RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Exploring the Genetic Architecture of Circulating 25-Hydroxyvitamin D

Linda T. Hiraki,[1,2]* Jacqueline M. Major,[3] Constance Chen,[1] Marilyn C. Cornelis,[4] David J. Hunter,[1,2,5] Eric B. Rimm,[2,4,5] Kelly C. Simon,[4,5] Stephanie J. Weinstein,[3] Mark P. Purdue,[3] Kai Yu,[3] Demetrius Albanes,[3] and Peter Kraft[1,2]

[1]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts; [2]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts; [3]Division of Cancer Epidemiology and Genetics National Cancer Institute, NIH, Bethesda, Maryland; [4]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts; [5]Channing Laboratory Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts

Received 24 April 2012; Revised 25 September 2012; accepted revised manuscript 28 September 2012.
Published online 7 November 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21694

ABSTRACT: **The primary circulating form of vitamin D is 25-hydroxy vitamin D (25(OH)D), a modifiable trait linked with a growing number of chronic diseases. In addition to environmental determinants of 25(OH)D, including dietary sources and skin ultraviolet B (UVB) exposure, twin- and family-based studies suggest that genetics contribute substantially to vitamin D variability with heritability estimates ranging from 43% to 80%. Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) located in four gene regions associated with 25(OH)D. These SNPs collectively explain only a fraction of the heritability in 25(OH)D estimated by twin- and family-based studies. Using 25(OH)D concentrations and GWAS data on 5,575 subjects drawn from five cohorts, we hypothesized that genome-wide data, in the form of (1) a polygenic score comprised of hundreds or thousands of SNPs that do not individually reach GWAS significance, or (2) a linear mixed model for genome-wide complex trait analysis, would explain variance in measured circulating 25(OH)D beyond that explained by known genome-wide significant 25(OH)D-associated SNPs. GWAS identified SNPs explained 5.2% of the variation in circulating 25(OH)D in these samples and there was little evidence additional markers significantly improved predictive ability. On average, a polygenic score comprised of GWAS-identified SNPs explained a larger proportion of variation in circulating 25(OH)D than scores comprised of thousands of SNPs that were on average, nonsignificant. Employing a linear mixed model for genome-wide complex trait analysis explained little additional variability (range 0–22%). The absence of a significant polygenic effect in this relatively large sample suggests an oligogenetic architecture for 25(OH)D.** Genet Epidemiol 37:92–98, 2013. © 2012 Wiley Periodicals, Inc.

KEY WORDS: genome-wide association; heritability; polygenic score; vitamin D

Introduction

Vitamin D is a hormone essential for normal growth and development. The primary circulating form, 25-hydroxy vitamin D (25(OH)D), is a modifiable trait with well documented effects in bone, muscle, and mineral homeostasis, that has also been linked with a growing number of diseases ranging from malignancy, most notably colorectal cancer to autoimmune diseases including multiple sclerosis [Holick, 2004; Munger et al., 2006]. Although vitamin D can be derived from dietary sources, it is mainly produced by skin ultraviolet B (UVB) exposure. To obtain the biologically active metabolite 1, 25-dihydroxyvitamin D3 ($1,25(OH)_2D$), vitamin D undergoes a series of hydroxylation steps primarily in the liver and kidney. $1,25(OH)_2D$, is the most biologically active metabolite of vitamin D but its levels are under tight homeostatic regulation and thus 25(OH)D is considered the best measure of vitamin D status [Holick, 2004].

*Correspondence to: Linda Hiraki, Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, Bldg II, Room 200, Boston, MA 02115. E-mail: lindahiraki@mail.harvard.edu

In addition to environmental determinants of circulating vitamin D levels, twin- and family-based studies suggest that genetics contribute substantially to vitamin D variability with heritability estimates ranging from 43% to 80% [Hunter et al., 2001; Orton et al., 2008; Shea et al., 2009; Wjst et al., 2007]. These estimates are often derived from comparisons of the higher intraclass correlation (ICC) observed in monozygotic (MZ) twins to that of dizygotic twins (DZ), with this difference providing the first impression of the magnitude of genetic influence [Hunter et al., 2001; Orton et al., 2008; Shea et al., 2009].

There have been two published genome-wide association studies (GWAS) of circulating 25(OH)D concentrations [Ahn et al., 2010; Wang et al., 2010]. These studies identified associated variants in the group-specific component gene (*GC*) that encodes the vitamin D binding protein; in the gene encoding *CYP2R1* (cytochrome P450, family 2, subfamily R, polypeptide 1) involved in hydroxylation of vitamin D3 to 25(OH)D; in *CYP24A1*, which encodes 24-hydroxylase involved in the degradation of $1,25(OH)_2D$; in the gene encoding *DHCR7* (7-dehydrocholesterol (7-DHC) reductase),

which converts 7-DHC to cholesterol, thereby removing the substrate from the synthetic pathway of vitamin D3, a precursor of 25-hydroxyvitamin D3, and *NADSYN1* (nicotinamide adenine dinucleotide (NAD) synthetase) [Ahn et al., 2010; Wang et al., 2010]. However, these single nucleotide polymorphisms (SNPs) collectively explain only a fraction of the heritability in 25(OH)D estimated by twin- and family-based studies. As we describe further below, we estimate that the percentage of residual variance in circulating 25(OH)D explained by SNPs in these gene regions is approximately 5%, comparable to prior studies [Ahn et al., 2010; Signorello et al., 2011; Sinotte et al., 2009; Wang et al., 2010].

A major limitation of prior GWAS of vitamin D is that the effect sizes of individual alleles are often small, such that the predictive value of a single variant of small effect on circulating 25(OH)D levels is negligible [Evans et al., 2009]. In an effort to include loci with small effects that may not reach genome-wide significance in GWAS, genome-wide scores calculated by including such SNPs have demonstrated improved discriminative accuracy for complex disease affection status in bipolar disorder, schizophrenia, coronary heart disease, type I and type II diabetes, and multiple sclerosis [Bush et al., 2010; Evans et al., 2009; Purcell et al., 2009]. The challenge in employing this method is choosing the optimal threshold for deciding which loci to include in the calculation of the genome-wide score. Too liberal a threshold is likely to incorporate noise and hence obscure any true signal, whereas an overly strict threshold may cause one to disregard loci that genuinely contribute to disease risk. A solution to overcome this problem is to fit all the SNPs from the genome-wide platform simultaneously. The software tool, genome-wide complex trait analysis (GCTA) employs linear mixed models to fit the effects of all SNPs as random effects [Yang et al., 2011]. This approach provides an estimate of the percentage variance explained by all the SNPs genotyped and tagged by the genome-wide platform [Yang et al., 2010].

We hypothesized that by using genome-wide data in the form of (1) a polygenic score comprised of hundreds or thousands of SNPs that do not individually reach GWAS significance, and (2) a linear mixed model in GCTA, we could predict variance in measured circulating 25(OH)D beyond that explained by known genome-wide significant 25(OH)D-associated SNPs.

Materials and Methods

We conducted a GWAS of prospectively collected 25(OH)D in five studies with a total of 5,575 individuals (Table I). The five GWAS were: a case-control study of breast cancer (BRCA) and a case-control study of type 2 diabetes (T2D), both nested within the Nurses' Health Study (NHS) [Willett et al., 1987], a case-control study of myocardial infarction (MI) nested within the Health Professionals Follow-up Study (HPFS) [Giovannucci et al., 2008], and previously selected case-control sets from prior studies which included subjects from the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC) (1994) and the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) [Hayes et al., 2005].

NHS began in 1976 when 121,700 female US registered nurses 30–55 years of age were enrolled and are contacted every 2 years by questionnaire with response rates exceeding 90% per cycle [Colditz et al., 1986]. Between 1989 and 1990, 32,826 women between the ages of 43 and 69 provided blood samples. Of those, we included participants with GWAS data and plasma 25(OH)D from nested case-control studies of breast cancer (n = 870) [Willett et al., 1987] and T2D (n = 721) [Qi et al., 2010]. Controls were matched to cases on age and month and year of blood draw.

HPFS began in 1986 with 51,529 US male health professionals, ages 40–75 years, providing baseline demographic and medical information, updated every 2 years [Koushik et al., 2006; Rimm et al., 1992]. Blood samples were collected from 18,225 of the HPFS participants from 1993 to 1995. We included in our analysis 1,245 HPFS participants with GWAS data and predisease diagnosis plasma 25(OH)D from a case-control study of MI [Giovannucci et al., 2008].

The ATBC study was a randomized, placebo-controlled cancer prevention intervention trial of supplementation with alpha-tocopherol, beta-carotene, or both conducted in southwestern Finland from 1985 to 1993 [Group, 1994]. Participants were all 50–69 year old male smokers. Fasting blood samples used for 25(OH)D measurement were ascertained at study entry and whole blood collection (DNA source) was done between 1992 and 1993. GWAS and circulating 25(OH)D levels data were available for 1,372 men [Ahn et al., 2010].

The PLCO study was a large, randomized multicenter trial designed to evaluate the effectiveness of cancer screening and examine early markers of cancer. Approximately 155,000 male and female US residents aged 55–74 at baseline were recruited in 1993 to 2001 and all screening-arm subjects provided nonfasting blood samples at baseline [Hayes et al., 2005]. A total of 1,316 Caucasian subjects with GWAS data and serum measurements of 25(OH)D from a nested case-control study of prostate cancer were included in these analyses [Ahn et al., 2008].

25(OH)D concentrations for BRCA plasma samples were measured by radioimmunoassay (RIA) [Hollis 1997] in three batches, two in Dr. Michael Holick's laboratory at Boston University School of Medicine (coefficients of variation (CVs) 8.7–17.6%) and a third in Dr. Bruce Hollis' laboratory at the Medical University of South Carolina in Charlestown, SC (CV 8.7%) [Ahn et al., 2010; Bertone-Johnson et al., 2005]. For the T2D study, plasma levels of 25(OH)D were measured in the Nutrition Evaluation Laboratory in the Human Nutrition Research Center on Aging at Tufts University (CV 8.7%) by rapid extraction followed by equilibrium I-125 RIA procedure as specified by the manufacturer's procedural documentation and analyzed by gamma counter (Cobra II, Packard) [Ahn et al., 2010]. 25(OH)D concentrations for the HPFS plasma samples were measured by RIA in Dr. Hollis' lab [Giovannucci 2005; Hollis et al., 1993]. For ATBC

Table I. Characteristics of GWAS cohorts

Cohort	*N* (cases/controls)	Location	% female	Years of blood draw	Type of specimen	Median (interquartile range) of 25(OH)D (ng/mL)	Vitamin D assay (CV)
Nurses Health Study-Cancer Genetic Markers of Susceptibility (BRCA) [Lango Allen et al., 2010; Willett et al., 1987]	870 (440/430)	USA	100	1989–1990	Plasma	32.0 (24.0–40.9)	RIA (8.7–17.6%)
Nurses Health Study-Type 2 diabetes (T2D) [Willett et al., 1987]	772 (349/373)	USA	100	1989–1990	Plasma	21.8 (17.0–27.5)	RIA (8.7%)
Health Professionals Follow-up Study (HPFS) [Giovannucci et al., 2008]	1,245 (413/832)	USA	0	1993–1995	Plasma	23.5 (18.7–28.9)	RIA (11.5%)
Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC) [The ATBC Study Group 1994; Gallicchio et al., 2010]	1,372 (869/503)	Finland	0	1985–1988	Serum	14.9 (9.7–21.2)	CLIA (9.3–12.7%)
Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial (PLCO) [Gallicchio et al., 2010; Hayes et al., 2005]	1,316 (714/602)	Multicenter, USA	3	1993–2001	Serum	22.7 (18.0–28.0)	CLIA (8.1%)

Coefficients of variation (CVs); radioimmunoassay (RIA); competitive chemiluminescence immunoassay (CLIA).

and PLCO serum 25(OH)D levels were measured by competitive chemiluminescence immunoassay (CLIA) at Heartland Assays, Inc. by Dr. Ron Horst, with CVs in the blinded duplicate quality-control samples of 9.3% (intrabatch) and 12.7% (interbatch) for ATBC [Ahn et al., 2010; Gallicchio et al., 2010] and 8.1% for PLCO.

GWAS genotyping was performed using the Illumina 550K (or higher density) platform in the BRCA, PLCO, and ATBC studies and the Affymetrix 6.0 platform in the T2D and HPFS studies. Quality-control assessment of genotypes including sample completion rates, SNP call rates, concordance rates, deviation from Hardy-Weinberg proportions in control DNA and final sample selection for associational analysis are described in detail elsewhere [Ahn et al., 2010; Easton et al., 2007; Hunter et al., 2007; Qi et al., 2010; Thomas et al., 2009]. Genotypes for markers contained on the Illumina 550K platform and not the Affymetrix 6.0 platform were imputed using a Markov chain based Haplotyper (MACH) and HapMap phase II CEU reference panel (Rel 22), with retention of only high-quality SNPs with a MACH-r2 of 0.95 or greater in all data sets. After restricting to SNPs either directly genotyped or imputed to Illumina, we had a total of 359,928 SNPs eligible for inclusion in a polygenic score.

The percentage of variance explained by the 25(OH)D GWAS identified SNPs from the gene regions encoding *GC, CYP2R1, CYP24A1,* and *DHCR7/NADSYN1* was estimated among subjects from the BRCA, T2D, HPFS, PLCO, and ATBC studies. Residual sums of squares for the score itself (r^2 = variance of the model/total variance) were estimated by linear regression for models containing the four SNPs alone; the SNPs along with 25(OH)D-associated covariates (age, body mass index (BMI), case-control status, region of residence, season of blood draw, quartiles of vitamin D supplement intake, quintiles of vitamin D intake from food sources, and two eigenvectors from principal components); and covariates alone. These sums of squares were then used to estimate the proportion of the variance explained by the SNPs after removing variation explained by other covariates, in the form of the adjusted r^2 (adjusted r^2 = [SSresid(covar) – SSresid(covar + score)]/SSresid(covar)).

To create the polygenic scores we first conducted GWAS of 25(OH)D in each of the five studies separately. GWAS adjusted for age, case-control status, sex, BMI (kg/m^2), season of blood collection, vitamin D dietary intake, vitamin D supplement intake, and eigenvectors to control for population stratification, with the BRCA, T2D, HPFS, and PLCO cohorts additionally adjusting for US region of residence. These individual study results were then meta-analyzed using an inverse-variance weighted approach, as implemented in the software METAL [Ahn et al., 2010; Willer et al., 2010].

Using the results of the meta-analysis, SNPs were ranked by decreasing significance (i.e., ascending *P*-value) for the association with 25(OH)D. An in-house Python script that references the HapMap Rel.23a (NCBI B36) database was used to remove those SNPs in linkage disequilibrium (LD > 0.2). The number of SNPs meeting specified *P*-value cutoffs ($P \leq 1.0 \times 10^{-7}$, $\leq 1.0 \times 10^{-5}$, $\leq 1.0 \times 10^{-3}$, ≤ 0.01, ≤ 0.05, ≤ 0.1, ≤ 0.15, ≤ 0.2, ≤ 0.5) was determined. PLINK software [Purcell et al., 2007] was then used to construct nine polygenic scores based on the SNPs corresponding to each cutoff value. The polygenic score is an allelic scoring system involving specified SNPs and identified risk alleles (allele associated with increased circulating 25(OH)D), to assign a single quantitative index of genetic risk to each subject. A polygenic score assumes an additive and identical allelic effect for each variant (not dominant or recessive), and no gene-gene interactions.

Fivefold cross-validation of the cohorts was completed testing the performance of nine potential polygenic scores corresponding to specified *P*-value cutoffs. To conduct cross-validation, training sets were created by meta-analyzing four of the five cohorts using METAL software [Willer et al., 2010], leaving the remaining cohort as the testing set. The polygenic score was constructed using the method described above, and the association between the score and circulating 25(OH)D is estimated in the testing set. Of the nine potential polygenic score models, the model with the highest associated r^2 value averaged across the testing sets was taken to be the model explaining the most variance in circulating 25(OH)D.

Table II. Proportion of variation in 25(OH)D explained by GWAS-identified SNPs

Model	BRCA (R^2, %)	T2D (R^2, %)	HPFS (R^2, %)	ATBC (R^2, %)	PLCO (R^2, %)	Average (R^2, %)
SNPs[a]	5.04	3.16	6.58	1.95	3.80	4.93
Covariates[b]	25.77	15.26	13.05	26.53	16.17	18.03
SNPs + covariates[b]	30.32	18.05	18.33	28.63	20.89	22.23
SNPs adjusted R^2	6.12	3.29	6.08	3.42	6.42	5.16

[a] SNPs include rs2282679 (*GC*), rs2060793 (*CYP2R1*), rs3829251 (*DHCR7/NADSYN1*), and rs6013897 in BRCA, T2D, and HPFS, and proxy (r^2 = 1.0) rs17217119 in ATBC and PLCO (*CYP24A1*).

[b] Covariates include age, case-control status, BMI (kg/m^2), season of blood collection, vitamin D dietary intake, vitamin D supplement intake, US region of residence and eigenvectors from principle components for population stratification.

We repeated the analyses applying the MACH r^2 threshold for imputation of 0.3, resulting in a total of 506,671 SNPs eligible for inclusion in a polygenic score. From these SNPs we removed those SNPs in linkage disequilibrium (LD > 0.8) and excluded those SNPs within 1 Mb of the four known genome-wide significant SNPs.

SAS/STAT® software (SAS Institute Inc., Cary, NC) was used to conduct a linear regression analysis in each of the five testing sets for each of the nine polygenic score models, and estimate the proportion of the total variation in circulating 25(OH)D due to the score in the crude analysis. We also estimated the proportion of the variance explained by the score after removing variation explained by other covariates including age, BMI, case-control status, region of residence, season of blood draw, quartiles of vitamin D supplement intake, quintiles of vitamin D intake from food sources, and two eigenvectors from principal components.

Individual cohorts were analyzed with software for linear mixed model analysis using the GCTA software tool (0.91.0). We used genotyped SNPs meeting quality control measures (BRCA: 529,423 SNPs; T2D: 678,082 SNPs; HPFS: 697,899 SNPs; ATBC: 530,324 SNPs; PLCO: 514,636 SNPs) to create a genetic relationship matrix, which uses the restricted maximum likelihood (REML) method to fit a linear mixed model to estimate the variance in residual 25(OH)D explained by additive genetic matrix created from these SNPs. The covariates listed above are included as fixed effects and we meta-analyzed the results from the five cohorts employing an inverse-variance weighted approach.

Results

Table I summarizes the characteristics of the cohorts included in the analyses. The percentage of variance explained by previously published 25(OH)D GWAS-identified SNPs from the gene regions encoding *GC* (rs2282679), *CYP2R1* (rs2060793), *CYP24A1* (rs6013897), and *DHCR7/NADSYN1* (rs3829251) was 4.9% in a crude model and 5.2% after conditioning on other known 25(OH)D-associated covariates. The known 25(OH)D-associated covariates themselves explained approximately 18% of the variance observed in circulating 25(OH)D (Table II).

The polygenic analysis assumes additive and equal contribution of each locus to the polygenic score. Crude models containing the polygenic score alone demonstrated that polygenic scores comprised of genome-wide significant SNPs (P-value < 10^{-7} in the training set) explained the largest variance in circulating 25(OH)D across all testing sets. Averaging across all testing sets, the models comprised of one or two SNPs with P-value < 10^{-7}, explained an average of 2.4% of the variance in 25(OH)D (Table III). Previously published GWAS-identified 25(OH)D-associated SNPs were contained in each of these scores, with rs2282679 in the *GC* region contained in every score, and rs10500804 in *CYP2R1* contained in the score generated for the ATBC testing set.

The known 25(OH)D-associated covariates in the five studies, explain approximately 16.0% of the variance observed in circulating 25(OH)D. After conditioning on the known 25(OH)D-associated covariates, the average of the residual variance explained in circulating 25(OH)D across all testing sets showed that a score comprised of 1–2 SNPs explained only an additional 2.6% of the variance in circulating 25(OH)D (Table III).

When we applied an r^2 threshold for imputation of 0.3, removed those SNPs with an LD > 0.8 and those SNPs within 1 Mb of the known genome-wide significant SNPs, we lost all SNPs with a P-value $\leq 10^{-7}$. In this case, the best performing score across all testing sets was that comprised of SNPs with a P-value $\leq 10^{-2}$ that explained approximately 0.27% of the variance observed in circulating 25(OH)D, and 0.45% after conditioning on the genome-wide significant SNPs and 25(OH)D-associated covariates (Table IV).

By employing a linear mixed model to estimate the variance in residual 25(OH)D explained by an additive genetic matrix created from all the SNPs genotyped and tagged by the genome-wide platform, we observed that the proportion of the variance explained in circulating 25(OH)D ranged from 8.8% (95% CI: 0–54.4%) in ATBC to 51.8% (95% CI: 0–100%) in BRCA. Meta-analysis of the five cohorts resulted in an genetic effect estimate of the proportion of additional variance in 25(OH)D of 8.9% (95% CI: 0–21.7%) across the cohorts after conditioning on known 25(OH)D-associated covariates.

Discussion

In these analyses of circulating 25(OH)D, we observed that a polygenic score comprised of the fewest number of SNPs explained a larger proportion of variance in circulating 25(OH)D, compared with polygenic scores comprised of thousands of SNPs. The proportion of variance in circulating 25(OH)D explained by the polygenic score comprised of one or two of the genome-wide significant SNPs was comparable to that explained by all four GWAS-identified SNPs, at approximately 5%. By employing a linear mixed model utilizing all the SNPs genotyped and tagged by the genome-wide platform simultaneously, there was no statistically significant improvement in the proportion of variation in 25(OH)D explained by utilizing whole genome SNPs simultaneously.

Table III. Proportion of variation in circulating 25(OH)D explained by crude polygenic score

	HPFS testing			T2D testing set			BRCA testing set			ATBC testing set			PLCO testing set				
Cutoff	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	Avg R^2 (%)	Avg adjusted R^2 (%)[a]
$\leq 10^{-7}$	1	3.91	<0.0001	1	1.73	0.0004	1	2.75	<0.0001	2	1.07	0.0001	1	2.61	3.57×10^{-9}	2.41	2.58
$\leq 10^{-5}$	3	1.53	<.0001	3	1.04	0.006	3	1.33	0.0007	9	0.29	0.0477	7	0.50	0.0103	0.94	1.08
$\leq 10^{-3}$	243	0.69	0.004	253	0.29	0.148	248	0.06	0.4736	262	0.02	0.6085	235	0.11	0.2221	0.23	0.30
$\leq 10^{-2}$	2,127	0.49	0.013	2,137	0.35	0.11	2,159	0.06	0.4872	2,156	0.02	0.5935	2,053	0.09	0.2743	0.20	0.16
≤0.05	9,233	0.12	0.226	9,301	0.05	0.559	9,310	0.16	0.2448	9,289	0.00	0.8948	8,881	0.28	0.0563	0.12	0.07
≤0.1	16,933	0.26	0.073	16,904	0.19	0.239	16,915	0.14	0.2778	16,942	0.01	0.7515	16,518	0.07	0.3452	0.13	0.09
≤0.15	23,903	0.54	0.010	23,961	0.20	0.234	23,983	0.31	0.1013	24,090	0.00	0.7949	23,279	0.33	0.0370	0.28	0.15
≤0.2	30,396	0.63	0.005	30,381	0.25	0.181	30,414	0.35	0.0791	30,562	0.04	0.4695	29,656	0.19	0.1183	0.29	0.14
≤0.5	61,847	0.004	0.022	61,762	0.0003	0.644	61,781	0.31	0.1028	62,121	0.00	0.8989	61,228	0.05	0.3987	0.16	0.16

[a] Adjusted R^2 conditional on covariates include age, case-control status, BMI (kg/m^2), season of blood collection, vitamin D dietary intake, vitamin D supplement intake, and eigenvectors from principle components for population stratification, with the BRCA, T2D, HPFS, and PLCO cohorts additionally adjusting for US region of residence.

Table IV. Proportion of variation in circulating 25(OH)D explained by crude polygenic score following imputation and removal of previously known vitamin D-related SNPs

	HPFS testing			T2D testing set			BRCA testing set			ATBC testing set			PLCO testing set				
Cutoff	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	No. of SNPs	R^2 (%)	*P*-value	Avg R^2 (%)	Avg adjusted R^2 (%)[a]
$\leq 10^{-7}$	0	-	-	0	-	-	0	-	-	0	-	-	0	0.00	-	-	-
$\leq 10^{-5}$	2	0.01	0.753	1	0.01	0.848	3	0.11	0.327	6	0.02	0.650	6	0.02	0.812	0.03	0.08
$\leq 10^{-3}$	445	0.002	0.086	459	0.03	0.616	453	0.01	0.750	455	0.12	0.201	451	0.06	0.606	0.08	0.19
$\leq 10^{-2}$	4,256	0.04	0.481	4,264	1.00	0.007	4,327	0.25	0.139	4,131	0.01	0.708	4,196	0.01	0.385	0.27	0.45
≤0.05	20,600	0.04	0.486	20,726	0.06	0.505	20,739	0.36	0.078	20,451	0.00	0.879	20,424	0.00	0.703	0.09	0.16
≤0.1	40,702	0.15	0.166	40,700	0.001	0.386	40,745	0.14	0.278	40,543	0.01	0.677	40,505	0.04	0.845	0.08	0.19
≤0.15	60,345	0.26	0.072	60,434	0.19	0.246	60,564	0.39	0.066	60,477	0.03	0.498	60,228	0.02	0.456	0.18	0.22
≤0.2	80,050	0.41	0.023	80,058	0.28	0.153	375,160	0.38	0.069	80,118	0.07	0.332	79,843	0.02	0.575	0.23	0.28
≤0.5	195,368	0.007	0.003	195,500	0.26	0.1707	195,575	0.22	0.165	196,225	0.01	0.665	195,418	0.00	0.591	0.25	0.24

[a] Adjusted R^2 conditional on four 25(OH)D GWAS SNPs (rs2282679, rs3829251, rs2060793, rs6013897) and covariates include age, case-control status, BMI (kg/m^2), season of blood collection, vitamin D dietary intake, vitamin D supplement intake, and eigenvectors from principle components for population stratification, with the BRCA, T2D, HPFS, and PLCO cohorts additionally adjusting for US region of residence.

- Value missing.

Circulating 25(OH)D is a complex trait for which family studies have estimated heritability ranging from 43% to 80% [Hunter et al., 2001; Orton et al., 2008; Shea et al., 2009; Wjst et al., 2007]. Further elucidation of the genetic architecture of this complex trait beyond environmental determinants of 25(OH)D, has the potential for identifying those at risk of vitamin D insufficiency. It may also provide a useful proxy for lifetime vitamin D exposure that may be applied in instrumental variable analyses investigating the association of vitamin D and common, complex diseases. However, as we demonstrated, known GWAS-associated SNPs explain only a fraction of the observed variance in circulating 25(OH)D at about 5%.

Prior studies have observed polygenic architecture in a number of complex phenotypes such as height, bipolar disorder, schizophrenia, coronary heart disease, type I and type II diabetes, multiple sclerosis, and rheumatoid arthritis [Bush et al., 2010; Evans et al., 2009; Lango Allen et al., 2010; Purcell et al., 2009; Stahl et al., 2012], but many of these estimates based on GWAS data still fall short of the estimates of heritability derived from family-based studies. We did not find that a polygenic score comprised of hundreds or thousands of SNPs explained variability in circulating 25(OH)D over and above that explained by oligogenetic models comprised of SNPs previously identified in GWAS. Even when applying more liberal thresholds for imputation and removing SNPs in LD, once the genome-wide significant SNP regions were excluded, the remaining SNPs did not improve our estimates of the variance in circulating 25(OH)D. Our results are consistent with previous studies of breast, prostate, and pancreatic cancer which found that polygenic risk scores explained only a small fraction of variation in disease risk [Machiela et al., 2011; Pierce et al., 2012; Witte and Hoffmann, 2011].

The linear mixed model estimates of the percent variance in 25(OH)D explained by GWAS-tagged SNPs were highly variable due to the relatively small sample size [Zaitlen and Kraft, 2012]. The meta-analysis of these estimates from our cohorts (8.9%) was consistent with the estimate of percent variance explained by GWAS-significant SNPs alone (4.9%); the 95% confidence interval for the percent variance explained by tag-SNPs measured in GWAS excluded effects larger than 25%.

Potential limitations of this study include the fact that we have restricted our SNPs for inclusion to directly genotyped or imputed Illumina SNPs. The polygenic models assume an additive allelic effect and does not account for possible gene-gene interactions. The study was also limited by small sample size. Despite employing techniques to utilize SNPs that do not

individually reach genome-wide significance, sample sizes on the order of 10,000s of subjects may be required to obtain estimates with reasonable confidence [Daetwyler et al., 2008; Lango Allen et al., 2010; Yang et al., 2010]

This is the first study of 25(OH)D that employed a variety of methods utilizing common SNP data to refine the estimate of variability explained by genetics. Although known GWAS-associated SNPs account for the largest proportion of variance in circulating 25(OH)D, it is still only a small fraction of the genetic variance expected based on family studies. This may indicate that the majority of the genetic effect for circulating 25(OH)D is due to rare variants, structural variants other than SNPs, epistasis, or gene-environment interaction [Maher, 2008; Manolio et al., 2009]. It may also reflect biases in the estimates of heritability [Zuk et al., 2012]. Acknowledging the limitations of utilizing common SNP information, our findings still provide a useful proxy for circulating 25(OH)D, unconfounded by environmental determinants of 25(OH)D, for future studies of complex diseases.

Acknowledgments

We thank P. Soule for assistance, and we thank the participants in the Nurses' Health Studies (NHS) and Health Professionals Follow-up Study (HPFS). This work was supported by the Canadian Institute of Health Research (Health Professionals Fellowship Award to L.H.). The NHS Cancer Genetic Markers of Susceptibility breast cancer GWAS was supported by N01-CO-12400. The NHS/HPFS type 2 diabetes GWAS (U01HG004399) is a component of a collaborative project that includes 13 other GWAS funded as part of the Gene-Environment Association Studies (GENEVA) under the National Institutes of Health (NIH) Genes, Environment, and Health Initiative (GEI). Genotyping was performed at the Broad Institute of the Massachusetts Institute of Technology and Harvard, with funding support from the NIH GEI (U01HG04424). The NHS/HPFS CHD GWAS was supported by HL35464 and CA55075 from the NIH with additional support for genotyping from Merck/Rosetta Research Laboratories, North Wales, PA. The NHS is supported by NIH grants P01CA087969 and 5U01HG004399–2, and the HPFS is supported by NIH grant P01CA055075. The ATBC work was supported by the Intramural Research Program of the National Cancer Institute at the National Institutes of Health. Additionally, this research was supported by U.S. Public Health Service contracts N01-CN-45165, N01-RC-45035, N01-RC-37004, and HHSN261201000006C from the National Cancer Institute, Department of Health and Human Services. The authors declare no conflict of interest.

References

Ahn J, Peters U, Albanes D, Purdue MP, Abnet CC, Chatterjee N, Horst RL, Hollis BW, Huang WY, Shikany JM and others. 2008. Serum vitamin D concentration and prostate cancer risk: a nested case-control study. *J Natl Cancer Inst* 100(11):796–804.

Ahn J, Yu K, Stolzenberg-Solomon R, Simon KC, McCullough ML, Gallicchio L, Jacobs EJ, Ascherio A, Helzlsouer K, Jacobs KB and others. 2010. Genome-wide association study of circulating vitamin D levels. *Hum Mol Genet* 19(13):2739–2745.

The ATBC Cancer Prevention Study Group. 1994. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Ann Epidemiol* 4(1):1–10.

Bertone-Johnson ER, Chen WY, Holick MF, Hollis BW, Colditz GA, Willett WC, Hankinson SE. 2005. Plasma 25-hydroxyvitamin D and 1,25-dihydroxyvitamin D and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* 14(8):1991–1997.

Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance MA, Haines JL. 2010. Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. *Am J Hum Genet* 86(4):621–625.

Colditz GA, Martin P, Stampfer MJ, Willett WC, Sampson L, Rosner B, Hennekens CH, Speizer FE. 1986. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Am J Epidemiol* 123(5):894–900.

Daetwyler HD, Villanueva B, Woolliams JA. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3(10):e3395.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R and others. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447(7148):1087–1093.

Evans DM, Visscher PM, Wray NR. 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 18(18):3525–3531.

Gallicchio L, Helzlsouer KJ, Chow WH, Freedman DM, Hankinson SE, Hartge P, Hartmuller V, Harvey C, Hayes RB, Horst RL and others. 2010. Circulating 25-hydroxyvitamin D and the risk of rarer cancers: design and methods of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. *Am J Epidemiol* 172(1):10–20.

Giovannucci E. 2005. The epidemiology of vitamin D and cancer incidence and mortality: a review (United States). *Cancer Causes Control* 16(2):83–95.

Giovannucci E, Liu Y, Hollis BW, Rimm EB. 2008. 25-hydroxyvitamin D and risk of myocardial infarction in men: a prospective study. *Arch Intern Med* 168(11):1174–1180.

Group TAS. 1994. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol* 4(1):1–10.

Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P, Reding D, Gelmann EP, Rothman N, Pfeiffer RM and others. 2005. Methods for etiologic and early marker investigations in the PLCO trial. *Mutat Res* 592(1–2):147–154.

Holick MF. 2004. Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. *Am J Clin Nutr* 80(6 Suppl):1678S–1688S.

Hollis BW. 1997. Quantitation of 25-hydroxyvitamin D and 1,25-dihydroxyvitamin D by radioimmunoassay using radioiodinated tracers. *Methods Enzymol* 282:174–186.

Hollis BW, Kamerud JQ, Selvaag SR, Lorenz JD, Napoli JL. 1993. Determination of vitamin D status by radioimmunoassay with an 125I-labeled tracer. *Clin Chem* 39(3):529–533.

Hunter D, De Lange M, Snieder H, MacGregor AJ, Swaminathan R, Thakker RV, Spector TD. 2001. Genetic contribution to bone metabolism, calcium excretion, and vitamin D and parathyroid hormone regulation. *J Bone Miner Res* 16(2):371–378.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A and others. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 39(7):870–874.

Koushik A, Kraft P, Fuchs CS, Hankinson SE, Willett WC, Giovannucci EL, Hunter DJ. 2006. Nonsynonymous polymorphisms in genes in the one-carbon metabolism pathway and associations with colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 15(12):2408–2417.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S and others. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317):832–838.

Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. 2011. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet Epidemiol* 35(6):506–514.

Maher B. 2008. Personal genomes: the case of the missing heritability. *Nature* 456(7218):18–21.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.

Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A. 2006. Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. *JAMA* 296(23):2832–2838.

Orton SM, Morris AP, Herrera BM, Ramagopalan SV, Lincoln MR, Chao MJ, Vieth R, Sadovnick AD, Ebers GC. 2008. Evidence for genetic regulation of vitamin D status in twins with multiple sclerosis. *Am J Clin Nutr* 88(2):441–447.

Pierce BL, Tong L, Kraft P, Ahsan H. 2012. Unidentified genetic variants influence pancreatic cancer risk: an analysis of polygenic susceptibility in the PanScan Study. *Genet Epidemiol* 36(5):517–524.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256):748–752.

Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Pare G and others. 2010. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum Mol Genet* 19(13):2706–2715.

Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. 1992. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am J Epidemiol* 135(10):1114–1126; discussion 1127–36.

Shea MK, Benjamin EJ, Dupuis J, Massaro JM, Jacques PF, D'Agostino RB, Sr., Ordovas JM, O'Donnell CJ, Dawson-Hughes B, Vasan RS and others. 2009. Genetic and non-genetic correlates of vitamins K and D. *Eur J Clin Nutr* 63(4):458–464.

Signorello LB, Shi J, Cai Q, Zheng W, Williams SM, Long J, Cohen SS, Li G, Hollis BW, Smith JR and others. 2011. Common variation in vitamin D pathway genes predicts circulating 25-hydroxyvitamin D Levels among African Americans. *PLoS One* 6(12):e28623.

Sinotte M, Diorio C, Berube S, Pollak M, Brisson J. 2009. Genetic polymorphisms of the vitamin D binding protein and plasma concentrations of 25-hydroxyvitamin D in premenopausal women. *Am J Clin Nutr* 89(2):634–640.

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FA and others. 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* 44(5):483–489.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K and others. 2009. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* 41(5):579–584.

Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, Berry D, Kiel DP, Streeten EA, Ohlsson C, Koller DL and others. 2010. Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* 376(9736):180–188.

Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17):2190–2191.

Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE. 1987. Moderate alcohol consumption and the risk of breast cancer. *N Engl J Med* 316(19):1174–1180.

Witte JS, Hoffmann TJ. 2011. Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer. *OMICS* 15(6):393–398.

Wjst M, Altmuller J, Braig C, Bahnweg M, Andre E. 2007. A genome-wide linkage scan for 25-OH-D(3) and 1,25-(OH)2-D3 serum levels in asthma families. *J Steroid Biochem Mol Biol* 103(3–5):799–802.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW and others. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.

Zaitlen N, Kraft P. 2012. Heritability in the genome-wide association era. *Hum Genet* 131(10):1655–1664.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 109(4):1193–1198.