Towards Precision Medicine management of vitamin D inadequacy: Development of ancestry-specific polygenic scores and investigation of genetic-environmental interactions affecting vitamin D concentrations

By

Kathryn Elizabeth Hatchell

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Population Health Sciences - Epidemiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: January 30, 2019

The dissertation is approved by the following members of the final oral committee:

Corinne D. Engelman, Associate Professor, Population Health Sciences
Scott J. Hebbring, Research Scientist, Marshfield Clinic – Center for Human Genetics
Qiongshi Lu, Assistant Professor, Biostatistics and Medical Informatics
Julie A. Mares, Professor, Nutritional Sciences
Paul E. Peppard, Associate Professor, Population Health Sciences

Abstract

Vitamin D inadequacy affects around 50% of adults in the United States and is associated with numerous adverse health outcomes. Vitamin D blood concentration [25(OH)D] has strong environmental and genetic predictors that may determine how much vitamin D intake is required to reach optimal 25(OH)D. Despite large genome-wide association studies (GWASs), only a small portion of the genetic factors of 25(OH)D has been discovered. The goal of this research is to uncover a fuller set of genetic factors and gene-by-environment interactions, that could be useful for prediction of vitamin D inadequacy, personalized vitamin D supplementation, and prevention of vitamin D associated morbidity and mortality. Using subsets of participants of European (n=9,569) and African ancestry (n=2,761), ancestry-specific polygenic scores (PGSs) were created using PRSice and validated with analyses performed in SAS. Overall SNP heritability and that accounted for by the PGS and published GWAS findings were calculated in GCTA and compared. Finally, interactions between the PGS and environmental predictors of 25(OH)D, available UV radiation and vitamin D intake, were investigated. Findings show that participants with high genetic risk had 25(OH)D that was 1.9-4.7 ng/ml lower than those with lowest genetic risk ($p=0.15$ to $3.2x10^{-13}$); requiring an additional 317 to 783 IU of vitamin D intake to maintain equivalent 25(OH)D. In European-ancestry participants who reached IOM vitamin D intake guidelines, the proportion of participants achieving adequate 25(OH)D increased as genetic risk decreased (70.4 vs 83.8 in the highest and lowest risk categories, respectively; $p=4.1x10^{-11}$); providing further evidence that those with high genetic risk require more vitamin D intake to reach adequate 25(OH)D. Where sample size allowed, heritability estimation showed that the PGS explains more heritability than do prior GWAS findings (3.7% vs 1.5%). Additionally, available UV radiation and vitamin D intake were shown to interact with PGS and influence 25(OH)D. Of note, due to limited minority group data, African-ancestry analyses were generally underpowered. PGSs are a powerful predictive tool that, in tandem with assessment of environmental predictors, UV radiation and vitamin D intake, could be leveraged for personalized vitamin D supplementation to prevent the negative downstream effects of vitamin D inadequacy.

Acknowledgements

First and foremost, thank you to my Committee Chair and primary mentor, Dr. Corinne Engelman, who provided me my formal introduction to the field of genetic epidemiology and led me to Precision Medicine. Your hands-on approach fostered my growth and development in genetic epidemiology and beyond. Thank you for kindly pushing me to go out of my comfort zone and for modeling how to maintain work-life-balance as a woman in science.

Thank you to all my committee members for their insightful comments and generous mentoring. To Dr. Qiongshi Lu who helped me troubleshoot numerous statistical issues along the way; your time and expertise is appreciated immensely, thank you. Thank you to Dr. Scott Hebbring for your knowledge of human genetics, particularly on working with GWAS data, and a special thank you for making the trip from Marshfield for all my dissertation committee meetings. Thank you to Dr. Julie Mares for your expertise in nutrition and nutritional epidemiologic methods and for helping me to maximize the impact of my data. Thank you to Dr. Paul Peppard, not only for your guidance on the methods used in my dissertation, but for your engaging lecturing and teaching style.

To my lab mates - Burcu, Danny, Eva and Ruocheng – I have learned so much from you over these past five years through presentations, asking questions and sharing code—thank you. I will miss you all and am excited to see what the future brings for each of you— I have no doubt you will all do great things.

To the Social Science Computing Cooperative (SSCC) – I, quite literally, could not have done this dissertation without the storage and computing power of the SSCC, so thank you.

Additionally, to the entire SSCC Helpdesk staff, especially Dan Bongert, thank you for answering my never-ceasing barrage of coding-related questions.

To my first funding source, the *County Health Rankings*, thank you for fostering my growth as an epidemiologist, for priceless hours of coding practice and for solidifying my understanding of the multidimensional determinants of health. A special thank you to Matt Rodock, my first (and best) SAS instructor, thank you for being so patient and building me a foundation to grow my coding skills on.

Thank you to the Computational Informatics in Biology and Medicine (CIBM) traineeship for funding the last two years of my dissertation and for providing me ample opportunities to network and build upon my computational skills.

Thank you to the Population Health Sciences department for providing me a solid statistical and epidemiological base to develop my dissertation questions from. Special shout-outs to Quinn Fullenkamp who always managed to make me laugh—even through my most stressful of semesters and Dr. Ron Gagnon for always having an open door for statistical questions. To my cohort—Anna, Natalie, Christine and Yousra—thank you for sharing your brilliance and motivating me.

To my family—my first teachers and biggest supporters—thank you for supporting me through the past 24 years of education. I am forever indebted to you for fostering my learning and teaching me the value of hard work. Mom and Dad, I dedicate my dissertation to you for your years of encouragement and support. Michael and Thomas, years of competition with you has made me a resilient and motivated person, so thank you ☺ Finally, to my Aunt Susi who, whether she remembers or not, for leading me to this field.

To my friends, near and far, thank you for your generous hearts and years of friendship— especially: KARMELS (or more accurately “ARMELS”), Paige Andrews (who went through this right with me) and Melanie Haak (who has been there since my first day of pre-school). Thanks for keeping me sane by listening to my rants and entertaining me with Snapchats as my code ran. To my boyfriend, Jason, who is always there to listen to my frustrations and delight in my successes. Thanks for always pointing out the bigger picture, to this incredibly detail oriented human ☺

Finally, I want to thank all the research participants who provided data to ARIC, MESA and WHI—your priceless contribution to science allows for medical advances and breakthroughs.

Table of Contents

List of tables and figures

Appendix Contents

Chapter 1: Overview

1. Introduction and literature review

1.1 Vitamin D

1.1.1 Scope of the health burden

Approximately half of the United States adult population is affected by vitamin D inadequacy, defined by 25-hydroxyvitamin D concentration [25(OH)D] ≤20ng/ml. Those of racial and/or ethnic groups with darker skin have an even higher prevalence with estimates for African American adults being over 80% and estimates for Hispanic adults being over 60% (1-3). Vitamin D inadequacy is a major public health burden because of the relationship it has with negative downstream health effects. Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, migraines, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-9). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Furthermore, some clinical trials have shown that vitamin D and calcium supplementation are important in the prevention of fractures and cardiovascular risk factors (17-20).

1.1.2 Environmental and genetic predictors

Vitamin D concentration is complex trait with numerous environmental and genetic predictors. The active form of vitamin D, 25(OH)D, is influenced by sun exposure and vitamin D intake through diet and supplements. In total, environmental exposures are estimated to account for 2-23% of the variance observed in vitamin D concentrations (21). Environment in tandem with personal characteristics such as age and measures of adiposity (BMI, waist circumference, etc.) have been observed to account for up to 32% of variance in vitamin D concentrations (21).

The genetic contribution to vitamin D concentration is estimated to explain 20-40% of the variance seen in serum concentrations (1, 22). However, most studies to date have been in related participants of European ancestry, therefore further study is warranted. Additionally, some evidence indicates that season may affect the genetic contribution to vitamin D concentrations by altering the amount of vitamin D produced upon exposure to sunlight (23, 24). Studies have also suggested a difference in vitamin D heritability by gender (25, 26). Several studies have investigated the SNP level genetic impact on vitamin D concentrations and a handful of genome-wide association studies (GWAS) have been carried out (1, 25, 27, 28). To date, only about 1-8% of variation in vitamin D concentrations can be explained through genetic effects and little exploration has gone into teasing out differing genetic contributions by race, gender and/or season (25, 29).

1.1.3 Genome-wide association studies

To date, four large-scale GWAS studies have been performed; three in participants of European ancestry and one multi-ethnic meta-analysis of participants from European, African or Hispanic ancestry (1, 25, 27, 28). Loci in group specific component (vitamin D binding protein) gene (*GC*) and NAD synthetase 1 gene (*NADSYN1*) / 7-dehydrocholesterol reductase gene (*DHCR7*) were found to have associations with vitamin D concentration in all three studies (1, 27, 28). Additionally, loci in vitamin D 25-hydroxylase gene (*CYP2R1*) and vitamin D 24-hydroxylase gene (*CYP24A1*) have significant associations with vitamin D concentrations in those of European ancestry (1, 28). Loci near kinesin family member 4B gene (*KIF4B*) were found to be strongly associated with vitamin D concentrations only in those of African ancestry (27). *GC* transports the vitamin D metabolites in the blood. *DHCR7* catalyzes the conversion of 7-dehydrocholosterol in the skin to previtamin D_3, a precursor to vitamin D_3. *CYP2R1* codes for a cytochrome P450 enzyme that hydroxylates vitamin D_2/D_3 to 25(OH)D. *CYP24A1* codes for another cytochrome P450 enzyme that degrades 25(OH)D to an inactive metabolite, 24,25-dihydroxyvitamin D. *KIF4B* is an expression quantitative trait locus (eQTL) for another nearby gene, *FAXDC2* which codes for the fatty acid hydroxylase domain-containing protein 2, which is involved in cholesterol and steroid biosynthesis; metabolites upstream of vitamin D activation (27). While GWAS studies contribute important biologic understanding into vitamin D concentrations, to date, only about 1-8% of variability can be explained with findings from GWAS studies (25, 29).

1.1.4 Knowledge gap

1.1.4.1 *Missing heritability*

Missing heritability is pervasive problem in human genetics. GWAS studies are a standard analytic approach, however GWAS studies miss much of the heritability of complex traits because they do not capture heritability due to rare variants or SNPs with weak effects (i.e. SNPs that do not meet the stringent GWAS threshold due to interactions (either genetic-by-genetic, or genetic-by-environmental interactions)). Given that vitamin D concentration is a complex trait with numerous genetic and environmental predictors, much of the heritability is likely attributable to SNPs with weak effects or that function through interaction. Therefore, methods beyond GWAS are necessary to discern predictors of vitamin D concentration allowing for personalized prevention of vitamin D inadequacy.

1.1.4.2 *Missing minority group studies*

In addition to the missing heritability problem, the field of human genetics, and prominently the field of vitamin D, vastly understudies minority populations as shown in Figure 1.1 (30). Additionally, oftentimes when minority groups are studied, there is a lack of insight and significant findings due to small sample size and poorly developed methods for mixed ancestral populations (30). Given this, increased research and proper methods in minority groups are crucial for better understanding of underlying health conditions and disparities they may produce. This is particularly important in the field of vitamin D where prevalence of vitamin D inadequacy is much higher in minority groups. This is because darker skin has natural sunblock and therefore absorbs less of the sun's rays which downstream lead to less of the active form of vitamin D, as depicted in Figure 1.2.

Figure 1.1 Persistent racial bias in genomics

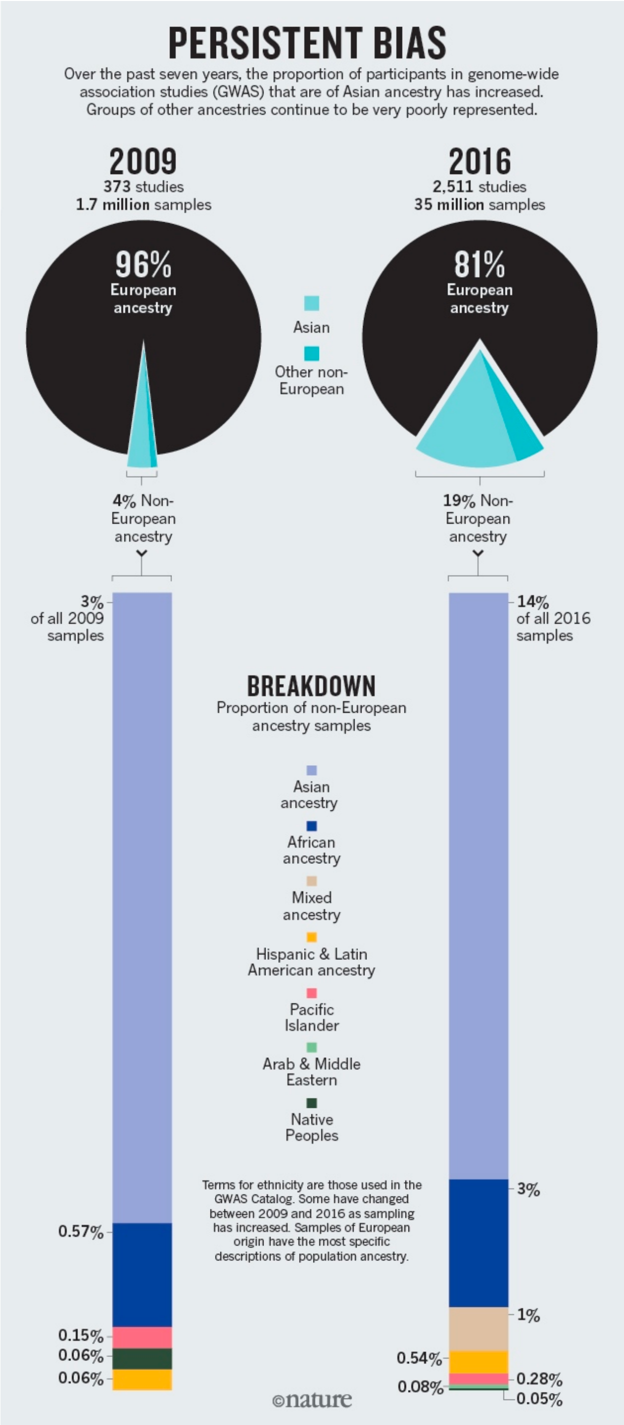


Figure 1.1 shows the evasive racial bias in genomics. In 2009, 96% of genomics studies were in samples of European ancestry. In 2016, this had dropped to 81%, a sign of improvement, but populations of non-European ancestry are still vastly understudied, potentially accentuating disparities in health outcomes between races.

Figure 1.2 Vitamin D metabolic pathway

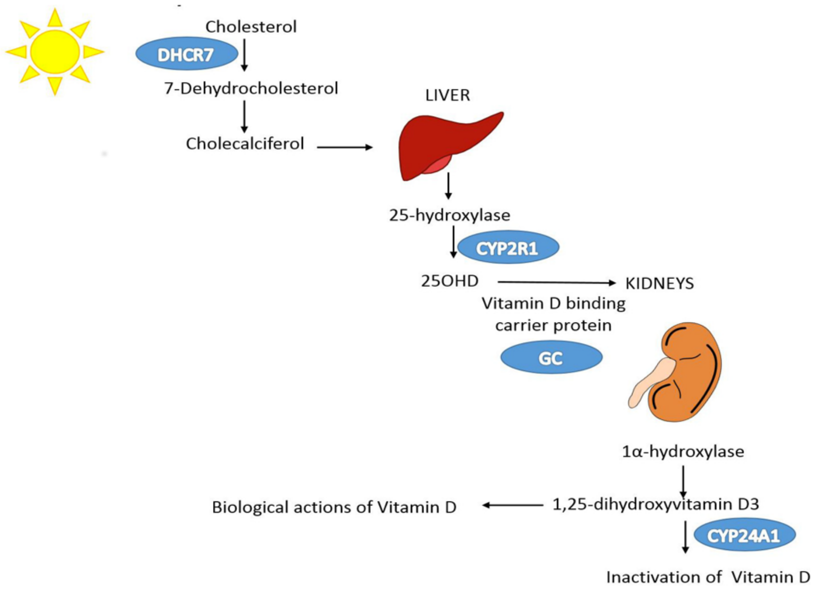


Figure 1.2 shows the pathway of vitamin D activation which starts with absorption of the sun's rays. The gene *DHCR7* is then responsible for production a vitamin D precursor which is transported to the liver where *CYP2R1* converts it to 25(OH)D. Then *GC* carries 25(OH)D to the liver where the active form of vitamin D is transported and used for biological functions. Finally, *CYP24A1* is responsible for inactivation of vitamin D.

1.1.4.3 Precision Medicine approach

Given the vast differences seen in prevalence of vitamin D inadequacy by race, the genetic and environmental contribution to 25(OH)D levels and the public health implications that vitamin D inadequacy has, applying a precision medicine approach to vitamin D supplementation is a natural fit. Precision medicine is defined by the NIH to be "an emerging approach for disease treatment and prevention that considers individual variability in genes, environment, and lifestyle for each person." To incorporate this approach, genetic and environmental factors need to be considered in tandem. To date there is a void of vitamin D research that considers interactions between a realm of genetic factors (beyond what the stringent GWAS threshold elucidates) and environmental exposures. Polygenic scores (PGSs) are one method that can be used to capture genetic predictors missed by GWAS studies. Subsequently, the interactions between PGS and environmental predictors of 25(OH)D concentrations can be investigated to move toward precision medicine management of vitamin D inadequacy.

1.1.6.1 Background, Justification and Feasibility

Introduction. To date, only a handful of vitamin D genes have been discovered through published vitamin D concentration GWASs; these genes include: *A2BP1*, *ANO6/ARID2, CYP2R1*, *CPY24A1, DAB1*, *DHCR7*, *GC*, *GPR114, HTR2A* and *KIF4B* (1, 28, 31). The collection of SNPs uncovered in these studies does not account for the full heritability of vitamin D concentrations. This is in large part because GWASs use a very stringent p-value that prohibits the discovery of SNPs that have small effect size or function through interaction. The former can be addressed by relaxing the very stringent genome-wide significance threshold (32). This is a valid approach when the goal is phenotype prediction, as shown by the 9-fold increase in phenotypic variance of human height captured using a less stringent cut-off (33-35).

Development of a PGS allows for capturing the additive effect of multiple SNPs to better predict a phenotype. The use of PGSs can lead to early detection of disease and future disease risk, which promotes the development of preventive and personalized action to combat undesirable health conditions. PGSs have been shown to predict Alzheimer's disease before the onset of symptoms that would result in a clinical diagnosis (32, 36). PGS have also been correlated with risk of coagulation deficiencies, such as activated partial thromboplastin time (aPTT), where an increased risk score increases blood clotting time as shown in Figure 1.3 (37).

Prediction of blood clotting time is essential knowledge for personalized dosing of antifibrinolytic drugs which promote blood clotting. The PGS approach holds promise for early, accurate prediction of risk of future disease onset, including vitamin D inadequacy, which can be used to proactively prevent or treat a health condition.

Figure 1.3 Relationship between PGS and blood clotting time

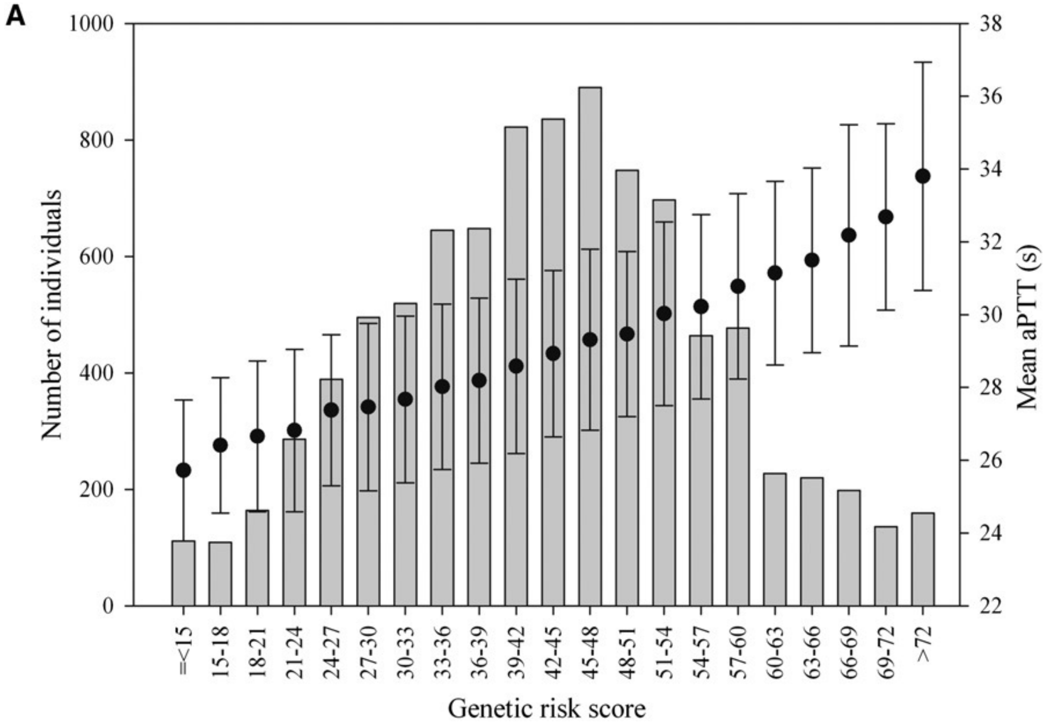


Figure 1.3 shows that as the genetic risk score (PGS) increases, so does mean aPTT (blood clotting time), as is shown by the black circles and standard error bars. This histogram shows that the number of subjects with each PGS is normally distributed (37).

Combining the above approaches, through leveraging the results from prior GWASs and creating a PGS (using a less stringent p-value cutoff) in an independent subset of the full sample (n=1,057 of N=13,684), my first aim will be to obtain the *objective* of expanding on ancestry specific findings to uncover a more complete set of SNPs that account for a larger proportion of the heritability of vitamin D and can be used to predict vitamin D inadequacy. Specifically, to attain the objective, I will test the ***working hypothesis*** that the set of SNPs that best distinguishes between adequate and inadequate vitamin D levels will include more SNPs (from more genes) than previous GWASs. The *rationale* for this is that to date, the top SNPs in or near genes discovered by GWAS (*GC*, *DHCR7* and *CYP2R1*) are estimated to account for only 1-8% of the variation in vitamin D concentrations, therefore, alternative approaches, such as relaxing the p-value cut-off to create a PGS, are necessary to account for the remaining 92-99% (1, 23, 25). When this research is completed, it is my *expectation* that the results will improve prediction of vitamin D inadequacy by race, which is crucial given the difference in risk of vitamin D inadequacy by race (38, 39).

1.1.7 Heritability

1.1.7.1 Definition and calculation

Heritability, technically defined, is the proportion of phenotypic variance that is explained by genetic variance, where phenotypic variance includes genetic and environmental variance (40). Classically, heritability is calculated in related subjects where the proportion of shared genotypes is known (41). However, this approach tends to overestimate heritability due to shared environment of the related subjects. To avoid this overestimate, GCTA will be used to give an estimate of SNP heritability of vitamin D concentrations in unrelated participants.

1.1.7.2 GCTA and SNP Heritability

Introduction. GWASs allow for the discovery of SNPs that are associated with various phenotypes. However, while GWASs may identify numerous SNPs, the proportion of heritability explained by those SNPs is small. One potential explanation for this is that the threshold used in GWASs is often too stringent to detect all relevant SNPs, especially those that have a weaker effect, such as those that have a low correlation (linkage disequilibrium) with the underlying, but not genotyped, functional SNP and those that function through an interaction (42). The software tool, GCTA, can determine the amount of phenotypic variance that a set of SNPs accounts for, and from this GCTA can calculate the SNP heritability (or total genotypic variance) of a phenotype (43). In aim 2, I will investigate SNP heritability of vitamin D concentrations. The *objective* of this aim is 3-fold. First, the SNP heritability of vitamin D concentrations will be determined by inputting all SNPs into the GCTA model; second, the heritability accounted for by SNPs previously identified through GWAS will be discerned; and third, the contribution that the SNPs included in the PGS in Aim 1 add to the heritability of vitamin D concentrations will be determined. To attain the objective of this aim I will test the *working hypothesis* that SNPs included in the PGS in Aim 1 will increase the proportion of heritability accounted for by SNPs currently reported in the literature. My *approach* to testing the working hypothesis will be done using GCTA. GCTA has been used for a variety of phenotypes with high levels of success. When used to study height, which is estimated to be 60-70% heritable, GCTA accounted for 56% of the phenotypic variance (35, 44). This was a vast improvement over the 5% of the heritability that was repeatedly and reliably shown from GWAS studies (35, 44). A similar level of success was attained when BMI was studied. Estimates from the literature suggest that BMI is 30-40%

heritable and GCTA was able to account for 27% of the phenotypic variance (44). Like height and BMI, vitamin D concentrations are a complex trait, affected by many genes and environmental factors, notably sun exposure and vitamin D intake. Vitamin D concentrations have been shown to be heritable, with heritability estimates ranging from 28% to 80% in Caucasian and African American populations; with most estimates in the 20-40% range (22, 31). Given the wide range in these estimates, a more accurate estimation of the heritability of vitamin D concentrations is needed and can be achieved using GCTA. The *rationale* for this aim is that successful completion of the proposed research will result in a SNP heritability estimate for vitamin D concentrations, as well as a quantification of how much heritability is currently accounted for. When the proposed research in Aim 2 is completed, it is my *expectation* that the results will guide where further studies should focus their efforts, i.e. on capturing more of the heritability or if sufficient heritability is explained, working to create personalized vitamin D dosing based on the genes discovered.

1.1.8 Gene-by-environment interactions

Vitamin D inadequacy is a complex phenotype affected by many genetic and environmental determinants. While attention has been paid to the genetic determinants through GWA and candidate gene studies and, separately, to the environmental determinants of vitamin D concentration, much less attention has been paid to how environmental factors interact with genetic factors. Vitamin D concentration is influenced by sun exposure and dietary intake. One study reports that vitamin D intake through diet and supplement use accounts for 1-8% of the variation in vitamin D concentrations between individuals, and that sun exposure accounts for 1-15% of the variation, however acknowledges that these are likely underestimated due to measurement error (21, 45-47). The same study reported an interaction between two *GC* SNPs and vitamin D intake and sun exposure, where the genetic effect was stronger in those with higher intake of vitamin D and with more available UV radiation (i.e. in summer) (21). Furthermore, another study created a PGS for vitamin D inadequacy using top SNPs in *GC* and *CYP2R1* (48). The PGS was found to be inversely correlated with serum vitamin D concentrations, where having fewer risk alleles increased serum vitamin D concentrations. Therefore, in Aim 3 the ***objective*** is to investigate how genetic risk of vitamin D inadequacy (measured via the PGS created in Aim 1) is altered by quartile of vitamin D intake or available UV radiation. To attain the objective of this aim, I will test the ***working hypothesis*** that the genetic effect on vitamin D inadequacy will vary by level of vitamin D intake and available UV radiation; where there will be an increased genetic effect on vitamin D concentrations with increased dietary intake of vitamin D and in summer (i.e. with higher available UV radiation). Additionally, those with a higher PGS (more risk alleles) will require higher levels of vitamin D intake and/or available UV radiation to achieve adequate vitamin D concentrations. Given that vitamin D inadequacy is a complex disease which has many contributing SNPs and environmental factors, I will test the working hypothesis using a linear modeling ***approach*** that stratifies the PGS effect by environmental factors (vitamin D intake and available UV radiation), followed up by adding interaction terms to the model (where evidence of an interaction exists). In regards to vitamin D concentrations, it is crucial to investigate gene-environment interactions as the risk inferred by genetic factors alone is not enough to predict one’s risk of inadequate vitamin D concentrations. Creating a PGS and stratifying by level of environmental factors is one way to quantify the effect genetics, in tandem with the environment, has on vitamin D inadequacy. Preliminary data show effect modification of the association between a PGS, based on a small set of SNPs, and 25(OH)D blood concentrations by both season of blood draw and quartile of dietary vitamin D intake. In this sample of 200 subjects of European descent from the Survey of the Health of Wisconsin (SHOW), the PGS ranged from 0 to 13.3. When stratified by season, the effect of each one unit increase in the PGS on decreased 25(OH)D blood concentrations is larger in summer (β=-1.1 ng/ml; p<.0001) compared to winter (β=-0.44 ng/ml; p=.29; PGS*season interaction p=0.25). When stratified by quartile of dietary vitamin D intake, the effect of each one unit increase in the PGS on decreased 25(OH)D blood concentrations is larger as dietary vitamin D intake increases: β=-0.38 (p=.35), -0.65 (p=.18), -0.82 (p=.07) and -1.27 (p=.006) ng/ml for the 1st, 2nd, 3rd and 4th quartiles of intake, respectively (PGS* dietary intake interaction p=0.03) as shown by Figure 1.4. These findings replicate and expand the findings of a previous study in an independent cohort and have important implications for both study design and precision medicine (21). It is my ***expectation*** that upon the completion of this aim, patients can be grouped by level of susceptibility to vitamin D inadequacy, which could help inform screening and treatment of vitamin D inadequacy based on genetic and environmental factors.

Figure 1.4: Effect modification of PGS due to environmental sources of vitamin D

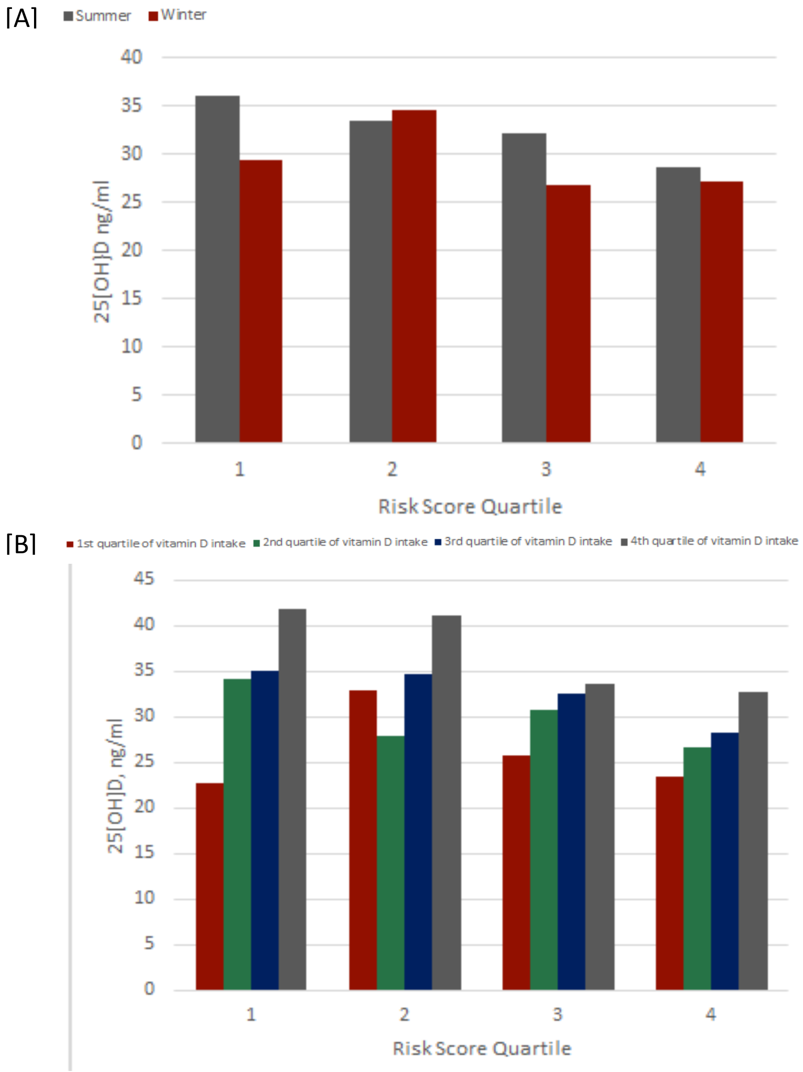


Figure 1.4 demonstrates that when environmental sources of vitamin D are increased (UV index in panel [A] and dietary vitamin D intake in panel [B]). Increasing risk score has a consistent effect on decreasing levels of vitamin D concentrations. This effect is much more muted at lower levels of environmental sources of vitamin D.

2. Significance

Personalized vitamin D supplementation is crucial because the current one-size fits all treatment regimen is not effective for all patients as shown in Figure 1.5 (49). In one clinical trial, when given the same dose, some patients experienced an increase in serum vitamin D concentrations, while others experienced a decline. For example, the individuals noted in the *rectangles* in Figure 1.5 (a/b and c/d) had similar baseline 25(OH)D concentrations, but markedly differing 25(OH)D response to supplementation. Consequently, many individuals (~26%; *n*=24 noted in the *light gray* shaded region) did not achieve a potential “target” 25(OH)D of 30 ng/mL, while 9 (~10%) achieved values above 50 ng/mL (noted in the *dark gray* shaded region) and 2 reached concentrations of ~60 ng/mL (indicated by cross); all 3 of these regions are associated with negative health impacts. Therefore, personalized treatment, has the promise of making healthcare more cost effective, by treating vitamin D inadequacy efficiently and preventing downstream morbidity and mortality.

Figure 1.5: Variability in response to vitamin D supplementation

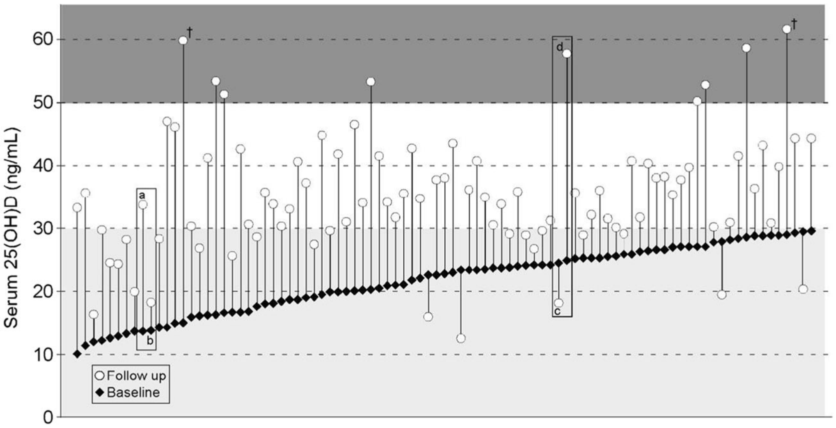


Figure 1.5 shows variability in response to vitamin D supplementation among 91 subjects who received 2,300-2,500 IU of oral vitamin D_3 for 4 months. All subjects were ≥90% compliant with vitamin D supplementation (49).

Vitamin D inadequacy is a pervasive health problem. To date, through GWA and candidate gene studies, only a handful of genes (*A2BP1*, *ANO6/ARID2, CYP2R1*, *CPY24A1, DAB1*, *DHCR7*, *GC*, *GPR114, HTR2A* and *KIF4B*) have been found to be associated with vitamin D concentrations. For personalized vitamin D dosing to be most effective, more genetic determinants that account for the moderate heritability of vitamin D need to be determined (50, 51). Additionally, how the environment works in tandem with genetics to alter vitamin D concentrations is poorly understood. A study in a sample of 1,204 women of European descent established an interaction between genetics and environment. Specifically, the study found two *GC* SNPs and four *CYP2R1* SNPs that had differing affects by season and level of vitamin D intake (21). I will expand on the findings of this paper through creating a PGS that accounts for a more expansive set of SNPs and testing for interactions with the environment. Understanding gene-environment interactions that influence vitamin D concentrations can lead to novel, accurate and effective personalized treatment approaches (21, 22). The contribution of the proposed research will be a more complete set of vitamin D SNPs which account for more of the heritability of vitamin D inadequacy as well as discerning gene-environmental interactions that affect vitamin D concentrations. *This is a significant contribution as it will allow for effective, efficient personalized dosing for vitamin D inadequacy.* Personalized treatment of vitamin D inadequacy could decrease morbidity in many realms, such as autoimmune disease, heart disease and cancer (1, 4, 5, 10-15).

3. Innovation

Discovery of vitamin D SNPs to date has been done through GWA and candidate gene studies. While these methods have led to discoveries in or near genes such as *CYP2R1*, *CPY24A1, DHCR7* and *GC* (as shown in Figure 1.2) they have not been able to identify the full gamut of vitamin D related SNPs, especially those of low frequency or small effect size (1, 28, 31, 42). In general, little exploration has been done in the realm of gene-environment interactions through which vitamin D SNPs act. *The proposed research is innovative because of the methodology it uses as well as the multi-ethnic nature of the sample.* My methodology takes advantage of the large GWAS that have been performed in European and African ancestry populations (27). With results from previously conducted GWASs, I will create a PGS that accounts for a more complete set of vitamin D SNPs, by using a less stringent p-value cutoff than what is used in previous GWASs (32). Using GCTA, I will calculate the SNP heritability of vitamin D and the proportion of the heritability explained by previously reported GWAS significant SNPs and by the PGS. Additionally, I will investigate modification of the PGS effect on vitamin D concentrations by environmental factors, expanding on the preliminary work done in a small subset of European women (21). Of note, the sizable African American population (N=3,896) will allow for a multi-ethnic approach to be utilized. Vitamin D inadequacy is pervasive among all races, but even more so in those with darker skin, which acts as natural sun block (1, 2). In totality, this has the potential to shift and personalize current vitamin D supplementation practices. Specifically, establishing a more complete set of vitamin D SNPs and learning how these SNPs interact with the environment to influence vitamin D concentrations in an ancestry specific way will enable clinicians to tailor vitamin D supplementation to the individual, making treatment more efficient and effective.

Chapter 2: Approach and Methods

1. Approach

1.1 Sample

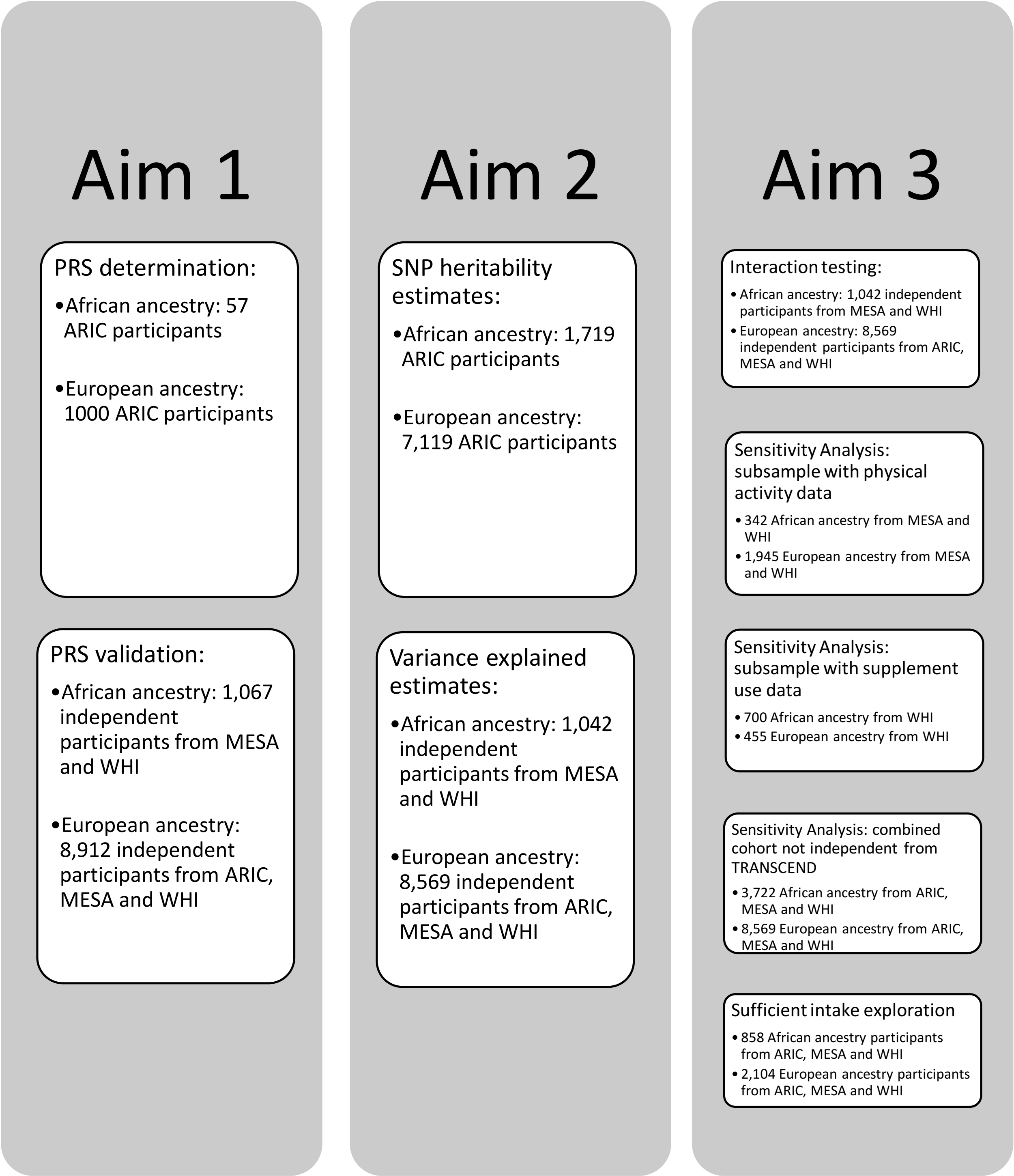
Following the success that was found through using a multi-cohort approach for other complex traits (i.e. height, diabetes, schizophrenia, PTSD), participants come from three national cohort studies: Atherosclerosis Risk in Communities (ARIC), the Multi-ethnic Study of Atherosclerosis (MESA) and the Women's Health Initiative (WHI) (52). ARIC is a prospective study of men and women ages 46-70 years. Participants are recruited in Forsyth, NC; Jackson, MS; Minneapolis, MN and Washington County, MD. Serum vitamin D was measured for particular ancillary studies of ARIC at visit 2 (1990-1992). MESA is a prospective study of men and women ages 44-84 who were recruited by Columbia University, Johns Hopkins University, Northwestern University, University of Minnesota, University of California at Los Angeles and Wake Forest University. Serum vitamin D was measured at MESA exam 1 (July 2000-August 2002). The WHI is a study which consists of various clinical trials as well as observational studies of women. Women participating in WHI were recruited from 40 clinical centers throughout the United States. Serum vitamin D was measured as part of the Calcium and Vitamin D (CaD) Trial. Participants were included if they had the minimum set of variables: genome wide data, serum vitamin D, age, sex, BMI and season of blood draw (these variables account for the minimum set (model 1); N=13,684). Participants with data on vitamin D intake will also be leveraged for analyses using the full set (model 2); N=13,015. In total, the sample size is 13,684 (9,905 European ancestry and 3,779 African ancestry). European ancestry participants come from ARIC (n=7,455), MESA (n=1,995) and WHI (n=455). African ancestry participants come from ARIC (n=1,903), MESA (n=1,176) and WHI (n=700). The data used in these analyses were collected under guidelines from the relevant institutional review boards and all participants provided informed consent, including consent for use of genetic data. Table 2.1 shows cohort level characteristics for the participants for both the minimum set (model 1) and full set (model 2) models. Data cleaning for phenotypic data included winsorizing 25(OH)D in the MESA sample to the 99th percentile, 63.7 ng/mL in those of European ancestry and 49.0 ng/mL in those of African ancestry. Additionally, 2 participants in MESA had 25(OH)D values equal to 0 and were given half of the minimum value detected per field standards; this equated to a value of 1.9 ng/mL in the European ancestry participant and 1.1 ng/mL in the African ancestry participant. In the WHI sample, participants with 25(OH)D values above the level of detection (150 ng/mL) were removed from the sample; this included 68 participants of European ancestry and 119 participants of African ancestry. 25(OH)D values were also winsorized in the WHI sample to approximate the lower and upper fences (58.6 ng/mL and 67 ng/mL for European and African ancestries, respectively) (53). Figure 2.1 shows samples used for separate analyses by Aim.

Data was requested and received for Coronary Artery Risk Development in Adults (CARDIA) and Framingham Heart Study (FHS). However, these data were not used in the analysis. Data from CARDIA was not used because 25(OH)D data was not in the dbGaP dataset, as it was collected in an ancillary study. An independent data request was attempted, but since the ancillary study has ended, no data requests are being granted. FHS data was not used because due to the small sampling region and corresponding re-identification issues associated with FHS, participant location cannot be disclosed. This barred development of available UV radiation, a variable integral to the analysis.

Table 2.1: Full sample characteristics

Cohort	ARIC model 1	ARIC model 2	MESA model 1	MESA model 2	WHI model 1	WHI model 2
N	9,358	8,838	3,171	3,022	1,155	1,155
% Female	56.0%	55.5%	53.2%	53.0%	100.0%	100.0%
% White	79.7%	80.6%	62.9%	64.1%	39.4%	39.4%
% Black	20.3%	19.5%	37.1%	35.9%	60.6%	60.6%
Mean Age (SD) [years]	56.9 (5.7)	57.0 (5.7)	62.4 (10.2)	62.5 (10.2)	63.7 (7.6)	63.7 (7.6)
Mean BMI (SD) [kg/m^2]	27.8 (5.3)	27.8 (5.3)	28.7 (5.5)	28.6 (5.4)	30.7 (6.4)	30.7 (6.4)
Mean 25(OH)D (SD) [ng/ml]	24.5 (8.9)	24.6 (8.9)	25.9 (11.6)	26.1 (11.5)	19.0 (13.7)	19.0 (13.7)
Mean available UV radiation (SD) [units]	5.4 (2.6)	5.4 (2.5)	4.7 (2.3)	4.7 (2.3)	5.4 (2.5)	5.4 (2.5)
Mean dietary intake (SD) [IU]	221.4 (145.5)	221.4 (145.5)	170.5 (163.5)	170.5 (163.5)	164.5 (137.4)	164.5 (137.4)

Figure 2.1: Sample used by aim



Aim 1

PRS determination:

- African ancestry: 57 ARIC participants
- European ancestry: 1000 ARIC participants

PRS validation:

- African ancestry: 1,067 independent participants from MESA and WHI
- European ancestry: 8,912 independent participants from ARIC, MESA and WHI

Aim 2

SNP heritability estimates:

- African ancestry: 1,719 ARIC participants
- European ancestry: 7,119 ARIC participants

Variance explained estimates:

- African ancestry: 1,042 independent participants from MESA and WHI
- European ancestry: 8,569 independent participants from ARIC, MESA and WHI

Aim 3

Interaction testing:

- African ancestry: 1,042 independent participants from MESA and WHI
- European ancestry: 8,569 independent participants from ARIC, MESA and WHI

Sensitivity Analysis: subsample with physical activity data

- 342 African ancestry from MESA and WHI
- 1,945 European ancestry from MESA and WHI

Sensitivity Analysis: subsample with supplement use data

- 700 African ancestry from WHI
- 455 European ancestry from WHI

Sensitivity Analysis: combined cohort not independent from TRANSCEND

- 3,722 African ancestry from ARIC, MESA and WHI
- 8,569 European ancestry from ARIC, MESA and WHI

Sufficient intake exploration

- 858 African ancestry participants from ARIC, MESA and WHI
- 2,104 European ancestry participants from ARIC, MESA and WHI

Figure 2.1 shows the sample used for each analysis by Aim. In Aim 1, ARIC was used for PGS determination because WHI and MESA had generalizability issues due to either being all women or having sparse genotyping. PGS was only validated in samples independent of the summary statistics used to calculated the PGS. In Aim 2, only ARIC had large enough sample sizes to calculate heritability estimates and variance explained estimates were calculated using a combined cohort of all samples independent of the PGS determination and summary statistics. Aim 3 analysis was a combined cohort analysis of all samples independent of the PGS determination and summary statistics, plus various sensitivity analyses.

1.2 Data

1.2.1 Dietary Data

Dietary data were collected via questionnaire. Each study used their own questionnaire. WHI used the Food Frequency Questionnaire supplemented with interview questions, ARIC and MESA both used their own implementation of a food intake questionnaire. From the questionnaire data, each study created a derived variable of typical vitamin D intake (measured in IU or mcg). All values were converted to IU for analysis. Additionally, WHI collected data on vitamin D supplement use. The sum of vitamin D intake from food and supplements was calculated and used for supplemental and sensitivity analyses, otherwise dietary intake alone was used.

1.2.2 Serum vitamin D data

Serum vitamin D concentration was measured by the studies using different assay types. WHI used the DiaSorin LIASON chemiluminescence, MESA used LCMS and ARIC used Tandem Mass Spec (Quest Labs). To control for differences in vitamin D concentrations due to different assays, vitamin D concentrations were converted to z-scores within studies for combined analyses. For cohort specific analyses, 25(OH)D values were natural log transformed as vitamin D values as vitamin D concentrations were non-normally distributed.

1.2.3 Genomic data

Genotyping for ARIC was performed with the AffymetrixGenome-WideHuman SNPArray 6.0 chip. Genotyping for WHI differed by sub-study (Supplemental Table 1), but used

AffymetrixGenome-WideHuman SNPArray 6.0 chip, Illumina HumanOmni1-Quad v1-0 B or the Illumina MEGA Consortium 15063755 B2 array. Genotyping for MESA was done using a 50K Human Gene Focused Panel. All cohorts followed the same quality control pipeline which included: removing participants with sex-mismatch, removing samples with a low call rate (<95%), removing SNPs with a low call rate (<95%), removing SNPs not in Hardy-Weinberg Equilibrium (using a Bonferroni adjusted p-value), removing SNPs with a low minor allele frequency (<0.002; chosen to yield a minor allele count >7 in both ancestries) and removing related participants (first degree relatives, with identity-by-descent >0.36) (54, 55). Principal component analysis (PCA) was performed using PLINK to remove ethnic outliers; this included removal of one participant from ARIC, seven participants from MESA and seven participants from WHI. PCA plots of samples by cohort can be found in Supplemental Figures 1-9. Genetic data was imputed using the Michigan Imputation Server and phased with Eagle v2.3 (56-58). European samples were imputed to the Haplotype Reference Consortium and African samples were imputed to the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA); SNPs with imputation quality less than <0.8 or minor allele frequency <0.001 were removed (57). Specific cut-offs used and final participant and SNP sample sizes are summarized in Supplemental Table 1 and Supplemental Figures 10-11.

1.2.4 UV Index data

Based on the month of blood draw and location, participants were assigned continuous available UV radiation values. Available UV radiation values assigned were an average of daily UV-index for the month prior to blood draw (the relevant exposure period). UV radiation data come from the National Weather Service Climate Prediction Center historical database. When available, UV radiation values corresponded to the exact location and year of the participants blood draw. When exact cities or years were not available, averages across locations and/or years were used. See Supplemental Tables 2-4 for specific month, year and location values used. Descriptive statistics for UV radiation by site and month are also presented in Supplemental Tables 5-7.

1.3 Aim 1

1.3.1 Analytical methods and expected outcomes

Results from 16,124 European-ancestry and 8,541 African-ancestry individuals included in the TRANSCEN-D ancestry specific meta-analyses were leveraged for determination of a fuller set of SNPs and computation of the PGS in an independent sample; this was done in an ancestry-specific manner (27). The ancestry-specific PGS was calculated using PRSice Version 2 (59). First a subset of samples was used to determine the p-value cutoff (and corresponding set of SNPs) that explained the most variability in serum vitamin D concentrations. The European ancestry sub-sample for PGS development included 1,000 randomly selected participants from the European ARIC cohort (N=7,455). Samples were randomly selected using SAS. All participants were assigned a random number using the RANUNI function with seed 1 in SAS; the RANUNI function returns a number that is generated from the uniform distribution on the interval (0,1) using a prime modulus multiplicative generator with modulus 2^{31}- and multiplier 397204094 (60). Participants were then sorted in ascending order by the random number; the first 1000 samples were selected. The African ancestry sub-sample for PGS development included 57 participants from the African ARIC cohort (N=1,960) that were independent from the TRANSCEN-D ARIC sample used for the summary statistics in PGS calculation.

PRSice computes a weighted risk score of independent SNPs; the weight was determined using results from the TRANSCEN-D meta-analysis by converting the z-scores from METAL into effect sizes using the deterministic relationship: β = z / $\sqrt{(2p(1-p)*N}$, where p is the allele frequency for the SNP, and N is the total sample size from TRANSCEN-D. The PGS was calculated for each participant as follows: PGS$_i$=$\sum_1^n \beta * C_i$, where *i* represents the participant whose PGS is calculated by summing over *n* SNPs. β is the risk allele effect size from the TRANSCEN-D meta-analysis for each SNP and *C* is the individual's count of risk alleles for that SNP. If participants were missing any SNPs, they were imputed to have the mean number of alleles for the SNP. SNPs included in the PGS were independent (r^2 thresholds< 0.5 and <0.2 were tested) and had a p-value at or below a threshold. The threshold was determined by iterative testing of the PGS over a range of p-values ($5x10^{-4}$ to 0.5, with interval increments of $5x10^{-4}$) to find the PGS that best associated with log[25(OH)D], while controlling for age, sex, BMI and available UV radiation (32). Sensitivity analyses including intake in the model were performed.

Once the optimal PGS was determined, a PGS was computed for the remaining 8,905 independent participants of European ancestry and 1,067 independent participants of African ancestry.

1.3 Aim 2

1.3.1 Analytic methods and expected outcomes

First, Genome-Wide Complex Trait Analysis (GCTA) was used to calculate the SNP heritability of vitamin D concentration. This was done by calculating the variance in the phenotype jointly explained by all the SNPs; GCTA fits all inputted SNPs simultaneously as random effects in a linear mixed model (61). Next, the proportion of phenotypic variance that SNPs previously reported in the literature account for was calculated. Finally, heritability accounted for by the SNPs included in the PGS (Aim 1) was calculated.

Specifically, GCTA utilizes genomic-relatedness-based restricted maximum-likelihood (GREML), similar to restricted maximum likelihood analysis (REML) and relies on the genetic relationship matrix (GRM) of SNPs from unrelated samples (Figure 2.2) (35, 62). Heritability is calculated as the proportion of phenotypic variance that is due to additive genetic effects (51). In GCTA, the GRM is calculated specific to the set of SNPs used as input. The method involves three steps: (1) calculate a linkage-disequilibrium (LD) score between SNPs, (2) stratify SNPs based on their LD, (3) compute the GRM with the stratified SNPs (43, 62). After the GRM is generated, phenotypic variance explained by the SNPs is calculated using REML, with the SNP effects treated as random effects (63). This two-stage approach is done to reduce the computational burden and to get more precise heritability estimates. Models were adjusted for age, gender, BMI and available UV radiation. Models stratified by gender were used to account for the difference in variances of certain covariates (i.e. BMI) by gender; which are not fully adjusted for in the REML model (61). Models stratified by available UV radiation were also used. Heritability estimates were computed: (1) globally with all SNPs as input, (2) with only

SNPs previously reported in the literature and (3) with SNPs included in the novel PGS discerned in Aim 1. Heritability estimates can be inflated with ancestrally mixed populations as contributions of rare variants which are not tagged in the general population play a role (64). Therefore, all methods were performed by ancestry - European and African - to avoid inflation of the estimates (64). Expected outcomes were that SNPs currently reported in the literature account for a small fraction of the heritability of vitamin D concentrations and that more of the heritability would be captured through the ancestry-specific PGS calculated in Aim 1.

Figure 2.2: Example histogram of values in a GRM

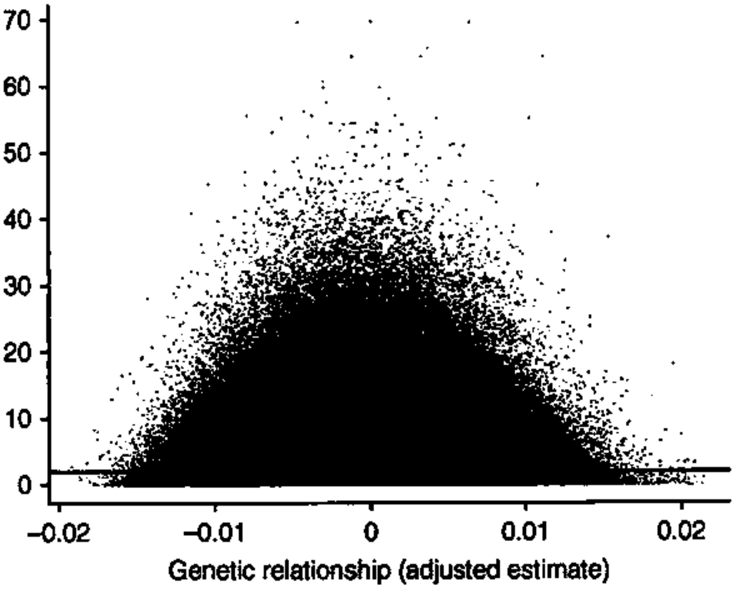


Figure 2.2 shows mean relatedness is given a value of zero. Subjects who are less related to each other than the average relatedness in the sample have a negative value, and conversely, subjects who are more related to each other have a positive value. Extreme values are typically around -0.02 and 0.02 (35).

Closely related individuals pose a problem for GCTA, as GCTA calculates heritability for unrelated individuals. Prior to all analyses, related individuals (IBD > 0.38) from the non-family studies (ARIC, MESA and WHI) were excluded. Given that GCTA only uses conventionally unrelated subjects, the heritability calculated may differ from that calculated in a pedigree analysis as it is strictly genetic and does not included shared environmental effects and it might not capture all causal variants that a pedigree analysis may capture (i.e. if they were not genotyped) (61). To reduce bias that could result from using GCTA to calculate heritability, the GRM was generated using all SNPs to capture the largest number of causal variants (61).

1.4 Aim 3

1.4.1 Analytic methods and expected outcomes

Ancestry specific stratified analysis investigated the relationship between PGS (computed in Aim 1) and environmental predictors of 25(OH)D in combined samples of participants from ARIC, MESA and WHI using SAS (version 9.4). Models were executed in samples stratified by quartile of available UV radiation or vitamin D dietary intake. All models controlled for age, sex, BMI and cohort. In models exploring the effect of the PGS by quartile of available UV radiation, models also controlled for vitamin D dietary intake; and vice versa. Sensitivity analyses included models that also controlled for physical activity (met-hours/week) or vitamin D supplement use; sensitivity analyses were carried out in a larger, combined cohort as well as smaller cohorts with required data. From the stratified analysis, results indicative of interaction informed further models. One-Degree of Freedom (DF) and 2-DF models were investigated; 1-DF models test only the relevant interaction term and 2-DF models test both the relevant interaction term and the main effect terms. Relevant interaction terms were the PGS and interaction with either available UV radiation or vitamin D intake. All 1-DF and 2-DF models controlled for age, sex, BMI, cohort and vitamin D intake/available UV radiation. Further analyses were performed in those who achieved Institute of Medicine (IOM) vitamin D intake guidelines (600 IU/day for those 1-70 years old and 800 IU/day for those over 70) to investigate differences in proportion of those reaching adequate 25(OH)D concentrations (≥ 20 ng/ml) by quartile of genetic risk.

In accordance with previous findings, it was expected that the effect of the PGS on vitamin D concentrations would vary by level of vitamin D intake and season; specifically, it was expected that the PGS would have a larger effect on vitamin D concentrations with increased dietary intake of vitamin D and in summer (i.e. with higher available UV radiation) (21). We also expected that those with a higher PGS will require higher doses of vitamin D (through diet, supplements or available UV radiation) to achieve adequate concentrations.

2. Specific aims

Vitamin D inadequacy, as defined by a 25-hydroxyvitamin D [25(OH)D] concentration less than 20 ng/mL, affects more than 50% of adults in the United States. Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, migraines, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-9). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Furthermore, some clinical trials have shown that vitamin D and calcium supplementation are important in the prevention of fractures and cardiovascular risk factors (17-20). Clinical trials of vitamin D alone have found that increasing vitamin D intake may lower risk of cancers, diabetes and depression and may reduce inflammation and improve lung function in patients with cystic fibrosis (7, 20, 65-69). Recent results reported from the VITAL trial showed null associations between vitamin D supplementation and cancer or cardiovascular disease, however, study design limits the interpretability of these findings (70). Avoiding vitamin D inadequacy is important, however, vitamin D concentrations over 50 ng/mL have been associated with increased morbidity and mortality (3, 71). Clinical trials of vitamin D have shown that individual response to vitamin D supplementation is highly variable (49, 72). Vitamin D concentrations are influenced by genetic factors and genetic variants may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration (21, 22, 73, 74). Therefore, knowledge of the genetic determinants of vitamin D concentrations could be invaluable in prevention of vitamin D associated morbidity and mortality.

Several studies have uncovered vitamin D associated single nucleotide polymorphisms (SNPs) (1, 28, 31). However, these SNPs account for a small portion of the variation in vitamin D concentrations (1). Understanding the complete set of genetic factors that contribute to vitamin D concentrations and how they function together and with environmental factors will allow for personalized treatment of vitamin D inadequacy, instead of the current (and ineffective) one size fits all treatment regimen. To more fully understand the biology of vitamin D concentrations, I propose uncovering a more complete set of SNPs as well as interactions with the two vitamin D sources: sun exposure and dietary intake.

My ***long-term goal*** is to promote adequate vitamin D concentrations through personalized vitamin D supplementation based on an individual's genetic makeup and non-genetic characteristics. The ***overall objective*** of this project will be to elucidate a fuller set of

SNPs influencing vitamin D concentrations through creation of a polygenic score (PGS), and to elucidate PGS*environment interactions that alter vitamin D concentrations. Cohort specific and combined analyses will be carried out in samples from Atherosclerosis Risk In Communities (ARIC), Multi-Ethnic Study of Atherosclerosis (MESA) and the Women's Health Initiative (WHI). My *central hypothesis* is that since vitamin D concentration is highly heritable, and only a small number of genes have been implicated, through inclusion of a fuller set of SNPs in a PGS, more of the variation in vitamin D concentration can be explained. Additionally, building off of previous work using a subset of WHI participants, I hypothesize that the genetic contribution to vitamin D inadequacy will be altered by environmental factors such as sun exposure and dietary intake of vitamin D. The *rationale* for my proposed research is that through discovery of the genetic and environmental influencers of vitamin D concentrations, researchers can more accurately predict and treat vitamin D inadequacy and prevent the downstream negative health effects. I will test my central hypothesis by executing the following *aims*:

Aim 1: Use results from existing ancestry specific meta-analyses of vitamin D genome-wide association studies (GWASs) to determine the optimal p-value threshold for the set of independent SNPs that explains the most variance in vitamin D concentrations in an independent sample and calculate the polygenic score (PGS) for this set of SNPs.

H1: The set of SNPs that best distinguishes between adequate and inadequate vitamin D levels will include more SNPs (from more genes) than previous GWASs.

Aim 2: Calculate the heritability of vitamin D concentration using Genome-Wide Complex Trait Analysis (GCTA) software under three scenarios: 1) globally (SNP heritability), 2) using the set of SNPs reported in the literature and 3) using the set of SNPs included in the PGS in Aim 1.

H2: The proportion of heritability accounted for by the set of SNPs from Aim 1 will be higher than that for the SNPs currently reported in the literature.

Aim 3: Examine modification of the relationship between the PGS and vitamin D concentration in analyses stratified by vitamin D intake and, separately, season of blood draw. If qualitative modification of the effect is observed, statistical interaction between the PGS and vitamin D intake and/or season of blood draw will be tested

H3a: The PGS effect on vitamin D concentration will vary by level of vitamin D intake and season; there will be a stronger PGS effect on vitamin D concentration with increased dietary intake of vitamin D and in summer.

H3b: Those with a higher PGS will require more vitamin D from diet and sun exposure to achieve optimal vitamin D concentrations.

This information could inform future vitamin D supplement dosing, tailoring it to a person’s genome instead of using the current one size fits all treatment regimen.

Chapter 3: Aim One

DEVELOPMENT OF ANCESTRY SPECIFIC POLYGENIC SCORES FOR VITAMIN D CONCENTRATION

Abstract

Background: Vitamin D inadequacy is a pervasive public health issue. Vitamin D concentrations are influenced by genetic factors which may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration. A polygenic score (PGS) can be used to summarize genetic determinants of vitamin D concentration and, potentially, to tailor vitamin D intake to prevent vitamin D inadequacy.

Methods: Using PRSice v2 and summary statistics from a multi-ethnic GWAS of 16,124 participants of European ancestry and 8,541 participants of African ancestry, optimally performing PGSs and model R^2 (tested using a linear model controlling for age, sex, body mass index (BMI) and available ultra-violet (UV) radiation) were calculated in an independent set of 1,000 participants of European ancestry and 57 participants of African ancestry. PGSs were then applied to an additional 8,905 participants of European ancestry and 1,067 participants of African ancestry.

Results: In the European ancestry determination and validation cohorts, the optimal PGS explained 1.3-2.1% of the variance in 25-hydroxyvitamin D [25(OH)D] concentrations, and the fully adjusted model explained 8-14% of the variance in 25(OH)D. In the African ancestry determination and validation cohorts, the PGS explained 0.01-4.4% of the variance in 25(OH)D, and the fully adjusted model explained 2.3-37% of the variance. Results showed that those with greater genetic risk have statistically significant and clinically meaningfully lower 25(OH)D concentrations.

Discussion: The variance explained by the PGS, while in line with what has been reported for other complex traits, captures only a modest portion of phenotypic variance. The association between the PGS and 25(OH)D concentrations indicates this PGS has the potential to be leveraged for personalized vitamin D supplementation.

Background

Vitamin D inadequacy, as defined by a 25-hydroxyvitamin D [25(OH)D] concentration less than 20 ng/mL, affects around 50% of adults in the United States (1-3). Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, migraines, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-9). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Furthermore, some clinical trials have shown that vitamin D and calcium supplementation are important in the prevention of fractures and cardiovascular risk factors (17-20). Clinical trials of vitamin D alone have found that increasing vitamin D intake may lower risk of cancers, diabetes and depression and may reduce inflammation and improve lung function in patients with cystic fibrosis (7, 65-69). Recent results reported from the VITAL trial showed null associations between vitamin D supplementation and cancer or cardiovascular disease, however, study design limits the interpretability of these findings (70). Avoiding vitamin D inadequacy is important, however, vitamin D concentrations over 50 ng/mL have been associated with increased morbidity and mortality (3, 71). Clinical trials of vitamin D have shown that individual response to vitamin D supplementation is highly variable (49, 72). Vitamin D concentrations are influenced by genetic factors and genetic variants may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration (21, 22, 73, 74). Therefore, knowledge of the genetic determinants of vitamin D concentrations could be invaluable in prevention of vitamin D associated morbidity and mortality.

Several studies have uncovered vitamin D associated single nucleotide polymorphisms (SNPs) (1, 25, 28, 31). To date, a handful of vitamin D genes have been discovered through published genome-wide association studies (GWASs) in those of European or African ancestry; these genes include: *A2BP1*, *ANO6/ARID2*, *CYP2R1*, *CPY24A1*, *DAB1*, *DHCR7*, *GC*, *GPR114*, *HTR2A*, *KIF4B*, *AMDHD1* and *SEC23A*. However, these SNPs account for a small portion of the variation in vitamin D concentrations; about 2.8% of the estimated 20-40% heritability (1, 22, 25). This is, in large part, because GWASs use a very stringent p-value that prohibits the discovery of SNPs that have a small effect size or function through interaction with other genetic or environmental factors. The former can be addressed by creating a polygenic score (PGS), which includes SNPs that did not meet the very stringent genome-wide significance threshold (32). Development of a PGS allows for capturing the additive effect of multiple SNPs across the genome, without requiring genome-wide significance, which in turn allows for better prediction of a phenotype.

A PGS includes a more complete set of genetic factors that contribute to a phenotype which can be used as a predictor, allowing for proactive prevention of a health condition for those at increased risk. For example, PGSs have been shown to predict Alzheimer’s disease before the onset of symptoms that would result in a clinical diagnosis (32, 36), at a time in the disease course where treatment may be more effective. PGS have also been correlated with risk of coagulation deficiencies, such as activated partial thromboplastin time (aPTT), where an increased risk score increases blood clotting time (37). The PGS can then be utilized for personalized dosing of antifibrinolytic drugs which promote blood clotting. The PGS approach holds promise for early, accurate prediction of risk of vitamin D inadequacy, which could then inform vitamin D supplementation. To date there is a research void in quantifying vitamin D inadequacy risk through the use of a PGS. Only a handful of studies have calculated genetic risk scores for vitamin D inadequacy using SNPs that reached the stringent GWAS p-value threshold, therefore missing much of the genetic contribution to the phenotype (22, 48, 75-77). Given that several studies have reported genetic dependent response to vitamin D supplementation, a comprehensive PGS holds predictive and preventive promise in relation to vitamin D inadequacy and downstream health outcomes (73, 78, 79).

Here, PGSs will be calculated using summary statistics from a multi-ethnic GWAS performed by the TRANSCEN-D, or the TRANS-ethniC Evaluation of vitamiN D, GWAS consortium (27). Summary statistics come from 16,124 participants of European ancestry and 8,541 participants of African ancestry. PGSs will be calculated, using PRSice v2, in an ancestry specific manner, allowing for better risk prediction of complex traits due to ancestry-specific heterogeneity in the underlying trait.

Methods

Participants

PGS development was done in an ancestry specific manner, using subsets of European- and African-ancestry samples from the Atherosclerosis in Communities Study (ARIC) as the target dataset. ARIC data was obtained through dbGaP Study Accession: phs000090.v4.p1. ARIC is a prospective epidemiologic study conducted in four U.S. communities: Wake Forest Baptist Medical Center, Winston-Salem, NC; University of Mississippi Medical Center, Jackson, MS; University of Minnesota, Minneapolis, MN; Johns Hopkins University, Baltimore, MD. ARIC includes a total of 15,792 participants of which 9,415 have all data required for this analysis (genomic data, 25(OH)D, age, sex, body mass index (BMI), geographic location and month of blood draw). Of these, 7,455 are of European ancestry; these participants were not included in the TRANSCEN-D GWAS (which is the base dataset for this analysis). A random sample of 1,000 participants were chosen from the 7,455 eligible participants for calculation of the optimal PGS p-value threshold; the remaining samples were used in validating the variance in 25(OH)D explained by the PGS and interaction testing with the PGS and environmental predictors of 25(OH)D (Aim 3). Samples were randomly selected by assigning a random variable using a random number generator in SAS and selecting the 1,000 largest values. For the African cohort, 1,960 participants of African ancestry were eligible. Samples were selected to be used for determining the optimal p-value cut-off based on independence from the samples used in the TRANSCEN-D GWAS meta-analysis. Accordingly, 57 participants of African ancestry from the 1,960 were used. See Table 3.1 for sample characteristics. Participant consent was previously obtained at each respective ARIC study site; IRB approval was granted for ARIC and this specific analysis.

Table 3.1 Optimal p-value determination sample characteristics

Cohort	ARIC European-ancestry set	ARIC African-ancestry set
N	1,000	57
% Female	53.2%	49.1%
Mean Age (SD) [years]	57.1(5.7)	55.6 (6.2)
Mean BMI (SD) [kg/m^2]	27.3 (4.8)	28.6 (5.7)
Mean 25(OH)D (SD) [ng/ml]	25.7 (8.7)	20.9 (7.8)
Available UV radiation (SD) [units]	5.0 (2.5)	7.1 (2.4)
Intake [IU]	219.2 (135.2)	221.2 (137.3)

After the optimal p-value threshold was determined and the PGS was calculated for each ancestry, the PGS was then applied to additional participants to validate performance of the risk score and to further investigate performance of the PGS (i.e. what percent of heritability is captured by the PGS [Aim 2] and how the PGS interacts with environmental factors to affect vitamin D concentrations [Aim 3]). For those of European ancestry, the PGS was applied to: 6,455 participants from ARIC, 1,995 participants from the Multi-ethnic Study of Atherosclerosis (MESA) and 455 participants from the Women's Health initiative (WHI). For those of African ancestry, the PGS was applied to: 367 participants from MESA and 700 participants from WHI. Participant characteristics for the full samples from each cohort can be found in Table 3.2. MESA is a prospective study of men and women ages 44-84 who were recruited from Columbia University, Johns Hopkins University, Northwestern University, University of Minnesota, University of California at Los Angeles and Wake Forest University. Serum vitamin D was measured at MESA exam 1 (July 2000-August 2002). MESA data was obtained through dbGaP Study Accession: phs000209.v13.p3. Women participating in WHI were recruited from 40 clinical centers in the United States. Serum vitamin D was measured as part of the Calcium and Vitamin D (CaD) Trial (80). WHI data was obtained through dbGaP Study Accession: phs000200.v11.p3. Participant consent was previously obtained at each respective MESA and WHI study site; IRB approval was granted for MESA, WHI and this specific analysis.

Table 3.2 Full sample characteristics

Cohort	Variable	European-ancestry	African-ancestry
ARIC	Sample size	7,455	57
	Age (SE) [years]	57.1 (5.7)	55.6 (6.2)
	BMI (SE) [kg/m^2]	27.3 (4.9)	28.6 (5.7)
	UV[1] (SE) [units]	5.0 (2.5)	7.1 (2.4)
	Intake[2] (SE) [IU]	222.8 (144.4)	219.1 (138.8)
	25(OH)D (SE) [ng/ml]	25.9 (8.8)	20.9 (7.8)
MESA	Sample size	1,995	367
	Age (SE) [years]	62.7 (10.3)	62.0 (10.4)
	BMI (SE) [kg/m^2]	27.8 (5.0)	30.3 (6.2)
	UV (SE) [units]	4.5 (2.3)	5.1 (2.2)
	Intake (SE) [IU]	189 (157)	161.8 (144.1)
	25(OH)D (SE) [ng/ml]	30.1 (10.9)	19.3 (8.8)
WHI	Sample size	455	700
	Age (SE) [years]	66.6 (6.8)	61.8 (7.4)
	BMI (SE) [kg/m^2]	29.9 (6.3)	31.2 (6.4)
	UV (SE) [units]	5.2 (2.5)	5.5 (2.6)
	Intake (SE) [IU]	420.9 (299.4)	308.8 (257.4)
	25(OH)D (SE) [ng/ml]	18.9 (10.7)	19.0 (15.4)

[1] available UV radiation
[2] vitamin D intake from diet
[3] MANOVA global test (performed in SAS (version 9.4) revealed differences in one or more variables by cohort, therefore cohort was adjusted for in all models that included multiple cohorts

Data Quality Control

Data cleaning for phenotypic data included winsorizing 25(OH)D in the MESA sample to the 99th percentile, 63.7 ng/mL in those of European ancestry and 49.0 ng/mL in those of African ancestry. Additionally, 2 participants in MESA had 25(OH)D values equal to 0 and were assigned half of the minimum value detected per field standards; this equated to a value of 1.9 ng/mL in the European-ancestry participant and 1.1 ng/mL in the African-ancestry participant. In the WHI sample, participants with 25(OH)D values far above the maximum level of detection (150 ng/mL), none of which had extreme vitamin D intake or sun exposure, were removed from the sample; this included 68 participants of European ancestry and 119 participants of African ancestry. 25(OH)D values were also winsorized in the WHI sample to approximate the lower and upper fences (58.6 ng/mL and 67 ng/mL for European and African ancestries, respectively) (53). All 25(OH)D values were log transformed to improve the normality of the distribution in each cohort.

Genotyping methods are described in publications by ARIC, MESA and WHI (81-85). Supplemental Table 1 gives information on the original genotyping platforms and SNP panels used by the studies. Quality control (QC) was done in an ancestry-specific manner for those of European and African ancestry. Ancestry was determined by self-report and confirmed with principal components analysis with 1000 genomes populations serving as anchoring populations. In summary, QC for each cohort removed: sex mismatches, samples and SNPs with high missingness (>5%), SNPs with low minor allele frequency (MAF<0.2%), and SNPs out of Hardy-Weinberg equilibrium (p<0.05/number of SNPs; Bonferroni adjusted cut-off). Datasets were then imputed using the Michigan Imputation Server (56, 58). European samples were imputed to the Haplotype Reference Consortium (HRC) and African samples were imputed to the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (58, 86). Post imputation QC included: removing SNPs with a low-quality score (<0.8) or MAF (<0.1%). Additionally, sample and SNP level missingness as well as HWE cutoffs were rechecked. Supplemental Figures 1-11, and Supplemental Table 1 give specifics on quality control for each cohort. QC was performed using PLINK v1.9 and vcfTools (87, 88).

Measurement of 25(OH)D

Serum vitamin D concentration was measured by the studies using different assay types. WHI used the DiaSorin LIASON chemiluminescence, MESA used liquid chromatography-mass spectrometry (LCMS) and ARIC used tandem mass spectrometry (MS/MS; Quest Labs). To control for differences in vitamin D concentrations due to different assays, vitamin D concentrations were converted to z-scores within studies for combined analyses.

Calculation of PGS

Optimal p-value cutoff and corresponding PGS were determined by calculating PGSs using numerous less stringent (compared to the GWAS threshold) p-value cutoffs and testing the association between the PGS and log[25(OH)D]. P-values were attained from TRANSCEN-D, an independent GWAS (27). TRANSCEN-D, or the TRANS-ethniC Evaluation of vitamiN D GWAS consortium, is a collaboration of 13 cohorts (9 of African ancestry, 3 of Hispanic ancestry and SUNLIGHT, a consortium of 15 European cohorts), which performed a multi-ethnic GWAS to uncover novel SNPs associated with vitamin D concentrations (1, 27). TRANSCEN-D included SNPS with MAF > 0.01 and tested them for association with log[25(OH)D] using an additive genetic model adjusting for age, sex, BMI, UV index and principal components 1-10. Ancestry-specific z-scores from TRANSCEN-D were converted to betas with the deterministic relationship: $\beta = z/(sqrt(2p(1-p)*N)$, where p is the reference allele frequency for the SNP (89). Another tuning parameter in PGS development is the linkage disequilibrium (LD) cutoff for clumping. SNPs need to be clumped to prevent SNPs in one correlated region from dominating the PGS. Here PGSs were calculated using two different LD cut-offs, $r^2 \geq 0.5$ or ≥ 0.2, keeping the

SNP with the strongest effect in the base dataset. SNPs in LD with one another were clumped, using the --clump-r2 option in PRSice v2. The LD cutoff that yielded the PGS that explains the most variance in 25(OH)D was used in downstream analyses. Given the small African ancestry sample, a reference panel (full ARIC African ancestry dataset with n= 1,900) was used to determine LD. PGSs were calculated in PRSice v2 which computes the sum of reference allele counts at each SNP weighted by the effect size (β) for that SNP from the TRANSCEN-D consortium (59).

Statistical analysis

To determine which set of SNPs to include in the PGS, SNPs at or below a given p-value threshold after clumping were included in the PGS, which was then tested in a regression model for association with log[25(OH)D]. P-value thresholds from $5x10^{-5}$ to 0.5 were tested, incrementing by $5x10^{-5}$ at each iteration. As the p-value threshold being tested increased, more SNPs were included in the PGS. The threshold where the PGS explained the most variance in the phenotype [log(25(OH)D] was selected as the most optimal PGS. The coefficient of determination, R^2, was the metric used to measure the proportion of phenotypic variance explained. Linear regression models were used to calculate the R^2 of a given PGS while controlling for participant age, sex, BMI and available UV radiation. Models which included multiple cohorts also adjusted for cohort. Available UV radiation for each participant was calculated based on their month of blood draw and location, using data from the National Weather Service Climate Prediction Center historical database. Participants were assigned UV radiation values of the average UV-index for the month prior to blood draw (the relevant exposure period); the UV radiation values ranged from 0.7 to 9.5. When available, UV radiation values corresponded to the exact location and year of the participant's blood draw. When exact cities or years were not available, averages across nearby locations and/or years were used. See Supplemental Tables 2-4 for specific month, year and location values used. Descriptive statistics for UV radiation by site and month are also presented in the appendix in Supplemental Tables 5-7.

A sensitivity analysis was performed including dietary intake in the model, as dietary intake is a leading predictor of 25(OH)D levels. However, with the inclusion of dietary intake in the model, the optimal PGS (and p-value cutoff) remained the same for the European cohort (0.00035), but performed much worse in the African-ancestry cohort, so intake was not included in the model to determine the optimal p-value cutoff. Results from this and other sensitivity analyses can be found in Supplemental Table 8. Additional sensitivity analyses were performed to ensure that the Randomized Controlled Trial study design of the WHI CaD trial was not biasing the results. There was no significant difference in 25(OH)D concentration between participants in the treatment arm compared to the placebo arm. Additionally, there was no significant difference in PGS*25(OH)D trend in WHI compared to the other cohorts.

Enrichment Analysis

After development of the PGS, SNPs were annotated to genes using the SNPnexus tool developed by the Barts Cancer Institute at the Queen Mary University of London and ancestry-specific enrichment analyses were performed using Gene Ontology. Genes were mapped to (1) biologic process and (2) molecular function, and Fisher's exact tests were performed to test for enrichment of processes or functions.

Results

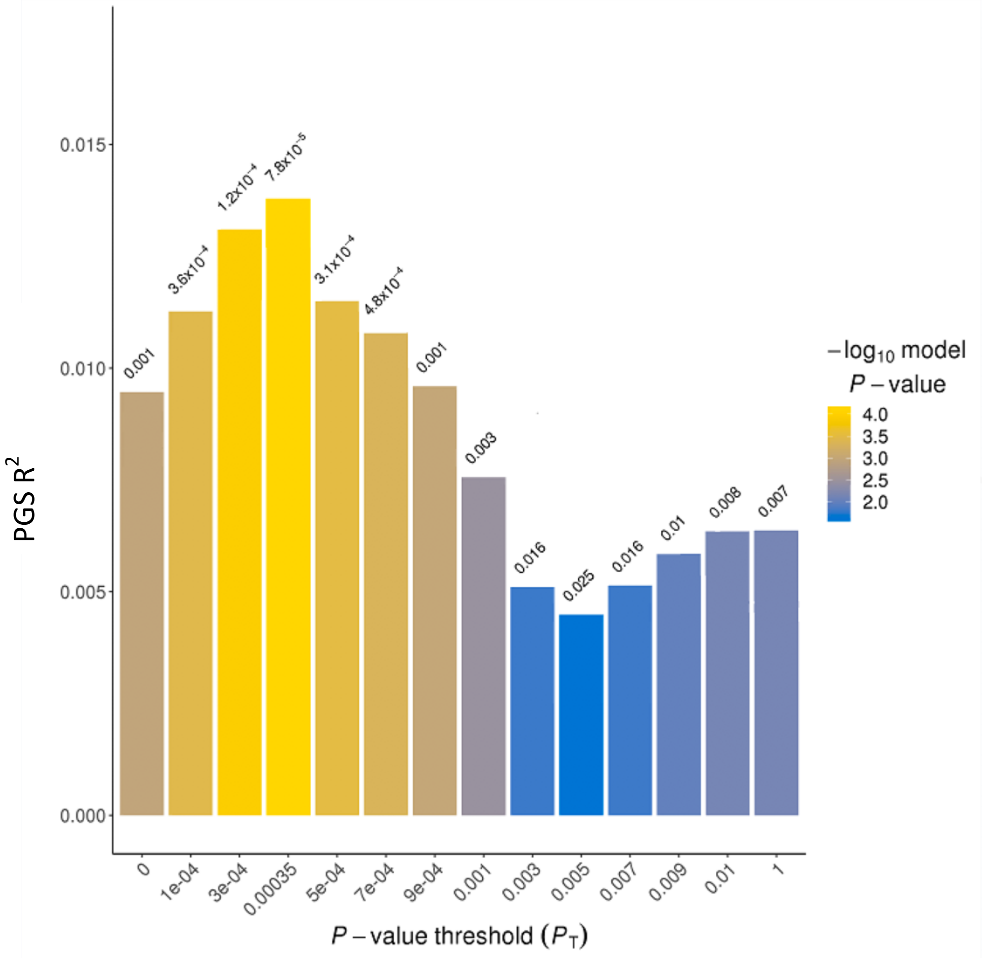
Table 3.3 shows statistics for the best performing PGS for each ancestry for the two LD cutoffs while controlling for age, sex, BMI and available UV radiation. In both ancestries, the PGS using the LD cut-off of 0.5 was more strongly associated with and explained more of the variance in 25(OH)D than did the PGS using the LD cut-off of 0.2. Therefore, this was the optimal PGS utilized going forward. For the European ancestry cohort, the optimal PGS explained 1.4% of the variance seen in log[25(OH)D] (p=0.00035), with the fully adjusted model explaining 12.9% of the variance seen in log[25(OH)D]. In the African ancestry cohort, the PGS explained 4.4% of the variance seen in log[25(OH)D] (p=0.06), with, the fully adjusted model explaining 37% of the variance seen in log[25(OH)D]. Of note, the optimally performing PGS in the African ancestry contained many more SNPs than that from the European cohort mostly due to the less stringent p-value cutoff, but also because a larger number of SNPs remained post clumping (850,697 vs 228,867) due to smaller LD blocks in the African-ancestry sample and more input SNPs from the GWAS summary statistics (8.4 million in the African ancestry vs 1.2 million in the European ancestry sample). Figures 3.1 and 3.2 depict the results visually, where a taller bar corresponds to a larger PGS R^2. Figure 3.1 shows a bar plot of the coefficient of determination (R^2) for TRANSCEN-D GWAS meta-analysis p-value thresholds from 0 to 1 at selected intervals in the European-ancestry sample; Figure 3.2 shows the same in the African-ancestry sample. To investigate biologic underpinnings of the PGS, ancestry specific enrichment analyses were performed. Many biological and molecular processes were enriched for using a Fisher's exact test and a false discovery rate (FDR) adjusted p-value. However, no novel categories or ancestry-specific trends emerged. Enrichment analysis results can be found in Supplemental Tables 9 and 10 and Supplemental Figures 12 and 13.

Table 3.3: Performance of optimal PGS in each ancestry

Ancestry	LD cutoff	p-value cut-off	PGS R^2 (model R^2)	p-value[a]	# SNPs
European (n=1,000)	0.5	0.00035	0.014 (0.129)	0.00008	341
	0.2	0.1142	0.011 (0.126)	0.0005	44,883
African (n=57)	0.5	0.01265	0.044 (0.37)	0.06	32,269
	0.2	0.01375	0.036 (0.366)	0.10	27,662

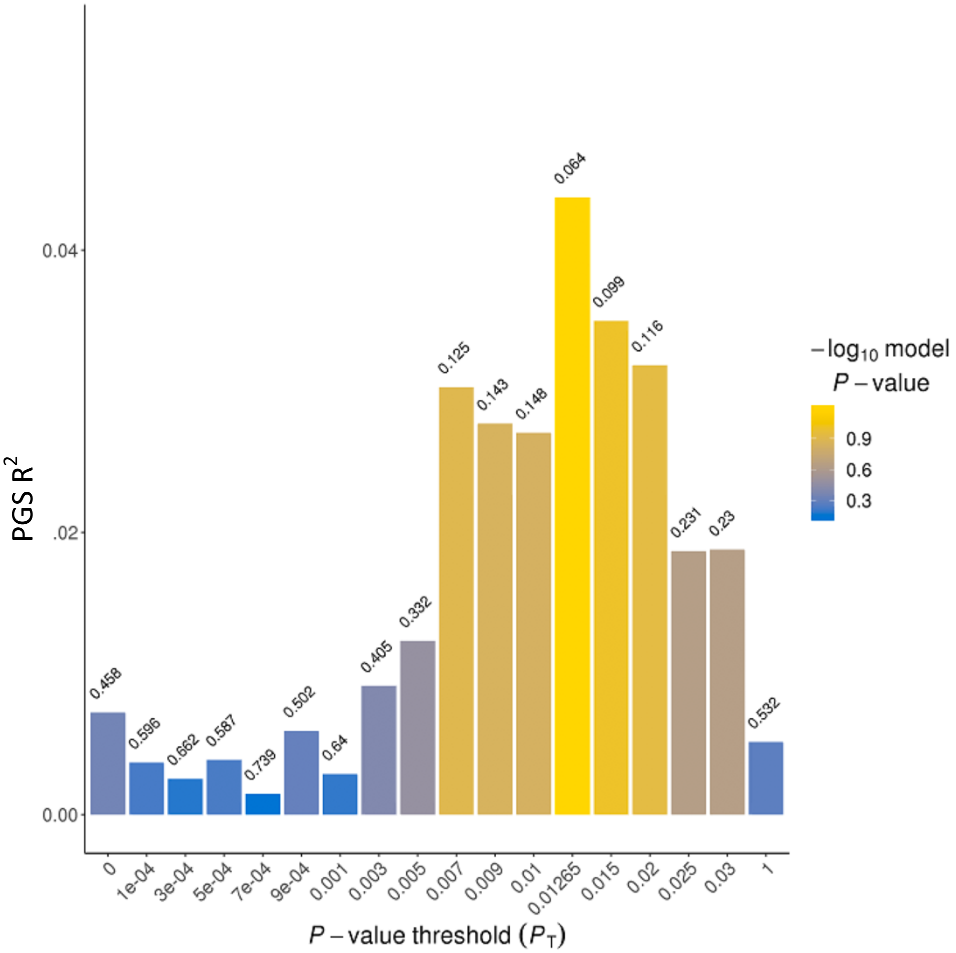
[a] p-value for association between PGS and log[25(OH)D]

Figure 3.1: PGS performance in those of European ancestry (n=1000; LD cutoff of 0.5)



In Figure 3.1, the bars represent the proportion of phenotypic variance explained by the PGS. The most optimally performing PGS has the tallest bar (p=0.00035) and explains 1.4% of the variance in log[25(OH)D]. The p-value cut-offs along the X-axis come from the TRANSCEN-D meta-analysis.

Figure 3.2: PGS performance in those of African ancestry (n=57; LD cutoff of 0.5)



In Figure 3.2, the bars represent the proportion of phenotypic variance explained by the PGS. The most optimally performing PGS has the tallest bar (0.01265) and explains 2.4% of the variance in log[25(OH)D]. The p-value cut-offs along the X-axis come from the TRANSCEN-D meta-analysis.

Tables 3.4 and 3.5 summarize the ancestry-specific optimal PGS when stratified by gender or available UV radiation. General trends showed that the PGS R^2 was typically greater in females (2.2% vs 2.1% in European ancestry and 9.6% vs 5.9% in African ancestry).

Table 3.4: Performance of European-ancestry optimal PGS stratified by gender or available UV radiation

	Strata	p-value cut-off	PGS R^2 (model R^2)	p-value[a]	# SNPs
Gender	Male (N=468)	0.1063	0.021 (0.123)	0.001	60,376
	Female (N=532)	0.0008	0.022 (0.118)	0.0003	705
UV Index	Bottom 50%[b] (N=520)	0.1109	0.017 (0.065)	0.003	62,678
	Top 50%[c] (N=480)	0.00065	0.019 (0.142)	0.0009	579

[a] p-value for association between PGS and log(25(OHOD)

[b] UV Index below 4.85

[c] UV Index above 4.85

Table 3.5: Performance of African-ancestry optimal PGS stratified by gender or available UV radiation

	Strata	p-value cut-off	PGS R^2 (model R^2)	p-value[a]	# SNPs
Gender	Male (N=29)	0.00005	0.059 (0.38)	0.14	195
	Female (N=28)	0.01365	0.096 (0.46)	0.05	34,570
UV Index	Bottom 50%[b] (N=28)	0.00005	0.049 (0.41)	0.18	195
	Top 50%[c] (N=29)	0.01365	0.027 (0.20)	0.39	34,570

[a] p-value for association between PGS and log(25(OHOD)

[b] UV Index below 8.2

[c] UV Index above 8.2

After the optimal, ancestry-specific PGS was discerned, the PGS was applied to independent participants from the remaining cohorts (Table 3.6). In European-ancestry cohorts, the optimal PGS explained a significant percent of the variance in 25(OH)D in the ARIC (1.3%) and MESA (2.1%) cohorts. The European ancestry-specific PGS had a lower and nonsignificant R^2 in the WHI cohorts, accounting for only 0.5% of the variance in 25(OH)D. In the African-ancestry cohorts, the optimal PGS only explained a significant percent of the variance in 25(OH)D (1.1%) in the WHI (consent group 2) cohort.

Table 3.6: Optimal PGS performance by cohort

	Cohort	N	PGS R^2 (model R^2)	p-value[a]	# SNPs
European	ARIC	6,462	0.013 (0.14)	$2.6x10^{-23}$	339
	MESA	1,995	0.021 (0.10)	$1.8x10^{-11}$	31
	WHI (consent group 1)	79	0.005 (0.12)	0.51	331
	WHI (consent group 2)	376	0.005 (0.08)	0.12	337
African	MESA	367	0.0005 (0.12)	0.42	1,246
	WHI (consent group 1)	65	0.0001 (0.023)	0.93	30,309
	WHI (consent group 2)	572	0.011 (0.035)	0.01	31,153
	WHI (consent group 3)	63	0.005 (0.097)	0.58	34,316

[a] p-value for association between PGS and log[25(OH)D]

[b] many fewer SNPs were included in the MESA PGS because of the sparse genotyping panel used

Figures 3.3 and 3.4 show ancestry-specific plots of 25(OH)D by decile or quintile of the PGS. Quintiles were used in the African ancestry due to the smaller sample size. Generally, it can be observed that those with greater genetic risk (lower PGS and decile) have lower 25(OH)D concentrations. For a clinically-based interpretation, in the European determination cohort (Figure 3.3, panel A; n=1,000), those with the highest genetic risk have vitamin D concentrations 4.0 ng/ml lower than those with the lowest risk ($p=1.3x10^{-3}$). In the European validation cohort (Figure 3.3, panel B; n=8,569), the trend persists as those with the highest risk have vitamin D concentrations 3.0 ng/ml lower than those with the lowest risk ($p=3.2x10^{-13}$). Figure 3.4 shows a similar trend, though nonsignificant, for those of African ancestry. In the African-ancestry determination cohort (panel A; n=57), individuals with the highest genetic risk have vitamin D concentrations 4.7 ng/ml lower than those with the lowest risk (p=0.23); in the validation cohort (panel B; n=1,042), those with the highest risk have vitamin D concentrations 1.9 ng/ml lower than those with the lowest risk (p=0.15). Raw data can be found in Supplemental Tables 11 and 12. Where sample size allowed (the combined European cohort, n=8,569), further exploration of the pattern was investigated for those in the 1st and 99th percentiles (Supplemental Table 11). These percentiles followed the linear trend as shown in Figure 3.3.

Figure 3.3: European ancestry-specific quantile plots

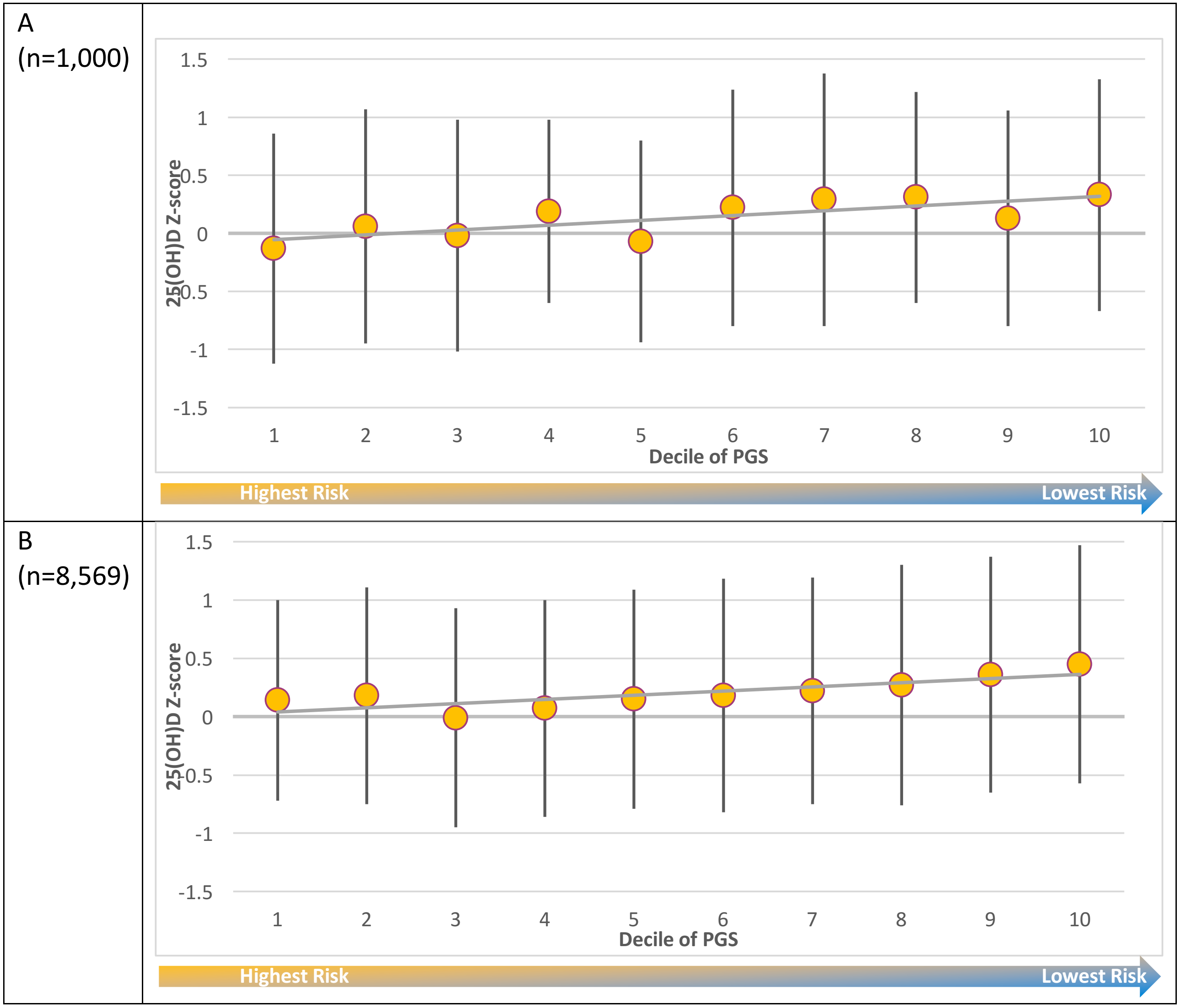


Figure 3.3 shows a visual representation of the association between PGS decile and normalized vitamin D concentrations in those of European ancestry. The x-axis is the PGS decile, where lower decile means more risk of low vitamin D concentrations. The y-axis is vitamin D concentrations (normalized for comparison between cohorts). Panel A is a plot for the subset of ARIC samples used to discern the optimal PGS (n=1000), panel B is a plot for the remaining independent samples (n=8,569). While the exact trend varies by plot, the general trend is that when PGS increases (i.e. lower genetic risk) 25(OH)D concentrations increase. In panel A, moving from the highest risk to the lowest risk decile increases vitamin D concentrations by 4.0 ng/ml. In panel B, moving from the highest risk to the lowest risk decile increases vitamin D concentrations by 3.0 ng/ml.

Figure 3.4: African ancestry-specific quantile plots

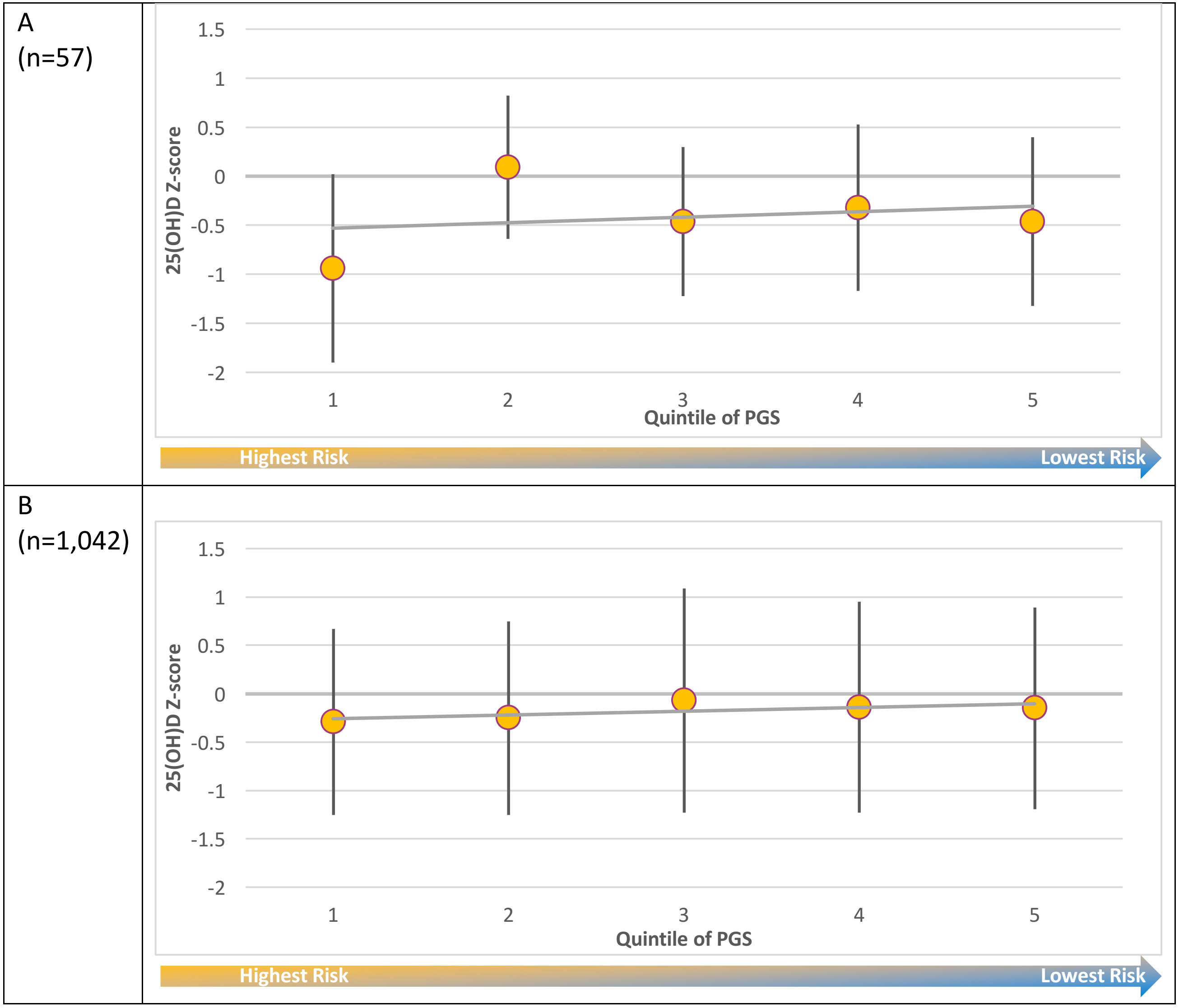


Figure 3.4 shows a visual representation of the association between PGS quintile and normalized vitamin D concentrations in those of African ancestry. The x-axis is the PGS quintile, where lower quintile means more risk of low vitamin D concentrations. The y-axis is vitamin D concentrations (normalized for comparison between cohorts). Panel A is a plot for the subset of ARIC samples used to discern the optimal PGS (n=57), panel B is a plot for the remaining independent samples (n=1,042). While the exact trend varies by plot, the general trend is that when PGS increases (i.e. lower genetic risk) 25(OH)D concentrations increase. In panel B, moving from the highest risk to the lowest risk quintile increases vitamin D concentrations by 4.7 ng/ml. In panel A, moving from the highest risk to the lowest risk quintile increases vitamin D concentrations by 1.9 ng/ml.

Discussion

Vitamin D inadequacy is a pervasive health problem, with a strong genetic basis. However, to date, much of the heritability of 25(OH)D remains unexplained. Furthermore, there is a tremendous gap in the research carried out in minority ancestries compared to European ancestry. Filling these knowledge gaps is critical in preventive care to manage 25(OH)D concentration and development of an ancestry-specific PGS is one way to address these gaps. To date, across all phenotypes, most PGS have been calculated in those of European ancestry. A handful of studies have begun to investigate ancestry-specific PGS, however, none of these approaches utilize an entirely ancestry-specific approach as was undertaken here (90-92). Given the underlying genetic difference between ancestries (i.e. different LD patterns and allele frequencies), an ancestry-specific approach is more appropriate. Here, the adjusted PGS for log[25(OH)D] accounted for between 0.5% and 2.1% of the variance in the phenotype for those of European ancestry and between 0.01% and 4.4% of variance in the phenotype for those of African ancestry. This performance aligns with what has been reported for PGS R^2 of other complex traits: triglycerides (2.3-2.8%), LDL (3.7-4.7%) and HDL (10.1-10.5%) (93, 94). Additionally, the full model accounted for between 8-14% and 2.3-37% of the variance in log[25(OH)D] for those of European and African ancestries, respectively. This is in line with what has been observed for other moderately heritable complex blood phenotypes. For example, the model R^2 values reported for triglycerides, LDL and HDL, are 3%, 5% and 10%, respectively, while controlling race and age (32, 93, 94). Additionally, the model presented here has an R^2 similar to what has been reported in the literature for models explaining vitamin D concentrations (22, 46, 95). Of note, multi-ethnic studies have reported higher model R^2 in African Americans compared to European Americans (96, 97).

The PGS R^2 captures only a modest portion of the phenotypic variance. This is the result of many concurrent influences. First, the PGS did not include rare variants. Rare variants (MAF $\leq$ 0.01) were not included in PGS determination as they were removed from TRANSCEN-D (the base set). Common SNPs account for only a small proportion of genetic variance in complex traits (98). Future PGSs that include rare variants will likely account for a greater portion of the variance. Additionally, the variance that the PGS can capture is limited by the input SNPs. In the best-case scenarios (i.e. densest chips), the overlap between the TRANSCEN-D summary statistics and the input dataset was 3,520,049 and 1,026,643 SNPs, for African and European ancestries, respectively. While over 1 million SNPs can be very informative, much of the genome is missing (and this was an even more drastic portion for samples with sparse genotyping, i.e. MESA). Thirdly, PRSice implements clumping which keeps only the SNP with the strongest association for SNPs in LD (r^2 >0.5 used here) in any given 500kb window. While this prevents SNPs in one correlated region from dominating the PGS, it also reduces the maximum variability that could be captured by a PGS.

Results from the determination of the optimal PGS revealed that the PGS and model explained more variance in log[25(OH)D] in those of African ancestry compared to those of European ancestry. This was not driven by different distributions in the outcome or predictor variables by ancestry (Supplemental Figure 14). This difference could be the result of more input SNPs for the African ancestry sample, which has two main causes: (1) there was a larger set of overlapping SNPs between base and target datasets inputted into PRSice and (2) the LD blocks are smaller in African ancestry populations, as expected, leading to fewer SNPs being removed via the clumping procedure in PRSice. Another reason the PGS and model R^2 (from Table 3.3) could be higher for those of African ancestry is that a small sample was used to determine the p-value cut-off which could have led to overfitting of the model. The results from the two larger African-ancestry cohorts (MESA and WHI (consent group 2) in Table 3.6 support this; the PGS and model R^2 from these cohorts are more in line with what is seen in the European ancestry.

The optimal p-value cutoff was noticeably less stringent for African ancestry compared to European ancestry (p-value of 0.01265 vs 0.00035). This is likely a result of smaller African ancestry sample in the base dataset (TRANSCEN-D; 8,541 African ancestry participants compared to 16,124 European ancestry participants). Small samples can lead to reduced power, larger p-values even in the presence of a true effect, and noise in the data, requiring a more lenient threshold to capture the true positive SNPs.

Within ancestries, the PGS R^2 differed by cohort. This is likely due to differing sample sizes and genotyping platforms used by the cohorts, where smaller samples and less dense genotyping chips had lower R^2. For example, in the European ancestry, the WHI sample R^2 was much lower than the other cohorts; the WHI cohorts had samples of 79 and 376 compared to 1,995 (MESA) and 6,462 (ARIC). In the African-ancestry cohorts, the PGS had poorer and somewhat disparate performance among cohorts, likely driven by small sample sizes; the only African-ancestry cohort with a significant p-value was the largest one (n=572). In addition to having a small sample size (n=367), the poor performance in MESA (R^2=0.05%), is likely due to sparse genotyping which led to many fewer SNPs post imputation (Supplemental Table 1).

The relationship between PGS and vitamin D concentrations was consistent across ancestries and cohorts; those with the lowest PGS (most risk) had lowest vitamin D concentrations. Moving from the highest to lowest quantile changed vitamin D concentrations by 1.9-4.7 ng/ml, a clinically meaningful difference. This information can be used to inform vitamin D supplementation in those with high genetic risk for vitamin D inadequacy. One study reported that for each additional 100 IU of vitamin D consumed, serum levels increased by 0.6 ng/ml (99). Using this conversion, here we see that compared to those with lowest genetic risk, those with highest genetic risk could require an additional 317 to 783 IU of vitamin D.

While this study reiterates the importance of capturing genetic risk using a PGS, which can be used for clinical predictions, the study does come with some limitations. To maintain sample independence from prior GWAS studies and for Aim 3, the sample size used in this analysis was relatively small, especially for the African-ancestry cohort. Future PGSs could be developed implementing the cross-prediction method developed by Mak et,al (100). This method allows and corrects for overlap between the base and target datasets, which would have allowed for a much larger African-ancestry sample (100). The sample size issues experienced for the African-ancestry cohort emphasize the importance of obtaining more diverse samples (i.e. in initiatives like All of Us). Through the TRANSCEN-D GWAS and the analysis here, nearly all of the publicly available African-ancestry samples have been exhausted and sample sizes for other racial/ethnic groups remains limited. Additionally, the genotyping performed did not capture rare variants, limiting the variance that could be captured by the PGS. However, this leaves room for future studies and replication that should be performed. Other future directions could include: quantifying differences in heritability accounted for (compared to using a stringent genome-wide significance threshold), quantifying the heritability that remains (through estimating SNP heritability), testing for interactions between the PGS and environmental factors and testing the relationship between the vitamin D PGS and other phenotypes (i.e. colorectal cancer risk, Multiple Sclerosis, Alzheimer's Disease, etc.) (101-103).

Conclusion

PGSs are a powerful predictive tool. This study calculated the most optimal PGS for vitamin D concentration that captures a moderate portion of the estimated variance of vitamin D concentration. Given the association between the optimal PGS and 25(OH)D concentrations, this PGS could be leveraged for personalized vitamin D supplementation, which could potentially prevent the negative downstream effects of vitamin D inadequacy.

Chapter 4: Aim Two

ANCESTRY SPECIFIC EXPLORATION OF 25(OH)D

HERITABILITY ESTIMATES

Abstract

Background: Vitamin D inadequacy affects about 50% of adults in the United States and is associated with numerous adverse health outcomes. Vitamin D concentrations are influenced by genetic factors and genetic variants may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration. A first step towards informed genetic research is calculating unbiased and valid heritability estimates from unrelated samples.

Methods: Using GCTA v1.26, a total of 8,838 participants (7,119 European ancestry and 1,719 African ancestry) were leveraged to estimate SNP heritability, heritability stratified by gender and available UV radiation, heritability accounted for by prior GWASs and heritability accounted for by the PGS in Aim 1 in an ancestry-specific manner.

Results: SNP heritability estimates were higher in those of African ancestry (32% vs. 22% in European ancestry) and in those with more available UV radiation (3% and 9% higher than those with low UV radiation in African and European ancestry, respectively). The PGS from Aim 1 explains more heritability than do SNPs from previous GWAS findings in European ancestry (17% vs. 7% of the SNP heritability).

Discussion: SNP heritability estimates for 25(OH)D in unrelated participants of European ancestry are on the low end of the range of estimates from related individuals. SNPs from previous GWAS only explain a small portion of 25(OH)D heritability in those of African and European ancestries. While the PGS from those of European ancestry in Aim 1 accounts for a larger portion of the total SNP heritability than do previous GWAS findings, a large portion of the heritability remains unexplained, promoting the need for further investigation into the genetic underpinnings of 25(OH)D.

Background

Vitamin D inadequacy, as defined by a 25-hydroxyvitamin D [25(OH)D] concentration less than 20 ng/mL, affects around 50% of adults in the United States (1-3). Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, migraines, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-9). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Vitamin D concentrations are influenced by genetic factors and genetic variants may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration (21, 22, 73, 74). Therefore, knowledge of the genetic determinants of vitamin D concentrations could be invaluable in prevention of vitamin D associated morbidity and mortality. An initial step is to determine the total genetic contribution (heritability) to vitamin D concentrations.

Heritability is a key metric for understanding the genetic component of a phenotype. Heritability, technically defined, is the proportion of phenotypic variance that is explained by genetic variance, where phenotypic variance includes genetic and environmental variance (40). Classically, heritability of vitamin D concentrations has been calculated using related individuals where the proportion of shared genotypes is estimated based on familial relationship (41). However, this approach tends to overestimate heritability due to attributing shared environment of the related individuals to heritability. Current estimates of heritability of vitamin D concentrations in related individuals are between 20-40% (1, 22, 25). Most of the heritability estimates to date have been calculated in populations of European ancestry. There have been two studies that reported 25(OH)D heritability in African ancestry. Engelman, et al. reported a heritability in 513 related samples to be 28% (31). Hansen, et al, reported 25(OH)D heritability to be 23% in 2,087 unrelated, older African Americans (95).

The heritability of vitamin D concentration has been reported to vary by gender and exposure to UV radiation (often categorized as 'season'), but no consensus has been reached (23, 24, 26). Additionally, these estimates come from related individuals without sufficient control for environmental exposures.

Although there is a void in overall and stratified ancestry-specific 25(OH)D heritability estimates in unrelated participants, there has been other genetic research regarding 25(OH)D. Several studies have uncovered vitamin D associated single nucleotide polymorphisms (SNPs) (1, 28, 31). To date, a handful of vitamin D genes have been discovered through published genome-wide association studies (GWAS) in those of European or African ancestry; these genes include: *AMDHD1*, *ANO6/ARID2*, *CYP2R1*, *CPY24A1*, *DHCR7*, *GC*, *HTR2A*, *KIF4B*, and *SEC23A* (1, 25, 28, 31). It is estimated that a subset of these genes capture 2.8% of the variance in 25(OH)D concentrations (1).

To better tease apart the underlying genetics of 25(OH)D, SNP heritability of vitamin D concentration was calculated in samples of unrelated individuals of European or African ancestry using Genome-wide Complex Trait Analysis (GCTA). Gender-specific heritability differences were investigated while controlling for available UV radiation and vitamin D intake, the two strongest environmental influencers of 25(OH)D concentrations, to remove variability in 25(OH)D explained by these environmental factors and avoid overestimating heritability.

Heritability estimates for 25(OH)D were also stratified by ancestry-specific median level of available UV radiation, a more precise measurement of sun exposure than the categorical variable 'season', while controlling for dietary vitamin D intake. Where sample size allowed, GCTA was used to independently calculate the heritability explained by genome-wide significant SNPs from recent GWAS meta-analyses, and (separately) the polygenic score (PGS) from Aim 1, in European- and African-ancestry populations.

Methods

Participants

SNP heritability estimates were calculated using eligible participants of either European or African ancestry from the Atherosclerosis in Communities Study (ARIC) obtained through dbGaP Study Accession: phs000090.v4.p1. ARIC includes 15,792 participants, of which 8,838 have data required for this analysis: genome-wide data, serum vitamin D, age, sex, BMI, month of blood draw and dietary intake. Serum vitamin D concentration was measured using tandem mass spectrometry (MS/MS; Quest Labs). Of these 8,838 participants 7,119 are of European ancestry and 1,719 are of African ancestry. ARIC is a prospective epidemiologic study conducted at four sites: Wake Forest Baptist Medical Center, Winston-Salem, NC; University of Mississippi Medical Center, Jackson, MS; University of Minnesota, Minneapolis, MN; Johns Hopkins University, Baltimore, MD. Serum vitamin D was measured at ARIC visit 2 (1990-1992). IRB approval and consent were obtained at each respective study site. Additionally, separate IRB approval was obtained for this specific study.

Data Quality Control

Genotyping methods are described elsewhere (81). Supplemental Table 1 gives information on the original genotyping platforms and SNP panels used. Quality control (QC) was done in an ancestry-specific manner for those of European and African ancestry. Ancestry was determined by self-report and confirmed with principal components analysis with 1000 genomes populations serving as anchoring populations (Supplemental Figures 3-4). In summary, QC for each cohort removed: sex mismatches, samples and SNPs with high missingness (>5%), SNPs with low minor allele frequency (MAF<0.2%), and SNPs out of Hardy-Weinberg equilibrium (p<0.05/number of SNPs; Bonferroni adjusted cut-off). Datasets were then imputed using the Michigan Imputation Server (56, 58). European samples were imputed to the Haplotype Reference Consortium (HRC) and African samples were imputed to the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (58, 86). Post imputation QC included: removing SNPs with a low-quality score (<0.8) or MAF (<0.1%). Additionally, sample and SNP level missingness as well as HWE cutoffs were rechecked. Supplemental Figures 3-4 and 10-11, and Supplemental Table 1 give specifics on quality control for each cohort. QC was performed using PLINK v1.9 and vcfTools (87, 88).

SNP Heritability Estimation

SNP heritability estimates were calculated using GCTA v1.26 (43). GCTA calculates SNP heritability as the proportion of phenotypic variance jointly explained by additive effects of a set of SNPs. All inputted SNPs are simultaneously fit as random effects in a linear mixed model (61). GCTA uses genomic-relatedness-based restricted maximum-likelihood (GREML), similar to restricted maximum likelihood analysis (REML) and relies on the genetic relationship matrix (GRM) based on SNPs from unrelated samples (35, 62). Heritability was estimated several ways: (1) SNP heritability by ancestry, (2) ancestry-specific SNP heritability by gender, (3) ancestry-specific SNP heritability stratified by ancestry-specific median level of available UV radiation, (4) ancestry-specific SNP heritability of the PGS computed in Aim 1 (where sample size allowed) and (5) ancestry-specific SNP heritability of previous GWAS findings. In each case, the model was adjusted for age, sex, BMI, available UV radiation and dietary vitamin D intake. In determination of the final model, models with and without dietary vitamin D intake were utilized (Supplemental Figures 15-16). Models without dietary intake tended to overestimate SNP heritability of vitamin D concentrations, particularly in the smaller sample (those of African ancestry). Therefore, all models used in Aim 2 were adjusted for dietary vitamin D intake.

SNP heritability and stratified SNP heritability estimates were calculated using all genotyped and imputed SNPs for both the European- and African-ancestry populations from ARIC; this was 8,315,761 and 9,335,785 SNPS, respectively. Extending from the methods used by the SUNLIGHT consortium, to estimate heritability captured by the PGS calculated in Aim 1, heritability was calculated two times; once using the clumped set of SNPs used to determine the PGS from Aim 1 (228,867 SNPs for European ancestry and 850,697 for African ancestry) and a second time removing the PGS SNPs from the clumped set of SNPs (228,526 SNPs for European ancestry and 818,428 for African ancestry). The difference in heritability estimates between these two models was then the heritability attributed to the PGS (25). Heritability could not be directly calculated from the SNPs in the PGS because one of the assumptions made by the GCTA modeling is an average null effect of the SNPs on the outcome. Of note, the

African ancestry sample size was too small for this analysis to be valid, so heritability attributed to the PGS was only calculated in those of European ancestry. In discerning the heritability captured by previous GWAS studies, heritability was calculated using a reduced set of SNPs: the full genotyped and imputed set with top GWAS findings (and SNPs in the surrounding LD block) removed (25, 104). The difference between this estimate and the overall heritability estimates was then the heritability attributed to previous GWAS findings (25). Top SNPs in *GC*, *CYP2R1*, *CYP24A1* and *DHCR7*/*NADSYN1* and SNPs in LD with these SNPs were removed (1, 27, 28). This was done separately for European and African ancestry samples. Additionally, a second heritability estimate was calculated that included recent novel findings. This included SNPs from *AMDHD1* and *SEC23A* in those of European ancestry and SNPs from *KIF4B*, *HTR2A* and *ANO6/ARID2* in those of African ancestry (25, 27). Table 4.1 summarizes the SNPs and LD blocks removed in each scenario. LD block size was determined using the Plots mode of the SNAP tool by the Broad (104). Of note, two SNPS (rs79666294 and rs6013897) were not found in the ARIC African ancestry imputed data. For SNP rs79666294, using the RAGGR tool by USC, SNP rs17570361 was found to be a good proxy (r^2 0.94). In the African-ancestry data, there was no proxy for rs6013897, so SNPs within the LD block of its position (52742479) were removed. All models were fit separately for European- and African-ancestry samples. T-tests for significant differences were performed using GraphPad Software, La Jolla, California, USA.

Table 4.1: Previous GWAS SNPs

SNP ID	Chromosome	Position[c]	Gene	EU LD block size	AFA LD block size
rs2282679	4	72608383	*GC*	1200 kb	2 kb
rs79666294	5	155047146	*KIF4B*	NA[a]	200 kb
rs10741657	11	14893332	*CYP2R1*	480 kb	300 kb
rs12785878	11	71456403	*NADSYN1/DHCR7*	120 kb	84 kb
rs719700	12	45635426	*ANO6/ARID2*	NA[a]	2 kb
rs10745742	12	95964751	*AMDHD1*	50 kb	NA[b]
rs1410656	13	46968386	*HTR2A*	NA[a]	28 kb
rs8018720	14	39086981	*SEC23A*	180 kb	NA[b]
rs6013897	20	54125940	*CYP24A1*	10 kb	4 kb

[a] Novel African ancestry SNP
[b] Novel European ancestry SNP
[c] Build 37

Results

Participant characteristics are summarized in Tables 4.2. Table 4.3 and Figure 4.1 show the overall SNP heritability estimates by ancestry, as well as SNP heritability estimates stratified by gender or ancestry-specific median of available UV radiation, within ancestry. Generally, SNP heritability is higher in the African-ancestry cohort compared to the European-ancestry cohort, though not significantly different. This can be seen in the overall estimate (32% vs 22%; p=0.49) as well as in each of the stratified estimates. While not significant, when stratified by gender, males, compared to females, demonstrate higher SNP heritability in those of European ancestry (36% vs 26%; p=0.47), but the opposite is true in those of African ancestry (37% vs 46%; p=0.86). When stratified by available UV radiation, heritability is higher in those of European ancestry with more available UV radiation (24% vs 15% those with lower UV radiation; p=0.51); in those of African ancestry, heritability is similar regardless of the available UV radiation.

Table 4.2 Participant characteristics

Cohort	Variable	European-ancestry	African-ancestry
ARIC	Sample size	7,119	1,719
	Age [years]	57.1 (5.7)	56.4 (5.8)
	BMI [kg/m^2]	27.3 (4.8)	30.2 (6.2)
	UV [units]	5.0 (2.5)	6.9 (2.3)
	Intake [IU]	222.8 (144.4)	215.7 (150.3)
	25(OH)D [ng/ml]	25.9 (8.8)	19.1 (7.1)

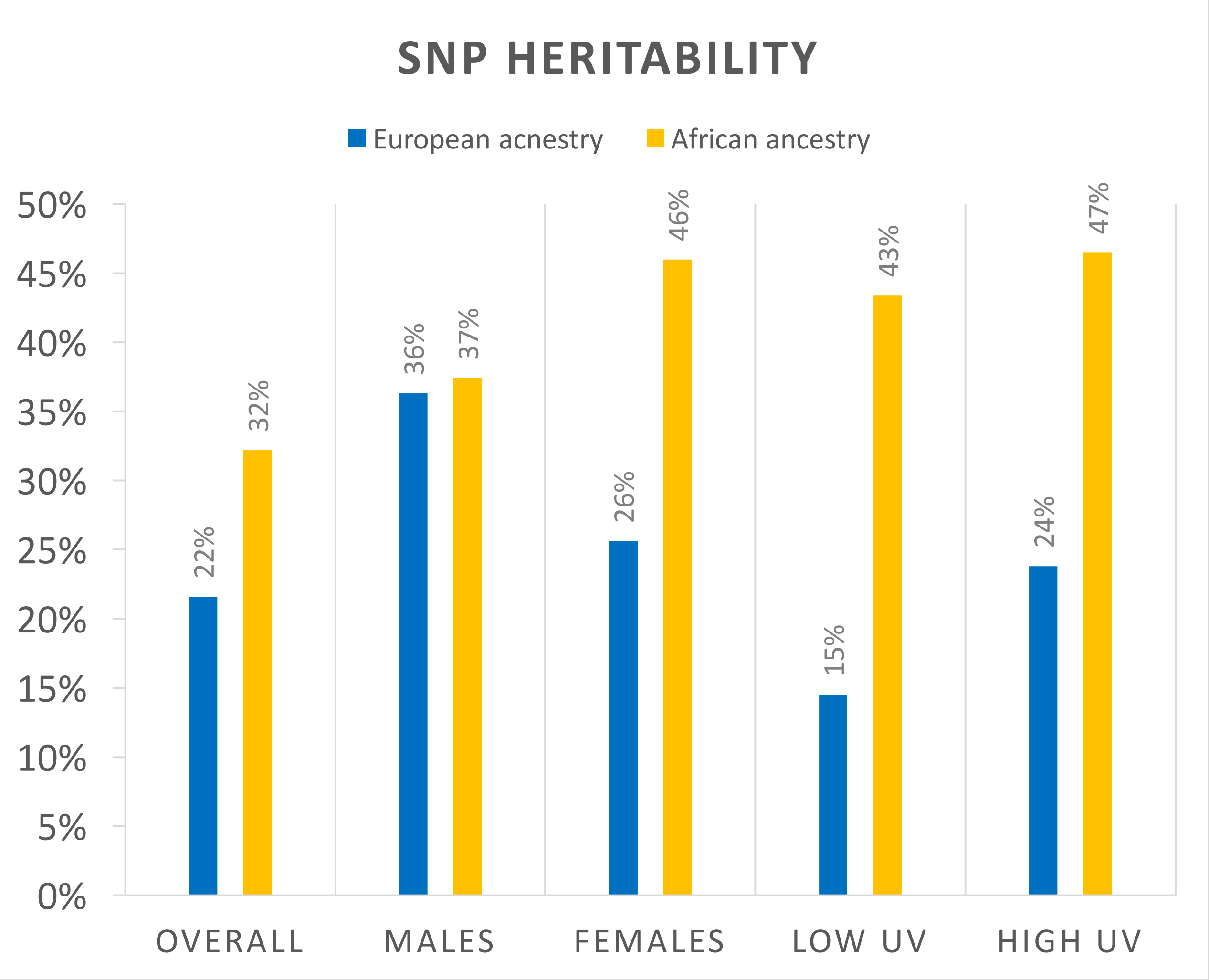
Table 4.3: Overall SNP heritability by race and stratified by gender and median available UV radiation

Ancestry	Partition	Sample Size	Heritability (SE)
European	Overall	7,119	22% (5.2)
	Males	3,301	36% (11.4)
	Females	3,818	26% (9.3)
	Low UV (<4.85)[b]	3,689	15% (9.4)
	High UV (>4.85)[b]	3,430	24% (10.0)
African	Overall	1,719	32% (17.8)
	Males	635	37% (38.1)
	Females	1,084	46% (30.4)
	Low UV (<8.1)[b]	910	43% (30.9)
	High UV (>8.1)[b]	809	47% (39.3)

[a] Model controlled for: age + sex + BMI +available UV radiation + vitamin D intake

[b] Available UV radiation was stratified by ancestry-specific median level

Figure 4.1: Overall SNP heritability by race and stratified by gender and median available UV radiation



Sample Size					
European Ancestry	7,119	3,301	3,818	3,689	3,430
African Ancestry	1,719	635	1,084	910	809

Figure 4.1 shows SNP heritability stratified by ancestry, gender and available UV radiation. In general, those of African ancestry or who have high available UV radiation are shown to have higher heritability, although not statistically significant (p=0.49 (African vs European ancestry), p=0.95 (high vs low UV within African ancestry) and p=0.50 (high vs low UV within European ancestry).

Figure 4.2 shows ancestry-specific SNP heritability estimates for all the SNPs, SNPs in the PGS computed in Aim 1 and SNPs discovered in previous GWAS studies (1, 27, 28). In those of European ancestry, the PGS from Aim 1 accounts for 17.1% (3.7/21.6*100) of the SNP heritability of 25(OH)D concentrations and previous replicated GWAS study findings (this includes SNPs from *CYP2R1*, *CYP24A1*, *DHCR7*/*NADSYN1* and *GC*) account for 6.9% (1.5/21.6*100) of the total SNP heritability (1, 27, 28). In those of African ancestry, these same top GWAS findings accounted for only 1.6% (0.5/32.2*100) of the total SNP heritability. Heritability accounted for remained unchanged when ancestry-specific novel findings were included in the heritability estimations (1, 25, 27, 28). African-ancestry sample size was too small to calculate heritability accounted for by the PGS in Aim 1.

Figure 4.2: SNP Heritability explained by all SNPs (overall), the PGS or previous GWAS findings

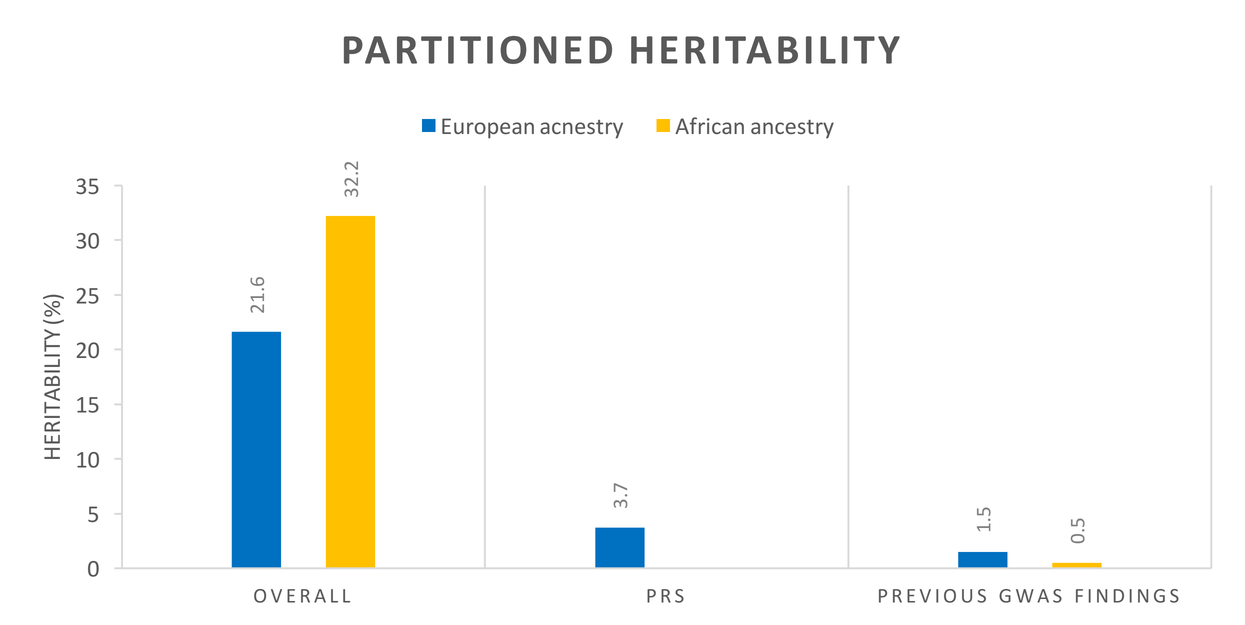


Figure 4.2 shows overall SNP heritability by ancestry as well as the heritability explained by the PGS from Aim 1 (European ancestry only) and by top GWAS findings. As shown here, the PGS from Aim 1 explains only a fraction of the total SNP heritability, but explains more heritability than top GWAS findings do in those of European ancestry.

Discussion

Vitamin D concentration [25(OH)D] is a complex trait with environmental and genetic predictors. A first step in discerning the genetic underpinnings of 25(OH)D is to have unbiased heritability estimates in non-related participants. Here, heritability of 25(OH)D was discerned in non-related participants of African or European ancestry and was investigated by participant gender and level of available UV radiation. Additionally, heritability accounted for by GWAS findings to date and a more comprehensive PGS was quantified, revealing that the PGS captures more of the genetic underpinnings than do GWAS findings.

The heritability estimates for those of European ancestry fall into the low range of estimates calculated by prior family-based studies (20-40%), whereas estimates for those of African ancestry are slightly higher than the estimate reported by Engelman, et al. (32% vs 28%) (1, 22). In regards to heritability estimates in unrelated samples, estimates presented here for those of African ancestry, are higher than those reported by Hansen, et al, in a sample of elderly African Americans (32% vs 23%) (95). These estimates build off what was reported by Hansen, et al, adding stratified estimates and including a wider age range of adult participants. However, small sample size (N=1,719, compared to N=2,087 in Hansen, et al.) could be biasing these estimates upward, therefore further study with larger samples is recommended.

When stratified by gender, in those of European ancestry males had higher heritably than females and in those of African ancestry, females had higher heritability than did males, although neither of these differences were statistically significant. Several studies have reported higher heritability in men (specifically in men of European or Asian ancestry), while reports of higher heritability in women are novel (24-26).

Differences in heritability by season have been reported by a handful studies in participants of European or Asian ancestry (23, 24, 26). There is mixed evidence on whether heritability of 25(OH)D concentrations are is higher in the summer or winter; studies from China and the United States report higher heritability in winter, while a study from Sweden reports higher heritability in the summer (23, 24, 26). However, none of the studies controlled for a full set of crucial confounders. The Chinese study did not control for vitamin D intake or BMI; the United States study did not control for age, sex or vitamin D intake; the Swedish study did not control for sex or BMI. Of note, the Swedish study was the only study to perform analyses controlling for vitamin D supplement use. Here, when stratified by available UV radiation, those of European ancestry with lower available UV radiation had lower heritability than those with higher available UV radiation; this aligns with what was reported by the Swedish study. In those of African ancestry, those with high and low available UV radiation had similar heritability of 25(OH)D. Given that those of African ancestry have darker skin, which absorbs less UV from the sun’s rays, it makes sense that heritability levels were similar for both strata of available UV radiation.

Oftentimes, as shown in Figure 4.1, the stratified estimates were higher than the overall estimates. Stratified models better account for the difference in variances of certain covariates by the stratified variable (i.e. BMI and sex), which are not fully adjusted for in the non-stratified model (61). Additionally, stratified samples are smaller which creates nosier and perhaps inflated estimates.

The PGS explained more of the SNP heritability than did previous GWAS findings, 17.1% compared to 6.9% in those of European ancestry (sample size was too small for PGS heritability calculations in those of African ancestry). However, neither the PGS nor previous GWAS findings explain a large portion of total SNP heritability, promoting the need for genetic studies with larger samples sizes and more dense SNP data that includes low frequency variants to fully understand the genetic determinants of vitamin D concentrations and, therefore, inform the most effective vitamin D supplementation practices. Of note, the estimate of heritability explained by previous GWAS findings presented here (1.5%) is less than the estimate presented by the SUNLIGHT consortium (2.8%) (105).This could be due to the fact that SUNLIGHT included all SNPs within +/- 500kb of the top GWAS findings, whereas here, smaller, more precise LD blocks surrounding the top GWAS SNP were removed (104). In those of European ancestry, the 25(OH)D heritability estimate for the PGS was 3.7%. As hypothesized, this is more than the heritability that the top 6 GWAS findings and LD blocks account for as reported by the SUNLIGHT consortium (2.8%) and as reported here (1.5%) (25).

This study contributes in-depth investigation into 25(OH)D heritability by ancestry, gender, available UV radiation, PGS and GWAS findings; adding important multi-ethnic research that teases apart genetic underpinnings of 25(OH)D concentrations to the literature. However, it is not without limitations. While GCTA allows for the calculation of heritability in unrelated participants, which avoids overestimation due to shared environment, it only accounts for additive SNP effects, potentially underestimating total heritability which also could include gene-by-gene and gene-by-environment interactions. Additionally, GCTA assumes SNPs are in linkage equilibrium which could lead to biased estimates and/or standard errors. While adjusting for available UV radiation is more precise than season, it is not a perfect proxy for actual UV radiation based on time spent outside. This study adds important investigation into

25(OH)D heritability of those of African ancestry, a group with very high prevalence of vitamin D inadequacy. The sample of African Americans utilized is relatively large, however, the sample size is limited due to a lower proportion of GWAS data from African Americans, which could have led to biased estimates produced by GCTA, promoting the need for further study (30).

Conclusion

As expected, SNP heritability estimates for 25(OH)D in unrelated participants of European ancestry are on the low end of the estimates from related individuals. Additionally, findings from previous GWAS only explain a small portion of 25(OH)D concentration heritability (5-7%) in those of African and European ancestries. While the PGS from those of European ancestry in Aim 1 accounts for a larger portion of the total SNP heritability (17%), a large portion of the heritability remains unexplained, promoting the need for further investigation into the genetic underpinnings of 25(OH)D.

Chapter 5: Aim Three

MULTI-ETHNIC ANALYSIS SHOWS GENETIC RISK AND ENIVRONMENTAL PREDCITORS INTERACT TO INFLUENCE 25(OH)D CONCENTRATION AND OPTIMAL VITAMIN D INTAKE

Multi-ethnic analysis shows genetic risk and environmental predictors interact to influence 25(OH)D concentration and optimal vitamin D intake

Hatchell, Kathryn E.[1]; Lu, Qionshi[2]; Mares, Julie A.[3]; Frazier-Wood, Alexis[4]; Engelman, Corinne D.[1]

[1]Department of Population Health Sciences, University of Wisconsin-Madison School of Medicine and Public Health, Madison, Wisconsin, 53706, USA
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison School of Medicine and Public Health, Madison, Wisconsin, 53706, USA
[3]Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, Wisconsin, 53706, USA
[4]USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, Houston, Texas, 77030, USA

Sources of Support:

ARIC
The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions.

Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle).

WHI
The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. This manuscript was not prepared in collaboration with investigators of the WHI, has not been reviewed and/or approved by the Women's Health Initiative (WHI), and does not necessarily reflect the opinions of the WHI investigators or the NHLBI.

WHI PAGE is funded through the NHGRI Population Architecture Using Genomics and Epidemiology (PAGE) network (Grant Number U01 HG004790). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination, was provided by the PAGE Coordinating Center (U01HG004801-01).

Funding support for WHI GARNET was provided through the NHGRI Genomics and Randomized Trials Network (GARNET) (Grant Number U01 HG005152). Assistance with phenotype

harmonization and genotype cleaning, as well as with general study coordination, was provided by the GARNET Coordinating Center (U01 HG005157). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004424).
The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession phs000200.v11.p3.

Funding for WHI SHARe genotyping was provided by NHLBI Contract N02- HL-64278.

KEH was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359). Computational resources were supported by a core grant to the Center for Demography and Ecology at the University of Wisconsin-Madison (P2C HD047873).

Conflict of Interest and Funding Disclosure:
Kathryn E. Hatchell- no conflicts of interest
Qiongshi Lu- no conflicts of interest
Julie A. Mares- no conflicts of interest
Alexis Frazier-Wood- no conflicts of interest
Corrine D. Engelman- no conflicts of interest

Corresponding Author:
Kathryn Hatchell
610 Walnut Street #1007F
Madison, WI 53726
khatchell@wisc.edu

Word Count: 2,562

Number of figures: 2
Number of tables: 1

Supplementary data submitted:
5 Tables and 4 Figures

Running Title: GxE interaction affects 25(OH)D & vitamin D intake

Abbreviations:

ARIC: Atherosclerosis Risk In Communities Study; BMI: body mass index; CAAPA: Consortium on Asthma among African-ancestry Populations in the Americas; CaD: Calcium and Vitamin D Trial; DF: degree of freedom; GWAS: genome-wide association study; HRC: Haplotype Reference Consortium; HWE: Hardy-Weinberg Equilibrium; IOM: Institute of Medicine; IU: international units; LCMS: liquid chromatography-mass spectrometry; MAF: minor allele frequency; MESA: Multi-Ethnic Study of Atherosclerosis; MET: metabolic equivalent; PGS: polygenic score; QC: quality control; RCT: randomized controlled trial; SE: standard error; SNP: single nucleotide polymorphism; TRANSCEN-D: TRANS-ethniC Evaluation of vitamiN D; UV: ultra-violet; VITAL: Vitamin D and Omega-3 trial; WHI: Women's Health Initiative

Abstract

Background: Vitamin D inadequacy affects almost 50% of adults in the United States and is associated with numerous adverse health effects. Vitamin D concentration [25(OH)D] is a complex trait with strong genetic and environmental predictors that work in tandem to influence 25(OH)D and may determine how much vitamin D intake is required to reach an optimal 25(OH)D concentration. To date, there has been little investigation into how genetics and environment interact to affect 25(OH)D.

Objective: Interactions between a polygenic score (PGS) and vitamin D intake (PGS*intake) or available ultra-violet (UV) radiation (PGS*UV) were evaluated in individuals of African or European ancestry.

Methods: Mega-analyses were performed using three independent cohorts (N=9,668; African ancestry n=1,099; European ancestry n=8,569). One-degree of freedom (DF) and 2-DF models were used to test for interaction. All models controlled for age, sex, body mass index (BMI), cohort and dietary intake/available UV. Additionally, in participants achieving Institute of Medicine (IOM) vitamin D intake recommendations, 25(OH)D was evaluated by level of genetic risk.

Results: The 2-DF PGS*intake, 1-DF PGS*UV and 2-DF PGS*UV models were statistically significant in participants of European ancestry ($p=3.3x10^{-18}$, $2.1x10^{-2}$ and $2.4x10^{-19}$, respectively), but not in those of African ancestry. In European-ancestry participants who reached IOM vitamin D intake guidelines, the percent of participants achieving adequate 25(OH)D increased as genetic risk decreased (71.7% vs 89.0% in the highest vs lowest risk categories; p=0.018).

Conclusions: Available UV radiation and vitamin D intake interact with genetics to influence 25(OH)D. Individuals with higher genetic risk may require more vitamin D exposure to maintain optimal 25(OH)D concentrations. Overall, the results showcase the importance of incorporating both environmental and genetic factors into analyses, as well as the potential for gene-environment interactions to inform personalized dosing of vitamin D.

Keywords: Gene-environment interaction, ancestry-specific, vitamin D, diet, polygenic risk score, African, European

Introduction

Vitamin D inadequacy, as defined by a 25-hydroxyvitamin D [25(OH)D] concentration less than 20 ng/mL, affects almost 50% of adults in the United States (1-3). Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, migraines, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-9). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Recent results from the Vitamin D and Omega-3 trial (VITAL) showed null associations between vitamin D supplementation and both cancer and cardiovascular disease, however, study design, including supplementation in individuals with adequate 25(OH)D concentrations, limits the interpretability of these findings (70).

Vitamin D concentration is a complex phenotype with genetic and environmental predictors that may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration (21, 22, 73, 74). Primary environmental predictors of 25(OH)D concentrations are vitamin D intake through diet and supplementation, and available ultraviolet (UV) radiation exposure. Therefore, knowledge of how genetic determinants of vitamin D concentrations interact with environmental predictors could be useful in the prevention of vitamin D associated morbidity and mortality. Understanding gene-by-environment interactions and how they affect 25(OH)D concentrations could be leveraged for downstream personalized supplementation for maintaining adequate vitamin D concentrations through a precision public health approach.

While attention has been paid to genetic determinants of vitamin D concentration through genome-wide association studies (GWAS) and, separately, to the environmental determinants, much less research has focused on how environmental factors interact with genetic factors. Investigating the effects of genetic or environmental predictors in isolation may miss much of the variance in 25(OH)D. Through GWAS, only 2.8% of the variance in 25(OH)D can be explained (25). Research has found that vitamin D intake through diet and supplement use accounts for 1-8% of the variation in vitamin D concentrations between individuals, and that sun exposure accounts for 1-15% of the variation (21, 45-47). One study in European ancestry women reported an interaction between two *GC* SNPs (the *GC* protein product transports the vitamin D metabolites in the blood) and both vitamin D intake and sun exposure, where the genetic effect was stronger, with more variance explained, in summer and in those with a higher intake of vitamin D (21). This same study reported preliminary evidence of differing genetic effect of a PGS (polygenic score), comprised of two SNPs, by level of vitamin D intake and season (21). Therefore, it is important to investigate gene-environment interactions as the risk inferred by genetic or environmental factors alone is not enough to predict risk of inadequate vitamin D concentrations.

Here, the interactions between a polygenic score (PGS; Hatchell, et, al, submitted) and vitamin D intake or available UV radiation will be tested using linear models in individuals of African or European ancestry. Additionally, to replicate findings from a previous study (21), in participants achieving Institute of Medicine (IOM) vitamin D intake guidelines, the percent reaching adequate (>20 ng/ml) 25(OH)D concentrations, stratified by level of genetic risk, will be determined. These results could help inform screening and treatment of vitamin D inadequacy based on genetic and environmental factors.

Methods

Participants

Analyses were performed in a sample of 8,569 participants of primarily European ancestry and 1,099 participants of primarily African ancestry who had data for the required variables: age, sex, body mass index (BMI), dietary intake of vitamin D, available UV radiation and genome-wide single nucleotide polymorphisms (SNPs). Participants were from Atherosclerosis Risk in Communities (ARIC), the Multi-ethnic Study of Atherosclerosis (MESA) and the Women's Health Initiative (WHI), and are independent of the GWAS meta-analysis, TRANS-ethniC Evaluation of vitamiN D, TRANSCEN-D, that provided the summary statistics used to calculate the PGS (52). ARIC is a prospective study of men and women ages 46-70 years. Participants are recruited in Forsyth, NC, Jackson, MS, Minneapolis, MN and Washington County, MD. Serum vitamin D was measured for particular ancillary studies of ARIC at visit 2 (1990-1992). ARIC data were obtained through dbGaP Study Accession: phs000090.v4.p1. MESA is a prospective study of men and women ages 44-84 who were recruited from Columbia University, New York, NY; Johns Hopkins University, Baltimore, MD; Northwestern University, Chicago, IL; University of Minnesota, Minneapolis, MN; University of California at Los Angeles, Los Angeles, CA and Wake Forest University, Winston-Salem, NC. Serum vitamin D was measured at MESA exam 1 (July 2000-August 2002). MESA data were obtained through dbGaP

Study Accession: phs000209.v13.p3. Women participating in WHI were recruited from 40 clinical centers in the United States. Serum 25(OH)D was measured as part of the Calcium and Vitamin D (CaD) Trial (80). WHI data were obtained through dbGaP Study Accession: phs000200.v11.p3. The data used in these analyses were collected under guidelines from the relevant institutional review boards and all participants provided informed consent, including consent for use of genetic data.

Calculation of available UV radiation

Available UV radiation was calculated based on participant month of blood draw and location. Participants were assigned continuous available UV radiation values that were an average of the UV-index for the relevant exposure period: the month prior to blood draw. UV data were obtained from the National Weather Service Climate Prediction Center historical database; the UV radiation values ranged from 0.7 to 9.5 UV index units. The methods are described in more detail elsewhere (Hatchell, et, al, submitted).

Measurement of 25(OH)D

Serum 25(OH)D concentrations were measured by the studies using different assays. WHI used the DiaSorin LIASON chemiluminescence, MESA used liquid chromatography-mass spectrometry (LCMS) and ARIC used tandem mass spectrometry (MS/MS; Quest Labs). To control for differences in vitamin D concentrations due to different assays, vitamin D concentrations were converted to z-scores within studies for combined analyses.

Data Quality Control

Quality control of phenotypic data included winsorizing 25(OH)D in the MESA and WHI samples to minimize the influence of outliers (53). In the WHI sample, participants with 25(OH)D values far above the maximum level of detection (150 ng/mL), none of which had extreme vitamin D intake or sun exposure, were removed from the sample; this included 68 participants of European ancestry and 119 participants of African ancestry. All 25(OH)D values were normalized by cohort to account for the different assays utilized.

Where available, physical activity was measured in metabolic equivalent (MET) hours per week. Physical activity was capped at 16 MET hours/day or 112 MET hours/week. Additionally, physical activity data were normalized by cohort to account for the different surveys utilized to acquire the data.

Genotyping and PGS Development

Genotyping methods are described in publications by ARIC, MESA and WHI (81-85). **Supplemental Table 1** gives information on the genotyping array used by the studies. Quality control (QC) was done in an ancestry-specific manner for those of European and African ancestry. Ancestry was determined by self-report and confirmed with principal components analysis using 1000 Genomes samples as anchoring populations (Supplemental Figures 1-9). In summary, QC for each cohort removed: sex mismatches, samples and SNPs with high missingness (>5%), SNPs with low minor allele frequency (MAF<0.2%), and SNPs out of Hardy-Weinberg Equilibrium (HWE) (p<0.05/number of SNPs; Bonferroni adjusted cut-off). Datasets were then imputed using the Michigan Imputation Server (56, 58). European-ancestry samples were imputed to the Haplotype Reference Consortium (HRC) and African-ancestry samples were imputed to the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (58, 86). Post imputation QC included: removing SNPs with a low-quality score (<0.8) or MAF (<0.1%). Additionally, sample and SNP level missingness as well as HWE cutoffs were rechecked. **Supplemental Figures 10 and 11**, and **Supplemental Table 1** give specifics on quality control for each cohort. QC was performed using PLINK v1.9 and vcfTools (87, 88).

Previously, an optimal PGS was determined in an ancestry-specific manner for those of European or African ancestries (Hatchell, et, al, submitted). PGSs were weighted using effect sizes from an independent multi-ethnic GWAS, TRANSCEN-D, the largest multi-ethnic vitamin D GWAS meta-analysis to date (27).

Statistical Analysis

One-degree of Freedom (DF) and 2-DF models were investigated; 1-DF models test only the relevant interaction term, while 2-DF models jointly test both the relevant interaction term and the PGS main effect term. Relevant interaction terms were the PGS interacting with either vitamin D intake (PGS*intake) or available UV radiation (PGS*UV). All 1-DF and 2-DF models controlled for age, sex, BMI, cohort, vitamin D intake and available UV radiation. All statistical analyses were performed using SAS (version 9.4). Further analyses were performed in those who achieved IOM vitamin D intake guidelines (600 IU/day for those 1-70 years old and 800 IU/day for those over 70) to investigate differences in the percent of those achieving adequate

25(OH)D concentrations ($\geq$20 ng/ml) by quartile of genetic risk. Statistical significance was determined by testing difference between two proportions.

Sensitivity analyses were performed in a subset of participants with physical activity data or vitamin D supplement use data (adjusting for these variables). All sensitivity analyses were performed in an ancestry-specific manner for European and African cohorts. Additional sensitivity analyses were performed to ensure that the randomized controlled trial (RCT) study design of the WHI CaD trial was not biasing the results.

Results

Participant characteristics of this sample can be found in **Table 5.1.** Gene-environment interactions for PGS*intake and PGS*UV were tested for with a 1-DF and 2-DF approach. Negative log p-values for the interaction terms are shown in **Figure 5.1**. The 2-DF PGS*intake, 1-DF PGS*UV and 2-DF PGS*UV models were statistically significant in participants of European ancestry (p=3.3x10^{-18}, 2.1x10^{-2} and 2.4x10^{-19}, respectively). In African-ancestry analyses, power was limited due to the smaller sample size, and no statistically significant interactions were discerned. Betas, standard errors and p-values for main effects and interaction terms can be found in **Supplemental Table 13**.

Additional sensitivity analyses were performed. Characteristics for participants used in sensitivity analyses can be found in **Supplemental Tables 14 and 15.** Sensitivity analyses controlling for physical activity showed the same pattern of significance for interaction terms, however, p-values were slightly attenuated due to smaller sample size (**Supplemental Figure 17**). Interaction terms were no longer significant in the sensitivity analyses that used the subsample with vitamin D supplement use, due to loss of power and small sample size (European ancestry n=455; African ancestry n=700). To ensure the RCT study design of WHI did not influence the results, sensitivity analyses were performed. There was no significant difference in 25(OH)D concentration between participants on the treatment arm compared to the placebo arm. Additionally, there was no significant difference in the association between the PGS and 25(OH)D in WHI compared to the other cohorts.

Next, in participants who reached IOM vitamin D intake dietary guidelines, the percent of participants achieving adequate 25(OH)D concentration by PGS quartile was calculated (**Figure 5.2)**. In those of European ancestry, as genetic risk decreased, those reaching optimal vitamin D concentrations increased (71.7% vs 89.0% in the highest and lowest risk categories, respectively). This is a statistically significant (p=0.018) and clinically meaningful difference. The trend persisted in those of African ancestry, however, the difference was not significant (p=0.28) due to small sample size. To ensure this trend was not solely driven by WHI participants (as this had already been published on), a sensitivity analysis removing WHI participants was performed. The trend remained; in those of European ancestry, as genetic risk decreased, the percent reaching optimal vitamin D concentrations increased (72.7% in the highest risk group and 88.6% in the lowest risk group; p-value for difference was 0.029) (**Supplemental Figure 18**). The African ancestry sample size was inadequate to perform the sensitivity analysis.

Discussion

Findings presented here build upon existing literature reporting that UV radiation and vitamin D intake modify the effect that genetics have on 25(OH)D concentrations (21, 106). Previously, in women of European ancestry, the genetic effects were reported to be stronger in summer and in women with high vitamin D intake (>400 IU/day) (21). Similarly, here, in men and women of European ancestry, as available UV radiation or vitamin D intake increased, the PGS had a larger effect on 25(OH)D (**Supplemental Table 13**). Results were not significant in participants of African ancestry, likely a reflection of the smaller sample size and subsequently reduced power.

Interaction results in those of European ancestry indicate that as available UV or vitamin D increases, so does the difference in 25(OH)D between those of lowest and highest risk. This implies that those with high genetic risk may require more vitamin D intake to reach and maintain optimal 25(OH)D concentrations. This trend proved true for those of European ancestry when looking at participants who reached IOM dietary guidelines for vitamin D intake. Fewer participants with high genetic risk reached optimal 25(OH)D concentrations (≥20 ng/ml). The trend was not significant in those of African ancestry due to small sample size. The small sample size used in this analysis or in the creation of the PGS could be limiting our power to detect an association. The results in the European ancestry sample are suggestive that a precision public health approach to achieve adequate blood levels of vitamin D may be more effective, tailoring intake recommendation to genetic risk.

While this study builds upon the novel interactions previously reported on by Engelman, et al, by including men and women of European and African ancestry, it is not without limitations. First while exploring gene-by-environment interactions that influence 25(OH)D concentrations in a multi-ethnic sample is novel, the relatively small size of the African-ancestry sample limited the power. To maintain independence from TRANSCEN-D, which provided ancestry-specific weights for the PGSs, and the PGS development sample, the sample size used in this analysis was relatively small, especially for the African-ancestry cohort (n=1,099), as nearly all of the publicly available African-ancestry samples with relevant data had been exhausted. This emphasizes that we, as a research community, need to include more individuals of African ancestry in our studies to better understand vitamin D requirements and other health outcomes and make ancestrally informed recommendations (i.e. in initiatives like All of Us). Additionally, while the use of available UV radiation is a substantial improvement from using season as a measure of UV exposure, it is not as good as the gold standard, but difficult to measure, 'actual UV radiation'; this measurement error could have limited power. Finally, vitamin D supplementation is a stronger predictor of 25(OH)D concentrations than vitamin D intake from food, which is generally in much lower amounts than those found in supplements. However, only the WHI study measured vitamin D supplement intake for the relevant visit. Therefore, only interactions involving dietary intake had adequate sample size to be investigated in this study, which could have led to the lack of a significant interaction being detected between the PGS and vitamin D intake in the 1-DF models. Nonetheless, findings here can guide future research in the quest for precision public health management of 25(OH)D inadequacy.

Conclusion

This research adds to the ongoing narrative deciphering the predictors of 25(OH)D concentrations (21). Levels of environmental sources of vitamin D (intake and UV radiation) are shown to affect 25(OH)D concentrations differently in those with low versus high genetic risk, reiterating the importance of well measured environmental factors in genetic analyses, as well as the importance of considering genetic risk when making recommendations on vitamin D intake. Moreover, genetic information can be utilized to inform personalized dosing of vitamin D to best achieve optimal 25(OH)D concentrations.

Statement of authors' contributions to manuscript.

K.H. and C.E. designed research; K.H. conducted research; K.H. analyzed data; Q.L. provided statistical guidance; A.W. provided MESA vitamin D intake data; J.M. provided nutritional insight; and K.H. wrote the paper. K.H. and C.E. had primary responsibility for final content. All authors read and approved the final manuscript.

Tables

Table 5.1: Sample characteristics

Cohort	Variable	European ancestry	African ancestry
ARIC	Sample size	6,178	57
	Age (SE) [years]	57.1 (5.7)	55.6 (6.2)
	% Female	54	49.1%
	BMI (SE) [kg/m^2]	27.3 (4.9)	28.6 (5.7)
	UV[1] (SE) [units]	5.1 (2.5)	7.1 (2.4)
	Intake[2] (SE) [IU]	223.3 (145.7)	221.2 (137.3)
	25(OH)D (SE) [ng/ml]	26.0 (8.8)	20.9 (7.8)
MESA	Sample size	1,936	342
	Age (SE) [years]	62.7 (10.3)	62.3 (10.4)
	% Female	53	51
	BMI (SE) [kg/m^2]	27.8 (5.0)	30.0 (5.9)
	UV (SE) [units]	4.5 (2.3)	5.1 (2.2)
	Intake (SE) [IU]	188.9 (157.2)	161.8 (144.1)
	25(OH)D (SE) [ng/ml]	30.1 (10.8)	19.5 (8.9)
WHI	Sample size	455	700
	Age (SE) [years]	66.6 (6.8)	61.8 (7.4)
	% Female	100	100
	BMI (SE) [kg/m^2]	29.9 (6.3)	31.1 (6.4)
	UV (SE) [units]	5.2 (2.5)	5.5 (2.6)
	Intake (SE) [IU]	192.3 (143.2)	146.4 (130.5)
	25(OH)D (SE) [ng/ml]	18.9 (10.7)	19.0 (15.4)

[1] available UV radiation
[2] vitamin D intake from diet
[3] MANOVA global test (performed in SAS (version 9.4) revealed differences in one or more variables by cohort, therefore cohort was adjusted for in all models that included multiple cohorts

Original to this manuscript.

Figure 5.1: Interaction test results from 1-DF and 2-DF models

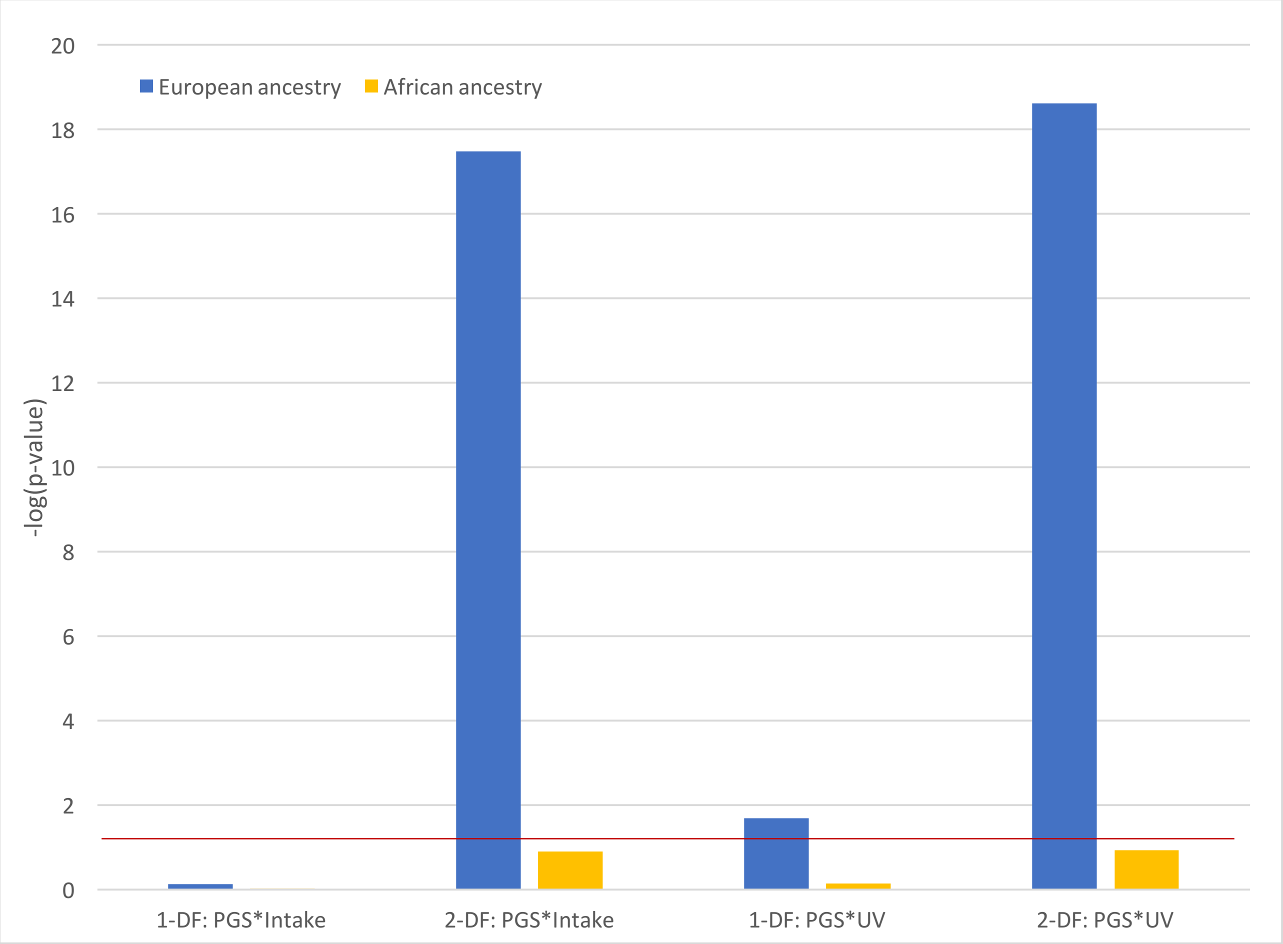
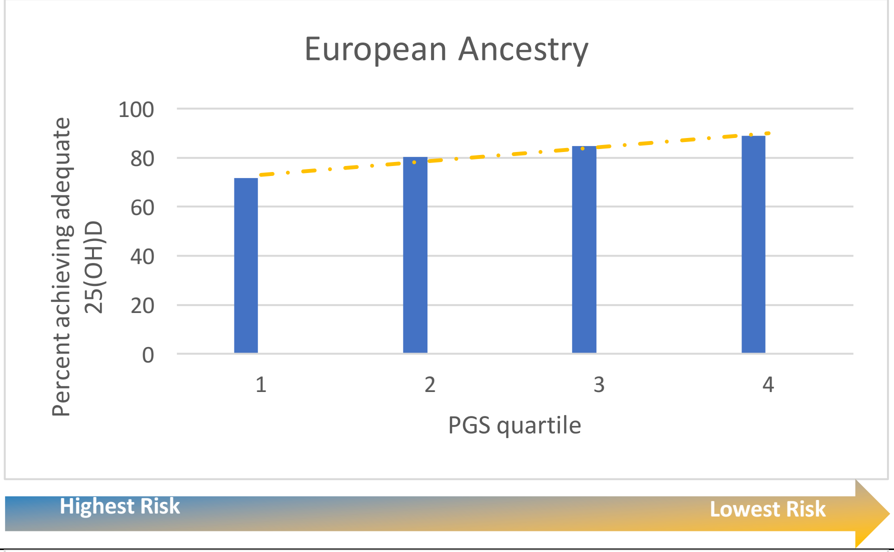


Figure 5.1 shows –log(p-values) for the 1-DF and 2-DF models of the PGS interaction; all models controlled for age, sex, BMI, cohort, vitamin D intake and available UV radiation. The red line denotes the p=0.05 significance cutoff. The 2-DF PGS*intake, 1-DF PGS*UV and 2-DF PGS*UV models were statistically significant in participants of European ancestry (p=3.3x10-18, 2.1x10-2 and 2.4x10-19, respectively). Original to this manuscript.

Figure 5.2: Percent achieving adequate 25(OH)D in those reaching IOM vitamin D intake guidelines by genetic risk

European-ancestry participants with supplement use data and who reached IOM vitamin D intake guidelines (n=184)



African-ancestry participants with supplement use data and who reached IOM vitamin D intake guidelines (n=17)

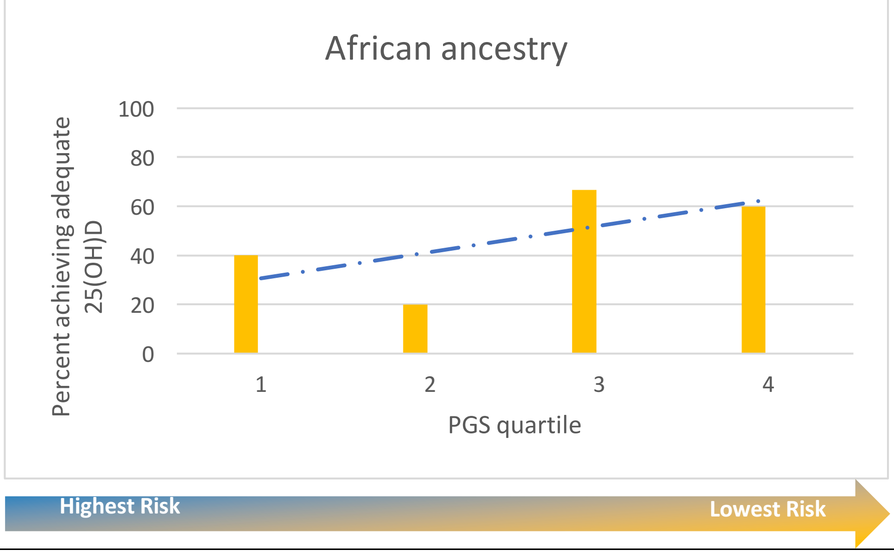


Figure 5.2 shows the percent of European- or African-ancestry participants reaching adequate 25(OH)D (20 ng/ml) by quartile of genetic risk. In those of European ancestry, as genetic risk decreased (higher PGS), those reaching optimal vitamin D concentrations increased. The difference in percent reaching adequate 25(OH)D between the two extreme quartiles was 17.3%; 71.7% of participants with the highest genetic risk and 89% of participants with the lowest risk reached adequate 25(OH)D. This is a statistically significant (p=0.018) and clinically meaningful difference. The trend was not significant in those of African ancestry. Original to this manuscript.

Chapter 6: Conclusion

Summary

Vitamin D inadequacy, as defined by a 25-hydroxyvitamin D [25(OH)D] concentration less than 20 ng/mL, affects about 50% of adults in the United States. Low vitamin D concentrations have been associated with increased risk of autoimmune diseases, hypertension, dyslipidemia, cardiovascular events, and cardiovascular mortality (1, 3-5). Additionally, recent Mendelian randomization studies have suggested a causal relationship between low vitamin D concentrations and increased risk of obesity, ovarian cancer, hypertension, lower cognitive function, multiple sclerosis, and all cause and cancer mortality (10-16). Furthermore, some clinical trials have shown that vitamin D and calcium supplementation are important in the prevention of fractures and cardiovascular risk factors, while vitamin D supplementation alone may lower risk of cancers, diabetes and depression, and may reduce inflammation and improve lung function in patients with cystic fibrosis (7, 17-19, 65-69). Recent results reported from the VITAL trial showed null associations between vitamin D supplementation and cancer or cardiovascular disease, however, study design limits the interpretability of these findings (70).

Avoiding vitamin D inadequacy is important, however, vitamin D concentrations over 50 ng/mL have been associated with increased morbidity and mortality (3). Clinical trials of vitamin D have shown that individual response to vitamin D supplementation is highly variable (72). Vitamin D concentrations are influenced by genetic factors and genetic variants may determine how much vitamin D intake is required to reach an optimal vitamin D blood concentration (21, 22, 73, 74). Therefore, knowledge of the genetic determinants of vitamin D concentration is invaluable in prevention of vitamin D associated morbidity and mortality. Several genome-wide association studies (GWASs) have uncovered vitamin D associated single nucleotide polymorphisms (SNPs) (1, 28, 31). However, these SNPs account for a small portion of the variation in vitamin D concentrations (1). Understanding the complete set of genetic factors that contribute to vitamin D concentrations and how they function together and with environmental factors may allow for personalized treatment of vitamin D inadequacy, instead of the current (and ineffective) one size fits all treatment regimen. The overall motivation for this research was to start to fill in the knowledge gaps surrounding the genetic landscape of 25(OH)D, such as missing heritability and missing minority group studies. The long-term goal is promoting adequate vitamin D concentrations through personalized vitamin D supplementation based on an individual's genetic makeup and non-genetic characteristics. To accomplish this, my goals were 3-fold: first, to capture more genetic variance in 25(OH)D concentration through development of a polygenic score (PGS); second, to calculate heritability for all genotyped SNPS, the PGS and previously reported vitamin D SNPs in a group of unrelated participants and third, to investigate the interaction between the PGS and environmental determinants of 25(OH)D, vitamin D intake and available UV radiation, all in an ancestry-specific manner for participants of African and European ancestry.

In Aim 1, a PGS was calculated in an ancestry-specific manner. Ancestry-specific PGSs were weighted by effect sizes from TRANSCEN-D, a multi-ethnic 25(OH)D GWAS, and were calculated in independent samples (27). Results showed that those with greater genetic risk have lower 25(OH)D concentrations, such that when compared to those with lowest genetic risk, those with highest genetic risk could require an additional 317 to 783 IU of vitamin D to maintain adequate 25(OH)D levels. In the European-ancestry determination and validation cohorts, the optimal PGS explained 1.3-2.1% of the variance in 25-hydroxyvitamin D [25(OH)D] concentrations, while the fully adjusted model explained 8-14% of the variance of the phenotype. In the African-ancestry determination and validation cohorts, the PGS explained 0.01-4.4% of the variance in 25(OH)D, and the fully adjusted model explained 2.3-37% of the variance. The variance explained by the PGS, while in line with what has been reported for other complex traits, captures only a modest portion of phenotypic variance. Stratified analysis showed that the PGS and model explained more phenotypic variance in those of African ancestry, women and those with higher levels of available UV radiation. The consistent association between the PGS and 25(OH)D concentrations indicates this PGS has the potential to predict risk of vitamin D inadequacy.

In Aim 2, SNP heritability was calculated in a sample of unrelated participants; SNP heritability estimates for the PGS and results from previous GWASs were also calculated. Heritability estimates made in samples of related participants, tend to overestimate heritability, due to shared environment, therefore estimates here were calculated in a sample of unrelated participants. Estimates were calculated by ancestry as well as stratified by gender and available UV radiation. In general, SNP heritability estimates were higher in those of African ancestry and in those with more available UV radiation. SNP heritability estimates for 25(OH)D in unrelated participants of European ancestry are on the low end of the range of estimates previously reported in related individuals, and the PGS from Aim 1 explains more heritability than do previous GWAS findings (17% vs 1.6-6.9%, respectively). Findings from previous GWAS only explain a small portion of 25(OH)D heritability in those of African and European ancestries.

While the PGS from those of European ancestry in Aim 1 accounts for a larger portion of the total SNP heritability than do previous GWAS findings, a large portion of the heritability remains unexplained, promoting the need for further investigation into the genetic underpinnings of 25(OH)D. Further studies could include both common and low frequency variants to account for more heritability (44).

Aim 3 investigated ancestry-specific (European and African) gene-environment interactions between the PGS and vitamin D intake and available UV radiation using 1-DF and 2-DF tests. The 2-DF PGS*intake, 1-DF PGS*UV and 2-DF PGS*UV models were statistically significant in participants of European ancestry (p=$3.3x10^{-18}$, $2.1x10^{-2}$ and $2.4x10^{-19}$, respectively). In African-ancestry analyses power was limited due to the smaller sample size, and no statistically significant interactions were discerned. Consistent with a previous study, interaction results indicate that as available UV or vitamin D intake increases, so does the difference in 25(OH)D between those of lowest and highest risk. This implies that those with high genetic risk may require more vitamin D intake to reach and maintain optimal 25(OH)D concentrations. This trend proved true for those of European ancestry when looking at participants who reached IOM dietary guidelines for vitamin D intake. Fewer participants with high genetic risk reached optimal 25(OH)D concentrations (≥20 ng/ml); 70.4% of those with highest risk compared to 83.8% of those with lowest risk (p=$4.1x10^{-11}$). The trend did not hold in those of African ancestry, likely due to small sample size. Overall, the results here reiterate the importance of well measured environmental factors in genetic analyses, as well as the importance of considering genetic risk when making recommendations on vitamin D intake.

Moreover, genetic information could be utilized to inform personalized dosing of vitamin D to best achieve optimal 25(OH)D concentrations.

Vitamin D concentration is a complex trait with strong underlying genetic and environmental predictors that work together in a complex way to influence 25(OH)D concentration. Here, a PGS was shown to be a useful metric for quantifying risk of vitamin D inadequacy, which could inform personalized vitamin D dosing to achieve and maintain an adequate 25(OH)D concentration. Additionally, the PGS explained more SNP heritability than did previous GWAS findings, giving insight into the complex genetic underpinnings of 25(OH)D. While the PGS explains more SNP heritability than do prior GWAS findings, a large portion remains unexplained. A portion of this missing heritability is likely due to interaction. Therefore, an investigation into genetic-by-environmental interactions was undertaken. Here, there was evidence of interactions between the PGS and both vitamin D intake and available UV radiation. While this research contributes new facets to understanding the complexity of 25(OH)D, much of the heritability remains unexplained, warranting further research, such as incorporating whole genome sequence data which captures rare variants.

Strengths

This research has many strengths which increase the value of its contribution to the existing literature. This study exhausted nearly all of the publicly available data with vitamin D and genetic information. This resulted in a large sample size for European ancestry, which is useful in maintaining adequate power and for improving generalizability of the findings. In addition to genetic data, this research has very rich environmental measures across cohorts, i.e.

for dietary intake and available UV radiation. Finally, this research used a multi-ethnic cohort with ancestrally appropriate analyses. This is increasingly important in genetic research given the disparate research done in minority groups compared to those of European ancestry. The majority of genetic research to date is done in populations of European ancestry where findings do not necessarily translate to other ancestries, which only perpetuates health disparities.

Limitations

Of course, this research comes with limitations as well. To maintain independence from TRANSCEN-D, which provided ancestry-specific weights for the PGSs, the sample size used in this analysis for the African-ancestry cohort was relatively small. The sample size issues experienced for the African-ancestry cohort emphasize the importance of obtaining more diverse samples (i.e. in initiatives like All of Us). Through the TRANSCEN-D GWAS meta-analysis and the analysis here, nearly all of the publicly available African-ancestry samples with relevant data have been exhausted and sample sizes for other racial/ethnic groups remain limited. Additionally, while the set of SNPs used to create the PGS was over 9 million for African ancestry and over 8 million for European ancestry, it did not include rare variants which could have limited the heritability that the PGS would explain. Finally, while this study had rich environmental data, the variables were still imperfect. Using available UV radiation is much more precise than ‘season’ (which has been used in past gene-environment interaction analyses), however, is not as good as the gold standard, ‘actual UV radiation’. Additionally, while all studies included vitamin D dietary intake data, only WHI had vitamin D supplement use data which limited utility of some analyses.

Future research

Future research stemming from this dissertation could include performing a SCAN test of SNPs found to be related to vitamin D, i.e. included in PGS or implied in GWAS studies, to discern if there is a clustering of SNPs in a protein space. Another project related to GWAS could be a conditional analysis, where given candidate genotypes are conditioned on to see if signal is removed, suggesting the variant could be functional. A third project could include using a Mendelian Randomization approach to test the association between the vitamin D PGS and other downstream outcomes linked to vitamin D concentrations, such as multiple sclerosis or colorectal cancer. Finally, as was reported on by the Kardia group from the University of Michigan, to maximize the potential of the data and to preserve sample size, PGS could be created on the full sample, skipping the training and testing steps, and using all genotyped (not imputed) SNPs along with corresponding betas from TRANSCEN-D (107).

Additionally, after replication in a larger cohort that includes minority participants and rare variants, translation of the findings could be investigated with a clinical trial. As evidenced by data reported here, the IOM recommendation for daily vitamin D intake is not enough for those with high genetic risk for low 25(OH)D to maintain adequate 25(OH)D. A clinical trial would determine if personalized supplementation can counter genetic predisposition to low 25(OH)D. I would propose a randomized clinical trial (RCT), in participants with GWAS data and inadequate 25(OH)D at baseline. Participants would either be in the control arm which receives the IOM recommendation or in the treatment arm which receives a personalized dose based on their ancestrally appropriate PGS. Participants would maintain their dosing for 6-months. At 6-months, if personalized supplementation counters genetic predisposition to low 25(OH)D, more participants with high genetic risk in treatment arm should have adequate 25(OHD compared to the control arm. This could prompt changes in IOM recommended dose based on genetic predisposition.

References

1. Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, Berry D, et al. Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet. 2010;376(9736):180-8. doi: 10.1016/S0140-6736(10)60588-0. PubMed PMID: PMC3086761.
2. Forrest KYZ, Stuhldreher WL. Prevalence and correlates of vitamin D deficiency in US adults. Nutrition Research.31(1):48-54. doi: 10.1016/j.nutres.2010.12.001.
3. Medicine Io. Dietary Reference Intakes for Calcium and Vitamin D. Ross AC, Taylor CL, Yaktine AL, Del Valle HB, editors. Washington, DC: The National Academies Press; 2011. 1132 p.
4. Holick MF. High Prevalence of Vitamin D Inadequacy and Implications for Health. Mayo Clinic Proceedings.81(3):353-73. doi: 10.4065/81.3.353.
5. Holick MF. Vitamin D Deficiency. New England Journal of Medicine. 2007;357(3):266-81. doi: 10.1056/NEJMra070553.
6. Mirhosseini N, Vatanparast H, Kimball SM. The Association between Serum 25(OH)D Status and Blood Pressure in Participants of a Community-Based Program Taking Vitamin D Supplements. Nutrients. 2017;9(11):1244. doi: 10.3390/nu9111244. PubMed PMID: PMC5707716.
7. Arshi S, Fallahpour M, Nabavi M, Bemanian MH, Javad-Mousavi SA, Nojomi M, et al. The effects of vitamin D supplementation on airway functions in mild to moderate persistent asthma. Annals of Allergy, Asthma & Immunology. 2014;113(4):404-9. doi: https://doi.org/10.1016/j.anai.2014.07.005.
8. Song T-J, Chu M-K, Sohn J-H, Ahn H-Y, Lee SH, Cho S-J. Effect of Vitamin D Deficiency on the Frequency of Headaches in Migraine. Journal of Clinical Neurology (Seoul, Korea). 2018;14(3):366-73. doi: 10.3988/jcn.2018.14.3.366. PubMed PMID: PMC6031995.
9. Kheiri B, Abdalla A, Osman M, Ahmed S, Hassan M, Bachuwa G. Vitamin D deficiency and risk of cardiovascular diseases: a narrative review. Clinical Hypertension. 2018;24:9. doi: 10.1186/s40885-018-0094-4. PubMed PMID: PMC6013996.
10. Kunutsor SK, Burgess S, Munroe PB, Khan H. Vitamin D and high blood pressure: causal association or epiphenomenon? European Journal of Epidemiology. 2014;29(1):1-14. doi: 10.1007/s10654-013-9874-z.
11. Afzal S, Brøndum-Jacobsen P, Bojesen SE, Nordestgaard BG. Genetically low vitamin D concentrations and increased mortality: mendelian randomisation analysis in three large cohorts. BMJ. 2014;349. doi: 10.1136/bmj.g6330.
12. Mokry LE, Ross S, Ahmad OS, Forgetta V, Smith GD, Leong A, et al. Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. PLoS Med. 2015;12(8):e1001866. doi: 10.1371/journal.pmed.1001866.
13. Ong J-S, Cuellar-Partida G, Lu Y, Fasching PA, Hein A, Burghaus S, et al. Association of vitamin D levels and risk of ovarian cancer: a Mendelian randomization study. International Journal of Epidemiology. 2016;45(5):1619-30. doi: 10.1093/ije/dyw207.
14. Vimaleswaran KS, Cavadino A, Berry DJ, LifeLines Cohort Study i, Jorde R, Dieffenbach AK, et al. Association of vitamin D status with arterial blood pressure and hypertension risk: a mendelian randomisation study. The lancet Diabetes & endocrinology. 2014;2(9):719-29. doi: 10.1016/S2213-8587(14)70113-5. PubMed PMID: PMC4582411.

15. Kueider AM, Tanaka T, An Y, Kitner-Triolo MH, Palchamy E, Ferrucci L, et al. State- and trait-dependent associations of vitamin-D with brain function during aging. Neurobiology of Aging. 2016;39:38-45. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2015.11.002.
16. Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, Hiraki LT, et al. Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. PLoS Medicine. 2013;10(2):e1001383. doi: 10.1371/journal.pmed.1001383. PubMed PMID: PMC3564800.
17. Weaver CM, Alexander DD, Boushey CJ, Dawson-Hughes B, Lappe JM, LeBoff MS, et al. Calcium plus vitamin D supplementation and risk of fractures: an updated meta-analysis from the National Osteoporosis Foundation. Osteoporosis International. 2016;27(1):367-76. doi: 10.1007/s00198-015-3386-5.
18. Bischoff-Ferrari HA, Willett WC, Wong JB, Giovannucci E, Dietrich T, Dawson-Hughes B. Fracture prevention with vitamin d supplementation: A meta-analysis of randomized controlled trials. JAMA. 2005;293(18):2257-64. doi: 10.1001/jama.293.18.2257.
19. Boonen S, Lips P, Bouillon R, Bischoff-Ferrari HA, Vanderschueren D, Haentjens P. Need for Additional Calcium to Reduce the Risk of Hip Fracture with Vitamin D Supplementation: Evidence from a Comparative Metaanalysis of Randomized Controlled Trials. The Journal of Clinical Endocrinology & Metabolism. 2007;92(4):1415-23. doi: 10.1210/jc.2006-1404.
20. Schnatz PF, Jiang X, Aragaki AK, Nudy M, O'Sullivan DM, Williams M, et al. Effects of Calcium, Vitamin D, and Hormone Therapy on Cardiovascular Disease Risk Factors in the Women's Health Initiative: A Randomized Controlled Trial. Obstetrics and gynecology. 2017;129(1):121-9. doi: 10.1097/AOG.0000000000001774. PubMed PMID: PMC5177479.
21. Engelman CD, Meyers KJ, Iyengar SK, Liu Z, Karki CK, Igo RP, et al. Vitamin D Intake and Season Modify the Effects of the GC and CYP2R1 Genes on 25-Hydroxyvitamin D Concentrations. The Journal of Nutrition. 2013;143(1):17-26. doi: 10.3945/jn.112.169482. PubMed PMID: PMC3521459.
22. Engelman CD, Fingerlin TE, Langefeld CD, Hicks PJ, Rich SS, Wagenknecht LE, et al. Genetic and Environmental Determinants of 25-Hydroxyvitamin D and 1,25-Dihydroxyvitamin D Levels in Hispanic and African Americans. The Journal of Clinical Endocrinology and Metabolism. 2008;93(9):3381-8. doi: 10.1210/jc.2007-2702. PubMed PMID: PMC2567851.
23. Snellman G, Melhus H, Gedeborg R, Olofsson S, Wolk A, Pedersen NL, et al. Seasonal Genetic Influence on Serum 25-Hydroxyvitamin D Levels: A Twin Study. PLOS ONE. 2009;4(11):e7747. doi: 10.1371/journal.pone.0007747.
24. Karohl C, Su S, Kumari M, Tangpricha V, Veledar E, Vaccarino V, et al. Heritability and seasonal variability of vitamin D concentrations in male twins. The American Journal of Clinical Nutrition. 2010;92(6):1393-8. doi: 10.3945/ajcn.2010.30176. PubMed PMID: PMC2980965.
25. Jiang X, O'Reilly PF, Aschard H, Hsu Y-H, Richards JB, Dupuis J, et al. Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. Nature Communications. 2018;9(1):260. doi: 10.1038/s41467-017-02662-2.
26. Arguelles LM, Langman CB, Ariza AJ, Ali FN, Dilley K, Price H, et al. Heritability and Environmental Factors Affecting Vitamin D Status in Rural Chinese Adolescent Twins. The Journal of Clinical Endocrinology & Metabolism. 2009;94(9):3273-81. doi: 10.1210/jc.2008-1532.

27. Hong J, Hatchell KE, Bradfield JP, Andrew B, Alessandra C, Chao-Qiang L, et al. Trans-ethnic Evaluation Identifies Novel Low Frequency Loci Associated with 25-Hydroxyvitamin D Concentrations. The Journal of Clinical Endocrinology & Metabolism. 2018:jc.2017-01802-jc.2017-. doi: 10.1210/jc.2017-01802.

28. Ahn J, Yu K, Stolzenberg-Solomon R, Simon KC, McCullough ML, Gallicchio L, et al. Genome-wide association study of circulating vitamin D levels. Human Molecular Genetics. 2010;19(13):2739-45. doi: 10.1093/hmg/ddq155. PubMed PMID: PMC2883344.

29. Wang X, Zhu H, Snieder H, Su S, Munn D, Harshfield G, et al. Obesity related methylation changes in DNA of peripheral blood leukocytes. BMC Med. 2010;8. doi: 10.1186/1741-7015-8-87.

30. Popejoy ABaF, Stephanie M. Genomics is failing on diversity. Nature. 2016(538):161-4. doi: doi:10.1038/538161a.

31. Engelman CD, Meyers KJ, Ziegler JT, Taylor KD, Palmer ND, Haffner SM, et al. Genome-wide association study of vitamin D concentrations in Hispanic Americans: The IRAS Family Study. The Journal of steroid biochemistry and molecular biology. 2010;122(4):186-92. doi: 10.1016/j.jsbmb.2010.06.013. PubMed PMID: PMC2949505.

32. Mormino EC, Sperling RA, Holmes AJ, Buckner RL, De Jager PL, Smoller JW, et al. Polygenic risk of Alzheimer disease is associated with early- and late-life processes. Neurology. 2016. doi: 10.1212/wnl.0000000000002922.

33. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. PLOS Genetics. 2013;9(3):e1003348. doi: 10.1371/journal.pgen.1003348.

34. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Current Opinion in Genetics & Development. 2015;33:10-6. doi: https://doi.org/10.1016/j.gde.2015.06.005.

35. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of heritability for human height. Nature genetics. 2010;42(7):565-9. doi: 10.1038/ng.608. PubMed PMID: PMC3232052.

36. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Medicine. 2017;14(3):e1002258. doi: 10.1371/journal.pmed.1002258. PubMed PMID: PMC5360219.

37. Tang W, Schwienbacher C, Lopez Lorna M, Ben-Shlomo Y, Oudot-Mellakh T, Johnson Andrew D, et al. Genetic Associations for Activated Partial Thromboplastin Time and Prothrombin Time, their Gene Expression Profiles, and Risk of Coronary Artery Disease. American Journal of Human Genetics. 2012;91(1):152-62. doi: 10.1016/j.ajhg.2012.05.009. PubMed PMID: PMC3397273.

38. Slominski A, Postlethwaite AE. Skin Under the Sun: When Melanin Pigment Meets Vitamin D. Endocrinology. 2015;156(1):1-4. doi: 10.1210/en.2014-1918. PubMed PMID: PMC4272394.

39. Shoenfeld N, Amital H, Shoenfeld Y. The effect of melanism and vitamin D synthesis on the incidence of autoimmune disease. Nat Clin Pract Rheum. 2009;5(2):99-105.

40. Wray NaV, P. Estimating trait heritability. Nature Education. 2008;1((1)):29.

41. Rogers AR. Quantitative characters II: heritability. Anth/Biol 5221: Human Evolutionary Genetics: University of Utah; 2017.

42. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29-. doi: 10.1186/1746-4811-9-29. PubMed PMID: PMC3750305.

43. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics. 2011;88(1):76-82. doi: 10.1016/j.ajhg.2010.11.011. PubMed PMID: PMC3014363.

44. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015;47(10):1114-20. doi: 10.1038/ng.3390 http://www.nature.com/ng/journal/v47/n10/abs/ng.3390.html - supplementary-information.

45. Shea MK, Benjamin EJ, Dupuis J, Massaro JM, Jacques PF, D'Agostino RB, Sr., et al. Genetic and non-genetic correlates of vitamins K and D. European journal of clinical nutrition. 2009;63(4):458-64. Epub 11/21. doi: 10.1038/sj.ejcn.1602959. PubMed PMID: 18030310.

46. Millen AE, Wactawski-Wende J, Pettinger M, Melamed ML, Tylavsky FA, Liu S, et al. Predictors of serum 25-hydroxyvitamin D concentrations among postmenopausal women: the Women's Health Initiative Calcium plus Vitamin D clinical trial. The American journal of clinical nutrition. 2010;91(5):1324-35. Epub 03/10. doi: 10.3945/ajcn.2009.28908. PubMed PMID: 20219959.

47. Sinotte M, Diorio C, Bérubé S, Pollak M, Brisson J. Genetic polymorphisms of the vitamin D binding protein and plasma concentrations of 25-hydroxyvitamin D in premenopausal women. The American Journal of Clinical Nutrition. 2009;89(2):634-40. doi: 10.3945/ajcn.2008.26445.

48. Nissen J, Rasmussen LB, Ravn-Haren G, Andersen EW, Hansen B, Andersen R, et al. Common Variants in CYP2R1 and GC Genes Predict Vitamin D Concentrations in Healthy Danish Children and Adults. PLOS ONE. 2014;9(2):e89907. doi: 10.1371/journal.pone.0089907.

49. Binkley N, Lappe J, Singh RJ, Khosla S, Krueger D, Drezner MK, et al. Can vitamin D metabolite measurements facilitate a “treat-to-target” paradigm to guide vitamin D supplementation? Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA. 2015;26(5):1655-60. doi: 10.1007/s00198-014-3010-0. PubMed PMID: PMC4412341.

50. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11. doi: 10.1038/nrg2809.

51. Lee Sang H, Wray Naomi R, Goddard Michael E, Visscher Peter M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. American Journal of Human Genetics. 2011;88(3):294-305. doi: 10.1016/j.ajhg.2011.02.002. PubMed PMID: PMC3059431.

52. Duncan LE, Ratanatharathorn A, Aiello AE, Almli LM, Amstadter AB, Ashley-Koch AE, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Molecular Psychiatry. 2017;23:666. doi: 10.1038/mp.2017.77 https://www.nature.com/articles/mp201777 - supplementary-information.

53. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean Journal of Anesthesiology. 2017;70(4):407-11. doi: 10.4097/kjae.2017.70.4.407. PubMed PMID: PMC5548942.

54. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73. doi: 10.1038/nprot.2010.116. PubMed PMID: PMC3025522.

55. Thompson EA.
Pedigrees and Relationships. Stat550: University of Washington; 2005.

56. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7. doi: 10.1038/ng.3656
http://www.nature.com/ng/journal/v48/n10/abs/ng.3656.html - supplementary-information.

57. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016;48(10):1279-83. doi: 10.1038/ng.3643. PubMed PMID: PMC5388176.

58. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016;48(11):1443-8. doi: 10.1038/ng.3679. PubMed PMID: PMC5096458.

59. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015;31(9):1466-8. doi: 10.1093/bioinformatics/btu848. PubMed PMID: PMC4410663.

60. Fishman GaM, LR. A statistical evaluation of multiplicative congruential generators with modulus. J Am Stat Assoc. 1982(77):129-36.

61. Gondro C, Werf Jvd, Hayes B. Genome-wide association studies and genomic prediction. 2013.

62. Visscher PM, Hemani G, Vinkhuyzen AAE, Chen G-B, Lee SH, Wray NR, et al. Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. PLoS Genetics. 2014;10(4):e1004269. doi: 10.1371/journal.pgen.1004269. PubMed PMID: PMC3983037.

63. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Alzheimer's Disease Genetics C. Alzheimer's Disease: Analyzing the Missing Heritability. PLoS ONE. 2013;8(11):e79771. doi: 10.1371/journal.pone.0079771. PubMed PMID: PMC3820606.

64. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392-406. doi: 10.1038/nrg.2016.27
http://www.nature.com/nrg/journal/v17/n7/abs/nrg.2016.27.html - supplementary-information.

65. Lappe JM, Travers-Gustafson D, Davies KM, Recker RR, Heaney RP. Vitamin D and calcium supplementation reduces cancer risk: results of a randomized trial. The American Journal of Clinical Nutrition. 2007;85(6):1586-91.

66. Jorde R, Sneve M, Figenschau Y, Svartberg J, Waterloo K. Effects of vitamin D supplementation on symptoms of depression in overweight and obese subjects: randomized double blind trial. Journal of Internal Medicine. 2008;264(6):599-609. doi: 10.1111/j.1365-2796.2008.02008.x.

67. Bertone-Johnson ER, Powers SI, Spangler L, Brunner RL, Michael YL, Larson JC, et al. Vitamin D intake from foods and supplements and depressive symptoms in a diverse population of older women. The American Journal of Clinical Nutrition. 2011;94(4):1104-12. doi: 10.3945/ajcn.111.017384. PubMed PMID: PMC3173027.

68. Pincikova T, Paquin-Proulx D, Sandberg JK, Flodström-Tullberg M, Hjelte L. Clinical impact of vitamin D treatment in cystic fibrosis: a pilot randomized, controlled trial. European Journal Of Clinical Nutrition. 2016;71:203. doi: 10.1038/ejcn.2016.259 https://www.nature.com/articles/ejcn2016259 - supplementary-information.

69. Barry EL, Peacock JL, Rees JR, Bostick RM, Robertson DJ, Bresalier RS, et al. Vitamin D Receptor Genotype, Vitamin D(3) Supplementation, and Risk of Colorectal Adenomas: A Randomized Clinical Trial. JAMA oncology. 2017;3(5):628-35. doi: 10.1001/jamaoncol.2016.5917. PubMed PMID: PMC5580351.

70. Manson JE, Cook NR, Lee IM, Christen W, Bassuk SS, Mora S, et al. Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. New England Journal of Medicine. 2018. doi: 10.1056/NEJMoa1809944.

71. Melamed ML, Manson JE. Vitamin D and cardiovascular disease and cancer: not too much and not too little? The need for clinical trials. Women's health (London, England). 2011;7(4):419-24. doi: 10.2217/whe.11.18. PubMed PMID: PMC4378570.

72. Aloia JF, Patel M, DiMaano R, Li-Ng M, Talwar SA, Mikhail M, et al. Vitamin D intake to attain a desired serum 25-hydroxyvitamin D concentration. The American Journal of Clinical Nutrition. 2008;87(6):1952-8.

73. Nimitphong H, Saetung S, Chanprasertyotin S, Chailurkit L-o, Ongphiphadhanakul B. Changes in circulating 25-hydroxyvitamin D according to vitamin D binding protein genotypes after vitamin D3 or D2supplementation. Nutrition Journal. 2013;12(1):39. doi: 10.1186/1475-2891-12-39.

74. Wjst M. Linking vitamin D, the microbiome and allergy. Allergy. 2017;72(3):329-30. doi: 10.1111/all.13116.

75. Shao B, Jiang S, Muyiduli X, Wang S, Mo M, Li M, et al. Vitamin D pathway gene polymorphisms influenced vitamin D level among pregnant women. Clinical Nutrition. doi: 10.1016/j.clnu.2017.10.024.

76. Zhang Z, He J-W, Fu W-Z, Zhang C-Q, Zhang Z-L. An analysis of the association between the vitamin D pathway and serum 25-hydroxyvitamin D levels in a healthy Chinese population. Journal of Bone and Mineral Research. 2013;28(8):1784-92. doi: 10.1002/jbmr.1926.

77. Chandler PD, Tobias DK, Wang L, Smith-Warner SA, Chasman DI, Rose L, et al. Association between Vitamin D Genetic Risk Score and Cancer Risk in a Large Cohort of U.S. Women. Nutrients. 2018;10(1):55. doi: 10.3390/nu10010055. PubMed PMID: PMC5793283.

78. Mazahery H, von Hurst PR. Factors Affecting 25-Hydroxyvitamin D Concentration in Response to Vitamin D Supplementation. Nutrients. 2015;7(7):5111-42. doi: 10.3390/nu7075111. PubMed PMID: PMC4516990.

79. Didriksen A, Grimnes G, Hutchinson M, Kjærgaard M, Svartberg J, Joakimsen R, et al. The serum 25-hydroxyvitamin D response to vitamin D supplementation is related to genetic factors, BMI, and Baseline levels2013.

80. Anderson GL, Manson J, Wallace R, Lund B, Hall D, Davis S, et al. Implementation of the women's health initiative study design. Annals of Epidemiology. 2003;13(9, Supplement):S5-S17. doi: https://doi.org/10.1016/S1047-2797(03)00043-7.

81. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, et al. The Gene, Environment Association Studies Consortium (GENEVA): Maximizing the Knowledge Obtained

from GWAS by Collaboration Across Studies of Multiple Conditions. Genetic epidemiology. 2010;34(4):364-72. doi: 10.1002/gepi.20492. PubMed PMID: PMC2860056.

82. Musunuru K, Lettre G, Young T, Farlow DN, Pirruccello JP, Ejebe KG, et al. Candidate Gene Association Resource (CARe): Design, Methods, and Proof of Concept. Circulation Cardiovascular genetics. 2010;3(3):267-75. doi: 10.1161/CIRCGENETICS.109.882696. PubMed PMID: PMC3048024.

83. Manichaikul A, Palmas W, Rodriguez CJ, Peralta CA, Divers J, Guo X, et al. Population Structure of Hispanics in the United States: The Multi-Ethnic Study of Atherosclerosis. PLoS Genetics. 2012;8(4):e1002640. doi: 10.1371/journal.pgen.1002640. PubMed PMID: PMC3325201.

84. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. American Journal of Epidemiology. 2011;174(7):849-59. doi: 10.1093/aje/kwr160. PubMed PMID: PMC3176830.

85. Anderson G. WHI Harmonized and Imputed GWAS Data [cited 2018 April 3]. dbGaP Study Accession: phs000746.v2.p3:[

86. Johnston HR, Hu Y-J, Gao J, O'Connor TD, Abecasis GR, Wojcik GL, et al. Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. Scientific Reports. 2017;7:46398. doi: 10.1038/srep46398
https://www.nature.com/articles/srep46398 - supplementary-information.

87. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. doi: 10.1093/bioinformatics/btr330.

88. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75. Epub 07/25. doi: 10.1086/519795. PubMed PMID: 17701901.

89. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics. 2016;48:481. doi: 10.1038/ng.3538
https://www.nature.com/articles/ng.3538 - supplementary-information.

90. Coram MA, Fang H, Candille SI, Assimes TL, Tang H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. American Journal of Human Genetics. 2017;101(2):218-26. doi: 10.1016/j.ajhg.2017.06.015. PubMed PMID: PMC5544393.

91. Vassos E, Di Forti M, Coleman J, Iyegbe C, Prata D, Euesden J, et al. An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. Biological Psychiatry. 2017;81(6):470-7. doi: 10.1016/j.biopsych.2016.06.028.

92. Márquez-Luna C, Loh P-R, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genetic Epidemiology. 2017;41(8):811-23. doi: 10.1002/gepi.22083.

93. Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DCM, Ritchie SJ, et al. Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. Molecular Psychiatry. 2016;21:1624. doi: 10.1038/mp.2015.225
https://www.nature.com/articles/mp2015225 - supplementary-information.

94. So H-C, Sham PC. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Scientific Reports. 2017;7:41262. doi: 10.1038/srep41262 https://www.nature.com/articles/srep41262 - supplementary-information.

95. Hansen JG, Tang W, Hootman KC, Brannon PM, Houston DK, Kritchevsky SB, et al. Genetic and Environmental Factors Are Associated with Serum 25-Hydroxyvitamin D Concentrations in Older African Americans. The Journal of Nutrition. 2015;145(4):799-805. doi: 10.3945/jn.114.202093. PubMed PMID: PMC4381765.

96. Signorello LB, Shi J, Cai Q, Zheng W, Williams SM, Long J, et al. Common Variation in Vitamin D Pathway Genes Predicts Circulating 25-Hydroxyvitamin D Levels among African Americans. PLoS ONE. 2011;6(12):e28623. doi: 10.1371/journal.pone.0028623. PubMed PMID: PMC3244405.

97. Murphy AB, Kelley B, Nyame YA, Martin IK, Smith DJ, Castaneda L, et al. Predictors of Serum Vitamin D Levels in African American and European American Men in Chicago. American journal of men's health. 2012;6(5):420-6. doi: 10.1177/1557988312437240. PubMed PMID: PMC3678722.

98. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461. doi: 10.1038/nature08494.

99. Cranney A, Horsley T, O'Donnell S, Weiler H, Puil L, Ooi D, et al. Effectiveness and safety of vitamin D in relation to bone health. Evidence Report/Technology Assessment. 2007(158):1-235. PubMed PMID: PMC4781354.

100. Mak T, Porsch RM, Choi SW, Sham PC. Polygenic scores for UK Biobank scale data. bioRxiv. 2018. doi: 10.1101/252270.

101. Littlejohns TJ, Henley WE, Lang IA, Annweiler C, Beauchet O, Chaves PHM, et al. Vitamin D and the risk of dementia and Alzheimer disease. Neurology. 2014;83(10):920-8. doi: 10.1212/WNL.0000000000000755. PubMed PMID: PMC4153851.

102. Alharbi FM. Update in vitamin D and multiple sclerosis. Neurosciences. 2015;20(4):329-35. doi: 10.17712/nsj.2015.4.20150357. PubMed PMID: PMC4727614.

103. Andersen SW, Shu X-O, Cai Q, Khankari NK, Steinwandel MD, Jurutka PW, et al. Total and Free Circulating Vitamin D and Vitamin D–Binding Protein in Relation to Colorectal Cancer Risk in a Prospective Study of African Americans. Cancer Epidemiology Biomarkers &amp;amp; Prevention. 2017;26(8):1242.

104. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938-9. doi: 10.1093/bioinformatics/btn564. PubMed PMID: PMC2720775.

105. Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinforma. 2009;10. doi: 10.1186/1471-2105-10-s1-s65.

106. Cheng T-YD, Millen AE, Wactawski-Wende J, Beresford SAA, LaCroix AZ, Zheng Y, et al. Vitamin D Intake Determines Vitamin D Status of Postmenopausal Women, Particularly Those with Limited Sun Exposure. The Journal of Nutrition. 2014;144(5):681-9. doi: 10.3945/jn.113.183541. PubMed PMID: PMC3985825.

107. Ware EB, Schmitz LL, Faul JD, Gard A, Mitchell C, Smith JA, et al. Heterogeneity in polygenic scores for common human traits. bioRxiv. 2017:106062. doi: 10.1101/106062.

Appendix

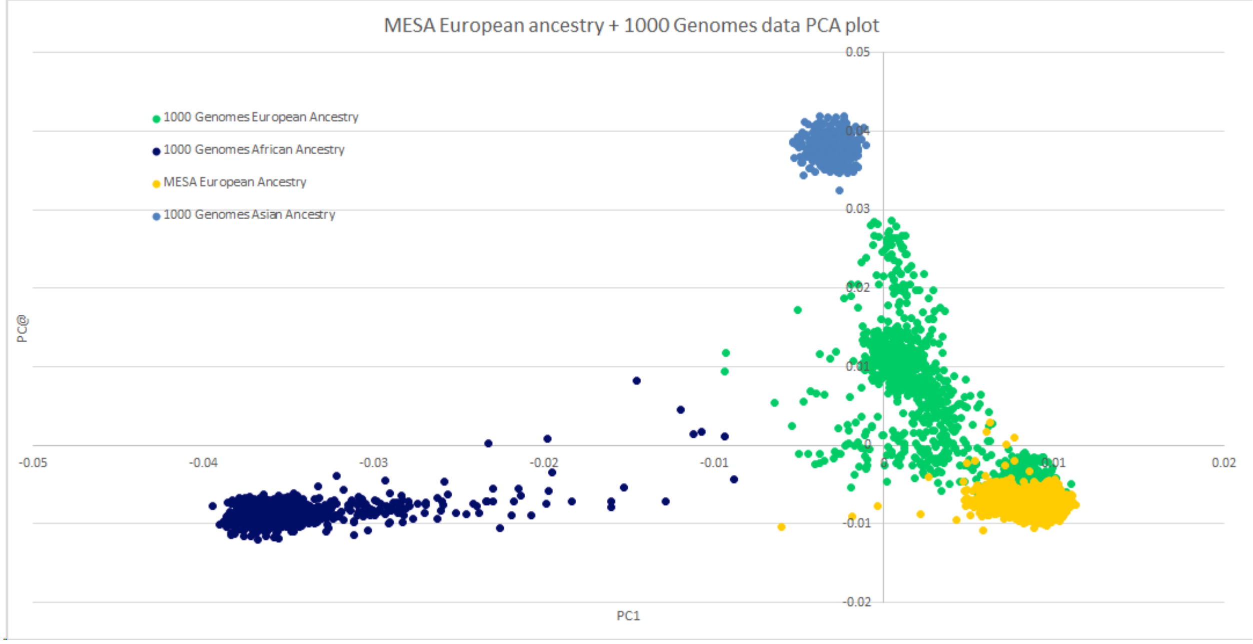
Supplemental Table 1: Genotyping information and quality control by cohort

Cohort	**Genotyping Array**	**Sample size**	**SNP quality control**				**Sample quality control**	
			Call rate	**Minor Allele Frequency**	**Hardy-Weinberg Equilibrium (post-imputation)**	**# of SNPs Passing QC**	**Call rate**	**Exclusions**
ARIC	AffymetrixGenome-WideHuman SNPArray 6.0	African ancestry: 1,908 European ancestry: 7,462	95%	0.002	African Ancestry: 6x10^{-8} (5x10^{-9}) European Ancestry: 6x10^{-8} (6x10^{-9})	African ancestry: 9,335,785 European ancestry: 8,315,761	95%	sex mismatch, relatedness, chromosomal abnormalities
MESA	Affymetrix 50K gene-focused molecular imprinted polymer array (CVDSNP55v1_A)	African ancestry: 1,176 European ancestry: 1,995	95%	0.002	African Ancestry: 1x10^{-6} (2x10^{-7}) European Ancestry: 1x10^{-6} (1x10^{-7})	African ancestry: 309,712 European ancestry: 455,155	95%	sex mismatch, relatedness
WHI (NHLBI cohort)	AffymetrixGenome-WideHuman SNPArray 6.0	African ancestry Consent group 1: 65 Consent group 2: 572	95%	0.002	Consent group 1: 6x10^{-8} (5x10^{-9}) Consent group 2: 6x10^{-8} (5x10^{-9})	Consent group 1: 9,551,098 Consent group 2: 9,997,380	95%	sex mismatch, relatedness, race mismatch
WHI (GARNET cohort)	Illumina HumanOmni1-Quad v1-0 B	European ancestry: Consent group 1: 86 Consent group 2: 443	95%	0.002	Consent group 1: 5x10^{-8} (5x10^{-9}) Consent group 2: 5x10^{-8} (4x10^{-9})	Consent group 1: 9,2037,621 Consent group 2: 9,722,526	95%	sex mismatch, relatedness, chromosome anomalies, race mismatch
WHI (PAGE cohort)	Illumina MEGA Consortium 15063755 B2 array	African ancestry: 63	95%	0.002	3x10^{-8} (5x10^{-9})	African ancestry: 9,641,566	95%	sex mismatch, relatedness, chromosome anomalies, race mismatch

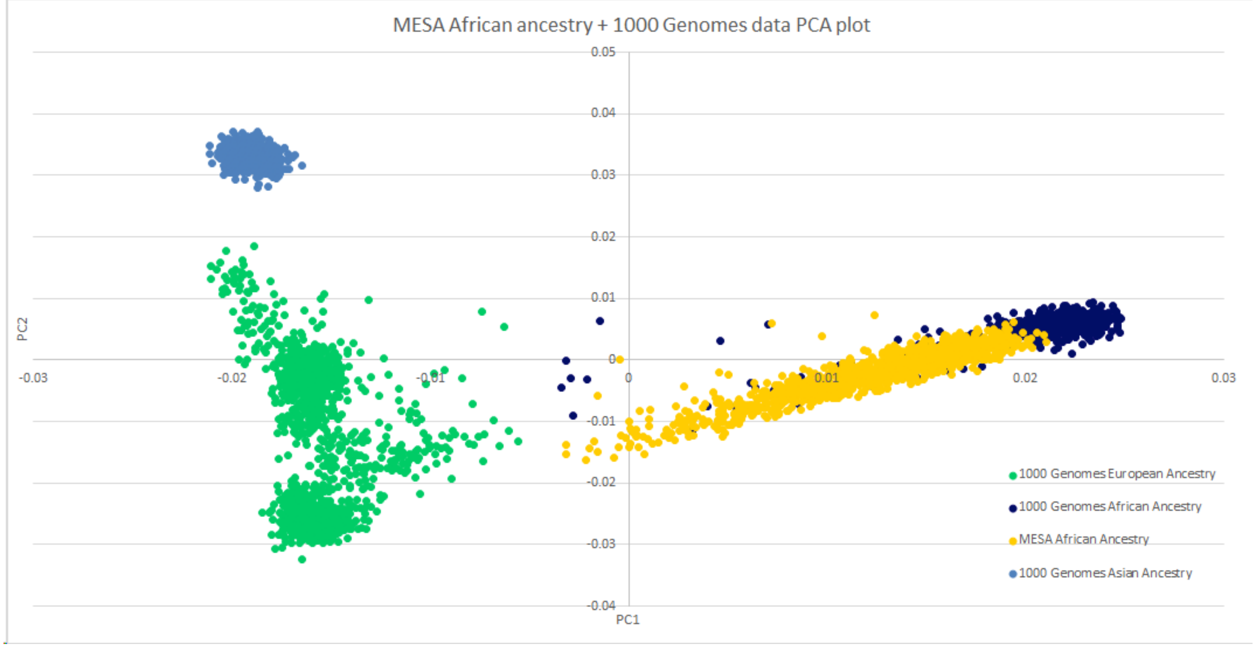
[a] All African-ancestry samples were imputed to CAAPA using Michigan Imputation Server

[b] All European-ancestry samples were imputed to HRC r1.1 2016 using Michigan Imputation Server

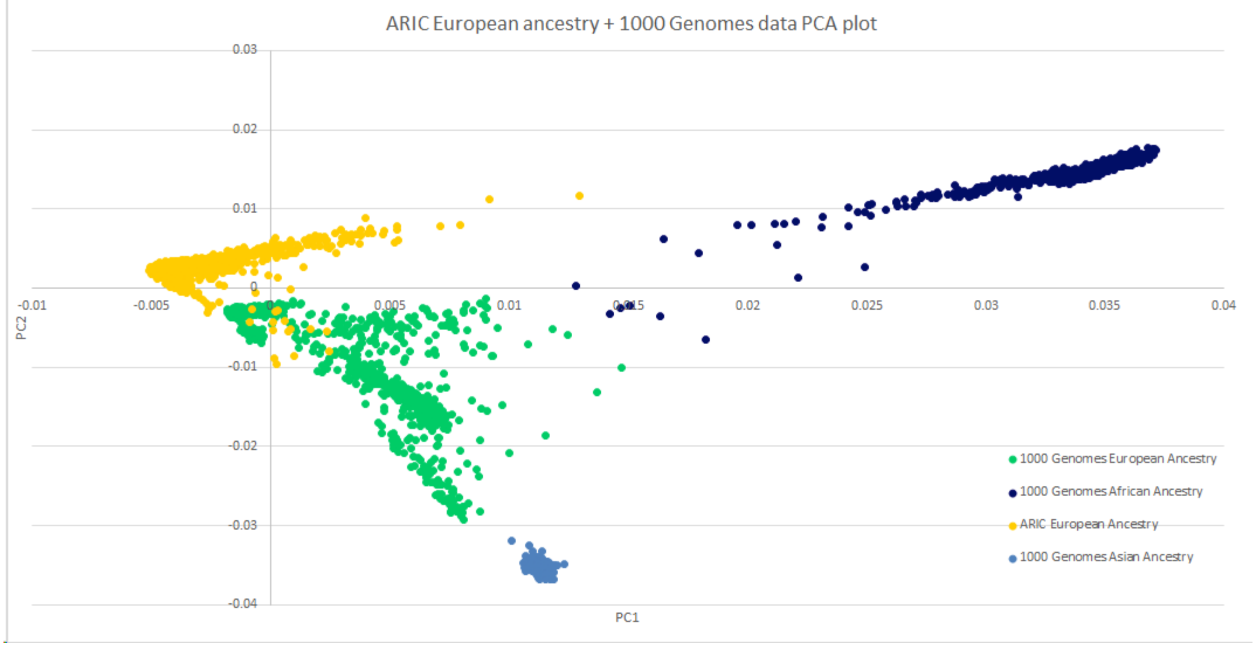
Supplemental Figure 1: Principle components plot for MESA European ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=1,995)



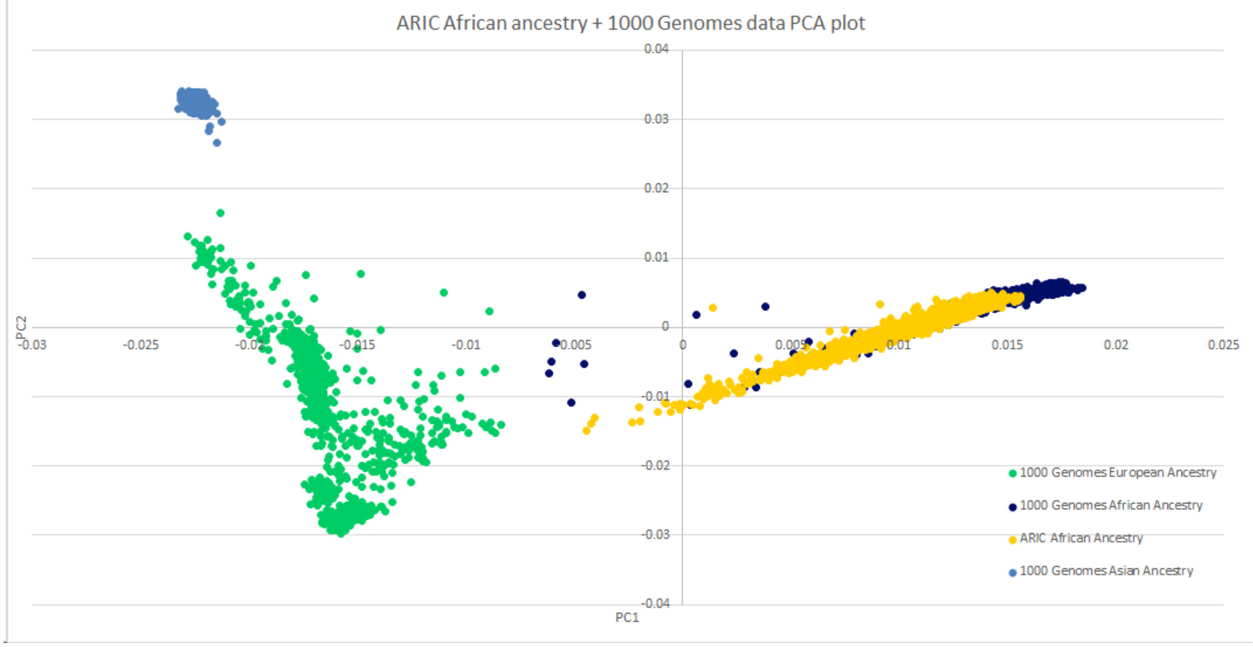
Supplemental Figure 2: Principle components plot for MESA African ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=1,176)



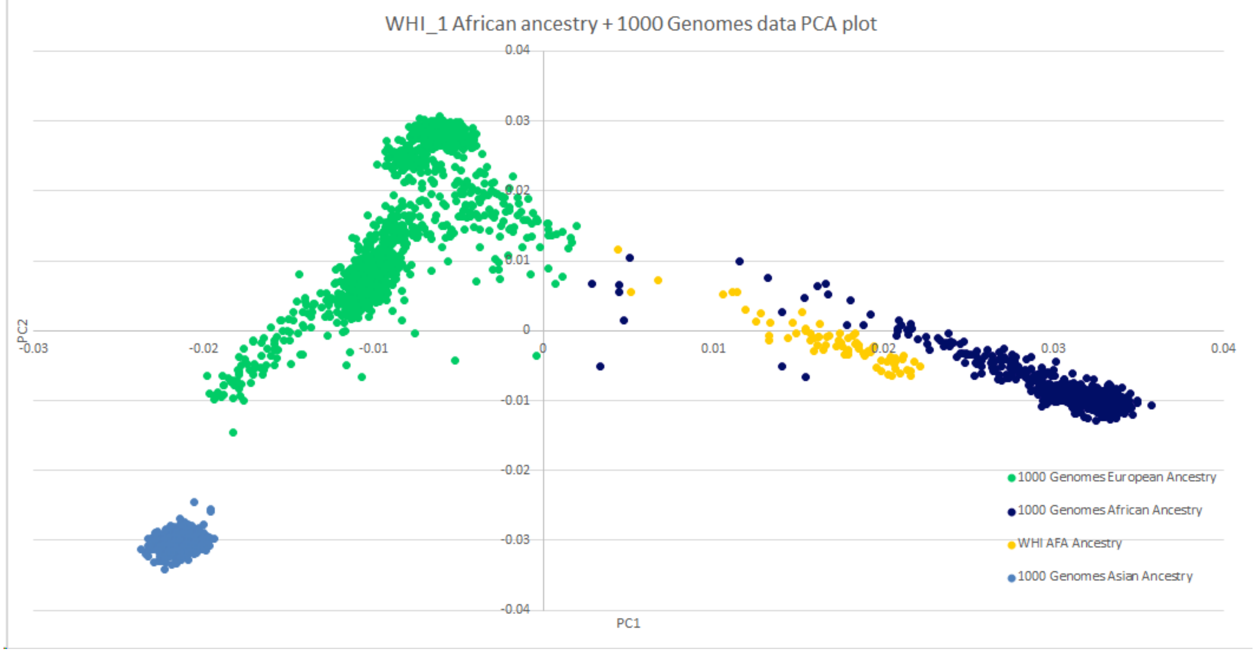
Supplemental Figure 3: Principle components plot for ARIC European ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=7,462)



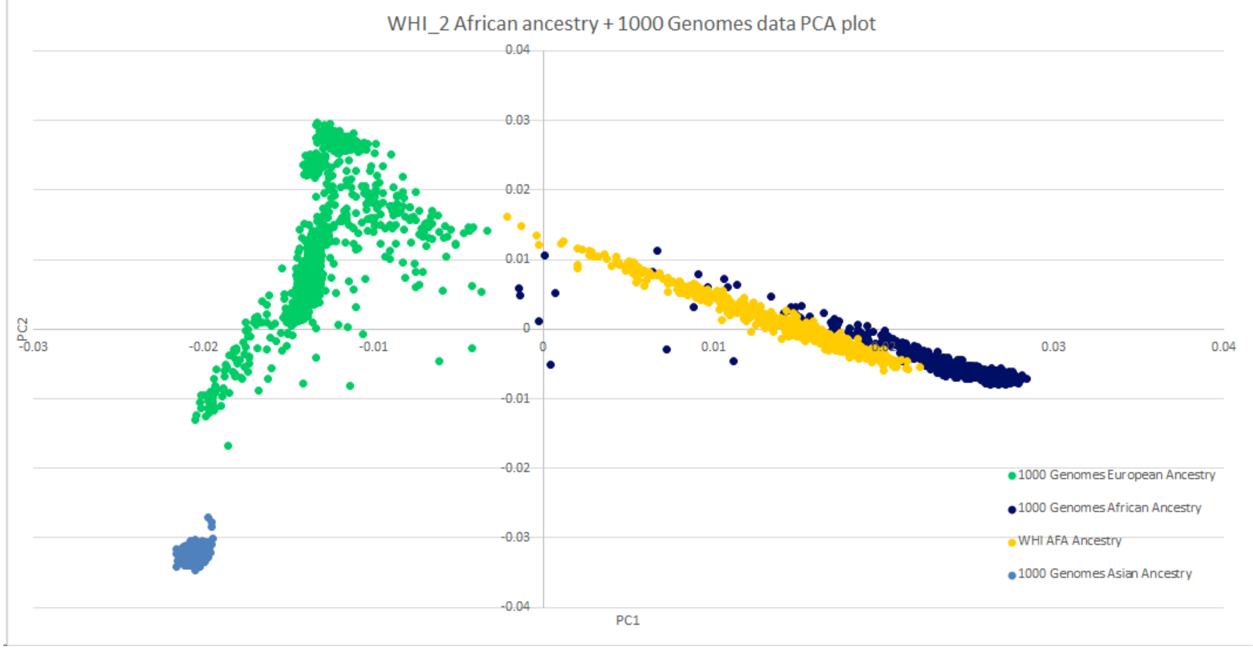
Supplemental Figure 4: Principle components plot for ARIC African ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=1,908)



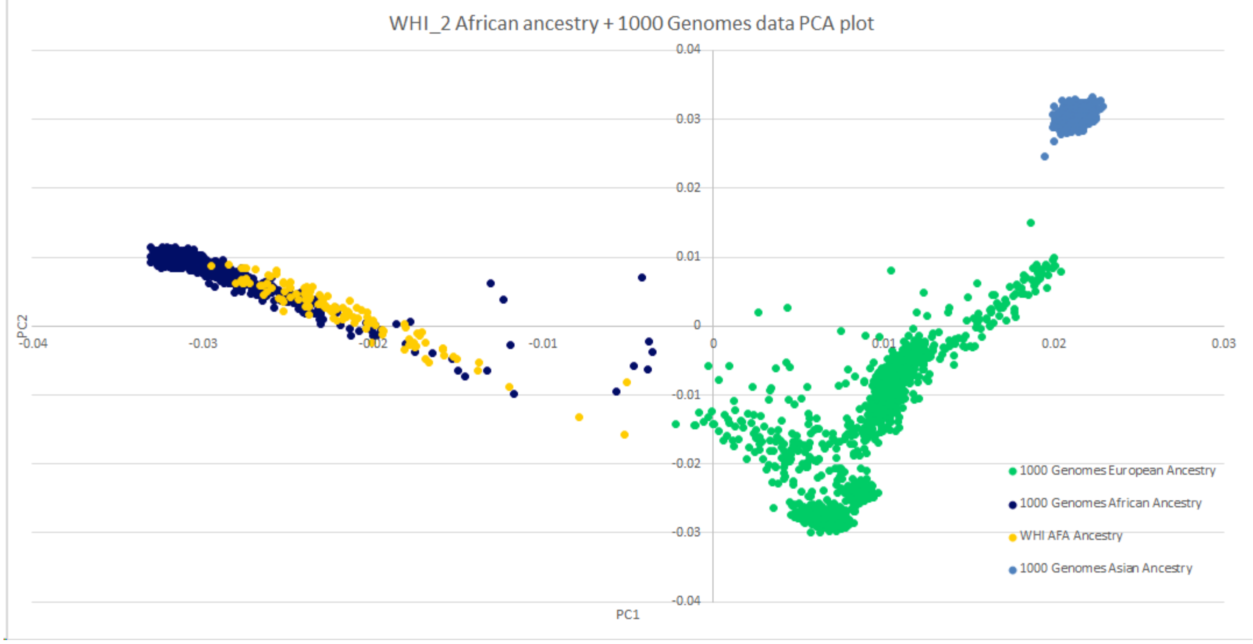
Supplemental Figure 5: Principle components plot for WHI subgroup 1 African ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=65)



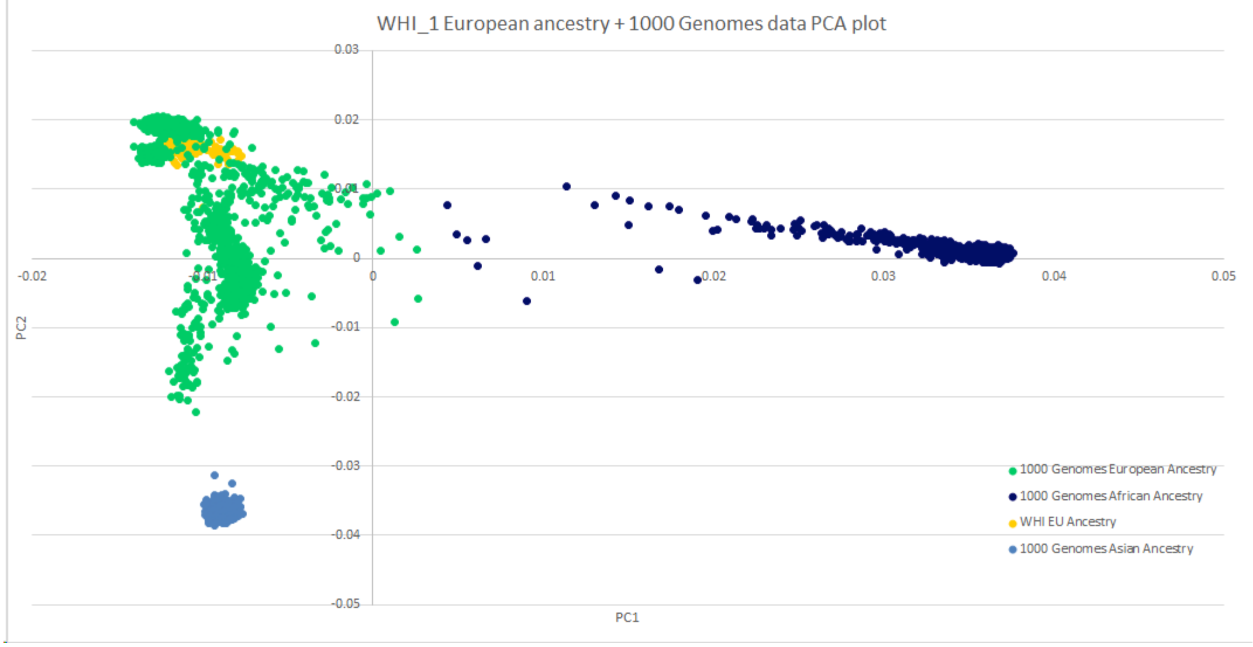
Supplemental Figure 6: Principle components plot for WHI subgroup 2 African ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=572)



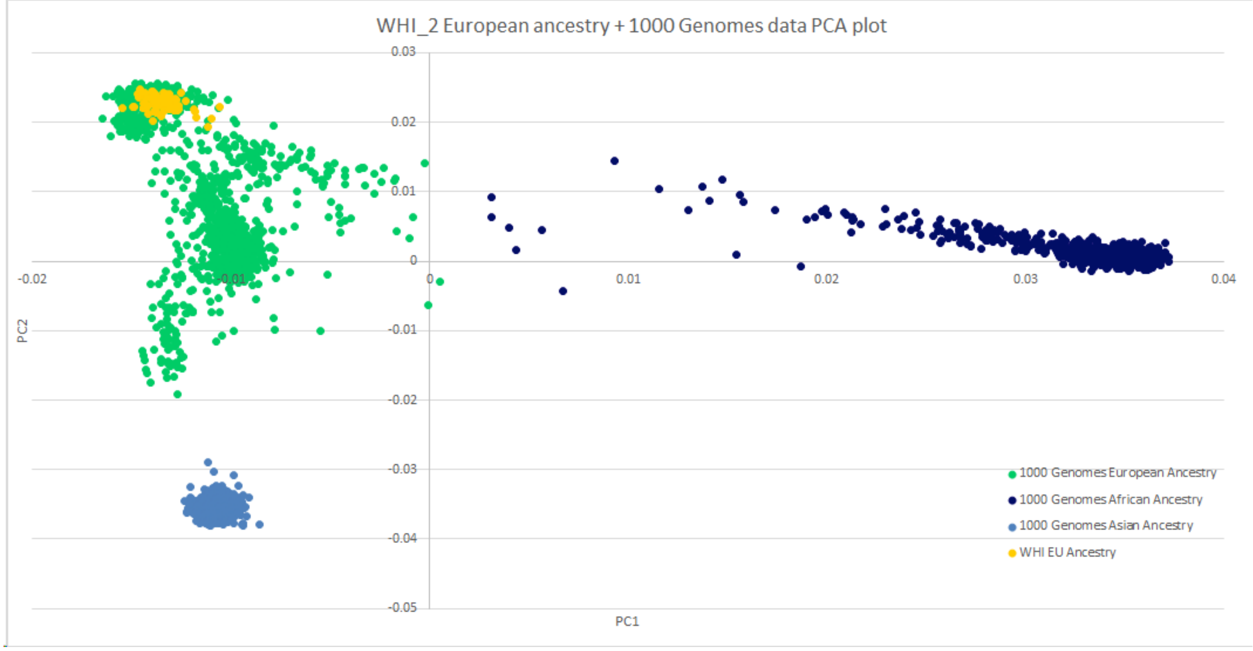
Supplemental Figure 7: Principle components plot for WHI subgroup 3 African ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=63)



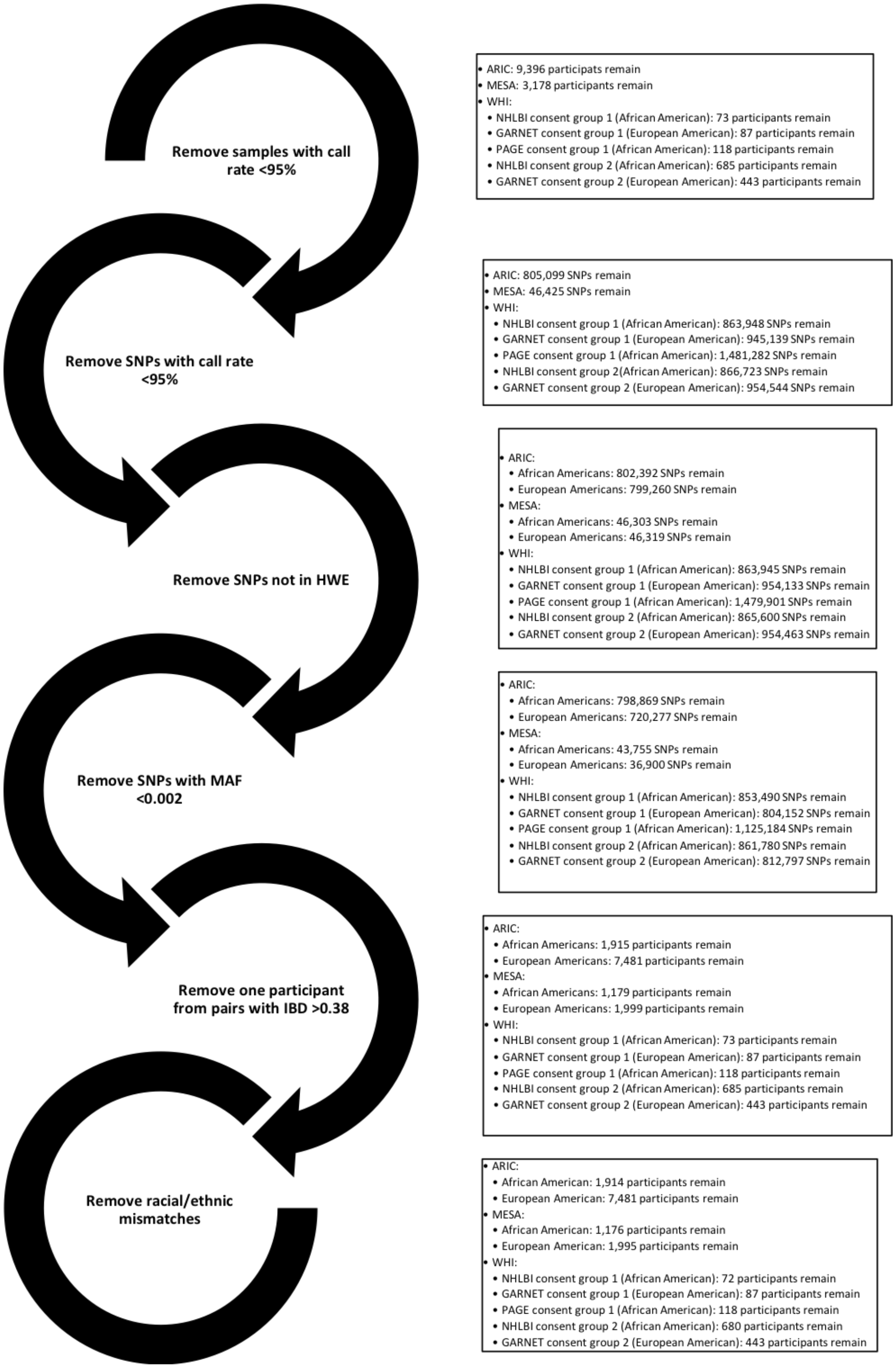
Supplemental Figure 8: Principle components plot for WHI subgroup 1 European ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=86)



Supplemental Figure 9: Principle components plot for WHI subgroup 2 European ancestry participants with 1000 Genomes European, African and Asian samples as anchoring populations (n=443)



Supplemental Figure 10**:** Pre-imputation quality-control process



Remove samples with call rate <95%

- ARIC: 9,396 participats remain
- MESA: 3,178 participants remain
- WHI:
  - NHLBI consent group 1 (African American): 73 participants remain
  - GARNET consent group 1 (European American): 87 participants remain
  - PAGE consent group 1 (African American): 118 participants remain
  - NHLBI consent group 2 (African American): 685 participants remain
  - GARNET consent group 2 (European American): 443 participants remain

Remove SNPs with call rate <95%

- ARIC: 805,099 SNPs remain
- MESA: 46,425 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 863,948 SNPs remain
  - GARNET consent group 1 (European American): 945,139 SNPs remain
  - PAGE consent group 1 (African American): 1,481,282 SNPs remain
  - NHLBI consent group 2(African American): 866,723 SNPs remain
  - GARNET consent group 2 (European American): 954,544 SNPs remain

Remove SNPs not in HWE

- ARIC:
  - African Americans: 802,392 SNPs remain
  - European Americans: 799,260 SNPs remain
- MESA:
  - African Americans: 46,303 SNPs remain
  - European Americans: 46,319 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 863,945 SNPs remain
  - GARNET consent group 1 (European American): 954,133 SNPs remain
  - PAGE consent group 1 (African American): 1,479,901 SNPs remain
  - NHLBI consent group 2 (African American): 865,600 SNPs remain
  - GARNET consent group 2 (European American): 954,463 SNPs remain

Remove SNPs with MAF <0.002

- ARIC:
  - African Americans: 798,869 SNPs remain
  - European Americans: 720,277 SNPs remain
- MESA:
  - African Americans: 43,755 SNPs remain
  - European Americans: 36,900 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 853,490 SNPs remain
  - GARNET consent group 1 (European American): 804,152 SNPs remain
  - PAGE consent group 1 (African American): 1,125,184 SNPs remain
  - NHLBI consent group 2 (African American): 861,780 SNPs remain
  - GARNET consent group 2 (European American): 812,797 SNPs remain

Remove one participant from pairs with IBD >0.38

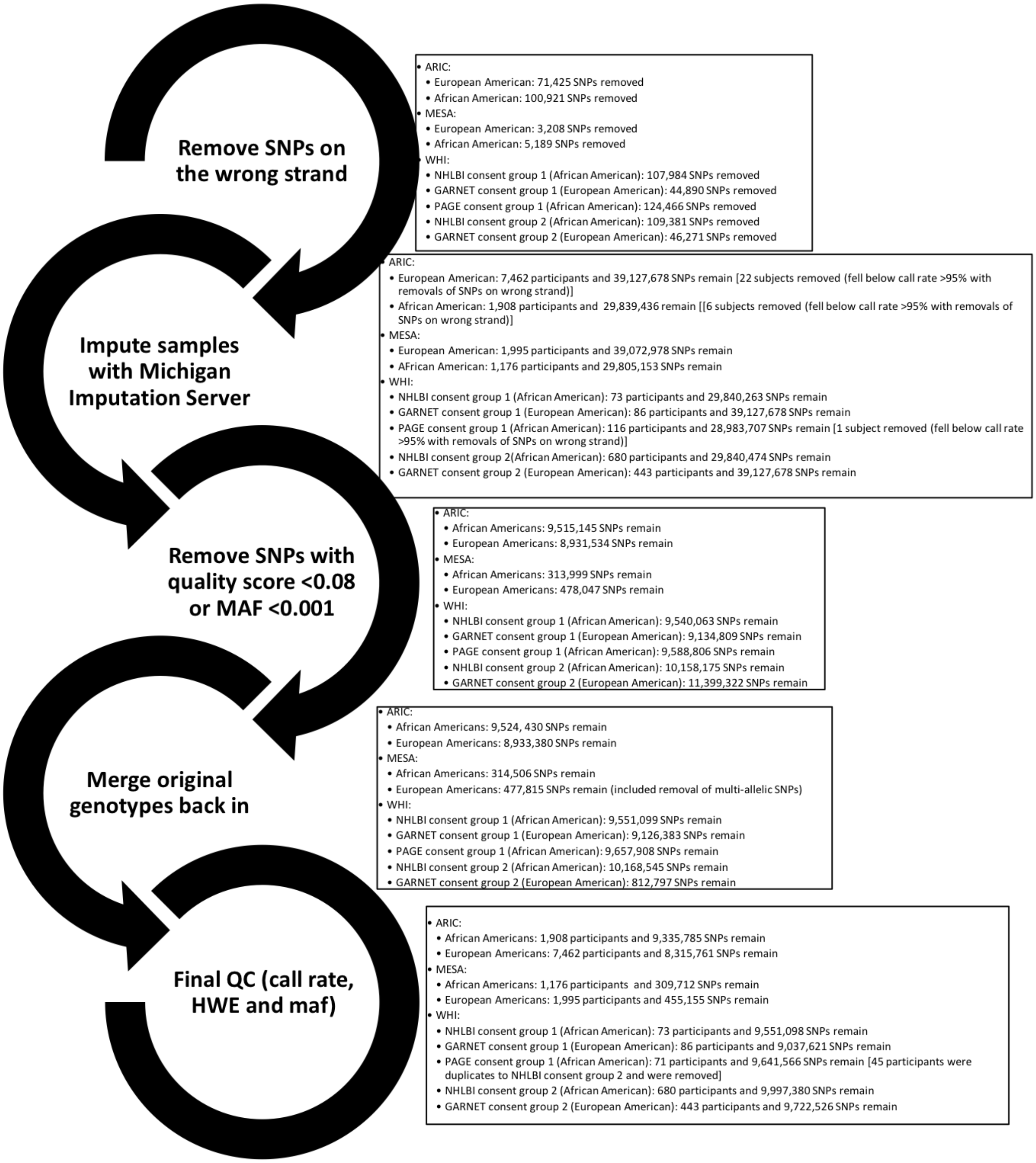
- ARIC:
  - African Americans: 1,915 participants remain
  - European Americans: 7,481 participants remain
- MESA:
  - African Americans: 1,179 participants remain
  - European Americans: 1,999 participants remain
- WHI:
  - NHLBI consent group 1 (African American): 73 participants remain
  - GARNET consent group 1 (European American): 87 participants remain
  - PAGE consent group 1 (African American): 118 participants remain
  - NHLBI consent group 2 (African American): 685 participants remain
  - GARNET consent group 2 (European American): 443 participants remain

Remove racial/ethnic mismatches

- ARIC:
  - African American: 1,914 participants remain
  - European American: 7,481 participants remain
- MESA:
  - African American: 1,176 participants remain
  - European American: 1,995 participants remain
- WHI:
  - NHLBI consent group 1 (African American): 72 participants remain
  - GARNET consent group 1 (European American): 87 participants remain
  - PAGE consent group 1 (African American): 118 participants remain
  - NHLBI consent group 2 (African American): 680 participants remain
  - GARNET consent group 2 (European American): 443 participants remain

Supplemental Figure 10 shows pre-imputation quality control steps and cutoffs used as well as corresponding sample sizes at each step.

Supplemental Figure 11: Quality-Control process starting at imputation



Remove SNPs on the wrong strand

- ARIC:
  - European American: 71,425 SNPs removed
  - African American: 100,921 SNPs removed
- MESA:
  - European American: 3,208 SNPs removed
  - African American: 5,189 SNPs removed
- WHI:
  - NHLBI consent group 1 (African American): 107,984 SNPs removed
  - GARNET consent group 1 (European American): 44,890 SNPs removed
  - PAGE consent group 1 (African American): 124,466 SNPs removed
  - NHLBI consent group 2 (African American): 109,381 SNPs removed
  - GARNET consent group 2 (European American): 46,271 SNPs removed

Impute samples with Michigan Imputation Server

- ARIC:
  - European American: 7,462 participants and 39,127,678 SNPs remain [22 subjects removed (fell below call rate >95% with removals of SNPs on wrong strand)]
  - African American: 1,908 participants and 29,839,436 remain [[6 subjects removed (fell below call rate >95% with removals of SNPs on wrong strand)]
- MESA:
  - European American: 1,995 participants and 39,072,978 SNPs remain
  - AFrican American: 1,176 participants and 29,805,153 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 73 participants and 29,840,263 SNPs remain
  - GARNET consent group 1 (European American): 86 participants and 39,127,678 SNPs remain
  - PAGE consent group 1 (African American): 116 participants and 28,983,707 SNPs remain [1 subject removed (fell below call rate >95% with removals of SNPs on wrong strand)]
  - NHLBI consent group 2(African American): 680 participants and 29,840,474 SNPs remain
  - GARNET consent group 2 (European American): 443 participants and 39,127,678 SNPs remain

Remove SNPs with quality score <0.08 or MAF <0.001

- ARIC:
  - African Americans: 9,515,145 SNPs remain
  - European Americans: 8,931,534 SNPs remain
- MESA:
  - African Americans: 313,999 SNPs remain
  - European Americans: 478,047 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 9,540,063 SNPs remain
  - GARNET consent group 1 (European American): 9,134,809 SNPs remain
  - PAGE consent group 1 (African American): 9,588,806 SNPs remain
  - NHLBI consent group 2 (African American): 10,158,175 SNPs remain
  - GARNET consent group 2 (European American): 11,399,322 SNPs remain

Merge original genotypes back in

- ARIC:
  - African Americans: 9,524, 430 SNPs remain
  - European Americans: 8,933,380 SNPs remain
- MESA:
  - African Americans: 314,506 SNPs remain
  - European Americans: 477,815 SNPs remain (included removal of multi-allelic SNPs)
- WHI:
  - NHLBI consent group 1 (African American): 9,551,099 SNPs remain
  - GARNET consent group 1 (European American): 9,126,383 SNPs remain
  - PAGE consent group 1 (African American): 9,657,908 SNPs remain
  - NHLBI consent group 2 (African American): 10,168,545 SNPs remain
  - GARNET consent group 2 (European American): 812,797 SNPs remain

Final QC (call rate, HWE and maf)

- ARIC:
  - African Americans: 1,908 participants and 9,335,785 SNPs remain
  - European Americans: 7,462 participants and 8,315,761 SNPs remain
- MESA:
  - African Americans: 1,176 participants and 309,712 SNPs remain
  - European Americans: 1,995 participants and 455,155 SNPs remain
- WHI:
  - NHLBI consent group 1 (African American): 73 participants and 9,551,098 SNPs remain
  - GARNET consent group 1 (European American): 86 participants and 9,037,621 SNPs remain
  - PAGE consent group 1 (African American): 71 participants and 9,641,566 SNPs remain [45 participants were duplicates to NHLBI consent group 2 and were removed]
  - NHLBI consent group 2 (African American): 680 participants and 9,997,380 SNPs remain
  - GARNET consent group 2 (European American): 443 participants and 9,722,526 SNPs remain

Supplemental Figure 11 shows quality control steps and cutoffs used as well as corresponding sample sizes at each step starting at the imputation phase

Supplemental Table 2: UV radiation values for ARIC

	Field Center			
Month of Visit	Wake Forest Baptist Medical Center, Winston-Salem, NC Recruitment in Forsyth County, NC	University of Mississippi Medical Center, Jackson, MS Recruitment in Jackson, MS	University of Minnesota, Minneapolis, MN Recruitment in Northwestern Minneapolis, MN	Johns Hopkins University, Baltimore, MD Recruitment in Washington County, MD
January	Month: December Year: average 1994-2002 Location: Raleigh, NC	Month: December Year: average 1994-2002 Location: Jackson, MS	Month: December Year: average 1994-2002 Location: Minneapolis, MN	Month: December Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
February	Month: January Year: average 1994-2002 Location: Raleigh, NC	Month: January Year: average 1994-2002 Location: Jackson, MS	Month: January Year: average 1994-2002 Location: Minneapolis, MN	Month: January Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
March	Month: February Year: average 1994-2002 Location: Raleigh, NC	Month: February Year: average 1994-2002 Location: Jackson, MS	Month: February Year: average 1994-2002 Location: Minneapolis, MN	Month: February Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
April	Month: March Year: average 1994-2002 Location: Raleigh, NC	Month: March Year: average 1994-2002 Location: Jackson, MS	Month: March Year: average 1994-2002 Location: Minneapolis, MN	Month: March Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
May	Month: April Year: average 1994-2002 Location: Raleigh, NC	Month: April Year: average 1994-2002 Location: Jackson, MS	Month: April Year: average 1994-2002 Location: Minneapolis, MN	Month: April Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
June	Month: May Year: average 1994-2002 Location: Raleigh, NC	Month: May Year: average 1994-2002 Location: Jackson, MS	Month: May Year: average 1994-2002 Location: R Minneapolis, MN	Month: May Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
July	Month: June Year: average 1994-2002 Location: Raleigh, NC	Month: June Year: average 1994-2002 Location: Jackson, MS	Month: June Year: average 1994-2002 Location: Minneapolis, MN	Month: June Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
August	Month: July Year: average 1994-2002 Location: Raleigh, NC	Month: July Year: average 1994-2002 Location: Jackson, MS	Month: July Year: average 1994-2002 Location: Minneapolis, MN	Month: July Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
September	Month: August Year: average 1994-2002 Location: Raleigh, NC	Month: August Year: average 1994-2002 Location: Jackson, MS	Month: August Year: average 1994-2002 Location: Minneapolis, MN	Month: August Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
October	Month: September Year: average 1994-2002 Location: Raleigh, NC	Month: September Year: average 1994-2002 Location: Jackson, MS	Month: September Year: average 1994-2002 Location: Minneapolis, MN	Month: September Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
November	Month: October Year: average 1994-2002 Location: Raleigh, NC	Month: October Year: average 1994-2002 Location: Jackson, MS	Month: October Year: average 1994-2002 Location: Minneapolis, MN	Month: October Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA
December	Month: November Year: average 1994-2002 Location: Raleigh, NC	Month: November Year: average 1994-2002 Location: Jackson, MS	Month: November Year: average 1994-2002 Location: Minneapolis, MN	Month: November Year: average 1994-2002 Location: Baltimore, MD and Pittsburgh, PA

*all visits occurred between 1990 and 1992; the National Weather Service Climate Prediction Center database starts with June 1994, therefore, the average for years 1994-2002 (the years in which vitamin D was collected for this project) was used

Supplemental Table 3: UV radiation values for MESA

	Site					
Month of visit	Wake Forest University, Winston-Salem, NC	Columbia University, New York, NY	Johns Hopkins University, Baltimore, MD	University of Minnesota, Minneapolis, MN	Northwestern University, Chicago, IL	University of California, Los Angeles, CA
January	Month: December Years: 2000-2002 Location: Raleigh, NC	Month: December Years: 2000-2002 Location: New York, NY	Month: December Years: 2000-2002 Location: Baltimore, MD	Month: December Years: 2000-2002 Location: Minneapolis, MN	Month: December Years: 2000-2002 Location: Chicago, IL	Month: December Years: 2000-2002 Location: Los Angeles, CA
February	Month: January Years: 2000-2002 Location: Raleigh, NC	Month: January Years: 2000-2002 Location: New York, NY	Month: January Years: 2000-2002 Location: Baltimore, MD	Month: January Years: 2000-2002 Location: Minneapolis, MN	Month: January Years: 2000-2002 Location: Chicago, IL	Month: January Years: 2000-2002 Location: Los Angeles, CA
March	Month: February Years: 2000-2002 Location: Raleigh, NC	Month: February Years: 2000-2002 Location: New York, NY	Month: February Years: 2000-2002 Location: Baltimore, MD	Month: February Years: 2000-2002 Location: Minneapolis, MN	Month: February Years: 2000-2002 Location: Chicago, IL	Month: February Years: 2000-2002 Location: Los Angeles, CA
April	Month: March Years: 2000-2002 Location: Raleigh, NC	Month: March Years: 2000-2002 Location: New York, NY	Month: March Years: 2000-2002 Location: Baltimore, MD	Month: March Years: 2000-2002 Location: Minneapolis, MN	Month: March Years: 2000-2002 Location: Chicago, IL	Month: March Years: 2000-2002 Location: Los Angeles, CA
May	Month: April Years: 2000-2002 Location: Raleigh, NC	Month: April Years: 2000-2002 Location: New York, NY	Month: April Years: 2000-2002 Location: Baltimore, MD	Month: April Years: 2000-2002 Location: Minneapolis, MN	Month: April Years: 2000-2002 Location: Chicago, IL	Month: April Years: 2000-2002 Location: Los Angeles, CA
June	Month: May Years: 2000-2002 Location: Raleigh, NC	Month: May Years: 2000-2002 Location: New York, NY	Month: May Years: 2000-2002 Location: Baltimore, MD	Month: May Years: 2000-2002 Location: Minneapolis, MN	Month: May Years: 2000-2002 Location: Chicago, IL	Month: May Years: 2000-2002 Location: Los Angeles, CA
July	Month: June Years: 2000-2002 Location: Raleigh, NC	Month: June Years: 2000-2002 Location: New York, NY	Month: June Years: 2000-2002 Location: Baltimore, MD	Month: June Years: 2000-2002 Location: Minneapolis, MN	Month: June Years: 2000-2002 Location: Chicago, IL	Month: June Years: 2000-2002 Location: Los Angeles, CA
August	Month: July Years: 2000-2002 Location: Raleigh, NC	Month: July Years: 2000-2002 Location: New York, NY	Month: July Years: 2000-2002 Location: Baltimore, MD	Month: July Years: 2000-2002 Location: Minneapolis, MN	Month: July Years: 2000-2002 Location: Chicago, IL	Month: July Years: 2000-2002 Location: Los Angeles, CA
September	Month: August Years: 2000-2002 Location: Raleigh, NC	Month: August Years: 2000-2002 Location: New York, NY	Month: August Years: 2000-2002 Location: Baltimore, MD	Month: August Years: 2000-2002 Location: Minneapolis, MN	Month: August Years: 2000-2002 Location: Chicago, IL	Month: August Years: 2000-2002 Location: Los Angeles, CA
October	Month: September Years: 2000-2002 Location: Raleigh, NC	Month: September Years: 2000-2002 Location: New York, NY	Month: September Years: 2000-2002 Location: Baltimore, MD	Month: September Years: 2000-2002 Location: Minneapolis, MN	Month: September Years: 2000-2002 Location: Chicago, IL	Month: September Years: 2000-2002 Location: Los Angeles, CA
November	Month: October Years: 2000-2002 Location: Raleigh, NC	Month: October Years: 2000-2002 Location: New York, NY	Month: October Years: 2000-2002 Location: Baltimore, MD	Month: October Years: 2000-2002 Location: Minneapolis, MN	Month: October Years: 2000-2002 Location: Chicago, IL	Month: October Years: 2000-2002 Location: Los Angeles, CA
December	Month: November Years: 2000-2002 Location: Raleigh, NC	Month: November Years: 2000-2002 Location: New York, NY	Month: November Years: 2000-2002 Location: Baltimore, MD	Month: November Years: 2000-2002 Location: Minneapolis, MN	Month: November Years: 2000-2002 Location: Chicago, IL	Month: November Years: 2000-2002 Location: Los Angeles, CA

*Month, but not specific year variables available for all participants, however, all visits occurred between 2000-2002, therefore, average for years 2000-2002 was used

Supplemental Table 4: UV radiation values for WHI

	Location								
Month of visit	Northeast (35-40 degrees N)	Northeast (>40 degrees N)	South (<35 degrees N)	South (35-40 degrees N)	Midwest (35-40 degrees N)	Midwest (>40 degrees N)	West (<35 degrees N)	West (35-40 degrees N)	West (>40 degrees N)
January	Month: December Locations: Raleigh, NC; Charleston, SC; Washington, DC	Month: December Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: December Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: December Locations: San Francisco, CA; Memphis, TN	Month: December Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: December Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: December Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: December Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: December Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
February	Month: January Locations: NC; Charleston, SC; Washington, DC	Month: January Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: January Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: January Locations: San Francisco, CA; Memphis, TN	Month: January Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: January Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: January Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: January Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: January Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
March	Month: February Locations: NC; Charleston, SC; Washington, DC	Month: February Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: February Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: February Locations: San Francisco, CA; Memphis, TN	Month: February Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: February Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: February Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: February Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: February Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
April	Month: March Locations: NC; Charleston, SC; Washington, DC	Month: March Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: March Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: March Locations: San Francisco, CA; Memphis, TN	Month: March Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: March Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: March Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: March Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: March Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
May	Month: April Locations: NC; Charleston, SC; Washington, DC	Month: April Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston,	Month: April Locations: Los Angeles, CA; Phoenix, AZ;	Month: April Locations: San Francisco, CA; Memphis, TN	Month: April Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: April Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL;	Month: April Locations: Phoenix, AZ; Los Angeles, CA;	Month: April Locations: San Francisco, CA; Denver, CO; Salt Lake City,	Month: April Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID

		MA; Portland, ME	Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL			Des Moines, IA; Bismarck, ND	Albuquerque, NM	UT; Las Vegas, NV	
June	Month: May Locations: NC; Charleston, SC; Washington, DC	Month: May Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: May Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: May Locations: San Francisco, CA; Memphis, TN	Month: May Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: May Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: May Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: May Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: May Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
July	Month: June Locations: NC; Charleston, SC; Washington, DC	Month: June Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: June Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: June Locations: San Francisco, CA; Memphis, TN	Month: June Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: June Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: June Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: June Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: June Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
August	Month: July Locations: NC; Charleston, SC; Washington, DC	Month: July Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: July Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: July Locations: San Francisco, CA; Memphis, TN	Month: July Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: July Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: July Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: July Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: July Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
September	Month: August Locations: NC; Charleston, SC; Washington, DC	Month: August Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: August Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: August Locations: San Francisco, CA; Memphis, TN	Month: August Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: August Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: August Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: August Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: August Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
October	Month: September Locations: NC; Charleston, SC; Washington, DC	Month: September Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA;	Month: September Locations: Los Angeles, CA; Phoenix, AZ; Houston,	Month: September Locations: San Francisco, CA; Memphis, TN	Month: September Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: September Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des	Month: September Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: September Locations: San Francisco, CA; Denver, CO; Salt Lake City,	Month: September Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID

		Portland, ME	TX; Atlanta, GA; Jacksonville, FL; Miami, FL			Moines, IA; Bismarck, ND		UT; Las Vegas, NV	
November	Month: October Locations: NC; Charleston, SC; Washington, DC	Month: October Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: October Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: October Locations: San Francisco, CA; Memphis, TN	Month: October Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: October Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: October Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: October Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: October Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID
December	Month: November Locations: NC; Charleston, SC; Washington, DC	Month: November Locations: New York, NY; Buffalo, NY; Burlington, VT; Boston, MA; Portland, ME	Month: November Locations: Los Angeles, CA; Phoenix, AZ; Houston, TX; Atlanta, GA; Jacksonville, FL; Miami, FL	Month: November Locations: San Francisco, CA; Memphis, TN	Month: November Locations: St. Louis, MO; Omaha, NE; Sioux Falls, SD	Month: November Locations: Milwaukee, WI; Minneapolis, MN; Chicago, IL; Des Moines, IA; Bismarck, ND	Month: November Locations: Phoenix, AZ; Los Angeles, CA; Albuquerque, NM	Month: November Locations: San Francisco, CA; Denver, CO; Salt Lake City, UT; Las Vegas, NV	Month: November Locations: Portland, OR; Seattle, WA; Billing, MT; Boise, ID

*blood draws for WHI were done in the years 1993-1999, with the exception of 1993 and 1994 (before the National Weather Service Climate Prediction Center database had started documenting data), the years used for UV radiation value match the year of visit. Those with visits in 1993 or 1994 are given the corresponding monthly average for the average of years 1995-2002.

Supplemental Table 5: Available UV radiation value descriptive statistics for ARIC (average from 1994-2002)

Field Center				
Month of blood draw	Wake Forest Baptist Medical Center, Winston-Salem, NC	University of Mississippi Medical Center, Jackson, MS	University of Minnesota, Minneapolis, MN	Johns Hopkins University, Baltimore, MD
January	2.1	2.7	0.9	1.4
February	2.5	3.2	1.1	1.8
March	3.7	4.7	1.9	2.8
April	5.7	6.8	3.4	4.5
May	7.1	8.2	4.8	6.1
June	8.1	8.9	6.2	7.3
July	8.7	9.1	7.5	8.3
August	8.9	9.5	7.8	8.4
September	8.3	9.1	6.7	7.5
October	6.6	7.7	4.7	5.7
November	4.5	5.5	2.5	3.6
December	2.8	3.5	1.3	2.0

Supplemental Table 6: Available UV radiation value descriptive statistics for MESA (average from 2000-2002)

	Site					
Month of visit	Wake Forest University, Winston-Salem, NC	Columbia University, New York, NY	Johns Hopkins University, Baltimore, MD	University of Minnesota, Minneapolis, MN	Northwestern University, Chicago, IL	University of California, Los Angeles, CA
January	1.9	1.1	1.5	0.7	0.9	2.3
February	2.5	1.6	1.9	1.1	1.5	2.8
March	3.8	2.6	2.9	2.0	2.5	4.1
April	5.9	4.4	4.8	3.3	4.0	5.9
May	6.9	5.4	6.3	4.6	5.3	7.3
June	7.7	6.4	7.1	5.6	6.2	8.2
July	8.4	7.6	8.2	6.9	7.5	8.9
August	8.1	7.2	8.0	7.3	7.7	9.1
September	7.6	6.5	7.3	6.1	6.7	8.9
October	6.0	5.0	5.7	4.3	4.9	7.3
November	4.3	3.0	3.7	2.1	2.6	4.5
December	2.5	1.6	2.0	1.1	2.4	2.9

Supplemental Table 7: Available UV radiation value descriptive statistics for WHI

Location									
Month of visit	Northeast (35-40 degrees N)	Northeast (>40 degrees N)	South (<35 degrees N)	South (35-40 degrees N)	Midwest (35-40 degrees N)	Midwest (>40 degrees N)	West (<35 degrees N)	West (35-40 degrees N)	West (>40 degrees N)
January	1993-4*:1.8	1993-4*: 1.1 1996: 1.2	1993-4*: 2.6 1999: 2.2	1993-4*: 2.0 1996: 2.2 1999: 1.5	1993-4*: 1.3 1999: 0.9	1993-4*: 1.0 1996: 1.1 1999: 0.7	1993-4*: 2.5 1999: 2.2	1993-4*: 1.8 1999: 1.4	1993-4*: 0.8 1996: 0.9
February	1993-4*: 2.2 1999: 2.1	1993-4*: 1.3 1999: 1.3	1993-4*: 3.1 1996: 3.3	1993-4*: 2.4 1996: 2.5	1993-4*: 1.6	1993-4*: 1.3	1993-4*: 2.9 1999: 2.9	1993-4*: 2.1 1999: 2.1	1993-4*: 1.1 1996: 1.1
March	1993-4*: 3.3	1993-4*: 2.2 1999: 2.2	1993-4*:4.5 1996: 4.5 1999: 4.5	1993-4*: 3.6 1996: 3.6	1993-4*: 2.6 1999: 2.5	1993-4*: 2.1 1999: 2.1	1993-4*: 4.2	1993-4*:3.2 1996: 3.3 1999: 3.3	1993-4*: 2.0
April	1993-4*: 5.1	1993-4*: 3.7 1996: 3.8	1993-4*: 6.4	1993-4*: 5.5 1996: 5.4	1993-4*: 4.3 1996: 4.2	1993-4*: 3.7 1996: 3.8 1999: 3.6	1993-4*: 6.1 1999: 5.9	1993-4*: 5.1	1993-4*: 3.4
May	1993-4*: 6.6	1993-4*: 5.2 1999: 4.2	1993-4*: 7.7 1996: 8.3 1999: 6.4	1993-4*: 7.0	1993-4*: 5.7	1993-4*: 5.2 1996: 5.6 1999: 4.4	1993-4*: 7.7	1993-4*: 6.6 1999: 5.4	1993-4*: 4.8
June	1993-4*: 7.7	1993-4*: 6.6 1999: 7.1	1993-4*: 8.5 1999: 7.4	1993-4*: 8.0 1996: 9.1	1993-4*: 7.2	1993-4*: 6.6 1996: 7.5 1999: 5.1	1993-4*: 8.7	1993-4*: 8.1 1999: 7.2	1993-4*: 6.3
July	1993-4*: 8.5	1993-4*: 7.8 1999: 6.6	1993-4*: 9.0 1996: 10.0	1993-4*: 8.8 1996: 9.5 1999: 7.5	1993-4*: 8.3	1993-4*: 7.8	1993-4*: 8.9	1993-4*: 8.8 1996: 9.9	1993-4*: 7.3 1996: 8.2
August	1993-4*: 8.7	1993-4*: 7.7 1996: 8.4 1999: 6.8	1993-4*: 9.1 1996: 9.9 1999: 8.0	1993-4*: 9.1 1999: 8.0	1993-4*: 8.6 1996: 9.1 1999:7 .5	1993-4*: 8.1 1996: 8.5	1993-4*: 9.2	1993-4*: 9.1	1993-4*: 7.7
September	1993-4*: 7.9 1996: 8.7	1993-4*: 6.7 1999: 5.3	1993-4*: 8.8 1999: 7.8	1993-4*:8.5 1996: 9.2	1993-4*: 7.6 1996: 8.3	1993-4*: 7.0 1996: 7.7 1999: 6.0	1993-4*: 9.2	1993-4*: 8.6 1999: 7.9	1993-4*: 6.7
October	1993-4*: 6.2 1999: 5.0	1993-4*: 5.0 1996: 5.5	1993-4*: 7.4 1999: 6.5	1993-4*: 6.7	1993-4*: 5.6 1996: 5.9 1999: 4.7	1993-4*: 5.0 1996: 5.3 1999: 4.0	1993-4*: 7.7 1996: 7.9 1999: 7.1	1993-4*: 6.7 1999: 6.0	1993-4*: 4.7 1996: 4.9
November	1993-4*: 4.1	1993-4*: 2.9 1996: 3.1 1999: 2.3	1993-4*: 5.2 1996: 5.5 1999: 4.5	1993-4*: 4.4 1996: 4.7 1999: 4.0	1993-4*: 3.3	1993-4*: 2.8 1996: 3.1	1993-4*: 5.1 1996: 5.2	1993-4*: 4.1	1993-4*: 2.6
December	1993-4*: 2.4	1993-4*: 1.6 1996: 1.7 1999: 1.2	1993-4*: 3.4 1996: 3.8 1999: 2.9	1993-4*: 2.7	1993-4*: 1.9 1996: 2.1	1993-4*: 1.5 1996: 1.7 1999: 1.3		1993-4*: 2.4	1993-4*: 1.3

Only location, month, year combos with observations in the final dataset are shown
*average of 1995-2002 (all data collected, since database started in late 1994)

Supplemental Table 8: Performance of models in determining optimal p-value cutoff

Ancestry	Model	Cut-off	PGS R^2 (model R^2)	p-value	# SNPs in PGS
European (n=1,000)	**No covariates**	0.00025	0.015 (0.015)	0.00009	251
	Age + sex	0.00025	0.016 (0.052)	0.00005	251
	Age + sex +BMI	0.00035	0.016 (0.065)	0.00005	341
	Age + sex + UV	0.00035	0.014 (0.106)	0.000001	341
	Age + sex + BMI +UV	0.00035	0.014 (0.129)	0.00008	341
	Age + sex + BMI +UV + intake	0.00035	0.013 (0.137)	0.00007	341
African (n=57)	**No covariates**	0.01265	0.081 (0.081)	0.024	32,269
	Age + sex	0.01265	0.107 (0.179)	0.008	32,269
	Age + sex +BMI	0.01265	0.08 (0.162)	0.03	32,269
	Age + sex + UV	0.01265	0.105 (0.273)	0.006	32,269
	Age + sex + BMI +UV	0.01265	0.044 (0.37)	0.06	32,269
	Age + sex + BMI +UV + intake	0.0072	0.011 (0.33)	0.47	19,261

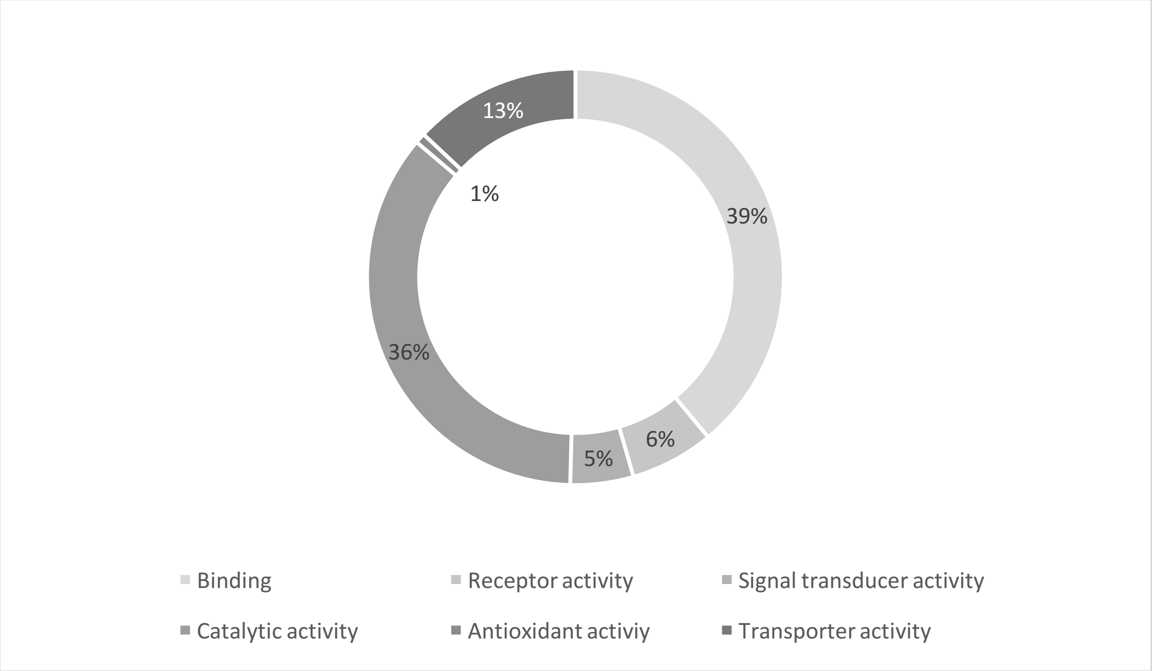
Supplemental Table 9: Top European ancestry Gene-Ontology biological process complete enrichment categories

GO biological process complete	Fold Enrichment	raw P value	FDR
regulation of cell communication by electrical coupling	42.75	6.21E-06	8.11E-03
regulation of anatomical structure morphogenesis	2.94	1.77E-06	4.63E-03
neuron differentiation	2.93	3.07E-06	5.35E-03
positive regulation of cell differentiation	2.65	5.74E-05	4.49E-02
positive regulation of developmental process	2.58	3.00E-06	5.87E-03
cell development	2.56	4.52E-07	7.07E-03
Neurogenesis	2.33	1.09E-05	1.22E-02
generation of neurons	2.32	2.45E-05	2.26E-02
positive regulation of multicellular organismal process	2.20	3.61E-05	2.98E-02
nervous system development	2.20	1.18E-06	6.15E-03

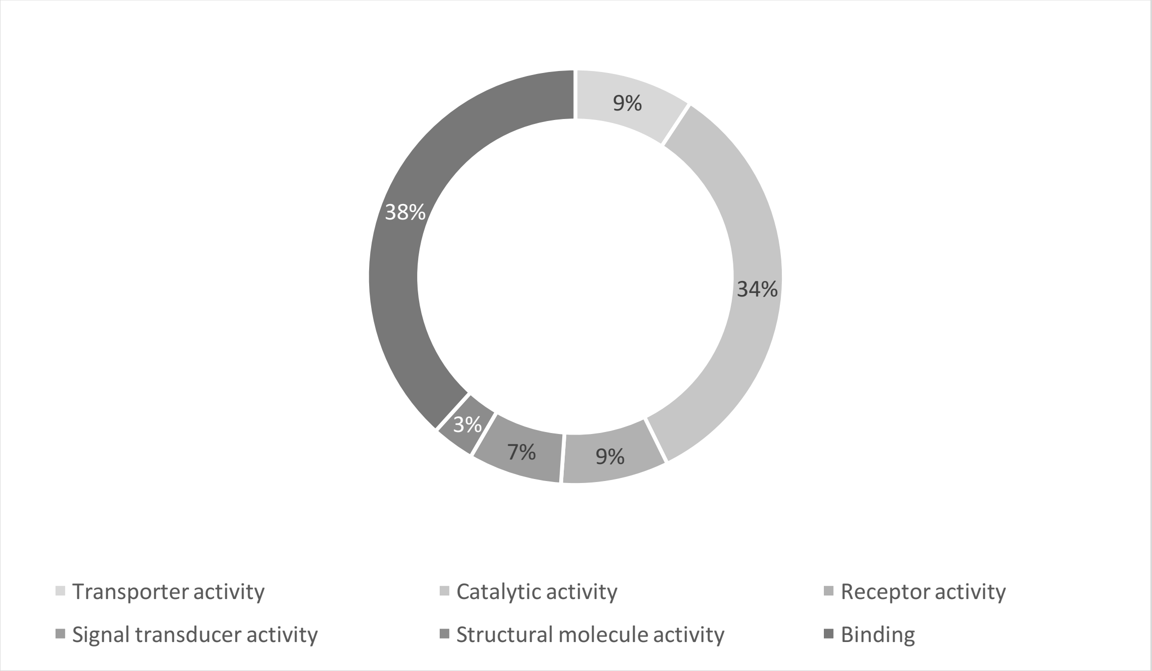
Supplemental Table 10: Top African ancestry Gene-Ontology biological process complete enrichment categories

GO biological process complete	Fold Enrichment	raw P value	FDR
cell-cell adhesion mediated by cadherin	2.88	7.73E-04	3.11E-02
cAMP metabolic process	2.85	1.17E-03	4.32E-02
cyclic nucleotide metabolic process	2.49	2.57E-04	1.28E-02
cardiac muscle cell contraction	2.37	1.16E-03	4.31E-02
glutamate receptor signaling pathway	2.32	6.10E-04	2.56E-02
regulation of action potential	2.28	5.64E-04	2.44E-02
cell communication involved in cardiac conduction	2.26	1.26E-03	4.63E-02
dendrite morphogenesis	2.24	1.96E-04	1.02E-02
calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	2.22	1.15E-03	4.31E-02
transmission of nerve impulse	2.22	7.60E-04	3.07E-02

Supplemental Figure 12: Top European ancestry PantherGO-Slim Molecular Function categories



Supplemental Figure 13: Top African ancestry PantherGO-Slim Molecular Function categories



Supplemental Table 11: 25(OH)D Z-scores by decile for those of European ancestry

Percentile											
Cohort	1	10	20	30	40	50	60	70	80	90	100	99
PRS determination (n=1,000)	NA	-0.13	0.06	-0.02	0.19	-0.07	0.22	0.29	0.31	0.13	0.33	NA
Combined (n=8,569)	0.01	0.14	0.18	-0.01	0.07	0.15	0.18	0.22	0.27	0.36	0.45	0.54

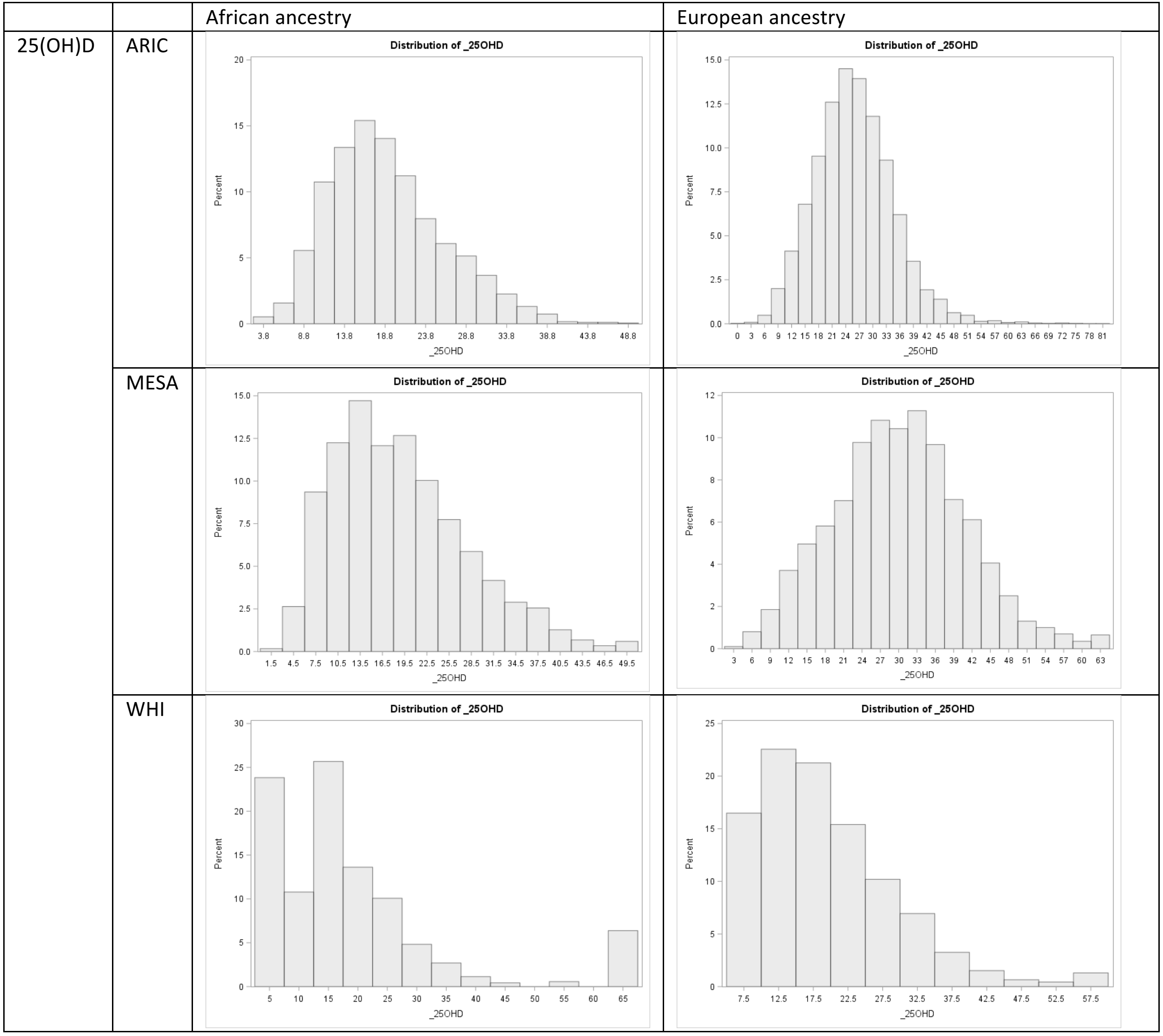
Supplemental Table 11 shows 25(OH)D Z-scores for those of European ancestry by deciles and for extreme percentiles (1st and 99th). As expected, as percentile increases, Z-scores do too.

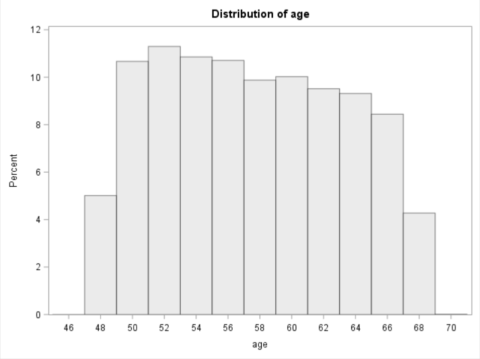
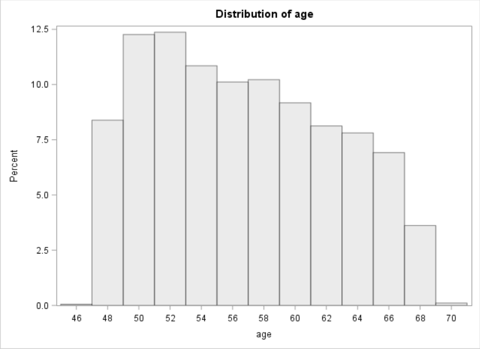
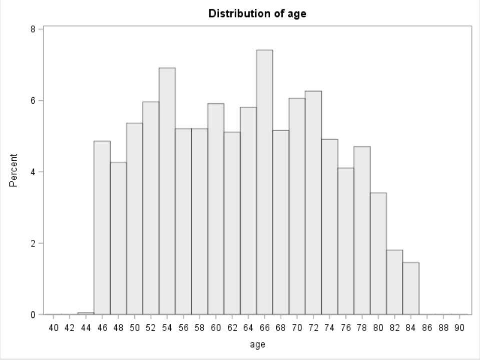
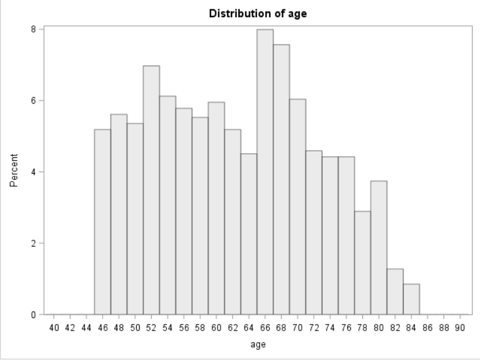
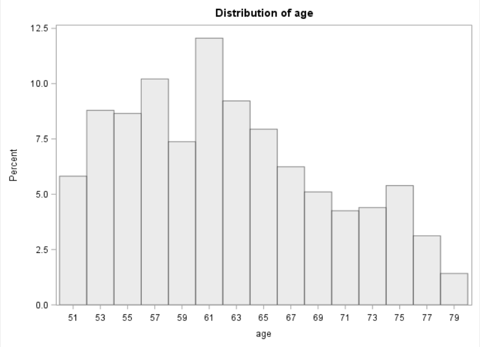
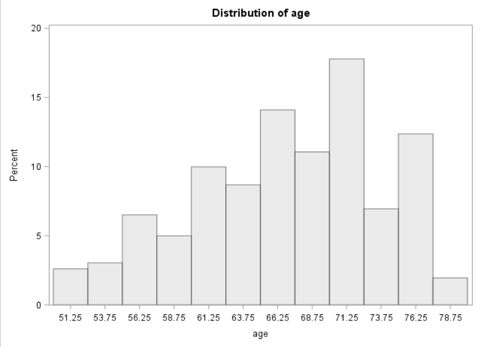
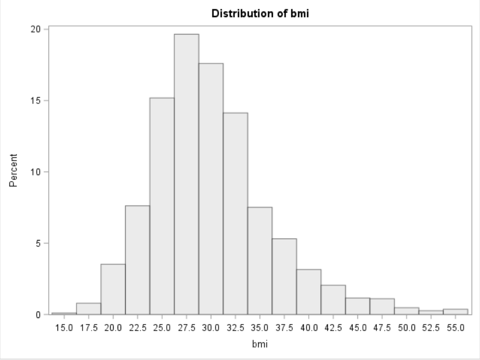
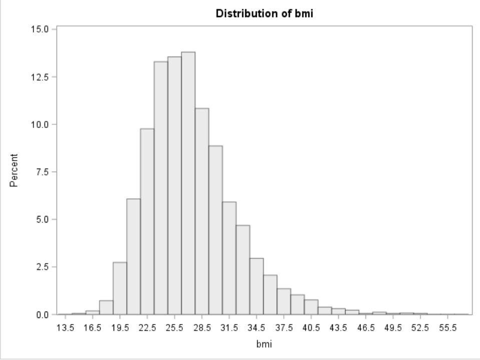
Supplemental Table 12: 25(OH)D concentrations by quintile for those of African ancestry

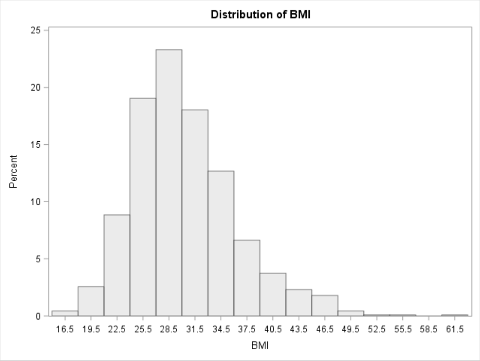
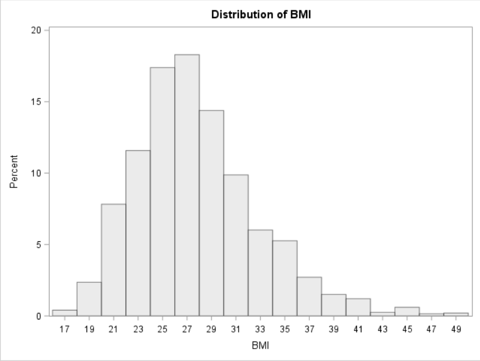
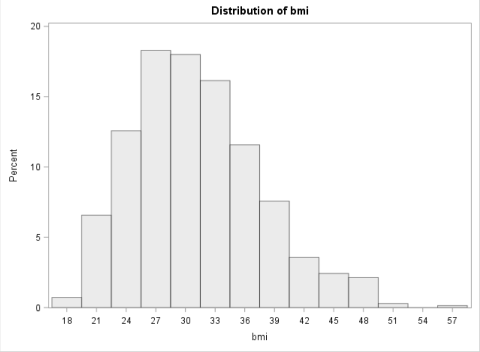
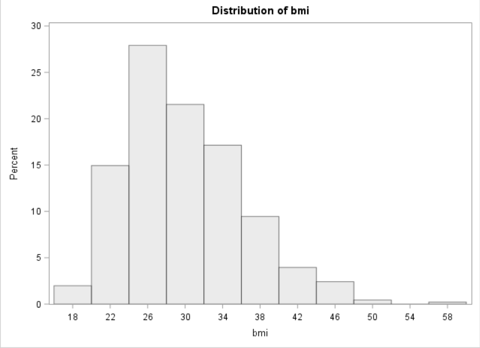
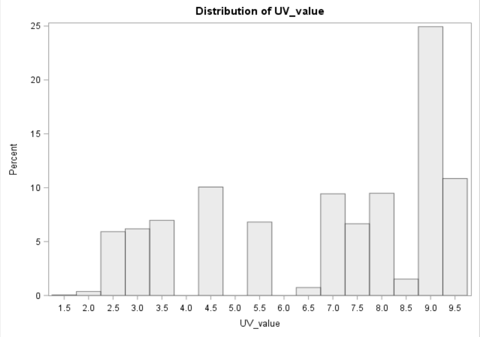
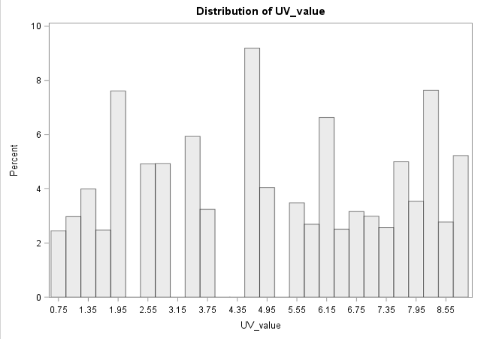
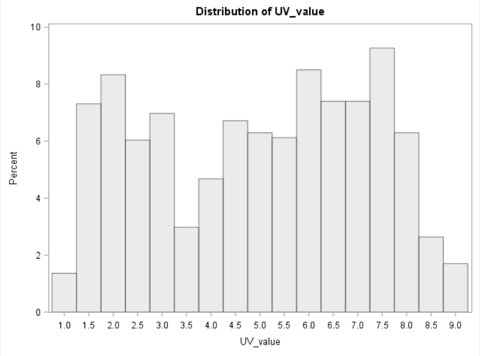
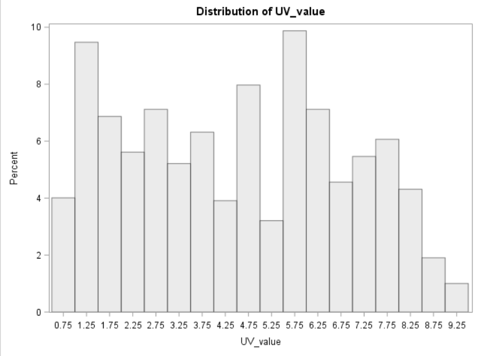
	Percentile				
Cohort	20	40	60	80	100
PRS determination (n=57)	-0.94	0.09	-0.46	-0.32	-0.46
Combined (n=1,042)	-0.29	-0.25	-0.07	-0.14	-0.15

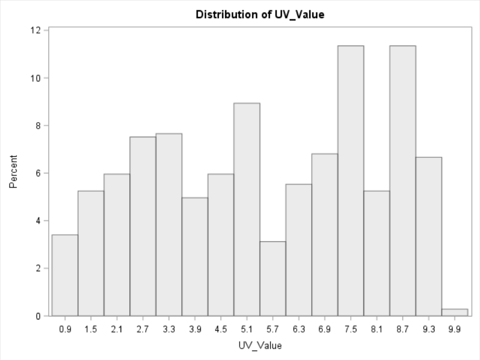
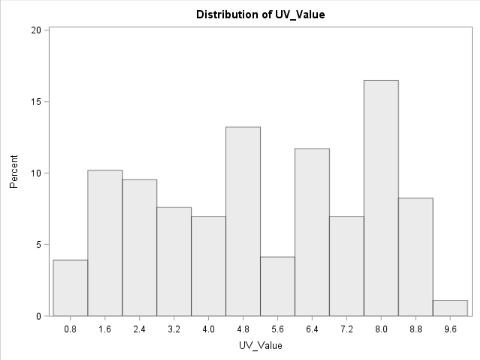
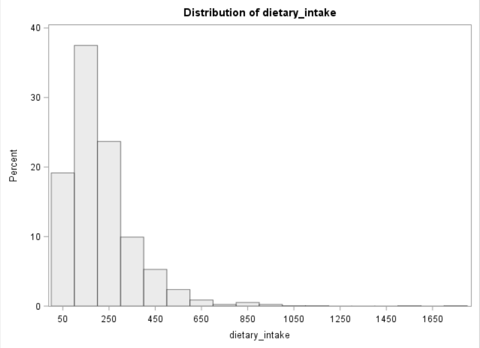
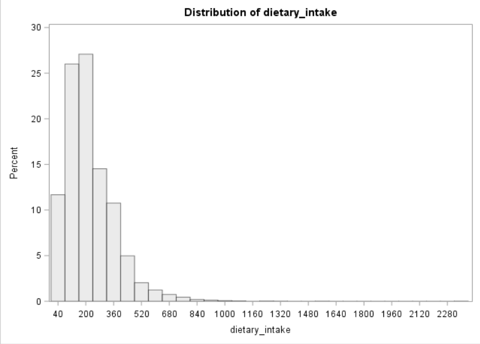
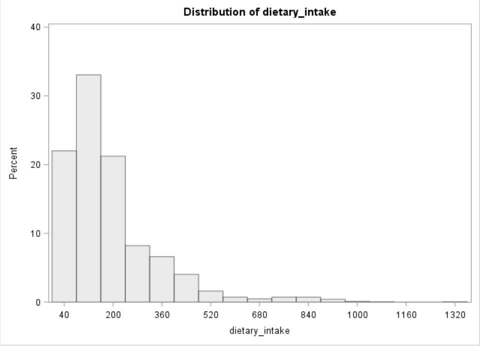
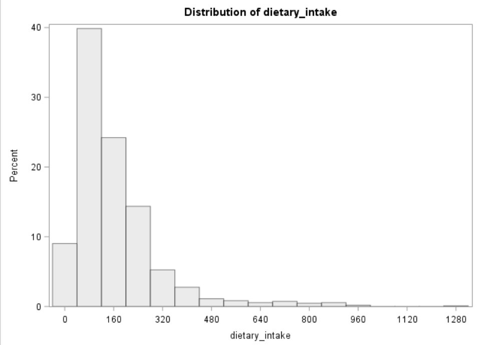
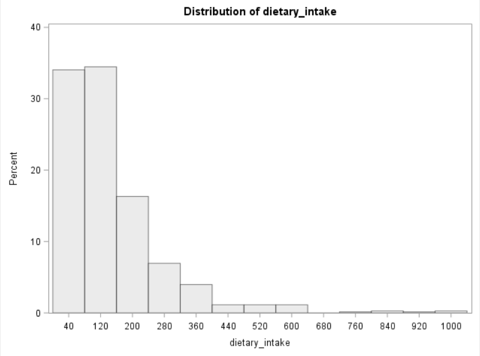
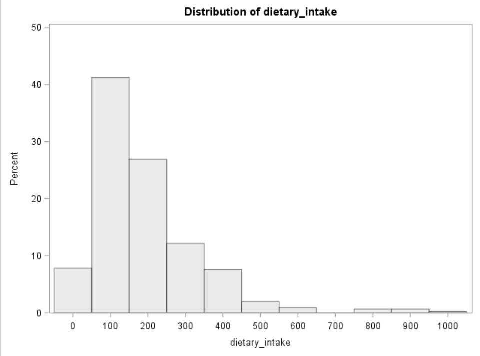
Supplemental Table 12 shows 25(OH)D Z-scores for those of European ancestry by quintile. As expected, as percentile increases, Z-scores do too.

Supplemental Figure 14: Histograms of key variables by cohort by ancestry



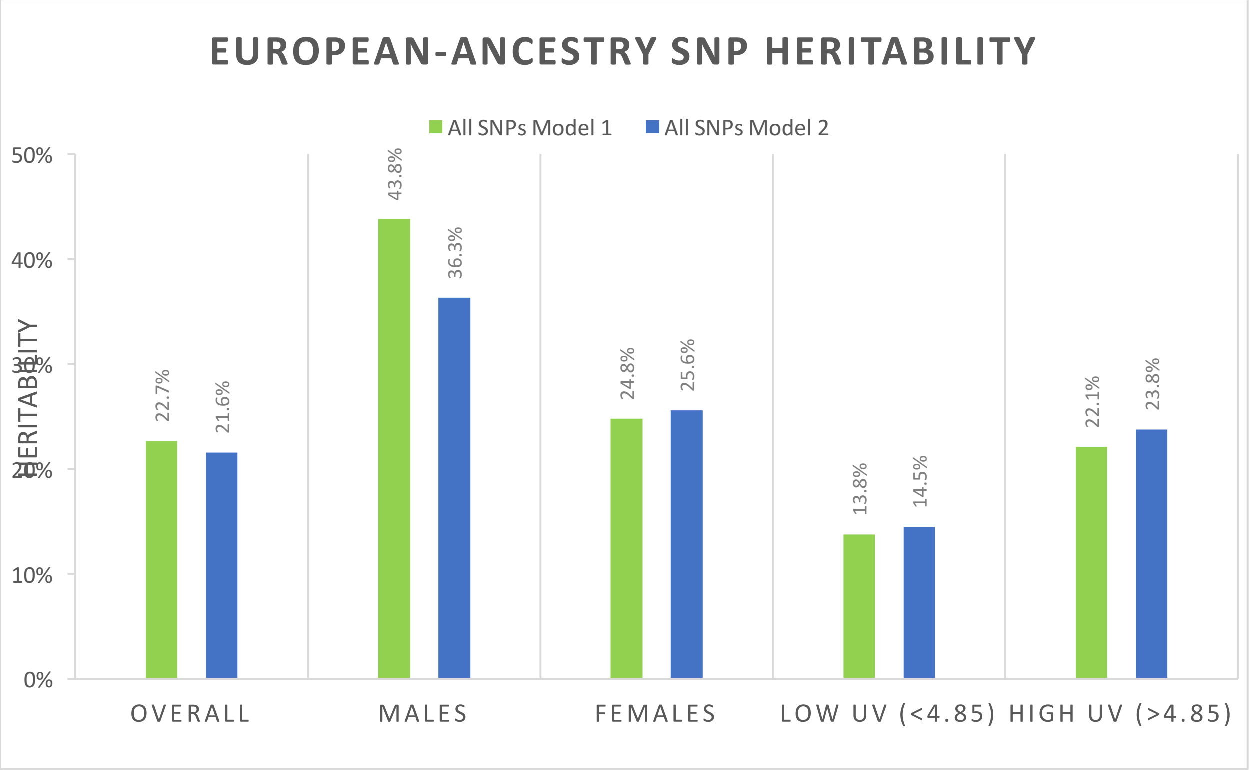
Age	ARIC		
	MESA		
	WHI		
BMI	ARIC		

<table>
<tr><td></td><td>MESA</td><td> </td><td> </td></tr>
<tr><td></td><td>WHI</td><td> </td><td> </td></tr>
<tr><td>Available UV radiation</td><td>ARIC</td><td> </td><td> </td></tr>
<tr><td></td><td>MESA</td><td> </td><td> </td></tr>
</table>

	WHI		
Dietary Intake	ARIC		
	MESA		
	WHI		

Available UV radiation appear to follow non-normal distributions. Available UV radiation is non-normal as the variable was calculated by month of blood draw, and blood draws were non-normally distributed throughout calendar time.

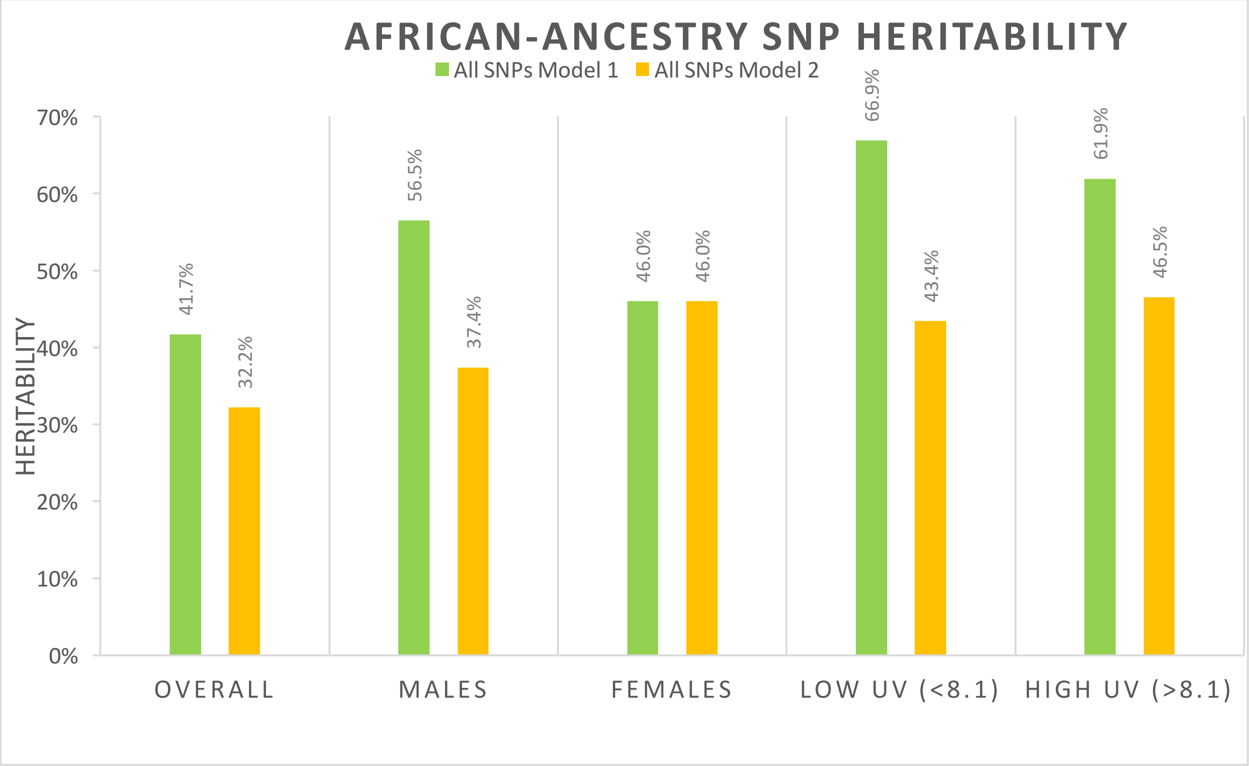
Supplemental Figure 15: European ancestry SNP Heritability



Sample Size				
7,119	3,301	3,818	3,689	3,430

Supplemental Figure 15 shows the estimated European ancestry SNP heritability for Model 1 and Model 2. Model 1 adjusts for age, sex, BMI and available UV radiation. Model 2 adjusts for age, sex, BMI, available UV radiation and dietary intake. Model 1 tends to produce biased estimates (i.e. overestimations), therefore model 2 was used for all Aim 2 analyses.

Supplemental Figure 16: African ancestry SNP Heritability



		Sample Size		
1,719	635	1,084	910	809

Supplemental Figure 16 shows the estimated African ancestry SNP heritability for Model 1 and Model 2. Model 1 adjusts for age, sex, BMI and available UV radiation. Model 2 adjusts for age, sex, BMI, available UV radiation and dietary intake. Model 1 tends to produce biased estimates (i.e. overestimations), therefore model 2 was used for all Aim 2 analyses.

Supplemental Table 13: Betas, standard errors and p-values for G*E interaction terms

	European Ancestry				**African Ancestry**			
Model	**PGS*UV**		**PGS*Intake**		**PGS*UV**		**PGS*Intake**	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Environmental main effect	0.096 (0.005)	<0.0001	0.11 (0.011)	<0.0001	0.07 (0.012)	<0.0001	0.13 (0.033)	0.0002
Genetic main effect	0.087 (0.038)	0.022	0.16 (0.018)	<0.0001	0.086 (0.071)	0.23	0.062 (0.031)	0.04
Interaction term	0.017 (0.0073)	0.021	0.0006 (0.018)	0.74	-0.0044 (0.012)	0.71	0.00042 (0.03)	.99

Supplemental Table 14: Characteristics of sub-sample with supplement use data

Cohort	**Variable**	**European-ancestry**	**African-ancestry**
WHI	**Sample size**	455	700
	Age (SE) [years]	66.6 (6.8)	61.8 (7.4)
	% Female	100	100
	BMI (SE) [kg/m^2]	29.9 (6.3)	31.2 (6.4)
	UV (SE)[1] [units]	5.2 (2.5)	5.5 (2.6)
	Intake (SE)[2] [IU]	420.9 (299.4)	308.8 (257.4)
	25(OH)D (SE) [ng/ml]	18.9 (10.7)	19.0 (15.4)

[1] available UV radiation

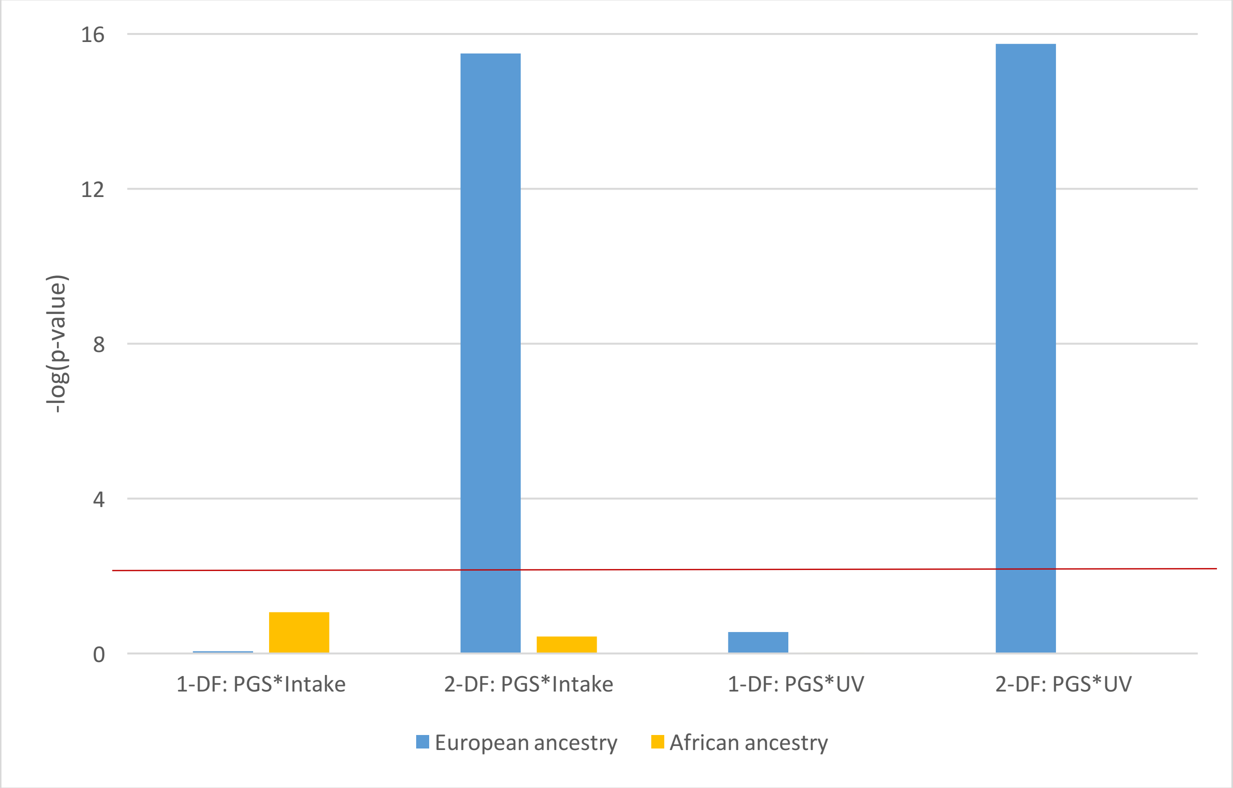
[2] vitamin D intake from diet and supplements

Supplemental Table 15: Characteristics of sub-sample with physical activity data

Cohort	Variable	European-ancestry	African-ancestry
MESA	Sample size	1,935	341
	Age (SE) [years]	62.7 (10.3)	62.3 (10.4)
	% Female	53%	52%
	BMI (SE) [kg/m^2]	27.8 (5.0)	30.1 (5.9)
	UV (SE)[1] [units]	4.5 (2.3)	161.4 (144.1)
	Intake (SE)[2] [IU]	188.9 (157.2)	25.7 (32.0)
	Physical activity (SE) [Met-hours/week]	26.0 (28.2)	24.8 (28.8)
	25(OH)D (SE) [ng/ml]	30.1 (10.9)	19.5 (8.9)
WHI	Sample size	436	363
	Age (SE) [years]	66.6 (6.8)	61.9 (7.6)
	% Female	100	100
	BMI (SE) [kg/m^2]	29.9 (6.3)	31.9 (6.4)
	UV (SE) [units]	5.2 (2.5)	5.5 (2.6)
	Intake (SE) [IU]	193.6 (144.7)	145.0 (119.4)
	Physical activity (SE) [Met-hours/week]	6.3 (9.5)	6.2 (11.0)
	25(OH)D (SE) [ng/ml]	18.9 (10.8)	16.8 (12.9)

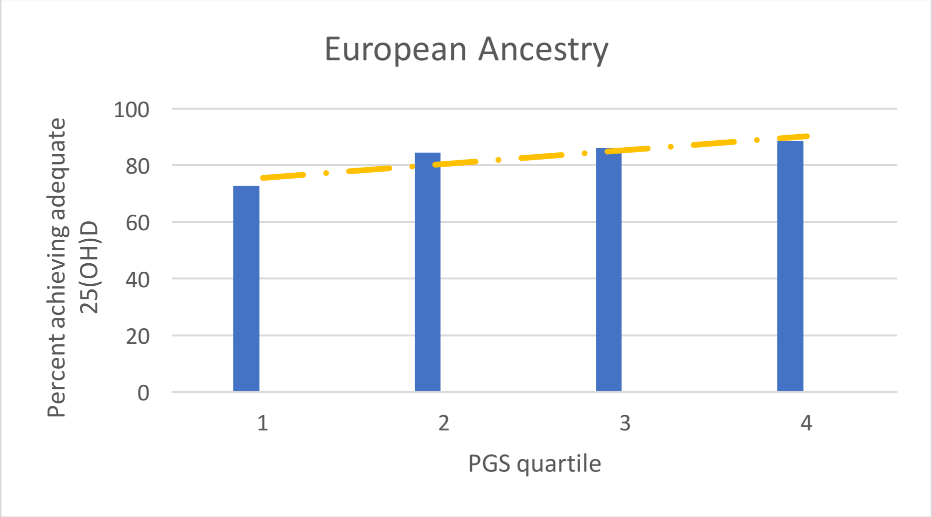
[1] available UV radiation
[2] vitamin D intake from diet

Supplemental Figure 17: Sensitivity analyses interaction test results from 1-DF and 2-DF models controlling for physical activity



Supplemental Figure 17 shows –log(p-values) for the 1-DF and 2-DF models of the PGS interaction; all models controlled for age, sex, BMI, cohort, physical activity, vitamin D intake and available UV radiation. The red line denotes the p=0.05 significance cutoff. The 2-DF PGS*intake and 2-DF PGS*UV models were statistically significant in participants of European ancestry (p=3.2x10^{-16}and 1.8x10^{-16}, respectively). Original to this manuscript.

Supplemental Figure 18: Percent achieving adequate 25(OH)D in those reaching IOM vitamin D intake guidelines by genetic risk in those of European ancestry independent of WHI

European ancestry participants with supplement use data and who reached IOM vitamin D intake guidelines (n=177)	

Supplemental Figure 18 shows the percent of European -ancestry participants who reach IOM vitamin D intake guidelines and achieved adequate 25(OH)D (20 ng/ml) by quartile of genetic risk. In those of independent from WHI (n=177), as genetic risk decreased (higher PGS), those reaching optimal vitamin D concentrations increased. The difference in percent reaching adequate 25(OH)D was 13.5%; 72.7% of participants with highest risk and 88.6% of participants with lowest risk reached adequate 25(OH)D. This is a statistically (p=0.028) and clinically significant difference. Original to this manuscript.

Description of cohorts

ARIC

"ARIC is a prospective epidemiologic study conducted in four U.S. communities. ARIC is designed to investigate the etiology and natural history of atherosclerosis, the etiology of clinical atherosclerotic diseases, and variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date.

ARIC includes two parts: the Cohort Component and the Community Surveillance Component. The Cohort Component began in 1987, and each ARIC field center randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community. A total of 15,792 participants received an extensive examination, including medical, social, and demographic data. These participants were reexamined every three years with the first screen (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth and last exam was in 1996-98. Follow-up occurs yearly by telephone to maintain contact with participants and to assess health status of the cohort.

In the Community Surveillance Component, currently ongoing, these four communities are investigated to determine the community-wide occurrence of hospitalized myocardial infarction and coronary heart disease deaths in men and women aged 35-84 years. Hospitalized stroke is investigated in cohort participants only. The study conducts community surveillance of inpatient heart failure (ages 55 years and older) and cohort surveillance outpatient heart failure events beginning in 2005. To date, the ARIC project has published 745 articles in peer-reviewed journals and other summary reports of ARIC data at various national and international scientific conferences and meetings (https://www.nhlbi.nih.gov/research/resources/obesity/population/aric.htm)."

MESA

"The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Approximately 38 percent of the recruited participants are white, 28 percent African-American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent (https://www.mesa-nhlbi.org/MESA_508TextOnly.htm)."

<u>WHI</u>

“The Women's Health Initiative (WHI) is a long-term national health study focused on strategies for preventing heart disease, breast and colorectal cancer, and osteoporotic fractures in postmenopausal women. Launched in 1993, the WHI enrolled 161,808 women aged 50-79 into one or more randomized Clinical Trials (CT), testing the health effects of hormone therapy (HT), dietary modification (DM), and/or calcium and Vitamin D supplementation (CaD) or to an Observational Study (OS). At the end of the initial study period in 2005, WHI Extension Studies (2005-2010, 2010-2020) continued follow-up of all women who consented.
This ground-breaking study changed the way health care providers prevent and treat some of the major diseases impacting postmenopausal women. Results from the WHI Hormone Trials have been estimated to have already saved $35.2 billion in direct medical costs in the US alone. To date, WHI has published over 1,400 articles and approved and funded 289 ancillary studies (https://www.whi.org/SitePages/WHI%20Home.aspx).”