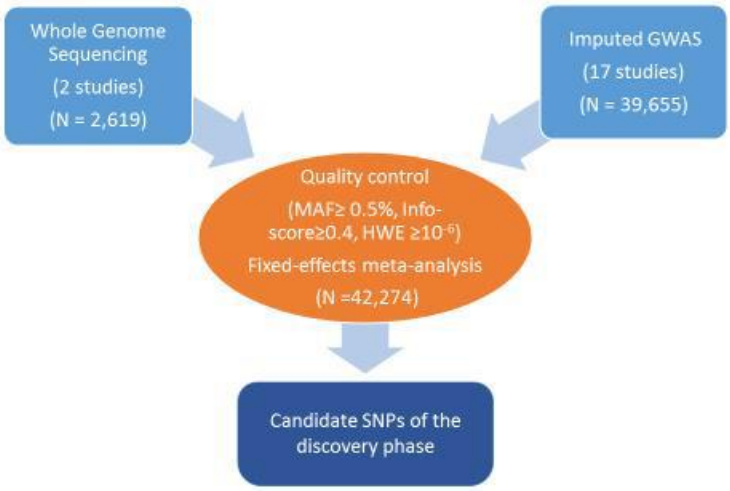
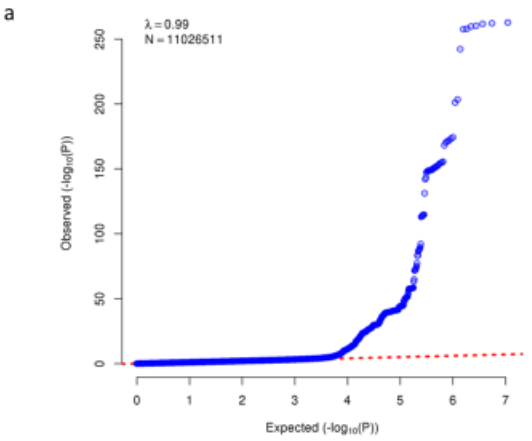


Figure 2: Discovery single-variant meta-analysis.

Legend: A. Quantile-quantile plot for the single SNV meta-analysis. B. Manhattan plot of the meta-analysis. The plot depicts variants with MAF > 0.5% across the 22 autosomes against the –log10 p-value from the meta-analysis of 19 cohorts, which included 42,274 individuals.



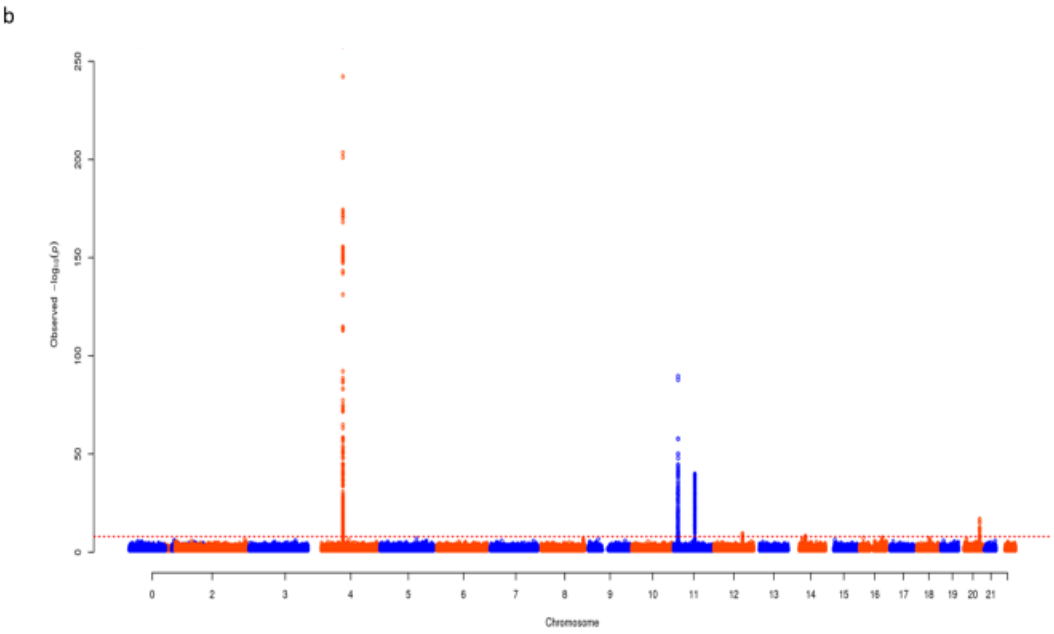
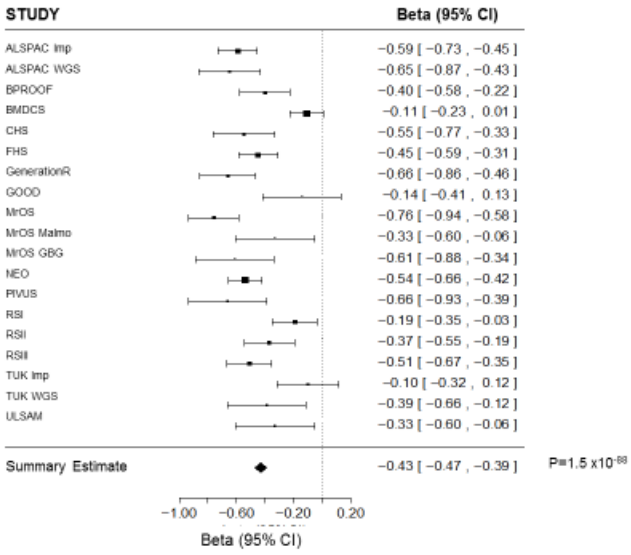


Figure 3: Forest Plot by Cohort for rs117913124 and Forest Plot of the rs117913124 and the Previously Described Common 25OHD-related Variants from Discovery Meta-analysis

Legend: A. Forest plot of estimates from all 19 studies for the low-frequency *CYP2R1* variant rs117913124 B. Forest-plot of the effect of the four common SUNLIGHT variants and of the *CYP2R1* low-frequency variant rs117913124 on log-transformed 25OHD levels.

a



b

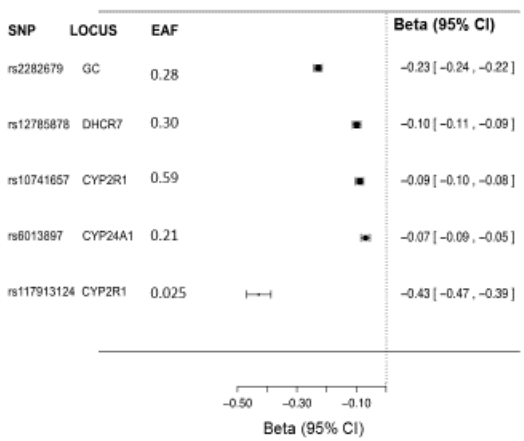


Figure 4: Association Signals from 11p.15.2

Legend: A. Snapshot from the UCSC genome-browser including the top low-frequency SNVs (see **Table 2**) and the lead common variant rs10741657 at the *CYP2R1* locus. The position of rs117913124 is highlighted in light blue. B. Regional disequilibrium plot showing the rs117913124 (purple dot), its perfect proxy rs11670203 (red dot) and the other genome-wide significant SNVs in the same locus (blue and green dots). The plot depicts SNVs within 1 Mb of a locus' lead SNV (x-axis) and their associated meta-analysis p value (-log10) (for more details see **Table S10**). SNVs are color coded according to r^2 with the lead SNV (labelled, r^2 calculated from UK10K whole genome sequencing dataset). Recombination rate (blue line), and the position of genes, their exons and the direction of transcription are also displayed (below plot). C. Linkage disequilibrium plot indicating the r^2 values between the SNVs of **Table 2** (top low-frequency variants) and between these low-frequency SNVs and the lead common variant (rs107416570) at the same *CYP2R1* locus (r^2 calculated from the 1000 Genomes dataset).

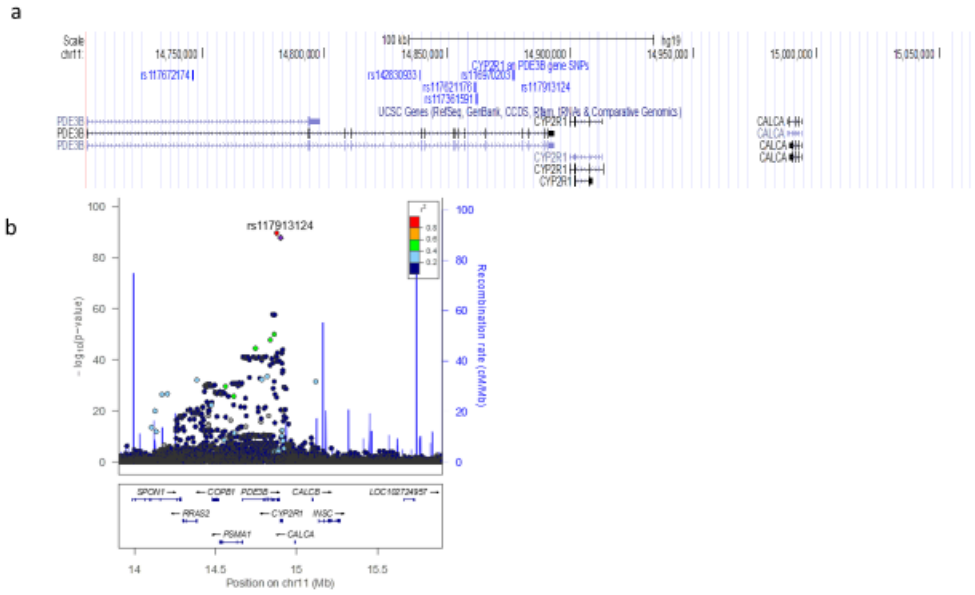


Figure 5: Effect of the rs117913124 on Vitamin D Insufficiency

Legend: Forest-plot of the effect of the low-frequency *CYP2R1* variant rs117913124 on vitamin D insufficiency in 4 studies.

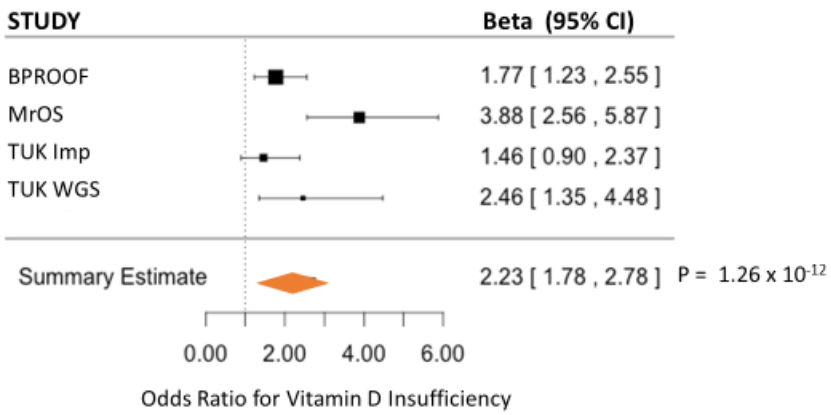


Figure 6: **Association Signals from Chromosomes 12 and 14**

Legend: Forest plots with A. estimates for the chromosome 12 common variant rs3819817 and B. estimates for the chromosome 14 common variant rs2277458 from all 19 studies of the meta-analysis where both variants were present.

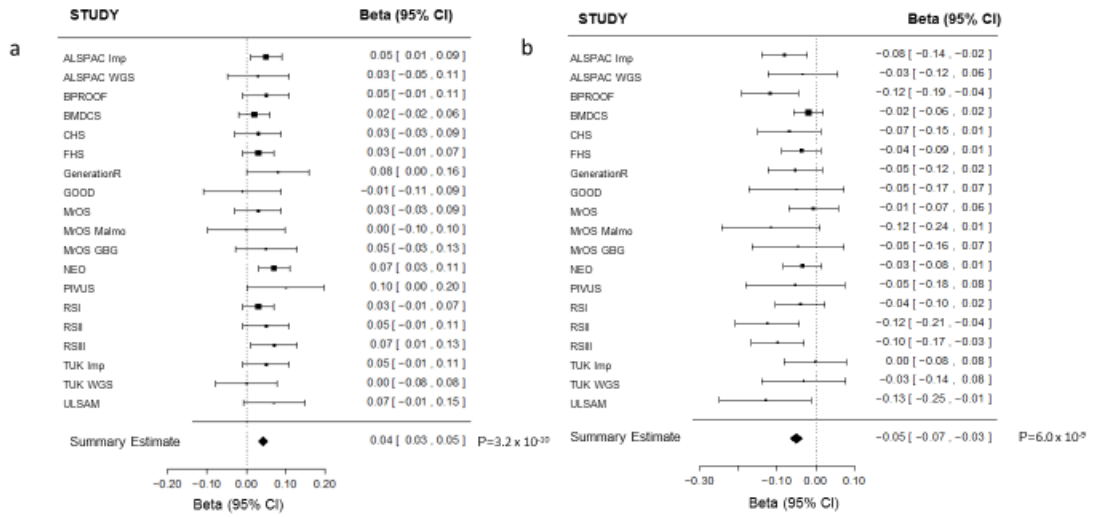


Figure 7: Schematic of the Vitamin D Metabolic Pathway

Legend: UVB: ultraviolet B rays.

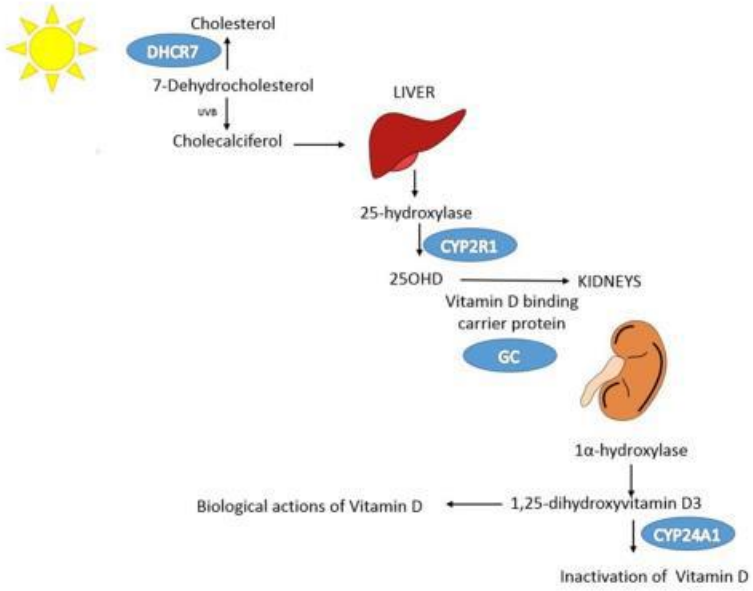


Table 1. Participating cohorts and number of DNA samples per cohort. WGS: Whole-Genome Sequenced

Study Acronym*	Imputed	WGS	TOTAL
ALSPAC	3,679	1,606	
TUK	1,919	1,013	
Generation R	1,442		
BPROOF	2,514		
FHS	5,402		
MrOS	3,265		
RSI	3,320		
RSII	2,022		
RSIII	2,913		
CHS	1,792		
BMDCS	863		
MrOS GBG	945		
GOOD	921		
MrOS Malmo	893		
PIVUS	943		
ULSAM	1,095		
NEO	5,727		
TOTAL	39,655	2,619	42,274

*For full names of the studies see Table S6

Table 2: Association results for genome-wide significant low-frequency variants from discovery 25OHD meta-analysis, before and after conditioning on the lead common *CYP2R1* SNP, rs10741657, and the lead low-frequency *CYP2R1* variant, rs117913124.

SNV	Chr	Position	EA*	EAF#	Candidate Gene	Function	Beta$	P-value		Beta$	P -value		Beta$	P-value	N
										Conditional on rs10741657			Conditional on rs117913124		
rs117913124	11	14900931	A	0.025	*CYP2R1*	exon 4 (synonymous codon)	-0.43	1.5×10^{-88}		-0.39	2.4×10^{-78}		NA	NA	41336
rs116970203		14876718	A	0.025	*CYP2R1 (nearest gene: PDE3B)*	Intron 11 variant	-0.43	2.2×10^{-90}		-0.40	3.3×10^{-80}		NA	NA	41138
rs117361591		14861957	T	0.014		Intron 11 variant	-0.44	9.1×10^{-51}		-0.40	2.2×10^{-44}		-0.05	0.017	38286
rs117621176		14861320	G	0.014		Intron 11 variant	-0.44	8.7×10^{-51}		-0.40	2.1×10^{-44}		-0.05	0.016	38273
rs142830933		14838760	C	0.014		Intron 5 variant	-0.44	1.4×10^{-48}		-0.40	1.7×10^{-42}		-0.05	0.03	37541
rs117672174		14746404	T	0.014		Intron 1 variant	-0.43	2.8×10^{-45}		-0.39	2.9×10^{-39}		-0.04	0.062	37209

*****Effect allele is the 25OHD decreasing allele

Effect allele frequency

$ Betas represent changes in standard deviations of the standardized log-transformed 25OHD levels

Table 3. Summary statistics results for the *CYP2R1* low-frequency variant, rs117913124, from 19 studies.

STUDY	25OHD measurement method#	N	Effect Allele A* Frequency	Beta$	Standard Error	P-value	Information score
ALSPAC Imputed	MS	3675	0.028	-0.59	0.07	$3.43x10^{-18}$	0.99
ALSPAC WGS	MS	1606	0.028	-0.65	0.11	$8.23x10^{-10}$	NA
BPROOF	MS	2512	0.027	-0.4	0.09	$4.99x10^{-6}$	0.97
BMDCS	MS	863	0.019	-0.11	0.06	0.058	0.98
CHS	MS	1581	0.022	-0.55	0.11	$5.15x10^{-7}$	0.88
FHS	CLIA	5402	0.021	-0.45	0.07	$2.32x10^{-10}$	0.97
GenerationR	MS	1442	0.033	-0.66	0.1	$1.78x10^{-6}$	1
GOOD	CLIA	921	0.028	-0.14	0.14	0.31	0.96
MrOS	MS	3265	0.018	-0.76	0.09	$5.63x10^{-16}$	0.96
MrOS Malmo	CLIA	893	0.033	-0.33	0.14	0.016	0.94
MrOS GBG	CLIA	945	0.026	-0.61	0.14	$7.87x10^{-6}$	1
NEO	MS	5727	0.025	-0.54	0.06	$2.73x10^{-19}$	1
PIVUS	CLIA	943	0.028	-0.66	0.14	$2.56x10^{-6}$	0.99
RSI	ECLIA	3320	0.025	-0.19	0.08	0.019	0.98
RSII	ECLIA	2022	0.033	-0.37	0.09	$2.38x10^{-5}$	0.99
RSIII	ECLIA	2913	0.027	-0.51	0.08	$4.61x10^{-10}$	0.98
TUK Imputed	CLIA	1919	0.021	-0.1	0.11	0.35	0.98
TUK WGS	CLIA	1013	0.025	-0.39	0.14	0.006	NA
ULSAM	MS	1095	0.025	-0.33	0.14	0.02	1

*****Effect allele is the 25OHD decreasing allele

MS: mass spectrometry, CLIA: chemiluminescence immunoassay, ECLIA: electrochemiluminescence immunoassay

$ Betas represent changes in standard deviations of the standardized log-transformed 25OHD levels

Table 4. Effect of different haplotype combinations of the low frequency (rs117913124) and the common (rs10741657) *CYP2R1* variants on 25OHD levels.

Results are based on individuals from the Twins UK Whole Genome Sequenced cohort (the first allele in each block is the rs117913124, the second allele is the rs10741657 for both chromatids). The two "AG" blocks in bold contain the 25OHD decreasing allele (A) of the low-frequency variant, which is always inherited with the 25OHD decreasing allele (G) of the common variant.

Haplotype		Beta$	P-value	N
rare/common*	rare/common*			
GA	**GA**	**-0.02**	**0.79**	**156**
AG	**GA**	**-0.49**	**0.02**	**23**
AG	**GG**	**-0.3**	**0.13**	**27**
GA	**GG**	**0.01**	**0.87**	**477**
GG	**GG**	**0.05**	**0.58**	**330**

*** The first allele in each chromatid corresponds to the low-frequency variant rs117913124; the second allele corresponds to the common variant rs10741657. 25OHD decreasing alleles appear in bold for both variants.**

$ Betas represent changes in standard deviations of the standardized log-transformed 25OHD levels

Table 5. Main findings of the GWAS meta-analysis

SNP	Chr	Candidate Gene	Effect allele	Effect allele frequency	Beta$	P-value	N
rs117913124	11	*CYP2R1*	A	0.025	-0.43	1.5×10^{-88}	41,336
rs3819817	12	*HAL*	C	0.45	0.04	3.2×10^{-10}	41,071
rs2277458	14	*GEMIN2*	G	0.21	-0.05	6.0×10^{-09}	39,746

$ Betas represent changes in standard deviations of the standardized log-transformed 25OHD levels, while controlling for age, sex, BMI and season of measurement